René Descartes

Discurso del método

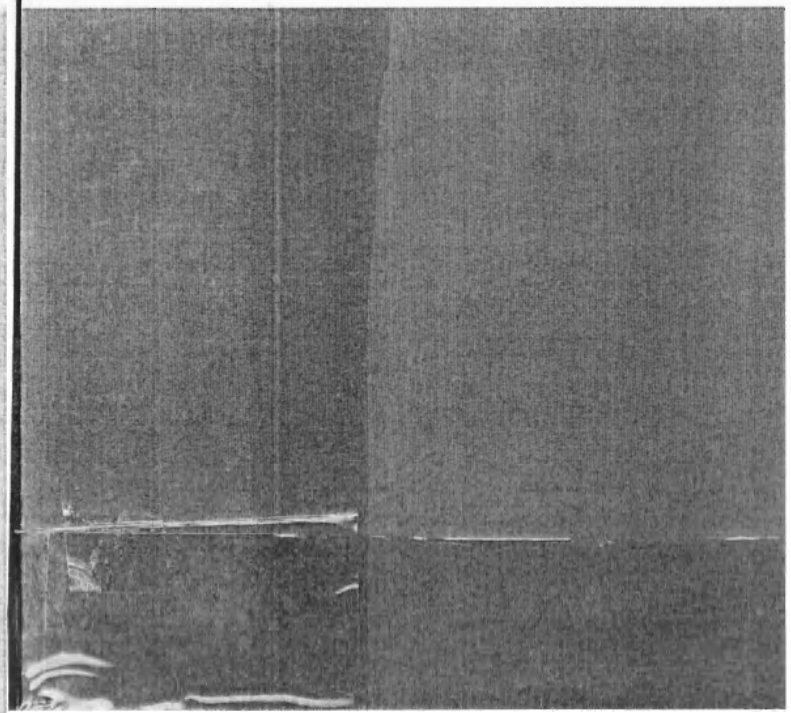
Estudio preliminar, traducción y notas de
Eduardo Bello Reguera

Tercer milenio

tecnos CLÁSICOS DEL PENSAMIENTO

René Descartes (1596-1650). «Héroe del pensamiento moderno», en expresión de Hegel, porque libera a la dama griega del caballero medieval conduciéndola por el nuevo camino de la modernidad.

Pero ¿no le llama Sartre —tras guiño de Kant— «sabio dogmático y buen cristiano»? Indudablemente, sabio. ¿Dogmático? «Descartes se atrevió al menos a enseñar a las buenas cabezas a sacudirse el yugo de la escolástica, de la opinión, de la autoridad; en una palabra, de los prejuicios y de la barbarie y, con toda esta rebelión cuyos frutos recogemos hoy, ha hecho a la filosofía un servicio más esencial quizá que todos los que ésta debe a los ilustres sucesores de Descartes» (D'Alembert).



Discurso del método

Colección
Clásicos del Pensamiento

fundada por Antonio Truyol y Serra

Director:
Eloy García

René Descartes

Discurso del método

Estudio preliminar, traducción y notas de
EDUARDO BELLO REGUERA

SEXTA EDICIÓN

Diseño de cubierta:
JV, Diseño gráfico, S. L.

TÍTULO ORIGINAL:
Discours de la méthode (1637)

1.ª edición, 1987
6.ª edición, 2006
Reimpresión, 2008

Reservados todos los derechos. El contenido de esta obra está protegido por la Ley, que establece penas de prisión y/o multas, además de las correspondientes indemnizaciones por daños y perjuicios, para quienes reprodujeren, plagiaren, distribuyeren o comunicaren públicamente, en todo o en parte, una obra literaria, artística o científica, o su transformación, interpretación o ejecución artística fijada en cualquier tipo de soporte o comunicada a través de cualquier medio, sin la preceptiva autorización.

© Estudio preliminar y notas, EDUARDO BELLO REGUERA, 1987
© EDITORIAL TECNOS (GRUPO ANAYA, S.A.), 2008
Juan Ignacio Luca de Tena, 15 - 28027 Madrid
ISBN: 978-84-309-4371-5
Depósito legal: M. 29.448-2008

Printed in Spain. Impreso en España por Fernández Ciudad, S. L.

ÍNDICE

ESTUDIO PRELIMINAR ..... *Pág.* IX

1. Algo más que un problema de método XI
2. El papel del método en la configuración de la ciencia moderna XXII
3. Descartes, ¿primer pensador moderno? XXXIX

Nota sobre la presente edición XLV
Bibliografía XLVII

DISCURSO DEL MÉTODO

PRIMERA PARTE 3
SEGUNDA PARTE 15
TERCERA PARTE 31
CUARTA PARTE 45
QUINTA PARTE 59
SEXTA PARTE 83

ESTUDIO PRELIMINAR

Por Eduardo Bello Reguera

A los trescientos cincuenta años de la publicación del *Discours de la méthode* (1637-1987), ¿se ha de considerar esta obra sólo como un documento *histórico*, particularmente significativo de las tensiones de una época, o más bien como la expresiva construcción *teórica* que inaugura la razón moderna? La tesis de la ruptura, según la cual no sólo el Renacimiento sino sobre todo la Época Moderna significan algo nuevo con relación al pasado medieval, subraya una y otra vez la segunda alternativa: Descartes ha fundado, definitivamente, el pensamiento moderno. Su discurso sobre el método entendido como camino de la razón teórica y práctica constituye el signo más inequívoco, la piedra fundamental del edificio del saber y de la cultura moderna. Frente a esta tesis, defendida por A. Koyré, E. Garin, M. Guéroult, S. Turró entre otros, la tesis de la continuidad se empeña en afirmar que, a pesar de Descartes, como a pesar de Copérnico, Kepler, Bruno, Galileo, Bacon, nada nuevo hay bajo el sol: la ciencia de Galileo y de Descartes perseguía una meta imposible e incluso falsa (A. C. Crombie) o se limitaba a recuperar las investigaciones del siglo XIV (P. Duhem) y, en filosofía, Descartes apenas se diferencia de un epígono del pensamiento medieval (E. Gilson).

El riesgo de unilateralidad de ambas tesis es bien patente. Por una parte, la afirmación de la originalidad del autor del *Discurso del método* hasta el punto de creer que su pensamiento ha brotado en su totalidad, como Minerva, de la cabeza de un Júpiter, ha hecho olvidar la presencia en éste y en otros textos de términos, supuestos y problemas de la tradición medieval. Por otra, el rastreo de tal presencia ha llevado a especialistas de dicha tradición, como Gilson, a sobrevalorar su papel en la construcción de la filosofía cartesiana, para difuminar en ella lo que tiene de originalidad y novedad, esto es, de modernidad.

Si este breve ensayo constituye un argumento más en favor de la tesis de la ruptura, no es porque en él se encuentre la simple repetición del enunciado: «Descartes es el primer pensador moderno», sino porque la validez de tal enunciado se infiere lógicamente del análisis de aquello que significa y especifica al *pensador moderno*. Ahora bien, para evitar el riesgo de unilateralidad señalado, tendremos en cuenta en este análisis las raíces históricas del pensamiento cartesiano, pero sin limitarnos a una de ellas, como hace E. Gilson[1], sino remitiéndonos a la pluralidad de las tradiciones teóricas en las que se inspira[2],

[1] *Études sur le rôle de la pensée médiévale dans la formation du système cartésien*, J. Vrin, Paris, 1930.

[2] S. Turró ha señalado el paradigma renacentista, sobre todo el mecanicismo y la nueva ciencia (*Descartes. Del hermetismo a la nueva ciencia*, Anthropos, Barcelona, 1985); E. Cassirer, el platonismo (*El problema del conocimiento*, FCE, México, 1953, vol. I, pp. 447-458) y el estoicismo (*Descartes, Corneille, Christine de Suède*, J. Vrin, París, 1942, pp. 71-87); A. Koyré subraya también el platonismo (*Études galiléennes*, Hermann, Paris, 1966; trad. esp. en Siglo XXI); D. M. Clarke reivindica el ideal aristotélico de demostración y certeza (*Descartes' Philosophy of Science*, Manchester University Press, 1982; trad. esp. en Alianza); P. Rossi vincula el saber moderno a las artes mecánicas y al nuevo ideal científico (*Los filósofos y las máquinas, 1400-1700*, Labor, Barcelona, 1966), L. Brunschvicg, señala la influencia de Montaigne (*Descartes et Pascal, lecteurs de Montaigne*, New York, 1944); y en general véase G. Rodis-Lewis, *L'Oeuvre de Descartes*, J. Vrin, Paris, 1971, 2 vols.).

así como a los acontecimientos históricos vividos y traducidos, antes que Hegel, a conceptos.

Desde esta perspectiva, la respuesta a la pregunta formulada al comienzo no puede ser dilemática. Si el *Discurso del método* aparece hoy día como expresiva construcción teórica de la razón moderna, no es sino sobre la base de las investigaciones acerca del momento histórico en el que ha sido elaborado, y a pesar de dichas investigaciones, es decir, a posar de los posibles intentos de desdibujar la originalidad cartesiana en el horizonte de su época histórica o de otra anterior.

Así, pues, nos vamos a referir en primer lugar a las tensiones histórico-culturales de comienzos del siglo XVII, cuyo eco se percibe claramente en los escritos de Descartes concretamente en el *Discurso*, poniendo de relieve sobre todo la tensión suscitada en torno a la nueva ciencia. Nos preguntaremos, en segundo lugar por el fundamento de la nueva ciencia, tal como aparece formulada en esta misma obra. Y, en tercer lugar, verificaremos el enunciado: «Descartes, primer pensador moderno».

1. ALGO MÁS QUE UN PROBLEMA DE MÉTODO

Que el *Discurso del método* es algo más que un tratado de metodología nadie lo pone en duda. En lo que ya no es tan fácil ponerse de acuerdo es en precisar qué es ese algo más, aun prescindiendo de los tres ensayos a los que el *Discurso* precedía como introducción[3]. Si nos parece insuficiente afirmar que se trata de una autobiografía intelectual, porque se corre el riesgo de trivializar y olvidar los auténticos problemas, la simple identificación de cada una de las seis partes de que se compone la obra tampoco basta para especificar ese algo más, aunque nos pone en camino. En tal identificación F. Alquié es más sobrio que el mismo Descartes, cuando escribe que «la primera es una historia, la segunda una lógica, la tercera una moral, la cuarta una metafísica, la quinta una exposición científica, la sexta una especie de llamamiento al público»[4].

[3] El 8 de junio de 1637 aparecía en Ley de la primera obra impresa de Descartes, con este título: *Discours de la méthode pour bien conduire sa raison, et chercher la vérité dans les sciences. Plus la Dioptrique, les Météores et la Géométrie qui sont des essais de cette méthode* (C. Adam y P. Tannery, vol. VI, p. XIII. En adelante se citará esta ed. mediante las iniciales AT, seguidas del vol. y de la p. correspondientes).

En la relación de Alquié se nos oculta lo esencial. No se nos dice que lo que Descartes se propone es establecer un nuevo fundamento tanto de la razón teórica como de la razón práctica[5]. Desde un nuevo fundamento epistemológico y metodológico (Parte II), se ha de entender la Parte I no tanto como un relato autobiográfico, sino más bien como una «consideración [crítica] sobre las ciencias»[6], como una crítica de los fundamentos del saber establecido. Desde el nuevo fundamento metafísico (Parte IV) se entenderá mejor no sólo el esbozo innovador de otra moral (Parte III) sino también el resumen de las teorías física y antropológica, el apunte para una teoría lingüística (Parte V), así como los proyectos de investigación en la ciencia de la naturaleza y las condiciones necesarias de su realización (Parte VI).

Si nos atenemos, en cambio, a aquella heterogénea relación, se nos plantea consecuentemente un problema de coherencia, que Alquié resuelve en términos de historia intelectual: el *Discurso del método* —escribe— es,

[4] Descartes, *Oeuvres philosophiques*, ed. de F. Alquié, Garnier, Paris, 1963, t. I, p. 550. En la ed. francesa (1637) Descartes pone los seis epígrafes al comienzo de la obra; en la latina (1644) figuran en el margen del comienzo de cada parte.

[5] Parte I, p. 13; II, p. 22; IV, p. 44 (AT, VI, pp. 10, 17 y 31).

[6] AT, VI, p. 3.

ante todo, «la historia de los pensamientos de Descartes. Y la coherencia que podemos descubrir en él es la de un relato, no la de un sistema[7]. Y, en efecto, en el *Discurso* resuena, como veremos, toda la historia de la evolución del pensamiento de Descartes: la de sus estudios en La Flèche (1606-1614), en la universidad de Poitiers (1614-1616) y los primeros años en el gran libro del mundo (1617-1619), la de los nueve años practicando el método (1619-1628) y la historia de ensayos y proyectos (1628-1637), así como la de las líneas principales de los escritos y tareas anunciados en el *Discurso y* realizados en el período 1638-1650[8].

Ahora bien, lo que se sigue omitiendo es que en este primer texto impreso resuena con igual fuerza la historia de la época. Si resolvemos el problema de la coherencia no en términos de biografía intelectual, sino de fundamentación del saber y del hacer, la atención del lector no ha de centrarse tanto en el entretenido y cauteloso relato, cuanto en las tensiones sociales y culturales —presentes en él— del primer tercio del siglo XVII, que constituyen la encrucijada incierta donde comienza y se constituye el camino de la razón moderna. Los problemas, las tensiones, las luchas, las contradicciones de la época son también ese «algo más» que encontramos en el *Discurso del método*. Es imposible comprender el sentido y la historia efectiva de éste, la *wirkliche Historie* del pensamiento del autor ignorando los problemas de la época y no teniendo en cuenta las raíces históricas de todo pensador.

La época histórica que Descartes traduce a pensamiento, época de crisis y de renacimiento, ha sido calificada por A. Koyré como tiempo de «incertidumbre y

[7] *Op. cit.*, p. 553.

[8] Cf. H. Gouhier, *Descartes. Essais sur le «Discours de la méthode», la métaphysique et la morale*, J. Vrin, Paris, 1973 (3e. ed.), pp. 23-24.

desarraigo»[9], como consecuencia de la ruptura de la unidad religiosa, de la unidad política, de la unidad cultural y cosmográfica en los siglos XV y XVI. Tales rupturas —duro y alborozado despertar del hombre renacentista y de la razón moderna— van precedidas y seguidas de fuertes tensiones, que designamos mediante los términos contrapuestos: tradición/renacimiento, reforma/contrarreforma, nobleza/burguesía, feudalismo/capitalismo, geocentrismo/heliocentrismo, teocentrismo/antropocentrismo, fe/razón, etc. Por una parte, la mayor pasión por el descubrimiento en inventores, geógrafos, viajeros, humanistas, artistas, científicos; por otra, la mayor estrategia de la censura: la inquisición y la tortura. De un lado, la libre interpretación defendida por la Reforma, como una ventana abierta en el claustro medieval a la libertad de pensamiento; de otro, la persecución para quien proyecta la libre discusión de los temas de actualidad (caso Pico de la Mirandola, 1486), la hoguera para quien se permite pensar por sí mismo sin atenerse a lo que dicta la *auctoritas* (caso Bruno, 1600), o la tortura mental, la condena y la cárcel para quien se empeña en interpretar el universo desde la razón científica moderna (caso Galileo, 1616, 1633).

G. Pico, manifiesto vivo del humanismo renacentista italiano[10], afirmó teóricamente la libertad sin ninguna barrera. Lutero, en cambio, sólo apuntalaba las del hombre exterior, mientras destruía la barrera espiritual del hombre interior. En ese poder de destrucción de la autoridad de la Iglesia, de toda mediación del hombre con Dios, se fundamenta una de las características de la libertad moderna: la independencia[11]. Independencia para interpretar

[9] «Entretiens sur Descartes», en *Introduction à la lecture de Platon*, suivi de..., Gallimard, Paris, 1962, p. 175.

[10] E. Garin, *La revolución cultural del Renacimiento*, Crítica, Barcelona, 1981, pp. 159-166; Pico de la Mirandola, *De la dignidad del hombre*, ed. de L. Martínez Gómez, Editora Nacional, Madrid, 1984.

[11] E. Fromm (1942), *El miedo a la libertad*, Paidós, Buenos Aires,

—libre interpretación— y para pensar. No cabe duda de que, al pensar con independencia, Lutero —como más tarde Bruno, Galileo, Descartes— liberó a la conciencia europea de la encorsetada autoridad de la tradición y de la Iglesia: libertad como liberación, como ruptura de la caja mágica que produce herejes, heterodoxos, librepensadores. Frente a los detentadores de la patente de la caja mágica, que también producía cazadores de brujas, torturadores y fanáticos luchadores religiosos —recuérdese la sangrienta noche de San Bartolomé[12]—, surgieron los artífices de la paz, los inventores de la tolerancia, los defensores de la libertad de conciencia, por la que lucharon en la forma de libertad religiosa en primer lugar los grupos minoritarios —como los calvinistas en Francia— perseguidos por igual en todos los Estados confesionales y las Iglesias oficiales. Al sintetizar el significado de la Reforma en este contexto, E. Fromm escribe: «Estamos convencidos de que la libertad religiosa constituye una de las victorias definitivas del espíritu de libertad»[13]. Pero no la definitiva, dado que a la independencia del pensar le es indispensable la libertad de la palabra.

A comienzos del siglo XVII, a los Bruno y a los Galileo se les prohíbe, con el grave tono inquisitorial, precisamente la libertad de la palabra[14]. Por eso Descartes,

1971, p. 104. Cf. M. Ballesteros, *La revolución del espíritu. Tres pensamientos en libertad*, Siglo XXI, Madrid, 1970.

[12] En la noche del 23 al 24 de agosto de 1572, con motivo de la boda de Enrique de Navarra con Margarita de Valois, fueron acuchillados los hugonotes de París. Bajo esta dura impresión, J. Bodin redactó su obra principal, *Six livres de la République* (1576), trad. esp. en Tecnos, 1985. El fanatismo religioso está presente en las guerras de religión de la época, como la de los campesinos alemanes (1524): cf. M. Lutero, *Escritos políticos*, ed. de J. Abellán, Tecnos, Madrid, 1986; o en la guerra de los Treinta Años (1618-1648) en la que participa Descartes. Cf. R. van Dülmen, *Los inicios de la Europa moderna*, Siglo XXI, Madrid, 1984, pp. 342- 383.

[13] *Op. cit.*, p. 138.

[14] G. Bruno fue quemado vivo en el Campo di Fiori (Roma), en 1600, tras ser detenido por la Inquisición de Venecia en 1592.

habiendo aplazado el ejercicio de ésta en 1633, procede con suma cautela cuatro años más tarde al decidirse a publicar el *Discurso del método*. Por una parte, omite hablar expresamente de los principios de su teoría física, con el fin de no perturbar a los «doctos, con los que no deseo indisponerme»[15]. Por otra, lo publica no en Francia, sino en Leyde (Holanda) y sin nombre de autor[16].

Francia, brazo ejecutor de la contrarreforma, como España, no era precisamente el oasis de la libertad de pensamiento, que en cierto modo comenzaba a brotar en los Países Bajos. En Francia se alzaba, en cambio, el torreón del absolutismo sobre cimientos tan *ad hoc* como la intransigencia y la censura. El tolerante rey Enrique IV, que consigue la permanencia de los hugonotes promulgando el Edicto de Nantes en 1598[17] es asesinado en 1610. A pesar de que la sexta ley fundamental exigía de Francia y de sus reyes la *catolicidad* de forma explícita y absoluta, el cardenal Richelieu, que había vinculado la política a su ética —«el primer deber del Estado es establecer el dominio de Dios»—, no duda, sin embargo, en entrar en alianza con la Suecia protestante en 1635, es decir, con los enemigos de la fe «por razón de Estado»: combatir al otro enemigo, la casa de Austria. Con todo, «las tensiones derivadas del *cauchemar des rébelions*, observa G. Barudio, urdidas en el interior por las frondas aristocráticas, los hugonotes y las intrigas de la Corte y agudizadas una y otra vez por las rebeliones campesinas y por la oposición parlamentaria, así como la presión del *«cauchemar des coalitions»* contra Francia, que Ri-

[15] Al comienzo de la Parte V; igualmente comienzo de la Parte VI.

[16] AT, VI, p. V. Esta cautela la repetirá Spinoza; pero en ambos casos el efecto no buscado va a ser la celebridad y la polémica.

[17] Edicto revocado por Luis XIV en 1685, provocando la emigración masiva de los hugonotes. Tal actitud intolerante contrasta con las llamadas a la tolerancia, a finales de esta década, por pensadores como J. Locke (*Carta sobre la tolerancia*, ed. de P. Bravo, Tecnos, Madrid, 1985).

chelieu veía siempre cercada, no permitían que el país viviera en paz, especialmente después de 1635 [...]. Mientras hubiera hugonotes y, por tanto, un Estado potencial dentro del Estado [...], mientras existiera la autonomía corporativa de la rica Iglesia, los señoríos de la nobleza y la compraventa de cargos para la burguesía, mientras las regiones, provincias y ciudades escapasen al centralismo real y los intelectuales estuviesen enzarzados en una polémica continua a pesar de la fundación de la *Académie française* en 1634, polémica de la que la lucha de Gassendi contra Descartes no es más que un ejemplo entre muchos, este reino sería difícil de mantener unido»[18]. Sobre todo cuando tal unión se pretende bajo el lema: «Sólo un señor, una fe, un bautismo, un Dios» *(Efes.)*.

¿Cuál es la posición de Descartes al respecto? En primer lugar, con relación al Estado y a la religión, no asume como en el campo de la ciencia el papel del revolucionario, ni siquiera el del reformador, tal como lo asumiera años antes La Boétie en su *Discurso de la servidumbre voluntaria o Contra el uno*[19]. *Descartes*, que mide siempre el margen de lo realmente posible no considera «razonable que un particular intentara reformar un Estado cambiándolo todo desde sus cimientos [...]. Mi propósito —precisa— no ha sido nunca otro que tratar de reformar mis propios pensamientos y construir sobre un terreno que sea enteramente mío»[20].

En segundo lugar, si paradójicamente Descartes excluye de la exigencia crítica y fundamentadora de la duda el problema del Estado y el tema de la religión

[18] *La época del absolutismo y la Ilustración: 1648-1779*, Siglo XXI. Madrid, pp. 84-85; *cauchemar* = pesadilla.

[19] La Boétie (1530-1563), amigo de Montaigne, escribe el libro hacia 1552, aunque no se publica hasta 1572 en París. Ha sido editado en Tecnos (Madrid, 1986) por J. M. Hernández-Rubio.

[20] Parte II, pp. 18-20 (AT, VI, pp. 13 y 15). Cf. el comienzo de la Parte III, primera máxima. Sobre el problema político, v. A. Negri, *Descartes político*, Feltrinelli, Milano.

¿dónde hemos de situar la tarea fundamentadora de la razón práctica? Evidentemente no allí donde él mismo se ha prohibido el paso, sino en la senda abierta en el terreno de la moral, como veremos.

Y, en tercer lugar, no cabe duda de que donde Descartes construye sobre un terreno originalmente suyo, donde lleva a cabo la creativa reforma del pensar, es en el espacio de la filosofía y de la ciencia, tal como lo expresa al final de la Parte II del *Discurso del método:* «Al darme cuenta de que todos los principios de las ciencias debían tomarse de la filosofía, en donde aun no hallaba ninguno cierto, pensé que era necesario ante todo que me propusiera investigarlos»[21].

¿Es este proyecto de reforma y fundamentación del saber y del hacer la salida de aquel tiempo de incertidumbre y desarraigo que describe A. Koyré? Si la Reforma había roto la certeza en Roma, como el descubrimiento de América había quebrado la certeza geográfica y la pólvora había hecho estallar la certeza y la seguridad del señor feudal, si los humanistas habían minado el edificio del saber medieval, derribando la clave de su bóveda, esto es, el principio de autoridad (Aristóteles, Sto. Tomás, la Biblia), recuperando de la antigüedad otros autores, otros textos, otras formas de vida, y si el heliocentrismo había roto la certeza cosmológica (aristotélica, ptolemaica y bíblica), el proyecto cartesiano contribuye indudablemente, en este marco, a la ruptura y destrucción del viejo mundo medieval y a la configuración de otro nuevo, el mundo moderno. Es muy importante tener presentes los dos aspectos de la única tarea fundacional. El primero, el de la crítica demoledora de la cultura recibida, está expresado sobre todo en la Parte I del *Discurso* y en la actitud de duda que problematiza de raíz todo saber. El segundo, en las demás partes del mismo y en el resto de su obra.

[21] Parte II, p. 29 (AT, VI, pp. 21-22).

El lugar primordial que ocupa el trabajo creativo en Descartes, le distancia precisamente de la actitud escéptica[22], generada por el sentimiento de incertidumbre y desarraigo tras la ruptura de toda certeza. Si nada es cierto, no se sabe nada, concluye F. Sánchez. Y Montaigne extrae la consecuencia: el hombre *no sabe* nada, porque *no es* sino finitud o nada[23]. Si Descartes es deudor del descubrimiento del yo de Montaigne, ello no significa que la duda cartesiana sea la fiel traducción del escepticismo que se encierra negativamente en el castillo del yo, por una parte, y, por otra, que en la afirmación de la *subjetividad* que sigue a la difícil senda de la duda está presente la herencia de los humanistas, quienes, mirándose con distanciamiento crítico en el espejo denominado «clásico», no sólo hacen renacer la cultura de la antigüedad, sino que, al reflexionar sobre los propios problemas, diseñan también una nueva imagen del hombre. Así, en el *Discurso de la dignidad del hombre*, con la fuerza expresivamente hercúlea de una pintura de Miguel Ángel, Pico «fijaba con mucha precisión el alcance subversivo de la nueva imagen del hombre»[24].

No menos importante, en la contribución de Descartes a la construcción de la razón moderna, es la interpretación de la naturaleza hecha por los pensadores del Renacimiento y muy especialmente la herencia científica de Copérnico, Kepler y Galileo, así como el atomismo que

[22] Parte IV, p. 46 (AT, VI, p. 32); III, p. 41 (AT, VI, p. 29).

[23] F. Sánchez, *Quod nihil scitur* (1576); trad. esp. en Aguilar. Los «ensayos» o experiencias de Montaigne datan de 1580 (I y II) y de 1588 (III): *Essais*, Gallimard, París; trad. esp. en Cátedra, Madrid, 1985-1987. Cf. R. H. Popkin, *Historia del escepticismo desde Erasmo hasta Spinoza*, FCE, Madrid, 1983.

[24] E. Garin (1973), *Medioevo y Renacimiento*, Taurus, Madrid, 1981, p. 75; E. Garin (1980), *El Renacimiento italiano*, Ariel, Barcelona, 1986; A. Heller (1978), *El hombre del Renacimiento*, Península, Barcelona, 1980.

conoce a través de Beeckman[25] y el proyecto de F. Bacon. Basta observar la Parte V del *Discurso*, donde Descartes resume su teoría física, o la Parte VI donde se proclama partidario de una filosofía práctica es decir, de un desarrollo aplicado del saber, tal como lo propone Bacon[26]. Ahora bien, ¿en qué se diferencia la filosofía de la ciencia de Descartes de la de Galileo, por ejemplo? ¿Qué es lo específico del proyecto cartesiano de la fundamentación del saber? ¿Coincide, acaso, con el de Bacon? ¿Qué lugar ocupa el discurso metodológico en el marco del proyecto de fundamentación? ¿No habría que comparar el *Novum organum* (1620) de Bacon y el *Dialogo sopra i due massimi sistemi del mondo* (1632) de Galileo con el *Discours de la méthode* (1637)?

Responderemos a algunas de estas preguntas al desarrollar, en seguida, el problema del fundamento de la ciencia moderna en Descartes. Pero, antes, sólo una pincelada más sobre el eco de la historia en el *Discurso*. Se trata, esta vez, de asomarse a la historia del pensamiento de Descartes resonando en el *Discurso*.

Es sabido que el texto publicado como introducción a los ensayos científicos de 1637 no fue escrito como una descripción coherente del método científico tal y como se planeó para las *Reglas*. Entre 1620 y 1636, observa Clarke, Descartes se ve comprometido de uno u

[25] Cf. S. Turró, *Descartes. Del hermetismo a la nueva ciencia*, Anthropos, Barcelona, 1985; A. Koyré, *Études galiléennes*, Hermann, Paris, 1966 (trad. esp. en Siglo XXI); D. M. Clarke, *Descartes' Philosophy of Science*, Manchester University Press, 1982 (trad. esp. en Alianza, Madrid, 1986). De guarnición en Breda (1618), Descartes ve un anuncio que planteaba a los científicos un problema de matemáticas: así conoció a Isaac Beeckman, atomista y decidido copernicano, con el que mantendrá larga correspondencia y amistad. Cf. E. Denisoff, *Descartes, premier théoricien de la physique mathématique*, Nauwelaerts, Louvain, 1970; S. Gankroger (ed.), *Descartes: Philosophy, Mathematics and Physic*, Harvester Press, Hassocks, 1980.

[26] Parte VI, p. 85 (AT, VI, pp. 61-62).

otro modo en la redacción de algo así como una biografía intelectual[27], un tratado de metafísica[28], las *Reglas*[29], *El Mundo*, un trabajo sobre meteorología y otro sobre óptica, junto con el ensayo en curso acerca de varios problemas de geometría. El texto de *El Mundo*, el primero preparado para su publicación, es abandonado en 1633, temeroso Descartes de correr la misma suerte que Galileo. Dos años más tarde decide publicar los ensayos sobre *Dióptrica y Meteoros*, precedidos de una corta introducción que persiste en la sección final de la Parte VI. Entre noviembre de ese mismo año y marzo de 1636, Descartes decide incluir en la publicación parte de sus trabajos de *Geometría*, lo que le obliga a escribir una nueva introducción. Pero en lugar de redactar una nueva, extracta del manuscrito de las *Reglas* las cuatro que son aplicables a cualquier disciplina, reordenadas en la Parte II del *Discurso*. «El texto final que resulta de tan curioso proceder —concluye Clarke— no es completamente coherente. Incluye una sección de la *Histoire de mon esprit* en la Parte I, y un resumen de las cuatro reglas principales del método cartesiano del Libro I de las *Reglas*. La Parte III es una pieza relativamente nueva sobre moral, mientras que la Parte IV reproduce parte de sus primeros trabajos sobre metafísica y fue incluida en el texto final debido a la presión del editor deseoso de completar el manuscrito[30]. La Parte V es un resumen de parte de su trabajo sobre física y meteorología que encontramos en

[27] «Acordaos de la *Histoire de votre esprit*. Es esperada por todos nuestros amigos» (Balzac a Descartes, 30 de marzo de 1628).

[28] La Parte IV sólo es un resumen del pequeño *Traité de métaphysique* (a Mersenne, 25 de noviembre de 1630), comenzado en 1629, y desarrollado más tarde en *Meditationes de prima philosophia*, 1641 (AT, VII, pp. 17 ss.); trad. esp. de V. Peña, Alfaguara, Madrid, 1977.

[29] Redactadas en 1628, son publicadas en 1701 (AT, X, pp. 359 ss.); ed. de J. M. Navarro Cordón, Alianza, Madrid, 1984.

[30] Descartes a Vatier, 22 de febrero de 1638.

Le Monde[31]. A pesar de que Descartes explica en la Parte V por qué abandona los planes de publicación de *Le Monde*, no obstante repite esta explicación al principio de la Parte VI y añade entonces el prefacio escrito originalmente para los dos ensayos sobre *Meteoros y Dióptrica*»[32].

A partir de 1637 Descartes desarrollará, sobre todo, problemas tratados en el *Discurso*, pero escasamente desarrollados: la metafísica, los principios de la filosofía, los principios de la física omitidos deliberadamente en la Parte V[33] y el tema moral, tanto en las cartas a Isabel y a Cristina de Suecia como en *Las pasiones del alma* (1649)[34].

La muerte de Descartes, acaecida en 1650, nos privó sin duda de otros resultados de sus proyectos de investigación inacabados. Pero nos queda, al menos, lo que considero esencial de su aportación: la reflexión sobre el fundamento de la ciencia y la conciencia moderna.

2. EL PAPEL DEL MÉTODO EN LA CONFIGURACIÓN DE LA CIENCIA MODERNA

Si hasta aquí nos hemos ocupado de ese «algo más» que hay en el *Discurso del método*, es necesario pregun-

[31] «Incluyo algunas cosas sobre metafísica, física y medicina en el primer Discurso para mostrar que (el método) alcanza a todo tipo de materias» (a Mersenne, abril, 1637).

[32] *La flosofía de la ciencia de Descartes*, Alianza, Madrid, 1986, pp. 190-191. Cf. G. Rodis-Lewis, *L'oeuvre de Descartes*, I, pp. 141-214; H. Gouhier, *Descartes. Essais sur le «Discours de la méthode»..., cit.*, pp. 55 ss.; G. Gadoffre, «Réflexions sur la genèse du *Discours de la Méthode*», *Revue de Synthèse*, New Series, 22, 1958, pp. 11-27.

[33] *Principia philosophiae*, 1644 (AT, XI-ii, pp. 21 ss.); trad. esp. de G. Quintás, Alianza, Madrid, 1995.

[34] *Les passions de l'âme*, 1649 (AT, XI, pp. 327 ss.); trad. esp. de J. A. Martínez Martínez y P. Andrade Boué, Tecnos, Madrid, 1997;

tar ahora: ¿Cuál es el papel del método en la configuración de la ciencia moderna? Anticipemos la respuesta de Heidegger, que es la nuestra: el método no es sino la «instancia fundamental» de la ciencia moderna[35].

Pero ¿no podría ser referida esta respuesta igualmente a la ciencia antigua, a la medieval o a la actual? Y, en cualquier caso, ¿es válida para la época moderna en general o solamente se adecua a la posición teórica de Descartes? Aunque la tesis de Heidegger está enmarcada en un breve comentario al discurso cartesiano del método en las *Regulae*, para la comprensión del sentido de éste como de la tesis de aquél es indispensable preguntarnos previamente por aquello que caracteriza a la ciencia moderna. Sólo entonces aparecerá con mayor nitidez el papel de Descartes en el proceso de su génesis y fundamentación.

Dando por supuesto que en los siglos XV se configura el paradigma de la ciencia moderna, que a su vez constituye uno de los fenómenos característicos de la época moderna[36], es necesario preguntar: ¿qué es lo que especifica a dicho paradigma? De otro modo ¿cuál es el rasgo esencial de la ciencia moderna? Caracterizar ésta, con relación a la ciencia antigua y medieval, diciendo

Lettres sur la morale, ed. de J. Chévalier, París, 1935; trad. esp. de E. Goguel, Buenos Aires, 1945.

[35] *Die Frage nach dem Ding*, Niemeyer, Tübingen, 1962; trad. esp.: *La pregunta por la cosa*, Alfa, Buenos Aires, 1975, p. 93.

[36] Th. S. Kuhn, *The structure of scientific revolutions*, University of Chicago Press, 1962 (trad. esp. en FCE); A. Koyré, *Études galiléennes*, *cit.* en nota 1, y *Études d'histoire de la pensée scientifique*, Gallimard, París, 1973 (trad. esp. en Siglo XXI), E. A. Burtt, *The metaphysical Foundations of modern physical Science*, Doubleday, New York, 1954 (trad. esp. en Ed. Sudamericana); N. W. Gilbert, *Renaissance concepts of method*, Columbia University Press, New York/London, 1960; E. M. Madden (ed.), *Theories of scientific method*, University of Washington Press, Seattle, 1960; R. S. Westfall, *The construction of modern science*, J. Wiley & Sons, New York, 1971; M. Heidegger, *Holzwege*, Klostermann, Frankfurt, 1950 (trad. esp.: *Sendas perdidas*, Losada, Buenos Aires, 1960, p. 68).

que la primera parte de los hechos y la segunda de conceptos generales y especulativos, sería olvidar que tanto en un caso como el otro se trata de hechos y de conceptos, y que lo decisivo es el modo en que los hechos son comprendidos y los conceptos aplicados. Tampoco es totalmente cierto que el experimento constituya lo específico de la ciencia moderna pues no sería difícil probar que está presente tanto en la época antigua como en la medieval; lo que hay que subrayar como importante no es, pues, el experimento en cuanto tal, sino el modo en que se proyecta el experimento para determinar conceptualmente los hechos. Y cuando se señala que la nueva ciencia se caracteriza por el cálculo y la medida, se olvida una vez más que la concepción pitagórica de que la realidad está esencialmente constituida por el número quedó formulada definitivamente en el *Timeo* de Platón, aunque luego fuera marginada; lo que no hay que olvidar aquí es que el problema radica también en el modo y en el sentido de los cálculos y mediciones. «Con las tres caracterizaciones nombradas de la ciencia moderna —ciencia de los hechos, ciencia experimental, y de la medición—, observa Heidegger, no hemos tocado el rasgo fundamental de la nueva posición intelectual [...]. Daremos un título a este carácter fundamental de la actitud intelectual moderna diciendo: la nueva exigencia del saber es exigencia *matemática*»[37].

Ahora bien, ¿en qué sentido se puede afirmar que la exigencia matemática es el rasgo fundamental de la nueva ciencia? ¿No cabría pensar más bien que, dado que los elementos citados ya están presentes en la ciencia griega,

[37] *La pregunta por la cosa*, *cit.*, pp. 64-65. Cf. L. Brunschvicg, «Mathématique et métaphysique chez Descartes», *Rv. de Mét. et de Mor.*, 34, 1927, pp. 277-324; J. Vuillemin, *Mathématiques et métaphysique chez Descartes*, PUF, Paris, 1960; J. L. Allard, *Le mathématisme de Descartes*, Editions de l'Université d'Ottawa, Ottawa, 1963; E. Denissoff, *Descartes, premier théoricien de la physique mathématique*, Nauwelaerts, Louvain, 1970.

lo que constituye lo especifico de la ciencia moderna no es sino su síntesis? «Sobre la base de tal síntesis la ciencia moderna desarrolla luego esa serie de descubrimientos que la cultura antigua ni podría sospechar y que han transformado el mundo»[38].

Tal vez el problema que habría que resolver, de acuerdo con esta tesis, es el problema de cómo se opera la síntesis. Tal vez la pregunta que encierra este problema coincide, en el fondo, con la pregunta por el sentido de la «exigencia matemática». ¿No es, acaso, esta exigencia la que traduce el enunciado de Galileo «la filosofía está escrita en este grandioso libro (que yo llamo universo) [...] en lenguaje matemático»?[39]. ¿No es esta misma exigencia matemática la que Descartes eleva a modelo de todo saber cierto?[40]. No es extraño pues, que A. Koyré coincida con Heidegger al preguntarse por «el papel desempeñado por las matemáticas en la constitución de la ciencia de lo real», esto es, la física moderna, evocando al mismo tiempo que la cuestión del papel y de la naturaleza de las matemáticas era el principal tema de discusión entre Aristóteles y Platón. No es extraño también que afirme Koyré, en base al mayor interés de Platón por las matemáticas, que «en la época galileana matematismo significa platonismo»[41], que el *Diálogo* de Galileo es una obra polémica y de batalla, apunta su máquina de guerra contra la ciencia y la filosofía tradicionales y combate la tradición aristotélica en nombre de otra filosofía, la de Platón[42], y que, sin embargo, «no es

[38] E. Severino (1984), *La filosofía moderna*, Ariel, Barcelona, 1986, p. 33.

[39] *Il Saggiatore*, *Opere*. Ed. Nazionale, Firenze, 1929-39, vol. VI, p. 232. Cf. *Discorsi e dimostrazione matematiche interno a due nuove scienze*, a cura di A. Carugo e L. Geymonat, Torino, 1958; trad. esp. en Editora Nacional.

[40] parte II del *Discurso*, pp. 26-27 (AT, VI, p. 19).

[41] *Estudios galileanos*, Siglo XXI, Madrid, 1980, pp. 265 y 271.

[42] *Ibid.*, pp. 200-202.

Galileo sino Descartes quien asegura la definitiva victoria del platonismo y desaloja al aristotelismo de las posiciones que había ocupado durante tanto tiempo»[43]. Volvamos a la pregunta: ¿en qué sentido se puede afirmar que la exigencia matemática es el rasgo fundamental de la nueva ciencia? Distinguiendo con Heidegger entre la «matemática» y lo «matemático», y asumiendo que aquélla sólo es una determinada configuración de lo matemático, una aproximación etimológica a la palabra nos indica que τά μαθήματα significa lo que se *puede aprender*, y por eso también lo que se puede enseñar[44]. Pero ¿cuál es el sentido auténtico de lo «matemático»? Desde hace mucho estamos habituados a pensar los números bajo lo matemático. ¿Por qué se consideran precisamente los números como algo matemático? Si μάθηοις quiere decir el aprender y μαθηματα lo aprendible, esto es, las cosas en cuanto las aprendemos, lo matemático no es sino aquello «de» las cosas que en verdad ya conocemos. Citemos el ejemplo que pone Heidegger. Vemos tres sillas y decimos: son tres. Lo que es «tres» no nos lo dicen ni las tres sillas, ni las tres manzanas, ni los tres gatos, ni cualesquiera otras tres cosas. Más bien podemos contar solamente tres cosas como tres, si conocemos ya el «tres». Por lo tanto, cuando concebimos el número tres como tal, sólo tomamos conocimiento explícito de algo que de alguna manera ya poseemos. Este tomar conocimiento es el verdadero aprender. El número es algo aprendible en sentido real, un μαθήμα, es decir, algo matemático. Por lo tanto, «lo matemático —concluye Heidegger— es aquella posición fundamental en la cual nos proponemos conocer las cosas en aquel modo en que ya nos son dadas, y

[43] *Ibid.*, p. 277. Cf. A. Koyré, «Galilée et Descartes», en *IX Congrès international de philosophie*, Paris, 1937, p. 41; W. R. Shea, «Descartes as a critic of Galileo», en *New perspectivas on Galileo*, Reidel, Dordrecht, 1978, pp. 139-159.

[44] *La pregunta por la cosa*, p. 65.

deben ser dadas. Por eso lo matemático es el presupuesto básico del saber de las cosas»[45].

Tal es el sentido de la leyenda que Platón puso a la entrada de su Academia: «Nadie que no haya comprendido lo matemático debe entrar aquí». Pero si éste es el sentido auténtico de lo matemático, aun es preciso mostrar de qué modo el rasgo fundamental del pensamiento y el saber modernos es matemático, de qué modo, en base a este presupuesto, la ciencia moderna parte de hechos, es experimental y ciencia de la medición.

Si tenemos en cuenta que en la época moderna la ciencia se constituye fundamentalmente estudiando el problema del movimiento —Kepler formula la ley de las órbitas de los planetas, Galileo la ley de caída de los cuerpos, Descartes la ley de inercia, Newton la ley de la gravedad—, esto es, estudiando problemas físicos, habría que preguntar con Einstein: ¿cómo puede ser que la matemática —un producto del pensamiento humano independiente de la experiencia— se adecue tan admirablemente a los objetos de la realidad empírica? En su opinión, las proposiciones de la matemática en tanto que se refieren a la realidad no son seguras, y en la medida en que no se refieren a ella alcanzan seguridad. De ahí que proponga que, para poder hacer alguna afirmación acerca de la realidad, «la geometría debe ser desprovista de su carácter meramente lógicoformal, mediante la coordinación de los objetos reales de la experiencia con los esquemas conceptuales vacíos de la geometría axiomática»[46].

Ahora bien, teniendo en cuenta la distinción heideggeriana entre la «matemática» como disciplina particular

[45] *Ibid.*, pp. 70-71.

[46] «Geometría y experiencia», Conferencia pronunciada ante la Academia Prusiana de las Ciencias el 27 de enero de 1921. Cf. J. Echeverría, «Nota sobre la Geometría de 1637 y el Método cartesiano», en *Descartes*, de V. Gómez Pin, Dopesa, Barcelona, 1979, Apéndice II.

y lo «matemático» como presupuesto básico del saber de las cosas, se trata de mostrar, no cómo opera la matemática, sino de qué modo lo matemático es lo que constituye y específica a la ciencia moderna como física matemática.

En la investigación y formulación de las leyes antes señaladas, lo matemático es el proceso que abre un ámbito o dominio del ente en el que se muestran las cosas, esto es, los *hechos*. Las determinaciones y enunciados prefijados en el proyecto de investigación son los *axiomas* —*Axiomata sive leges motus*, «Principios o leyes del movimiento», escribe Newton[47]—, entendidos como principios o proposiciones fundamentales. Según Heidegger, en el proceso de investigación que abre un ámbito o dominio del ente, el proyecto matemático axiomático, no sólo «profigura en *esquema fundamental* (*Grundris*) la estructura de cada cosa y de sus relaciones con toda otra cosa», sino también delimita el ámbito de la naturaleza de tal modo que los cuerpos sólo pueden ser cuerpos, los hechos sólo pueden ser hechos, «en tanto están incluidos y entretejidos en ese ámbito»[48].

Que el libro del universo esté escrito en lenguaje matemático significa, pues, que los cuerpos no tienen propiedades, ni fuerzas ocultas. Significa que los cuerpos naturales sólo son tal como se muestran en el ámbito del proyecto matemático, es decir, que los caracteres del lenguaje matemático —precisa Galileo— «son triángulos, círculos y otras figuras geométricas, sin los cuales es imposible entender una sola palabra»[49]. Pero ¿cómo acceder a los cuerpos, a los hechos? Todo depende del modo de cuestionar y determinar la naturaleza cognoscitivamente. Que los caracteres del lenguaje matemático

[47] *Philosophiae naturalis principia mathematica* (1687), William y J. Innys, London, 1726 (3d. ed.), título del cap. II. Trad. esp. de A. Escohotado, Tecnos, Madrid, 1987.

[48] *La pregunta por la cosa*, p. 85.

[49] *Il Saggiatore*, cit. p. 232.

sean tales significa, en primer lugar, que el mostrarse de los hechos o de los cuerpos está profigurado en el esquema fundamental, que el modo de acceso a los cuerpos o investigación de lo que se muestra en la experiencia está predeterminado en el plan del proyecto matemático. Significa, en segundo lugar, que «el cuestionar sólo puede ser formulado de tal manera que ponga de antemano las condiciones a las cuales la naturaleza debe responder de tal o cual manera. Sobre la base de lo matemático —observó a Heidegger— la experiencia se transforma en experimento en sentido moderno. *La ciencia es experimental sobre la base del proyecto matemático*»[50]. Y, en tercer lugar, sobre la base de este mismo proyecto, la *medición numérica* no sólo es una posibilidad de lectura del universo, sino también una exigencia tanto de la investigación de las leyes que traducen las relaciones entre los hechos, como de la regulación exacta y fiable de las condiciones del experimento.

Pues bien, sobre la base del proyecto matemático se han de entender, por una parte, la matemática como disciplina particular y, por otra, el método entendido como instancia fundamental de la ciencia moderna. Sobre lo primero, hay que tener en cuenta que «la fundamentación de la geometría analítica por Descartes, la fundamentación del cálculo de fluxiones por Newton, y la simultánea fundamentación del cálculo diferencial por Leibniz, todo esto, tan nuevo, matemático en sentido restringido, fue posible y ante todo necesario sobre la base del rasgo matemático fundamental del pensar en general»[51].

[50] *La pregunta por la cosa*, p. 85, el subrayado es mío. Cf. E. Denissoff, *Descartes, premier théoricien de la physique mathématique*, Nauwelaerts, Louvain, 1970. Un punto de vista opuesto sostiene, en cambio, A. Gewirtz, en «Experience and the non-mathematical in the cartesian methode», *Journal of the History of Ideas*, 2, 1941, pp. 183-210.

[51] *La pregunta por la cosa*, p. 86.

En cuanto al *papel del método* en la configuración de la ciencia moderna, si observamos la producción teórica de la época —el *Novum organum* (1620), el *Diálogo* (1632), y el *Discurso* (1637), así como los *Principia* (1687) de Newton[52]—, podemos inferir al menos dos conclusiones. Por una parte, en todas estas obras, sobre todo en las tres primeras, tiene lugar una crítica explícita de los métodos tradicionales, de modo que el fundamento de la verdad ya no tiene que ser referido a la *auctoritas*, a la tradición, a la revelación, o a la conclusión del silogismo aristotélico. La *parte destructiva*, en la que Bacon critica y refuta los ídolos o falsas nociones como causas del error de la filosofía tradicional[53], se corresponde con la tarea crítica, destructiva, que Salviati —portavoz de Galileo en el *Diálogo*— lleva a cabo de las posiciones teóricas del aristotélico Simplicio[54], y con la crítica que de la cultura heredada y de sus métodos hace Descartes en la Parte I del *Discurso;* tal crítica y rechazo de un modelo de verdad, de la tradición y de sus métodos, sólo es en estos casos la consecuencia negativa de otro modelo de verdad, de otra instancia metodológica. Esta otra instancia metodológica, por otra parte, es la que no sólo ha hecho triunfar el sistema copernicano, sino que al liberarlo del obstáculo tradicional ha hecho posible al mismo tiempo la configuración de la ciencia moderna[55].

[52] Cf. J. Marrades Millet, «Descartes, Newton y Hegel sobre el método de análisis y síntesis», *Pensamiento*, 164, 1985, pp. 393-404, sobre todo; E. M. Madden (ed.), *Theories of scientific Method*, University of Washington Press, Seattle, 1960; M. Malherbe y J. M. Pousseur (eds.), *F. Bacon. Science et Méthode*, J. Vrin, Paris, 1985.

[53] F. Bacon, *La gran restauración*, ed. de M. A. Granada, Alianza, Madrid, 1985, pp. 97-168.

[54] *Dialogo sopra i due massimi sistemi del mondo, Opere*, Ed. Nazionale, Firenze, vols. VII y VIII. Cf. L. Geymonat (1957), *Galileo Galilei*, Península, Barcelona, 1969, pp. 142-153.

[55] I. Lakatos (1978), *La metodología de los programas de investigación científica*, Alianza, Madrid, 1983, cap. 4 sobre todo.

Ahora bien, si convenimos en que lo «matemático» es la instancia fundamentante del nuevo método, la estimación de las posiciones teóricas de los pensadores señalados —Bacon, Galileo, Descartes, Newton— ya no es la misma.

El método pensado por Bacon podrá haber influido en la tradición filosófica denominada empirismo, podrá ser valioso aún en el ámbito de lo que se llamó historia natural. Pero Bacon, aun siendo antiescolástico, no ha sabido salirse del marco del pensamiento de Aristóteles, ni menos aún incorporar en el suyo el proyecto matemático moderno. Tal es el límite que se señala a su método desde Hume hasta nuestros días. Hume, por su doble condición de empirista e inglés, tenía todas las razones para estar prevenido en favor de Bacon; no obstante, escribe: «Es muy inferior a Galileo, su contemporáneo, y quizá también a Kepler. Bacon mostró de lejos la ruta de la verdadera filosofía; Galileo no sólo la mostró, sino que él mismo marchó por ella a grandes pasos. El inglés no tuvo ningún conocimiento de la geometría; el florentino resucitó esta ciencia sobresalió en ella y pasa por ser el primero que la aplicó, con los experimentos, a la física»[56].

Al señalar el punto débil del método de Bacon, Hume subraya dos notas de la innovación de Galileo: fue el primero que aplicó la *geometría* a la física por medio de los *experimentos*. Pero sólo desde la perspectiva heideggeriana, antes expuesta, según la cual la ciencia es experimental sobre la base del proyecto matemático, es posible comprender el debate suscitado sobre cuál de los dos elementos —el matemático o el experimental—es determinante en la constitución del método *practicado* por Galileo[57], así como los límites de dicho debate. Por

[56] Cit. por R. Blanché (1969), *El método experimental y la filosofía de la física*, FCE, México, 1972, p. 63. Cf. M. Malherbe y J. M. Pousseur (eds.), *op. cit.*, p. 115.

[57] Galileo, más que hacer una teoría o discurso del método, lo practica. Una pretendida teoría del método en Galileo exige tener en

una parte, A. Koyré y E. Cassirer defienden la preeminencia del elemento racional o matemático, relegando lo empírico a un papel secundario. Por otra, Della Volpe llama la atención sobre la función decisoria que Galileo atribuye a la verificación experimental. L. Geymonat, a su vez, argumenta en favor de una posición de equilibrio entre los dos elementos que, sin duda, están presentes en la práctica metodológica de Galileo entendida como instancia fundamental de la ciencia que desarrolla.

Según Koyré, lo que se propone Galileo no es tanto conocer el curso de los hechos cuanto las *esencias* que los fundamentan; por lo tanto, carecería de sentido querer partir de los hechos para llegar a ellas, pues corresponde, en cambio, a la matemática, y solamente a ella, captarlas directamente[58]. Considera Cassirer que Galileo, aun partiendo de la experiencia y terminando en ella, se propone en su método ante todo determinar los datos de la experiencia en *relaciones generales* de carácter no ya empírico sino conceptual[59]. Contra la interpretación de Cassirer sobre todo polemiza Della Volpe, subrayando la función decisoria que en el método de Galileo tiene la verificación experimental en tanto que instancia probatoria o desaprobatoria de determinados hechos de la experiencia[60]. L. Geymonat apoya a Della Volpe aun sin coincidir con él exactamente: por una parte, reivindica la función de la verificación experimental y su decisorio valor de *prueba* de la verdad de la hipótesis; por otra, afirma la relevancia de la función *lógico-instrumental* de la matemática, función lógica en tanto que expresión del rigor del razonamiento deductivo, función instrumental en tanto que posibilidad de medir cuantitativamente las

cuenta no sólo el *Dialogo*, sino todas sus obras, sobre todo *Discorsi*. Cf. R. Blanché, *op. cit.*, pp. 77-90; R. E. Butts y J. C. Pitt (eds.), *New perspectivas on Galileo*, Reidel, Dordrecht, 1978.

[58] *Estudios galileanos*, pp. 146-147.

[59] *El problema del conocimiento*, I, *cit.*, pp. 344 ss.

[60] *Logica come scienza positiva*, Messina, 1956, p. 226.

relaciones entre los hechos, entre las cosas. Para Geymonat, sin embargo, el problema del método en Galileo no es un problema fácil de resolver, sino complejo. De ahí las incertidumbres y las diferentes interpretaciones. Pero «lo singular es, sin embargo, que a pesar de estas incertidumbres, Galileo consiguió aportar al desarrollo de la conciencia metodológica de la ciencia una enorme contribución, cosa que se ven obligados a reconocer unánimemente todos los historiadores, sea cual sea su orientación [...]. Y consiguió dibujar —aunque con cierta inseguridad— el camino a través del cual se desarrollaría a continuación la ciencia moderna», dejando en sus escritos «una cosecha inagotable de observaciones metodológicas de la mayor actualidad»[61].

El límite de este debate es olvidar o ignorar que sólo *sobre la base de lo matemático* la experiencia se transforma en experimento en sentido moderno, que sólo sobre la base de lo matemático el proceso fundamental de investigación determina lo que se muestra, la experiencia, y al mismo tiempo el método o modo de proceder[62].

Descartes es, sin duda, quien mejor ha captado lo matemático como rasgo fundamental del pensar de su época. Cuando juzga la obra de Galileo, *Discorsi*, escribe: «Encuentro en general que filosofa mucho mejor que el vulgo en que abandona cuanto le es posible los errores de la Escuela, y trata de *examinar las materias físicas por razones matemáticas*. En esto estoy enteramente de acuerdo con él y considero que no hay otro medio de encontrar la verdad»[63]. Como se sabe, desarrolla esta posición teórica tanto en las *Regulae* como en el *Discours:* la primera, redactada enteramente antes de la lectura de Galileo; la segunda, en su mayor parte, como se ha indicado al final del punto 1.

[61] *Galileo Galilei, cit.*, pp. 212-213, también: pp. 201-212.

[62] M. Heidegger, «La época de la imagen del mundo», en *Sendas perdidas, cit.*, p. 70.

[63] A Mersenne, 11 de octubre de 1638. La cursiva es mía. En esta misma carta rechaza que deba algo a Galileo: «En cuanto a Galileo, le diré que nunca le he visto ni he mantenido ninguna comunicación con él, y que por lo tanto no podría haber tomado de él ninguna cosa». La formación matemática de Descartes proviene de la lectura de Clavius y, sobre todo, de su amistad con Beeckman desde 1618 (G. Rodis-Lewis, *L'Oeuvre de Descartes*, pp. 21 y 25-28). Beeckman, matemático, atomista, resueltamente copernicano, no sólo pone al corriente a Descartes acerca de la física de su tiempo, sino que, estando convaleciente éste, le lleva un libro de Galileo en agosto de 1634 (a Mersenne, 14 de agosto de 1634). Tal libro puede ser el *Trattato di maccaniche*, trad. por Mersenne ese mismo año o, tal vez, el *Dialogo* (1632), más difícil de localizar porque dio lugar a la condena de su autor; pero la trad. latina de esta obra se publica en Holanda en 1635. Para un estudio de sus respectivas posiciones, véase: F. Enriques, «Descartes et Galilée», *Rv. Mét. et de Mor.*, 1937, pp. 221-235; W. R. Shea, «Descartes as a critic of Galileo», en *New perspectivas on Galileo*, Reidel, Dordrecht, 1978, pp. 139-159.

Como observa Heidegger, en las *Regulae ad directionem ingenii* encontramos los principios de «una fundamentación de lo matemático para que se convierta en su totalidad en una norma para el espíritu investigador» y, en general, una «norma de *todo* pensar»[64]. Tal norma está desarrollada en Descartes ya como proyecto de una *mathesis universalis*, ya como discurso sobre el *método*, ya como problema que plantea otro modelo de *verdad*. No se trata de tres direcciones paralelas de su pensamiento. Las implicaciones recíprocas son constantes. Y, en cualquier caso, en la reflexión sobre éstos y otros problemas se fundamenta y acuña el moderno concepto de *ciencia*.

La *mathesis universalis* no es el método, ni siquiera es una aplicación particular del método. Es, a lo más, una propedéutica del método en el sentido de que el proyecto de la *mathosis* constituye el horizonte teórico «arcaico» o principal en el que emerge el discurso sobre el mé-

[64] *La pregunta por la cosa*, p. 92. Cf. nota 25. E. Denissoff, *Descartes, premier théoricien de la physique mathématique*, Nauwe laerts, Louvain, 1970.

todo[65]. La posibilidad de la *mathesis* está pensada desde el supuesto de la unidad de la ciencia (*Regla I*) y desde la concepción de que lo matemático determina la realidad como mensurable. Desde este doble supuesto, la *scientia universalis* constituye la «investigación general» desde la cual se fundamentan las ciencias particulares tales como la aritmética, la geometría, la astronomía, la música, la óptica, la mecánica y otras muchas[66]. La define Descartes cómo la ciencia del *orden* y de la *medida:* «Debe haber una cierta ciencia general que explique todo lo que puede buscarse acerca del orden y la medida no adscrito a una materia especial, y que es llamada, no con un nombre adoptado, sino ya antiguo y recibido por el uso, *Mathesis Universalis,* ya que en ésta se contiene todo aquello por lo que las otras ciencias son llamadas partes de la Matemática»[67]. Cuando Descartes escribe que *el método es necesario para la investigación de la verdad de las cosas* (título de la Regla IV), está pensando que el proceso de investigación abierto desde la *mathesis* en los diferentes dominios del ente no puede tener lugar al azar, sino siguiendo al mismo tiempo un determinado modo de proceder y un determinado modelo de verdad. Aunque el proyecto de una *mathesis* es abandonado al destino del manuscrito de las *Regulae,* la reflexión sobre la fundamentación del nuevo saber sigue vinculada estrechamente a aquella doble propuesta o exigencia en el *Discurso del método*[68].

[65] L. P. Weber, *La constitution du texte des «Regulae»*, SEDES, Paris, 1964, pp. 7-11. Mayor interrelación defiende J. L. Marion (*Sur l'ontologie grise de Descartes*, Vrin, París, 1975, pp. 55-59). Cf. L. J. Beck, *The method of Descartes: A study of the Regulae*, Clarendon Press, Oxford, 1952.

[66] *Reglas para la dirección del espíritu*, Alianza, Madrid, 1984, p. 86 (AT, X, p. 377).

[67] *Ibid.*, p. 86 (AT, X, p. 378). Diez años más tarde llamará «Mathematica pura» a una ciencia que engloba geometría, aritmética, mecánica, etc. (a Ciermans, 23 de abril de 1638). H. Scholz, *Mathesis Universalis*, Basel/Stuttgart, 1969; J. Vuillemin, *Mathématiques et métaphysique chez Descartes*, PUF, Paris, 1960, J. L. Coolidge, *A history of geometrical Method*, Clarendon Press, Oxford, 1940.

Si el rasgo específico de la ciencia moderna es la fundamentación del proceso de investigación en la reflexión sobre lo matemático que, a su vez, no sólo determina el objeto como lo mensurable, sino también el experimento y el modo de aproximación a las cosas (μέθοδος), no es necesario abrir el debate acerca del elemento predominante en el método cartesiano. Como en el caso de Galileo, Cassirer y Koyré estiman que es el factor racional-matemático el que predomina, al incorporar Descartes la tradición platónica que le llega a través de Clavius y se observa sobre todo en la física[69]. Por su parte, Blanché y Clarke, aun conscientes del papel que desempeña dicho factor en la filosofía de la ciencia cartesiana y, concretamente, en el discurso sobre el método, ponen de relieve la importancia que tiene la experiencia y el experimento en dicho discurso.

Descartes, sostiene Planché, «no dejó de interesarse en las observaciones y en los experimentos, ni de practicar él mismo el razonamiento experimental»[70]. El mérito de esta tesis consiste en haber mostrado un dato teórica y operativamente cierto, que la lectura racionalista ha marginado. Pero de ahí no creo que sea posible concluir, como hace Clarke, que Descartes sea un aristotélico innovador[71]. Tal conclusión, válida para F. Bacon como

[68] Parte I, p. 5 (AT, VI, p. 3); parte II, p. 22 (AT, VI, p. 17). *La recherche de la vérité par la lumière naturelle*, AT, X, pp. 495 ss.

[69] E. Cassirer, *El problema del conocimiento, I., cit.*, pp. 459-461, A. Koyré, *Estudios galileanos*, pp. 277-278.

[70] R. Blanché, *El método experimental y la filosofía de la física*, p. 108. Cf. G. Milhaud, «Descartes expérimentateur», en *Descartes savant*, J. Vrin, París, 1921; D. M. Clarke, *La filosofía de la ciencia de Descartes, cit.*, cap. 2. Véase la Parte V del *Discurso*, en la que describe la circulación de la sangre, y la Parte VI, en la que propone condiciones de la experiencia y de posibles experimentos (pp. 86-87; AT, VI, pp. 63-64).

[71] *Op. cit.*, cap. 8.

observa Blanché[72], no es defendible en el caso de Descartes, pues la reflexión sobre lo matemático, ausente en el autor del *Novum organum* es determinante en las *Regulae* y en el *Discurso*, como se ha mostrado. Es más: al argumentar Clarke que la exigencia de certeza, aspecto dominante de la ciencia cartesiana, está determinada por «el ideal aristotélico de demostración y de certeza»[73], capta perfectamente el rasgo fundamental de la ciencia moderna que otros han señalado, pero incurre en el error de atribuir la exigencia de certeza al ideal aristotélico, y no al *lenguaje* (Galileo) o *modelo* (Descartes) *matemático*, que predomina y configura la ciencia moderna y al mismo tiempo determina el modo de proceder de ella. Platón, Euclides, Arquímedes —y no Aristóteles— son los pensadores antiguos que inspiran los trabajos científicos de Tartaglia, Torricelli, Clavius, Kepler, Galileo, el círculo de Mersenne y Descartes, entre otros.

El ideal y la exigencia de certeza, por otra parte, no sólo está vinculado en Descartes al modelo matemático, sino que, además, es este modelo el que determina el nuevo criterio de verdad formulado en la primera regla del método, a saber, la *evidencia*. Tal vinculación aparece explícitamente enunciada ya en las *Regulae*: «Aquellos que buscan el recto camino de la verdad no deben ocuparse de ningún objeto del que no puedan tener una certeza igual a la de las demostraciones aritméticas y geométricas»[74]. La razón de ello la establece Descartes inmediatamente después como título de la Regla III: «[...] pues la ciencia no se adquiere de otra manera», sino a través de «lo que podamos intuir clara y evidentemente o deducir con certeza».

Ahora bien, lo más importante del nuevo criterio de verdad no es que constituya la piedra angular del nuevo

[72] *Op. cit.*, p. 62.
[73] *Op. cit.*, p. 207.
[74] *Reglas...*, p. 72 (AT, X, p. 366).

método de investigación, establecida ya por Descartes en 1619. Lo más importante es el papel «revolucionario» de la evidencia[75], en tanto que problematiza el modelo tradicional de verdad definido en términos de *adaequatio*, establece críticamente otro modelo, y abre la discusión que dominará la filosofía moderna incluso más allá de Kant. Más aún: el tratamiento crítico que Descartes da al problema de la verdad, en su discurso sobre el método y sobre la fundamentación del saber moderno, es uno de los méritos que le hacen acreedores al título de «primer pensador moderno».

Desde esta perspectiva, la pregunta por las reglas del método, por las operaciones de la mente (intuición, deducción) o por el papel del análisis y de la síntesis[76], o por problemas tales como el significado de la duda, la posible circularidad en la formulación del primer principio, etc., sólo tienen un interés particular, cuyo sentido último radica en la comprensión de la reflexión sobre lo matemático, de la tarea fundamentadora de la ciencia y del discurso sobre el método como modo de investigar la verdad de las cosas. A esta perspectiva nos conduce la observación de Heidegger, al comentar el último enunciado formulado como título de la Regla IV: «Esta regla no expresa el lugar común de que una ciencia deba tener también su método, sino que quiere decir que el procedimiento, esto es, el modo como estamos en general tras las cosas (μέθοδος), decide de antemano sobre lo que encontramos de verdadero en las cosas. El método no es

[75] J. L. Marion, *op. cit.*, p. 245. Cf. Husserl, *Méditations cartésiennes*, pp. 6-10 (trad. esp. en Tecnos).
[76] J. L. Marion, tras hacer la historia del método cartesiano, ofrece un esquema de las reglas metodológicas (*op. cit.*, cap. XIII). J. Marrades lleva a cabo un estudio clarificador de la función del análisis y de la síntesis en el método de Descartes desde la posición de Newton (*art. cit.*, pp. 393-404). Cf. J. Hintikka y U. Remes, *The method of analysis*, Reidel, Dordrecht, 1974; J. Vuillemin, «Trois philosophes intuitionnistes: Epicure, Descartes et Kant», *Dialectica*, 35, 1981, pp. 21-41.

una indumentaria de la ciencia entre otras, sino la instancia fundamental a partir de la cual se determina lo que puede llegar a ser objeto y cómo llegar a serlo [...]. Lo decisivo es la manera y el modo en que esta reflexión sobre lo matemático influenció la controversia con la metafísica tradicional (*prima philosophia*), y cómo a partir de esto se determinó el destino futuro y la figura de la filosofía moderna»[77].

3. DESCARTES, ¿PRIMER PENSADOR MODERNO?

Si la reflexión sobre lo matemático ha sido determinante no sólo en la controversia con la filosofía tradicional, sino también en el proceso de fundamentación de la ciencia y del pensamiento modernos, y si Descartes ha desempeñado, como hemos mostrado, un papel decisivo tanto en la controversia como en el proyecto de fundamentación, es lógico que los siglos posteriores —D'Alembert, Hegel, Husserl, Heidegger, Sartre, por ejemplo—, le hayan reconocido el título de «primer pensador moderno»[78].

Pero ¿no son igualmente filósofos modernos los Bruno, los Telesio, los Campanella? ¿No se argumenta, por otra parte, en favor de la tesis: Galileo, filósofo? Los primeros especulan, es cierto, sobre la naturaleza o el universo, pero no lo hacen desde la reflexión sobre lo

[77] *La pregunta por la cosa*, p. 93.

[78] D'Alembert, *Discurso preliminar de la Enciclopedia*, Aguilar, Buenos Aires, 1974, pp. 101-103; Hegel, «Vorlesungen über die Geschichte der Philosophie», en *Werke*, Verlag, Frankfurt, 1971, vol. XX, pp. 120-123 (trad. esp. en FCE, México, 1955, vol. III, pp. 252-254); E. Husserl, *Méditations cartésiennes*, pp. 1-3 (trad. esp. en Tecnos); M. Heidegger, *Sendas perdidas*, p. 78; J.-P. Sartre, «Questions de Méthode», en *Critique de la raison dialectique*, Gallimard, Paris, 1960, p. 17 (trad. esp. en Losada).

matemático, que constituye el rasgo fundamental del pensamiento moderno. El argumento de la tesis según la cual «la filosofía de Galileo, dispersa a través de toda su obra, se encuentra profundamente articulada, aunque no sistematizada, en torno a la idea clave de la matematización de lo real»[79], si bien tiene a su favor el elemento ausente en el naturalismo renacentista, olvida que en Galileo no están suficientemente tratados —algunos ni siquiera mencionados— los *principios metacientíficos* que constituyen los fundamentos no sólo de la ciencia, sino también de la filosofía moderna. Es más: si afirmamos que Descartes es el primer pensador moderno, no es sobre la base de su aportación científica —en lo que tal vez no es comparable a Galileo y haya sido superado por Newton— aunque la aplicación que hizo del álgebra a la geometría ha sido lo que ha inmortalizado su nombre, según D'Alembert[80]; tampoco debe ese título solamente al hecho de haber sido el filósofo que mejor ha tematizado los métodos, contenidos e ideales de la nueva ciencia[81]. Si lo decisivo es la manera y el modo en que la reflexión sobre lo matemático quebró la metafísica tradicional y, al mismo tiempo, alzó sobre cimientos nuevos la figura de la filosofía moderna, Descartes sólo es el primer pensador moderno en la medida en que tales cimientos hayan sido proyecto y realización exclusivamente suyas, es decir, en la medida en que lleven el sello de un modo de reflexión que ha determinado la figura y el destino del pensamiento moderno.

Ahora bien, ¿cuál es el modo peculiar de la reflexión cartesiana, que llega a constituirse en pensamiento de una época? No es otro sino el modo como establece los

[79] J. J. Ferrero Blanco, *Galileo Galilei, el filósofo*, Universidad de Deusto, Bilbao, 1986, p. 318.

[80] *Op. cit.*, p. 101. Cf. P. J. Davis y R. Hersh, *Descartes' Dream. The world according to mathematics*, the Harvester Press, Hassocks, 1986.

[81] S. Turró, *op. cit.*, p. 374.

principios axiomáticos sobre los cuales se fundamenta todo lo demás como consecuencia, esto es, las disciplinas científicas, pero también las no científicas. Cuando Descartes advierte que «todos los principios de las ciencias debían tomarse de la filosofía, en donde no hallaba ninguno cierto»[82], se propone ante todo descubrir y formular tales principios llamados a ser al mismo tiempo los cimientos de la ciencia y del pensar modernos. Como; observa Heidegger, la reflexión sobre los primeros principios del ser y del saber, que en Descartes tiene su máxima expresión en *Meditationes de prima philosophia*[83], no sólo significa volver a plantear desde su raíz el problema que dio lugar a la πρώτη φιλοσοφία sino diseñar con nueva figura la metafísica que desde Descartes llega hasta Nietzsche[84]. El innovador alcance de este proyecto es tal que Hegel no duda en afirmar que «René Descartes es un héroe del pensamiento moderno», porque «reconstruye la filosofía sobre los cimientos puestos ahora de nuevo al descubierto al cabo de mil años»[85].

Pero sería un error creer que el descubrimiento de tales cimientos se limita a mostrar una reiteración de lo ya dicho. Cuando el autor del *Discurso del método* publica el esbozo de sus *Meditaciones* nos invita a juzgar —y, en consecuencia, a comparar— sobre la solidez de tales cimientos o principios del nuevo pensar en los términos siguientes: «Sin embargo, con el fin de que se pueda apreciar si los fundamentos que he establecido son bastante firmes, me veo en cierto modo obligado a hablar de ellas»[86].

[82] *Discurso del método*, Parte II, p. 29 (AT, VI, 21-22). Cf. S. Gankroger (éd.), *Descartes: Philosophy, Mathematics and Physic*, Harvester Press, Hassocks, 1980.

[83] Cf. nota 28.

[84] M. Heidegger, «La época de la imagen del mundo», en *Sendas perdidas*, p. 78.

[85] *Op. cit.*, p. 123; trad. esp., p. 254.

[86] Parte IV, p. 44 (AT, VI, p. 31).

Y, en efecto, se puede pensar que la *razón*, como principio del saber moderno, no es sino el eco del λόγος griego, que Aristóteles establece como hilo conductor de las categorías. Se puede pensar que, cuando Descartes especifica lo humano por la razón, considerada ésta ya como facultad de juzgar[87], ya como habla, discurso o lenguaje[88], resuena en su propuesta el enunciado del Estagirita según el cual el hombre es «un ser vivo que tiene *logos*». Pero, al pensar de este modo, se olvida que la preeminencia del principio racional cartesiano no radica en sí mismo, sino en otro principio más fundamental, a saber, el *yo pienso*. Dado que «el pensar es el acto fundamental de la razón; ésta, la razón, es puesta ahora, con el *«cogito sum» expresamente* y de acuerdo a su propia exigencia, como *primer fundamento de todo saber* y como hilo conductor de todas las determinaciones de las cosas en general»[89].

La originalidad de Descartes al descubrir y formular este axioma no se menoscaba buscándole un precedente en San Agustín, ya que éste nunca elevó a primer principio de un sistema de pensamiento su enunciado: *si enim fallor, sum*. Como nos recuerda Hegel, «la Edad Media no tenía como principio el pensamiento libre, que parte de sí mismo»: sólo éste es la premisa o fundamento (*Grundlage*) de «una filosofía propia e independiente, que sabe que procede sustantivamente de la razón, y que la conciencia de sí es un momento esencial de la verdad»[90], el momento en el que la verdad se identifica con

[87] Parte I, p. 4 (AT, VI, p. 2). Cf. A. Álvarez Gómez, «Descartes: la razón, única guía del hombre», *Cuad. Salm. Filos.*, 12, 1985, pp. 19-43; P. A. Schouls, «Descartes and the autonomy of reason», *Journal of the History of Philosophy*, X: 3, 1972, pp. 307- 323.

[88] Parte V, pp. 79-80 (AT, VI, pp. 57-58). Cf. N. Chomsky, *Lingüística cartesiana*, Gredos, Madrid, 1978, pp. 15-26; E. Lledó, *Filosofía y lenguaje*, Ariel, Barcelona, 1970-74, pp. 173-207.

[89] M. Heidegger, *La pregunta por la cosa*, p. 96. El segundo subrayado es mío.

[90] *Op. cit.*, pp. 120-121; trad. esp., pp. 252-254. Cf. H. G. Frank-

la certeza. Es esta autocerteza del yo lo que constituye el fundamento axiomático desde el cual se entiende la revolución cartesiana y el modo de pensar moderno[91].

Más aún: la originalidad de Descartes aparece patente en el doble significado de este fundamento axiomático: *a*) si *epistemológicamente* constituye el primer principio de un sistema deductivo, ello supone el desplazamiento de otros principios —revelación, autoridad, tradición— del lugar preeminente del que gozaban en la época medieval, y la afirmación del pensamiento libre, esto es, de la razón libre de la teología[92]; supone, en fin, una redistribución de la *episteme*, como diría Foucault; *b*) desde una perspectiva *metafísica*, el fundamento axiomático significa que, con Descartes, no es Dios sino el hombre el que se constituye en sujeto, es decir, en aquel existente en el cual se funda todo lo existente a la manera de su ser y de su verdad.

Si, como señala Heidegger, «la metafísica funda una época al darle un fundamento de su figura esencial mediante una determinada interpretación de lo existente y mediante una determinada concepción de la verdad»[93], la figura esencial de la época moderna se alza sobre el cimiento axiomático del yo, del hombre transformado en sujeto (*subjectum* es la traducción de ὑποκείμενον), esto es, en el punto de referencia de la existencia como tal y de su verdad. Entender este yo elevado a *subjectum* como algo «subjetivo» significa olvidar la dimensión ontológica de la proposición: yo *pienso, luego existo*. Desde este yo elevado por Descartes a sujeto preeminente las cosas mismas se convierten en «objetos»[94].

furt, *Demons, dreamers and madmen: The defense of reason in Descartes' Méditations*, Bobbs, Merril, Indianapolis, 1970.

[91] E. Husserl, *op. cit.*, p. 6. Cf. E. A. Burtt, *The metaphysical foundations of modern science*, Doubleday, New York, 1954; trad. esp. Ed. Sudamericana.

[92] Hegel, *op. cit.*, 120; trad. esp., p. 252. La ciencia va a ser la gran beneficiada de esta liberación. La relación filosofía/religión en cambio, va a ser objeto de un incesante debate, que tiene sus momentos fuertes en la *Crítica de la razón pura* y en *La Religión dentro de los límites de la mera Razón* de Kant.

[93] «La época de la imagen del mundo», *cit.*, p. 68.

Desde el «yo pienso» como fundamento axiomático se configura, pues, otra imagen del mundo, la imagen moderna del mundo definida por el objetivar, es decir, por el *re-presentar*. El mundo moderno ya no es ni el mundo griego, percibido, ni el mundo del pensador medieval, creado. El mundo moderno, construido desde la reflexión sobre lo matemático y el mecanicismo atomista, es un mundo representado. De ahí que Descartes asuma la tesis de Galileo según la cual el gran libro del universo está escrito en lenguaje matemático. Pero con tal de dejar bien claro que dicho lenguaje ha de ser remitido, en último término, al axiomático principio normativo de todo saber: el yo *pienso*. De este modo, el hecho de que el mundo pase a ser imagen o representación «es exactamente el mismo proceso con el que el hombre pasa a ser *subjectum* dentro de lo existente»[95].

Pues bien, los principios —el «yo pienso» y el principio de no contradicción implicado en él—, que sólo surgen de la razón de acuerdo al rasgo fundamental matemático del pensar, se convierten en los principios del saber auténtico, es decir, de la filosofía en sentido estricto, de la metafísica en sentido moderno. Una metafísica que sólo es la raíz del árbol de la *filosofía*, cuyo tronco es la física y cuyas ramas son la mecánica, la medicina, la moral y las demás ciencias[96].

Con todo, el nacimiento de la filosofía moderna desde su raíz no se entiende sino como crítica radical —como

[94] *La pregunta por la cosa*, pp. 95-96.

[95] M. Heidegger, «La época de la imagen del mundo», *cit.*, p. 82. Cf. M. Foucault, *Les mots et les choses*, Gallimard, Paris, 1966, cap. III (trad. esp. en Siglo XXI).

[96] «Lettre-Préface» a la ed. francesa de *Les principes de la philosophie*, AT, IX- ii, p. 14. Cf. S. Gaukroger (ed.), *Descartes: Philosophy, Matemathics and Physic*, Harvester Press, Hassocks, 1980.

«duda» de todo— de la tradición y de sus propios fundamentos. Así, al menos, lo ha comprendido un ilustrado célebre y al mismo tiempo pensador matemático, D'Alembert: «Descartes se abrevió al menos a enseñar a las buenas cabezas a sacudirse el yugo de la escolástica, de la opinión, de la autoridad, en una palabra, de los prejuicios y de la barbarie y, con esta rebelión cuyos frutos recogemos hoy, ha hecho a la filosofía un servicio más esencial quizá que todos los que ésta debe a los ilustres sucesores de Descartes [...]. Si acabó por creer explicarlo todo, al menos comenzó por dudar de todo; y las armas de que nos servimos para combatirlo no dejan de pertenecerle porque las volvamos contra él»[97].

NOTA SOBRE LA PRESENTE EDICION

La traducción se ha hecho a partir de la edición de Charles Adam y Paul Tannery, *Oeuvres de Descartes*, Léopold Cerf, París, 1902, vol. VI. Este volumen contiene las dos ediciones del *Discours de la méthode* hechas por Descartes: la edición francesa (Leyde, 1637) y la latina (Amsterdam, 1644), cuya versión de E. de Courcelles fue revisada y corregida por el propio autor, como consta en las páginas 517 y 539. En la edición latina no se traduce el tercer ensayo, *Géometrie*. En numerosos casos hemos comparado ambas ediciones, indicando en nota los términos y expresiones coincidentes o aquellos en los que se aprecia un giro o matiz que permite precisar mejor el sentido en nuestro idioma. Se puede observar, en nuestra traducción, una constante fidelidad al texto —cuyo sentido se pretende aclarar mediante el aparato crítico—, sin renunciar por ello a la exigencia de estilo de que hace gala Descartes. Una particularidad de la

[97] D'Alembert, *Discurso preliminar de la Enciclopedia*, *cit.*, p. 103.

presente edición consiste en facilitar al lector la referencia a la original edición francesa, indicando al margen las páginas de ésta (1-78), y señalando mediante una barra / el lugar exacto donde empiezan y acaban dichas páginas. Como herramientas valiosas hemos de mencionar, además, las ediciones de E. Gilson y de F. Alquié, así como las traducciones de M. García Morente y de G. Quintás Alonso. Las referencias a las obras de Descartes remiten a la edición de Ch. Adam y P. Tannery que se citan mediante las siglas AT, seguidas del volumen y de la página correspondiente.

BIBLIOGRAFÍA

I. EDICIONES PRINCIPALES

Oeuvres de Descartes, par Ch. Adam et P. Tannery, Léopold Cerf Imprimeur-Editeur, París 1897-1913, 13 vols. Nouvelle présentation mise à jour par B. Rochot, CNRS-J. Vrin, 1964-1974. Reimpresión en Vrin, Paris, 1996, sólo 11 vols.

Oeuvres philosophiques de Descartes, textes établis, présentés et annotés par Ferdinand Alquié, Garnier, Paris, 1963 1973, 3 vols.

Discours de la méthode pour bien conduire sa raison, et chercher la vérítée dans les sciences. Plus la Dioptrique, les Météores et la Géométrie qui sont des essais de cette méthode, Jean Marie, Leyde, 1637.

Specimina Philosophiae: seu Dissertatio de Methodo recte regendae rationis, et veritatis in scientiis investigandae: Dioptrice, et Meteora, trad. de E. de Courcelles, L. Elzevier, Amsterdam, 1644.

Discours de la méthode pour bien conduire sa raison et chercher la vérité dans les sciences, par R. D. Nouvelle édition, augmentée des remarques du P. Poisson, Paris, 1724.

Discours de la méthode de Descartes. *Novum Organum* de Bacon. *Théodicée* de Leibnitz, fragments. Publiés en un seul volume, avec des notes, par A. Lorquet, Paris, 1840.

Discours de la méthode, texte et commentaire, R. D., par E. Gilson, Paris, 1925 (4e. éd., J. Vrin, 1966).

Discours de la méthode, R. D., *Oeuvres et lettres*, par A. Bridoux, Gallimard (Pléiade), Paris, 1937.

Discours de la méthode. Abhandlung über die Methode. Mit einem Vorwort von K. Jaspers und einem Beitrag über Descartes und die Freiheit von J. P. Sartre, Mainz, 1948.

Discours de la méthode. Précédé de Descartes, par G. Rodis-Lewis. Le discours de la méthode par J. Nabert. Et suivi de Descartes et son temps par E. Souriau, Paris, 1953.

Discours de la méthode, R. D., par D. Huisman, Nathan, Paris, 1981.

Entretien avec Burman, manuscrit de Göttingen, texte présenté, traduit et annoté par Ch. Adam, Boivin, Paris, 1937.

II. TRADUCCIONES AL CASTELLANO

Nos limitamos a mencionar las principales traducciones al español del *Discurso del método*, las de: J. C. García Borrón, Bruguera, Barcelona, 1968; R. Frondizi, Revista de Occidente, Madrid, 1974 —reed. en Alianza—; A. Rodríguez Huéscar, Aguilar, Buenos Aires, 1980; F. Alonso, Akal, Madrid, 1982; A. Gual Mir, Edaf, Madrid, 1982; H. Arnau Grass y J. M. Gutiérrez González, Alhambra, Madrid, 1983; E. Frutos, Planeta, Barcelona, 1984.

A continuación señalamos otras *fuentes* indispensables para conocer la posición teórica del autor del D.M. La cronología de las obras, indicada a la izquierda, expresa en principio la fecha final de su redacción, que coincide o no con la de su publicación.

1618: *Compendio de música*, introducción de A. Gabilondo, trad. de P. Flores y C. Gallardo, Tecnos, Madrid, 1992.
1628: *Reglas para la dirección del espíritu*, ed. J. M. Navarro Cordón, Alianza, Madrid, 1984.
1633: *El Mundo. Tratado de la luz*, ed. bilingüe de S. Turró, Anthropos, Barcelona, 1989.
1633: *Tratado del hombre*, ed. de G. Quintás Alonso, Editora Nacional, Madrid, 1980.
1633: *Du Foetus*, ed. bilingüe de P. Pardos, C. Vicén y A. Alonso, prólogo de J. L. Rodríguez, Prensas Universitarias de Zaragoza, Zaragoza, 1987.
1637: *Discurso del método. Dióptrica, Meteoros y Geometría*, prólogo, trad. y notas de G. Quintás Alonso, Alfaguara, Madrid, 1981.
1641-1647: *Meditaciones metafísicas, con Objeciones y Respuestas*, ed. de Vidal Peña, Alfaguara, Madrid, 1977.
1644: *Los principios de la filosofía*, ed. de G. Quintás Alonso, Ed. Reus, Madrid, 1995..
1643-1650: *Correspondencia con Isabel de Bohemia y otras cartas*, introducción de M. Cabot, traducción de M.ª T. Gallego, Alba, Barcelona, 1999.
1647: *Observaciones sobre el programa de Regius*, ed. de G. Quintás Alonso, Aguilar, Buenos Aires, 1980.
1649: *Las pasiones del alma*, ed. de J. A. Martínez Martínez y P. Andrade Boué, Tecnos, Madrid, 1997.

III. ESTUDIOS SOBRE *EL DISCURSO DEL MÉTODO*

ÁLVAREZ GÓMEZ, A.: «Para leer el Discurso del Método», en *Estudios sobre filosofía moderna y contemporánea*, Universidad: Centro de Estudios Metodológicos Interdisciplinares, León,

BECK, J. L.: *The Method of Descartes: A Study of the Regulae*, Clarendon Press, Oxford, 1952.

BELLO, E.: «La función de la experiencia en el discurso del método», en E. Ranch Sales y F.-M. Pérez Herranz (eds.), *Centenario de René Descartes (1596-1996)*, Universidad de Alicante, Alicante, 1997, pp. 109-123.
— «Descartes, lo matemático y la constitución del saber moderno», en M. Álvarez Gómez *et al.*, *La filosofía de Descartes y la fundación del pensamiento moderno*, SCLF, Salamanca, 1997, pp. 253-278.
CAHNÉ, P. A.: *Index du Discours de la méthode de R. Descartes* Edizioni dell'Ateneo, Roma, 1977.
CHAMIZO DOMÍNGUEZ, P. J.: «El *Discurso del Método* de Descartes como ensayo», *Aporía*, 4: 15-16, 1982, pp. 69-83.
CHAUVOIS, D.: *Descartes. La méthode et ses erreures en physiologie*, Ed. du Cèdre, Paris, 1966.
CHEVALIER, J.: «Le Discours de la Méthode», en *Archives de Philosophie: Autour du «Discours de la Méthode»* (1637-1937) XIII, 1937, pp. 1-13.
CLARKE, D. M.: «Ensayos metodológicos», en *La filosofía de la ciencia de Descartes*, Alianza, Madrid, 1986, pp. 174-204.
COOLIDGE, J. L.: *A History of Geometrical Method*, Clarendon Press, Oxford, 1940.
COSTABEL, P.: «Les *Regulae* et l'actualité scientifique de leur temps», *Les Études philosophiques*, 1976, pp. 405-415.
CROMBIE, A. C.: «Descartes on method and physiology», *Cambridge Journal*, 5, 1951, pp. 178-186.
DERRIDA, J.: «La langue et le discours de la méthode», *Recherches sur la philosophie et le langage*, 3, 1983, pp. 35-51, trad. esp. en *La filosofía como institución*, Granica, Barcelona, 1984, pp. 145-186.
DIJKSTERHUIS, E. J.: «La Méthode et les essais de Descartes», en *Descartes et le cartésianisme hollandais*, PUF et Ed. française d'Amsterdam, Paris, 1950.
ECHEVERRÍA, J.: «Nota sobre la Geometría de 1637 y el Método cartesiano», en *Descartes* (de V. Gómez Pin), Dopesa, Barcelona, 1979, Apéndice II, pp. 117-127.
GADOFFRE, G.: «Réflexions sur la genèse du *Discours de la Méthode*», *Revue de Synthèse*, New Series, 22, 1958, pp. 11-27.
GALLOIS, P.: «La méthode de Descartes et la médicine», *Hippocrate*, 6, 1938, pp. 65-77.
GEWIRTZ, A.: «Experience and the non-mathematical in the Cartesian methode», *Journal of the History of Ideas*, 2, 1941, 183-210.
GILBERT, N. W.: *Renaissance Concepts of Method*, Columbia University Press, New York/London, 1960.
GOLLIET, P.: «Le problème de la méthode chez Descartes», *Revue des sciences humaines*, 61, 1951, pp. 56-73.
GOUHIER, H.: *Descartes. Essais sur le «Discours de la méthode», la métaphysique et la morale*, J. Vrin, Paris, 1973 (3e. ed.).

HALL, T. S.: «Descartes' physiological method: position, principles, exemples», *Journal of the History of Biology*, 3, 1970, pp. 53-79.
HINTTIKA, J.: «A discourse on Descartes' method», en *Descartes. Critical and Interpretive Essays*, ed. M. Hooker, J. Hopkins University Press, Baltimore, 1978, pp. 74-88.
MARRADES MILLET, J.: «Descartes, Newton y Hegel sobre el método de análisis y síntesis», *Pensamiento*, 164, 1985, pp. 393-404.
MARTÍNEZ MARTÍNEZ, J. A.: *Razón y método en Descartes*, Universidad Complutense, Madrid, 1984.
MUÑOZ-ALONSO LÓPEZ, G.: *El legado de Descartes. Método y «Mathesis Universalis»*, Grupodis, Madrid, 1985.
PEÑA GARCÍA, V.: «Acerca de la "razón" en Descartes: reglas de la moral y reglas del método», *Arbor*, 112, 1982, 167-183.
ROTH, L.: «The Discourse on Method (1637-1937)», *Mind*, 46, 1937, pp. 32-43.
SCHERECKER, P.: «La méthode cartésienne et la logique», *Rv. de phil. de la Fr. et de l'étranger*, 123, 1937, 336-367.
SCHOLTZ, H.: *Mathesis Universalis*, Verlag, Basel/Stuttgart, 1969.
SERRUS, Ch.: *La Méthode de Descartes et son application à la métaphysique*, Alcan, Paris, 1933.
VUILLEMIN, J.: «Sur la difference et l'identité des méthodes de la métaphysique et des mathématiques chez Descartes et Leibniz...», *Archiv für Geschichte der Philosophie*, 43, 1961, pp. 239-303.
WEBER, J. P.: *La constitution du texte des Regulae*, Sedes, Paris, 1964.
XIRAU, R.: «Apuntes de una lectura de Descartes: primera parte del *Discurso del método*», *Dianoia*, 29, 1983, pp. 105-120.

IV. ESTUDIOS GENERALES

ALANEN, L.: *Studies in cartesian epistemology and philosophy of mind*, Helsinki, 1982.
ALQUIÉ, F.: *La découverte métaphysique de l'homme chez Descartes*, PUF, Paris, 1966.
ÁLVAREZ GÓMEZ, A.: «Descartes: la razón, única guía del hombre», *Cuad Salm. Filos.*, 12, 1985, pp. 19-43.
ANGELIS, E.: *El metodo geometrico nella filosofia del seicento*, Ed. F. Le Monnier, Firenze, 1964.
BACON, F.: *La gran Restauración*, ed. de M. A. Granada, Alianza, Madrid, 1985.
BELAVAL, Y.: *Leibniz, critique de Descartes*, Gallimard, Paris, 1960.
BELGIOIOSO, G. (ed.): *Cartesiana*, Congedo Editore, Lecce, 1992.
BLAKE, R. M.: «The role of experience in Descartes' theory of method», en *Theories of scientific Method*, ed. E. Madden, University of Washington Press, Seattle, 1960, pp. 75-103.

BLANCHÉ, R.: *La méthode expérimentale et la philosophie de la physique*, A. Colin, Paris, 1969; trad. esp. en FCE, México, 1972.
CASSIRER, E.: *Descartes, Corneille, Christine de Suède*, Vrin, Paris, 1942.
— *El problema del conocimiento, I*, FCE, México, 1953.
CHOMSKY, N.: *Lingüística cartesiana*, Gredos, Madrid, 1969.
CLARKE, D. M.: *Descartes' Philosophy of Science*, Manchester University Press, Manchester, 1982; trad. esp. en Alianza, Madrid, 1986.
COSTABEL, P.: *Demarches originales de Descartes savant*, J. Vrin, Paris, 1982 (reimp.).
COTTINGHAM, J.: *A Descartes Dictionary*, Basil Blakwell, Oxford, 1993.
DAMBSKA, I.: «Sur certains principes méthodologiques dans les *Principia Philosohiae* de Descartes», *Rv. de Mét. et Mor.*, 62 , 1957, pp. 57-66.
DENISSOFF, E.: *Descartes, premier théoricien de la physique mathématique*, Nauwelaerts, Louvain, 1970.
DICHESNEAU, F.: *Les modèles du vivant de Descartes à Leibniz*, Vrin, Paris, 1998.
DONEY, W. (ed.): *Descartes: A Collection of critical Essays*, Doubleday, New York, 1967.
DUQUE, F.: «Sentido del argumento ontológico en Descartes y Leibniz», *Pensamiento*, 42, 1986, pp. 159-180.
GARBER, D.: *Descartes' metaphysical physics*, Chicago University Press, Chicago, 1992.
GAUKROGER, A. (ed.): *Descartes: Philosophy, Mathematics and Physic*, Harvester Press, Hassoks, 1980.
GILSON, E.: *Études sur le rôle de la pensée médiévale dans la formation du système cartésien*, J. Vrin, Paris, 1930.
GLOUBERMAN, M.: *Descartes: the probable and the certain*, Rodopoi, Amsterdam, 1996.
GÓMEZ PIN, V.: *Descartes*, Dopesa, Barcelona, 1979.
GOUHIER, H.: *Les premières pensées de Descartes*, J. Vrin, Paris, 1958.
GRIMALDI, N.: «Sartre et la liberté cartésienne», *Rv. de Mét. et Mor.*, I, 1987, 67-88.
GROSHOLZ, E. R.: *Cartesian method and the problem of reduction*, Clarendon Press, Oxford, 1991.
GUEROULT, M.: *Descartes selon l'ordre des raisons*, Aubier, Paris, 1953, 2 vols.
HEIDEGGER, M.: «La época de la imagen del mundo», en *Sendas perdidas*, Losada, Buenos Aires, 1960, pp. 68-99.
HINTIKKA, J., y REMES, U.: *The Method of Analysis*, Reidel, Dordrecht, 1974.
HUARTE, J.: *Realidad y conocimiento*, vol. I, *Análisis crítico del razonamiento cartesiano*, Unión Editorial, Madrid, 1992.

HUSSERL, E.: *Méditations cartésiennes*, J. Vrin, Paris, 1969; trad. esp. en Tecnos, Madrid, 1985.
JASPERS, K.: *Descartes y la filosofa*, La Pléyade, Buenos Aires, 1973.
KOYRÉ, A.: *Études d'histoire de la pensée scientifique*, Gallimard, Paris, 1973, trad. esp. en Siglo XXI.
— «Entretiens sur Descartes», en *Introduction à la lecture de Platon*, suivi de..., Gallimard, Paris, 1962, pp. 163-229.
LAPORTE, J.: *Le rationalisme de Descartes*, J. Vrin, Paris, 1950.
LEROY, M.: *Descartes, le philosophe au masque*, Rieder, Paris, 1929, 2 vols.
LEVI, A.: *French moralist. The theory of the passions, 1585 to 1649*, Oxford, 1964.
LLEDÓ, E.: «Semántica cartesiana (Una lectura del *Discours de la méthode*)», en *Filosofía y lenguaje*, Ariel, Barcelona, 1970-74, pp. 173- 207.
LORENZO, J. de: *El racionalismo y los problemas del método*, prólogo de G. Bueno, Cincel, Madrid, 1985.
LYNES, J. W.: «Descartes Theory of Elements: From "Le Monde" to the "Principes"», *Journal of the History of Ideas*, 43, 1982, pp. 55-72.
MARION, J. L.: *Sur l'ontologie grise de Descartes*, Vrin, Paris, 1975.
— *Sur le prisma métaphysique de Descartes*, PUF, Paris, 1986.
MARTINET, M.: «Science et hypothèses chez Descartes», *Archives Internationales d'histoire des sciences*, 24, 1974, 319-339.
MAZZALOMO, U.: «Filosofía y matemática en Descartes», *Philosophia*, 41-42, 1979-80, pp. 63-92.
NICOLÁS, J. A., y FRÁPOLLI, M.ª J. (eds.): *Evaluando la modernidad. El legado cartesiano en el pensamiento actual*, Comares, Granada, 2001.
RABADE ROMEO, S.: *Descartes y la gnoseología moderna*, G. del Toro, Madrid, 1971.
RODIS-LEWIS, G.: *L'Oeuvre de Descartes*, J. Vrin, Paris, 1971, 2 vols.
— *La morale de Descartes*, Paris, 1957 (3e. éd., 1970).
— «Maîtrise des passions et sagesse chez Descartes», *Descartes*, Royaumont, pp. 212-243.
ROED, W.: *Descartes' Erste Philosophie*, Bouvier, Bonn, 1971.
ROSSI, P.: *Los filósofos y las máquinas, 1400-1700*, Labor, Barcelona, 1966.
ROTH, L.: *Spinoza, Descartes and Maimonides*, Russell-Russell, New York, 1963.
SARTRE, J.-P.: «La liberté cartésienne», en *Descartes 1596-1650*, introd. et choix par J. P. S., Paris/Genève, 1946, *Situations I*, Gallimard, Paris, 1947, pp. 289-308.
SCHOULS, P. A.: «Cartesian certainty and the "natural light"», *Australian Journal of Philos.*, 48, 1970, 116-119; «Descartes and autonomy of Reason», *Journal of the History of Philosophy*, X: 3, 1972, pp. 307-323.

—*The imposition of method: a study of Descartes and Locke*, Clarendon Press, Oxford, 1980.
SEARLE, J.: *Mentes, cerebros y ciencia*, trad. de M. L. Valdés, Cátedra, Madrid, 1985.
SPAEMANN, R.: «La moral provisoire de Descartes», *Archives de Philos.*, 35:3, 1972, pp. 353-369.
TOURNADE, G.: *L'orientation de la science cartésienne*, Vrin, Paris, 1982.
TURRÓ, S.: *Descartes. Del hermetismo a la nueva ciencia*, prólogo de E. Lledó, Anthropos, Barcelona, 1985.
VUILLEMIN, J.: *Mathématiques et métaphysique chez Descartes*, PUF, Paris, 1960.

V. REPERTORIOS BIBLIOGRÁFICOS

«Bulletin Cartésien», en *Archives de Philosophie*: aparece cada año en uno de los números de esta revista, desde 1970.
CHAPPEL, V., y DONEY, W.: *Twenty-five years of Descartes scholarship, 1960-1984. A bibliography*, Garalad, New York, 1987..
SEBBA, G.: *Bibliographia cartesiana. A critical guide to the Descartes litterature, 1800-1960*, M. Nijhoff, The Hague, 1964.

DISCURSO DEL MÉTODO

PARA DIRIGIR BIEN LA RAZÓN
Y BUSCAR LA VERDAD
EN LAS CIENCIAS

DISCOURS

DE LA METHODE

Pour bien conduire ſa raiſon, & chercher
la verité dans les ſciences.

PLUS

LA DIOPTRIQVE.

LES METEORES.

ET

LA GEOMETRIE.

Qui ſont des eſſais de cete METHODE.

A LEYDE
De l'Imprimerie de IAN MAIRE.
CIↃ IↃ C XXXVII.
Auec Priuilege.

AT VI 1 Si este discurso pareciera demasiado extenso para ser leído de una sola vez, podría dividirse en seis partes. En la primera se encontrarán diversas consideraciones relacionadas con las ciencias. En la segunda, las reglas principales del método que el autor ha indagado. En la tercera, algunas reglas de moral que ha extraído de este método. En la cuarta, las razones mediante las cuales prueba la existencia de Dios y del alma humana, que son los fundamentos de su metafísica. En la quinta, el orden de las cuestiones de física que ha investigado y, en particular, la explicación del movimiento del corazón y de algunas otras dificultades que conciernen a la medicina, y también la diferencia que hay entre nuestra alma y la de los animales. Y en la última, las cosas que cree necesarias para llegar, en la investigación de la naturaleza, más allá de donde se ha llegado, así como las razones que le han impulsado a escribir[1].

PRIMERA PARTE

El buen sentido[2] es la cosa mejor repartida del mundo, pues cada cual cree estar tan bien provisto

[1] Los enunciados de este índice temático, que en la edición francesa figuran al comienzo del *Discurso*, aparecen en la traducción latina al margen, al comienzo de cada una de las partes.

[2] La expresión francesa *bon sens* (ed. lat. *bona mens*) es sinónima aquí de *razón*, entendida como la capacidad de distinguir lo verdadero

2 de él, que / incluso los más descontentadizos en cualquier otra cosa, no suelen apetecer más del que ya tienen. En lo cual no es verosímil que todos se equivoquen; más bien esto muestra que la facultad de juzgar bien y de distinguir lo verdadero de lo falso, que es lo que propiamente se llama buen sentido o razón, es por naturaleza igual en todos los hombres; y, por lo tanto, que la diversidad de nuestras opiniones no proviene de que unos sean más racionales que otros, sino tan sólo de que dirigimos nuestros pensamientos por caminos diferentes, y no tenemos en cuenta las mismas cosas. No basta, pues, tener un buen ingenio[3], lo principal es aplicarlo bien. Las almas más eminentes son capaces de los mayores vicios, como también de las mayores virtudes; y los que caminan muy lentamente pueden llegar mucho más le-

de lo falso o facultad de juzgar. Al contrario, *bona mens* (trad. fr.: *bon sens)* significa también *sabiduría*, en sentido estoico. Así la Regla I habla de «de bona mente, sive de bac universali sapientia» (AT, X, p. 360), que tal vez retoma el tema de un trabajo inacabado, luego perdido. *Studium bonae mentis* (Baillet, *Vie de M. Des Cartes*, 1691, t. II, p. 406; AT, X, p. 191). Hay un punto de confluencia de los dos sentidos, dado que en ambos se trata de la razón como luz natural, de la razón como principio fundante tanto del saber teórico (*bon sens*) como del práctico (*bona mens* o sabiduría), de tal modo que la sabiduría (*sagesse*) no es sino el «poder de juzgar bien» desarrollado en su mayor grado de perfección posible, mediante el método que regula su uso [E. Gilson, «Commentaire historique» a su ed. del *Discours de la méthode*, J. Vrin, Paris, 1925 (3e. éd. 1947), p. 82, cf. *Conversación con Burman, cit.*, p. 175].

[3] La noción *esprit* (ed. lat.: *ingenium*), usada sobre todo para mostrar la oposición a la sustancia extensa, designa el pensamiento en general (*res cogitans*), tal como se encuentra, por ejemplo, en la *II Medit.*: «Así pues, hablando con precisión, no soy más que una cosa que piensa, es decir, un espíritu (*esprit*), un entendimiento o una razón» (AT, IX, p. 21). Se ha preferido, en la traducción el término *ingenio*, porque con él se obvia la referencia fácil a algo suprahumano que sugiere el término «espíritu», y se designa la disposición natural de la que todos están dotados en tanto que el hombre es *res cogitans*.

jos, si siguen siempre el camino recto, que los que corren, pero se alejan de él.

Por mi parte, nunca he presumido que mi ingenio fuese en algo más perfecto que el de los demás; hasta he deseado con frecuencia tener el pensamiento tan ágil, o la imaginación tan nítida, o la memoria tan amplia y viva, como otros lo tienen. Y no conozco otras cualidades, excepto éstas, que puedan contribuir a la perfección del ingenio; pues en lo que concierne a la razón, o al sentido, ya que es la única cosa que nos hace hombres, y nos distingue de los animales, quiero creer que está entera en cada uno de nosotros, y seguir en esto la común opinión de los filósofos[4], que dicen que sólo existen
3 diferencias de grado entre los / *accidentes*, y de ninguna manera entre las *formas* o naturalezas de los *individuos* de una misma *especie*.

Pero no me arredra afirmar que creo haber tenido una gran suerte, al encontrarme desde joven en ciertos caminos, que me han conducido a consideraciones y máximas, a partir de las cuales he llegado a formar un método, por medio del cual me parece que es posible aumentar gradualmente mi conocimiento y elevarlo poco a poco hasta el punto más alto, al que la mediocridad de mi ingenio y la brevedad de la vida puedan permitirle llegar. Pues de ese método he recogido ya tales frutos[5] que, si bien en los juicios que hago sobre mí mismo siempre tiendo a inclinarme hacia el lado de la desconfianza más que al de la presunción, y si bien al observar con mirada de filósofo las diferentes acciones y empresas de los hombres, no encuentro

[4] Según el uso frecuente de Descartes, se refiere en sentido restringido a los filósofos escolásticos.

[5] Los *Ensayos* a los que el *Discurso* sirve de introducción son ya resultados concretos de la aplicación del método en dominios particulares de la ciencia, sobre todo en geometría y en física.

casi ninguna que no me parezca vana e inútil, no deja de producirme una gran satisfacción el progreso que creo haber realizado ya en la investigación de la verdad, ni dejo de concebir tales esperanzas para el futuro que, si entre las ocupaciones propias de los hombres hay alguna que sea sólidamente buena e importante, me atrevo a creer que es la que yo he elegido.

Puede suceder, no obstante, que esté equivocado y apreciar acaso como oro y diamantes lo que no es sino un trozo de cobre o vidrio. Sé cuán expuestos estamos a equivocarnos en todo lo que nos afecta, y cuán sospechosos deben sernos también los juicios de los amigos cuando los pronuncian en nuestro favor. Con todo me gustaría mostrar en este
4 discurso / los caminos que he seguido, y representar así mi vida como en un cuadro, a fin de que cada uno pueda juzgar, y enterándome luego por rumor público de las opiniones emitidas, tendré un nuevo medio para instruirme, que añadiré a los que acostumbro a emplear.

No es, pues, mi propósito enseñar aquí el método que cada cual debe seguir para dirigir bien su razón, sino sólo mostrar de qué manera he procurado conducir la mía. Aquellos que se atreven a dar preceptos deben estimarse más hábiles que aquéllos a quienes se los dan, y si faltan en la cosa más mínima son censurables por ello. Pero como no propongo este escrito sino a modo de historia o, si se prefiere, de fábula, en la que junto a algunos ejemplos imitables se encontrarán tal vez algunos otros que sería razonable no seguir, espero que será útil para algunos sin ser nocivo para nadie, y que todos agradecerán mi franqueza.

Desde mi niñez fui habituado en el estudio de las letras, y como me persuadían que por medio de ellas se podía adquirir un conocimiento claro y se-

guro de todo cuanto es útil para la vida, sentía yo un vivísimo deseo de aprenderlas. Pero cambié por completo de opinión tan pronto como hube concluido mis estudios[6], al término de los cuales se acostumbra a entrar en el rango de los doctos. Pues me embargaban tantas dudas y errores que, habiendo intentado instruirme, me parecía no haber alcanzado otro resultado que el de haber descubierto progresivamente mi ignorancia. Y, sin, em-
5 bargo, / realizaba mis estudios en una de las escuelas más célebres de Europa[7], en donde pensaba yo que debía haber hombres sabios, si es que existían en algún lugar de la tierra. Había conseguido aprender allí todo lo que los compañeros aprendían; y no contento aún con las ciencias que nos enseñaban, hojeé cuantos libros pudieran caer en mis manos referentes a las que se consideran como las más curiosas y raras[8]. Conocía, además, los juicios que los otros hacían sobre mí, y no veía que se me considerase inferior a mis condiscípulos, aunque entre ellos hubiese ya algunos a quienes se destinaba a ocupar los puestos de nuestros maestros. En fin, nuestra época me parecía tan floreciente y fértil en destacados ingenios como haya

[6] El título de licenciado en derecho por la universidad de Poitiers cierra el período de «estudios» mencionado, que va de 1606 a 1616. Pero el «cambio» teórico es más significativo que ese final histórico, si entendemos por tal el distanciamiento crítico de Descartes respecto de la cultura establecida —la escolástica, la opinión, la autoridad—, como proceso inicial de una filosofía crítica tematizada en el *Discurso* en términos de «duda» y de búsqueda de un fundamento cierto del saber.

[7] El colegio de La Flèche fue fundado por los jesuitas en 1604, en un edificio donado por Enrique IV, de ahí el nombre de «Collège Royal». Descartes estudió en él de 1606 a 1614.

[8] Se trata no sólo de la química incipiente y de una parte de la óptica, la que permite ver cosas extraordinarias a través de espejos y lentes, sino también de la astrología, la quiromancia, la cábala, la magia, etc.

podido serlo cualquiera de las precedentes. Por todo esto llegué a tomarme la libertad de juzgar a los demás por mí mismo y de pensar que no había doctrina alguna en el mundo tal y como se me había prometido anteriormente.

Con todo, no dejaba de valorar los ejercicios que se practican en las escuelas. Sabía que las lenguas que en ellas se aprenden son necesarias para comprender las obras de la antigüedad; que la graciosa elegancia de las fábulas despierta el ingenio; que las acciones memorables de las historias lo exaltan, y que leídas con discreción contribuyen a la formación del juicio; que la lectura de todos los buenos libros es como una conversación con las gentes más distinguidas de los siglos pasados, que han sido sus autores, y hasta una conversación estudiada en la que no nos descubren sino lo más selecto de sus pensamientos; que la elocuencia posee una belleza y un poder de seducción incompara-
6 bles; que la poesía encierra / delicadezas y suavidades que arrebatan; que las matemáticas posibilitan sutilísimas invenciones que pueden contribuir mucho, tanto a satisfacer a los curiosos, como a facilitar todas las artes y a disminuir el trabajo de los hombres[9]; que los escritos que tratan de las costumbres contienen muchas enseñanzas y exhortaciones a la virtud que son muy útiles; que la teología enseña a ganar el cielo; que la filosofía proporciona el medio para hablar con verosimilitud de todas las cosas y hacerse admirar de los menos

[9] P. Rossi (*Los filósofos y las máquinas 1400-1700*, Labor, Barcelona, 1966) estudia el papel desempeñado por las artes mecánicas en la constitución del pensamiento moderno (pp. 102-103). Pero no es tanto la matemática aplicada a la técnica lo que atrae a Descartes en la reflexión metodológica, cuanto el modelo de certeza de la matemática pura (véase *Discurso*, Parte II).

sabios[10]; que la jurisprudencia, la medicina y las demás ciencias dan honores y riquezas a quienes las cultivan; y, en fin, que es bueno haberlas examinado todas, incluso las más supersticiosas y falsas, para conocer su justo valor y no dejarse engañar por ellas.

Pero creía también que ya había dedicado suficiente tiempo a las lenguas, e incluso a la lectura de los libros antiguos y a sus historias y fábulas. Pues es casi lo mismo conversar con gentes de otros siglos que viajar. Es conveniente saber algo sobre las costumbres de los diversos pueblos, para juzgar las del propio con mejor acierto, y no creer que todo lo que sea contrario a nuestras modas es ridículo y contra la razón, como suelen hacer los que nada han conocido. Pero cuando se dedica demasiado tiempo a viajar llega uno a sentirse extranjero en su país; y cuando se posee excesiva curiosidad por lo que se hacía en los siglos pasados se permanece generalmente muy ignorante de lo que ocurre en el presente. Además, las fábulas son causa de
7 imaginemos como posibles acontecimientos / que no lo son; y hasta las más fieles historias, si bien no cambian ni aumentan el valor de las cosas para hacerlas más dignas de ser leídas, al menos omiten casi siempre las circunstancias más vulgares y menos ilustres; de ahí proviene que lo demás no parezca tal como es, y que los que ajustan sus cos-

[10] La enseñanza de la filosofía en los colegios de jesuitas ocupa los tres últimos años. Comprendía: lógica, física, metafísica y moral, impartidas desde un punto de vista aristotélico-tomista. La escolástica tardía tuvo su mejor intérprete en el jesuita F. Suárez, *Disputationes metaphysicae*, Salamanca, 1597. No obstante esta filosofía especulativa de la Escuela, la posibilidad de una *filosofía práctica*, que contribuya al desarrollo de la técnica, es una consecuencia de la aplicación del método a determinados dominios del saber (véase *Discurso*, Parte VI). Cf. nota 11.

tumbres a los ejemplos que sacan de tales historias se exponen a caer en las extravagancias de los héroes de nuestras novelas y a concebir proyectos que superan sus fuerzas.

Estimaba mucho la elocuencia y era un enamorado de la poesía; pero pensaba que una y otra eran dones de la naturaleza más que frutos del estudio. Los que poseen excelente capacidad para razonar y con facilidad disponen con orden sus pensamientos, con el fin de hacerlos claros e inteligibles, siempre pueden persuadir mejor sobre aquello que proponen, aunque hablen la lengua inculta de los godos y jamás hayan estudiado retórica. Así como los que son capaces de las más agradables invenciones y saben expresarlas con el mayor ornato y dulzura, no dejarían de ser los mejores poetas aunque el arte poético les fuera desconocido.

Me deleitaba sobre todo en el estudio de las matemáticas, dada la certeza y evidencia de sus razonamientos; pero no me daba cuenta todavía de su verdadero uso[11] y, pensando que sólo eran aplicables a las artes mecánicas, me extrañaba de que, siendo sus cimientos tan firmes y sólidos, no se hubiese construido sobre ellos nada más elevado. Comparaba, en cambio, los escritos de los antiguos
8 paganos que tratan de las costumbres con palacios / de soberbia magnificencia, pero construidos sobre arena y barro. Exaltan en grado máximo las virtudes

[11] Como sostiene Einstein, en la medida en que las proposiciones de la matemática son *ciertas*, esto es, estrictamente axiomáticas, no se refieren a la realidad. Desde este modelo de certeza, a Descartes no le satisface la matemática aplicada, por ejemplo, al arte militar, a la técnica de las fortificaciones. Pero como Galileo, y luego Einstein, descubrirá que el verdadero uso o función de la matemática consiste en tratar de «examinar cuestiones de física por medio de razonamientos matemáticos» (A Mersenne, 11 de octubre de 1638). Cf. *Conversación con Burman, cit.*, pp. 177-179.

y las presentan como lo más estimable de cuantas cosas hay en el mundo; pero no nos enseñan suficientemente a conocerlas, y a menudo lo que designan con tan digno nombre no es sino insensibilidad, orgullo, desesperación o parricidio.

Respetaba nuestra teología y, como otro cualquiera, aspiraba a ganar el cielo; pero habiéndome enseñado, como algo muy seguro, que su camino no es menos accesible para los ignorantes que para los doctos y que las verdades reveladas, que a él conducen, están por encima de nuestra inteligencia, nunca me hubiera atrevido a someterlas a la fragilidad de mis razonamientos, pues pensaba que para emprender su examen y finalizarlo con éxito era necesario contar con alguna asistencia extraordinaria del cielo y ser algo más que hombre.

Nada diré sobre la filosofía, sino que, viendo que ha sido cultivada por los ingenios más relevantes que han existido desde hace siglos y que, sin embargo, nada hay en ella que no sea aún objeto de disputa y, por lo tanto, dudoso, no tenía yo la suficiente presunción para esperar alcanzar en ella algo mejor que los otros. Considerando, además, cuán diversas opiniones pueden darse referentes a una misma materia, defendidas por gentes doctas, si bien sólo una de ellas puede ser verdadera, estimaba casi como falso todo lo que no era más que verosímil[12].

[12] Lo «verosímil» es aquello que no depende de un razonamiento demostrativo. La crítica de la filosofía escolástica es manifiesta, tanto por el lugar que en ella ocupa la disputa, como por el carácter simplemente probable de las conclusiones a las que llega por medio de razonamientos silogísticos. El modelo matemático de certeza le permite identificar no sólo lo que es falso y verosímil, sino también la vacía erudición del relato de las opiniones en disputa; pero no le evita incurrir en la actitud escéptica, a lo Montaigne —que llevaba a Descartes a «abandonar enteramente el estudio de las letras»—, al menos mientras busca y encuentra, en 1619, otra verdad.

Y en cuanto a las demás ciencias, dado que to-
9 man sus principios de la filosofía, estimaba / que no se podía haber construido nada sólido sobre cimientos tan poco firmes. De ahí que ni el honor ni el provecho que prometen fueran razones suficientes para determinarme a aprenderlas; pues no me encontraba, gracias a Dios, en una situación tal que me viese obligado a hacer de la ciencia un oficio para mejorar mi fortuna; y aunque no profesara el desprecio de la gloria a lo cínico, apreciaba muy poco sin embargo aquélla que sólo se podía adquirir mediante falsos títulos. Y, en fin, en lo que se refiere a las vanas doctrinas, pensaba que ya conocía bastante bien su valor, para no dejarme engañar nunca más ni por las promesas de un alquimista, ni por las predicciones de un astrólogo, ni por las imposturas de un mago, ni por los artificios o la presunción de todos los que profesan saber más de lo que saben.

Por ello, tan pronto como la edad me permitió salir de la sujeción de mis preceptores, abandoné completamente el estudio de las letras. Y, tomando la decisión de no buscar otra ciencia que la que pudiera hallar en mí mismo o en el gran libro del mundo, dediqué el resto de mi juventud a viajar, a conocer cortes y ejércitos, a tratar con gentes de diversos temperamentos y condiciones, a recoger diferentes experiencias, a ponerme a mí mismo a prueba en las ocasiones que la fortuna me deparaba, y a hacer siempre tal reflexión sobre las cosas que se me presentaban, que pudiese obtener algún provecho de ellas. Pues me parecía que podría encontrar mucha más verdad en los razonamientos que cada uno hace en los asuntos que le atañen, y cuyo
10 resultado / puede serle inmediatamente su castigo si ha juzgado mal, que los que hace en su despacho un hombre de letras sobre especulaciones que no producen efecto alguno y que no tienen para él otra

consecuencia, acaso, que la de inspirarle tanta más vanidad cuanto más se apartan del sentido común, ya que habrá tenido que emplear mucho más ingenio y artificio para intentar hacerlas verosímiles. Y siempre sentía un deseo inmenso de aprender a distinguir lo verdadero de lo falso, para ver claro en mis acciones y andar con seguridad en esta vida.

Es cierto que, mientras me limitaba a considerar las costumbres de los otros hombres, apenas encontraba en ellas nada que me convenciera, y observaba casi tanta diversidad como había advertido antes entre las opiniones de los filósofos. De tal manera que el mayor provecho que obtenía de esto era que, al ver varias cosas que si bien nos parecen muy extravagantes y ridículas no dejan de ser por ello comúnmente aceptadas y aprobadas por otros grandes pueblos, aprendió a no creer nada con demasiada firmeza de todo lo que se me había persuadido únicamente por el ejemplo y la costumbre; y así me liberaba poco a poco de muchos errores, que pueden ofuscar nuestra luz natural y hacernos menos aptos para escuchar la voz de la razón[13]. Pero después de haber empleado varios años estudiando de este modo en el libro del mundo e intentando adquirir alguna experiencia, tomé un día la resolución de estudiar también en mí mismo y de emplear todas las fuerzas de mi ingenio en elegir los caminos que debía seguir. Lo cual me
11 dio mejor / resultado, me parece, que si no me hubiera alejado nunca de mi país ni de mis libros.

[13] «Escuchar la voz de la razón» es, para Heidegger, una de las claves de la filosofía moderna. Sin embargo, no es la *filosofía*, esto es, el «corresponder que sintoniza con la voz del Ser del ente» (*¿Qué es filosofía?* Narcea, Madrid, 1978, p. 67). Pero, si este corresponder como decir privilegiado es traducible a lenguaje, ¿no tiene que ver en ello la voz de la razón? (cf. M. Heidegger, *La pregunta por la cosa*, pp. 96-97). Cf. A. Álvarez Gómez, «Descartes: la razón, única guía del hombre», *Cuad. Salm. Filos.*, 12, 1985, pp. 19-43.

SEGUNDA PARTE

Me encontraba entonces en Alemania, a donde había ido con motivo de unas guerras que aún no han terminado[1]; y cuando volvía al ejército después de asistir a la coronación del emperador, el comienza del invierno me retuvo en un acuartelamiento en el que, no encontrando conversación alguna que me distrajera y no teniendo tampoco, por fortuna, preocupaciones o pasiones que me perturbaran, permanecía durante todo el día solo y encerrado junto a una estufa, donde disponía de la tranquilidad necesaria para entregarme a mis pensamientos[2]. Entre

[1] Se trata de la guerra de los Treinta Años, que finaliza en 1648 con la paz de Westfalia. Descartes ha dejado Holanda el 29 de abril de 1619 para asistir a la coronación de Fernando II como emperador, en Francfort, cuyas fiestas duraron del 20 de julio al 9 de septiembre de 1619. Con el apoyo de la Liga Católica, Fernando II lucha contra Federico V, que lidera la Liga Evangélica. ¿Por qué Descartes está al lado de la política contrarreformista del primero? ¿Por qué mantiene la extensa correspondencia con Elisabeth, hija del segundo, si además Descartes está enrolado en el ejército del duque de Baviera, Maximiliano, quien combate contra el elector palatino? Pero ¿cuál es la participación de Descartes en la guerra? ¿Se alista en el ejército por alguna causa concreta, o más bien para huir de los acontecimientos que tienen lugar en Holanda? Si tenemos en cuenta que Descartes se dedica, sobre todo, a sus reflexiones y que abandona definitivamente el ejército en 1620 para dedicarse a viajar, cabría concluir que la presencia de Descartes en el ejército ha tenido escasa significación. Para un estudio monográfico de esta actitud política, véase: A. Negri, *Descartes político*, Feltrinelli, Milano.

[2] El célebre descubrimiento de este momento, el principio de la unidad del saber, es consignado en el escrito *Olympica* (1620) en los siguientes términos: «10 de noviembre de 1619, cuando lleno de entusiasmo descubrí los fundamentos de una ciencia admirable» (AT, X, p. 179). El entusiasmo va acompañado, en la noche del 10 al 11, de tres sueños, cuyo relato e interpretación puede verse en: *Olympica*; G. Milhaud, *Descartes savant*, Alcan, Paris, 1921, cap. II; J. Maritain, «Le songe de Descartes», *Revue universelle*, 1920.

los cuales, uno de los primeros fue el darme cuenta de que a menudo no existe tanta perfección en obras compuestas de muchos elementos y realizadas por diversos maestros como en aquellas ejecutadas por uno solo. Así vemos que los edificios que un solo arquitecto ha comenzado y acabado suelen ser más bellos y mejor ordenados que aquellos otros que varios han tratado de reformar, utilizando antiguos muros que habían sido construidos para otros fines. Así sucede con esas viejas ciudades que, no habiendo sido al comienzo sino aldeas, se han convertido con el transcurso del tiempo en grandes urbes; están generalmente muy mal trazadas si las comparamos con esas plazas regulares que un ingeniero diseña según su fantasía sobre un terreno llano; pues, si bien considerando cada uno de sus edificios aisladamente se encuentra a menudo en ellos tanto o más arte que en los de las ciudades nuevas, sin embargo, al ver cómo están emplazados —aquí uno grande, allá uno pequeño— y cómo hacen las calles curvas y desiguales, / se diría que ha sido más bien el azar, y no la
12 voluntad de unos hombres que hacen uso de su razón, el que los ha dispuesto así. Y si se considera que, no obstante, siempre han existido funcionarios encargados de cuidar de que los edificios de los particulares contribuyan al ornato público, fácilmente se comprenderá lo difícil que es hacer algo perfecto cuando se trabaja sobre obras realizadas por otro. Del mismo modo, me imaginaba que los pueblos que han evolucionado poco a poco desde un estado semisalvaje a otro civilizado, elaborando sus leyes sólo cuando la incomodidad de los crímenes y pe-

leas les ha obligado, no pueden estar políticamente tan bien organizados[3] como aquellos que, desde el momento en que se reunieron por primera vez, han observado las constituciones de algún prudente legislador[4]. Como también es muy cierto que el gobierno de la verdadera religión, cuyas leyes Dios solo ha instituido, debe estar incomparablemente mejor regulado que cualquier otro. Pero hablando de las cosas humanas, pienso que si Esparta ha sido en otro tiempo muy floreciente no se debió a la bondad de cada una de sus leyes en particular, pues algunas eran muy extrañas y hasta contrarias a las buenas costumbres, sino al hecho de que siendo concebidas por un solo legislador, todas tendían a un mismo fin. De igual modo, pensaba que las ciencias expuestas en los libros, al menos aquéllas cuyas razones son sólo probables y carecen de toda demostración, habiendo sido elaboradas y paulatinamente engrosadas con las opiniones de muchas y diferentes personas, no están tan cerca de la verdad como los simples razonamientos que un hombre de buen sen-
13 tido[5] puede hacer, naturalmente, / acerca de las cosas que se presentan. Y también pensaba que, como hemos sido todos nosotros niños antes de ser hombres y hemos tenido que dejarnos regir por nuestros apetitos y por nuestros preceptores, con frecuencia contrarios unos a otros y que, tal vez, ni unos ni otros nos han aconsejado siempre lo mejor, es casi imposible que nuestros juicios sean tan puros y sólidos como lo habrían sido si, desde el momento de nacer, hubiéramos dispuesto del pleno uso de nuestra razón y nos hubiéramos guiado exclusivamente por ella.

[3] La expresión francesa *si bien policés* (ed. lat.: *tam bene instituta república)* se refiere al orden y a la organización de un pueblo. Tal sentido proviene del término *police* en el sentido en que lo define el *Dictionnaire de l'Académie* (1694): «el orden, las ordenanzas que se observan en un Estado, en una república, en una ciudad».

[4] Alusión a Licurgo, creador de la legislación espartana.

[5] Un hombre que sólo hace uso de la razón natural (ed. lat.: *sola ratione naturali utens*). Cf. J. Morris, «Descartes' natural light», *Journal of the History of Philosophy*, XI: 2, 1973, pp. 169-187.

Es cierto que no vemos que se derriben todas las casas de una ciudad con el único propósito de reconstruirlas de otra manera y de contribuir a un mayor embellecimiento de sus calles; pero se ve muchas veces que algunos particulares mandan derribar las suyas para edificarlas de nuevo, y hasta algunas veces se ven obligados a ello cuando sus viviendas amenazan ruina y cuando sus cimientos son poco firmes. Ante cuyo ejemplo llegué a persuadirme de que no sería en verdad sensato que un particular intentara reformar un Estado, cambiándolo todo desde sus fundamentos y derribándolo para hacerlo resurgir; como tampoco lo sería reformar el cuerpo de las ciencias, o el orden establecido en las escuelas para su enseñanza[6]. Pero, con relación a todas aquellas opiniones que hasta entonces había aceptado, no podía hacer nada mejor que proponerme de una vez abandonarlas, con el fin de sustituirlas luego bien por otras mejores o bien por
14 las mismas, pero después que las hubiera / sometido al juicio de la razón. Creí firmemente que, por este medio, acertaría a dirigir mi vida mucho mejor que si me limitase a edificar sobre antiguos cimientos y

[6] El planteamiento de la reforma del saber desde sus cimientos es tratado con suma cautela, sobre todo en lo que afecta al saber práctico, esto es, a las instituciones sociales como el Estado, la Iglesia y las instituciones docentes. Como observa G. Quintás, la suerte corrida por quienes habían defendido una reforma radical en este campo «desaconsejaba» ahora tal pretensión, como ya lo hiciera Montaigne (*Essais*, III, 9). Lo cierto es que el proyecto cartesiano de reconstruir el saber desde fundamentos nuevos afectará a dichas instituciones, pues la posibilidad crítica de someter también estas opiniones al «juicio de la razón» está abierta, como se dice a continuación.

me apoyase únicamente sobre los principios de los que me había dejado persuadir durante mi juventud, sin haber examinado nunca si eran verdaderos. Pues, si bien advertía en esto diversas dificultades, ni eran insolubles sin embargo, ni comparables a las que se encuentran en la reforma de las menores cosas que atañen a lo público. Estos grandes cuerpos políticos son muy difíciles de reconstruir, una vez que han sido derribados, e incluso de sostenerlos en pie cuando han sido fuertemente sacudidos, y sus caídas son necesariamente muy duras. Además, en lo que toca a sus imperfecciones, si las tienen, como la única diferencia que existe entre ellos es suficiente para asegurar que algunos las tienen, el uso las ha suavizado muchísimo sin duda; y hasta ha evitado o corregido insensiblemente muchas, que con la prudencia no hubiesen podido remediarse tan eficazmente. Y, en fin, tales imperfecciones son casi siempre más soportables para un pueblo habituado a ellas[7] que lo sería su cambio; como los caminos reales que serpentean entre montañas que, a fuerza de ser frecuentados, llegan a ser poco a poco tan lisos y cómodos que es preferible seguirlos que intentar un atajo más recto, escalando las rocas y descendiendo hasta los precipicios.

Por todo esto, de ningún modo podría estar de acuerdo con esos hombres de carácter lioso e inquieto, que no cesan de idear constantemente alguna nueva reforma, sin haber sido llamados a la administración de los asuntos públicos ni por su
15 nacimiento ni por su fortuna[8]. / Y si estimara que

[7] Descartes sigue aquí a Montaigne (*Essais*, III, 9) y a Charron (*De la sagesse*, II, 8, art. 7). La ed. lat. añade a «plus supportable»: *ab assuetis populis* (AT, VI, p. 547), cf. E. Gilson, *op. cit.*, p. 173.

[8] Lo que se pretende es alejar cualquier posible imagen de «reformador» político o social, con el fin de no ser importunado en su proyecto de reforma del saber, como puntualiza a continuación. Hay que notar, de paso, que el acceso a la administración de los asuntos públicos sólo era posible a través del «nacimiento» (nobleza) y de la «fortuna» (riqueza). Esta última era la puerta de acceso de la burguesía a los espacios de poder.

hay en este escrito la menor cosa que pudiera hacerme sospechoso de semejante locura, de ningún modo podría permitir o sufrir que fuera publicado. Mi propósito no ha sido nunca otro que intentar reformar mis propios pensamientos y construir sobre un terreno que sea enteramente mío. Y si, habiéndome complacido bastante mi obra, os la propongo como modelo, no significa esto que quiera aconsejar a nadie que lo siga. Aquellos a los que Dios ha distinguido con sus dones tendrán quizá proyectos más elevados; pero mucho me temo que éste resulte ya demasiado audaz para algunos. La mera resolución de liberarse de todas las opiniones de las que hemos sido empapados en otro tiempo, no es un ejemplo que todos deban seguir; el mundo casi se compone, en efecto, de dos tipos de personas a quienes este ejemplo no conviene en modo alguno. A saber, de los que, creyéndose más hábiles de lo que son, no pueden contener la precipitación de sus juicios ni tener bastante paciencia para conducir ordenadamente sus pensamientos; de ahí se desprende que, si alguna vez se hubieran tomado la libertad de dudar de los principios que han recibido y de apartarse del camino común, nunca llegarían a encontrar el sendero que hay que tomar para ir más recto y permanecerían extraviados toda su vida. Y, además, de aquellos que, teniendo la suficiente razón o modestia para apreciar que son menos capaces de distinguir lo verdadero de lo falso que otros hombres por los cuales pueden ser instruidos, deben más bien contentarse con seguir las opiniones de éstos que buscar por sí mismos otras mejores.

16 Y, en cuanto a mí, me hubiera encontrado sin duda entre estos últimos, si no hubiera tenido más que un solo maestro o no hubiera conocido las diferencias que siempre han existido entre las opiniones de los más doctos. Pero, habiendo aprendido desde el colegio que no se puede imaginar nada, por extraño e increíble que sea, que no haya sido dicho por alguno de los filósofos; y habiendo observado luego, en mis viajes, que todos aquellos cuyo sentimientos son muy contrarios a los nuestros no son por ello bárbaros ni salvajes, sino que muchos hacen tanto o más uso que nosotros de la razón; habiendo considerado también cuán diferente llegaría a ser un hombre que, con su mismo ingenio, hubiese sido criado desde su infancia entre franceses o alemanes en lugar de haber vivido siempre entre chinos o caníbales; y que hasta en las modas de nuestros trajes, lo que nos ha gustado hace diez años, y acaso vuelva a gustarnos dentro de otros diez, nos parece hoy extravagante y ridículo, de modo que son más bien la costumbre y el ejemplo los que nos persuaden, que conocimiento alguno cierto; y habiendo considerado, en fin, que la pluralidad de votos no es una prueba válida en favor de las verdades algo difíciles de descubrir, pues es mucho más verosímil que un hombre solo las descubra que todo un pueblo, no podía elegir a nadie cuyas opiniones me parecieran que debían ser preferidas a las de los demás, y me encontraba como obligado a emprender por mí mismo la tarea de conducirme.

Pero, al igual que un hombre que camina solo y en la oscuridad, tomé la resolución de avanzar
17 tan lentamente y de usar / tal circunspección en todas las cosas que, aunque adelantara muy poco, me guardaría muy bien al menos de tropezar y caer. Ni siquiera quise empezar a rechazar por

completo ninguna de las opiniones que en otro tiempo hubieran podido deslizarse en mi creencia sin pasar por el tamiz de la razón, hasta que no hubiese dedicado el tiempo suficiente a hacer el proyecto de la obra que me proponía, y a buscar el verdadero método para llegar al conocimiento de todas las cosas de las que mi ingenio fuera capaz.

Había estudiado un poco, cuando era más joven, de entre las partes de la filosofía, la lógica, y de las matemáticas, el análisis de los geómetras y el álgebra, tres artes o ciencias[9] que al parecer debían contribuir en algo a mi propósito. Pero, al examinarlas atentamente, advertí con relación a la lógica que sus silogismos y la mayor parte de sus preceptos sirven más para explicar a otro cuestiones ya sabidas o incluso, como el arte de Lulio,

[9] Descartes ha estudiado la *lógica* aristotélica y, como dice más adelante, «el análisis de los antiguos» y el «álgebra de los modernos». En cuanto al *análisis geométrico* se ha de tener en cuenta: 1) que su descubrimiento por los humanistas del Renacimiento se debe a helenistas matemáticos; 2) que el texto más antiguo y extenso es el del alejandrino Pappus, cuya traducción de Commandino (Pappi Alexandrini, *Mathematicorum Collectionum*, Lib. VII) ha podido leer Descartes; 3) Descartes considera, a diferencia de Pappus, que el análisis es superior a la síntesis, tanto como método de exposición como de invención (*Segundas Respuestas*, AT, VII, p. 155); 4) que el análisis geométrico no es lo mismo que el análisis propuesto como tercera regla del método. En cuanto al *álgebra*. Descartes la ha estudiado, sin duda, en el *Álgebra* de Clavios, S. J. (*Operum mathematicorum*, t. II, Moguntiae, 1612, pp. 1-181), del que ha tomado la notación, y no de Ramus o de Vieta; pero remonta su descubrimiento, más allá de los árabes, a los alejandrinos Pappus y Diophanto (cf. Regla IV; AT, X, p. 376), cf. E. Gilson, *op. cit.*, pp. 187-191. Si escribe «artes o ciencias», no es sino para evitar la inútil controversia secular acerca de la naturaleza de estas disciplinas (J. Sirven, *Les années d'apprentisage de Descartes*, Paris, 1928). Cf. A. Robert, «Descartes et l'Analyse des anciens», *Archives de philosophie*, XIII, 1937, pp. 221-245; J. L. Coolidge, *A history of geometrical methods*, Clarendon, Press, Oxford, 1940; J. Hintikka y U. Remes, *The method of analysis*, Reidel, Dordrecht, 1974.

para investigar las que desconocemos[10]. Y si bien contiene, en efecto, muchos preceptos que son muy buenos y verdaderos, hay sin embargo, mezclados con ellos, tantos otros perjudiciales o bien superfluos, que es casi tan difícil separarlos como sacar una Diana o una Minerva de un bloque de mármol en el que ni siquiera hay algo esbozado. En lo que concierne, por otra parte, al análisis de los antiguos y al álgebra de los modernos, además de que no se refieren sino a materias muy abstractas, que parecen carecer de todo uso, el primero está siempre tan circunscrito a la consideración de las figuras, que no permite ejercitar el entendi-
18 miento / sin fatigar excesivamente la imaginación; y en la segunda, hay que sujetarse tanto a ciertas reglas y cifras, que se ha convertido en un arte confuso y oscuro, bueno para enredar el ingenio, en lugar de una ciencia que lo cultive. Tal fue la causa por la que pensé que había que buscar algún otro método que, reuniendo las ventajas de los otros tres, estuviera exento de sus defectos»[11]. Y

[10] El propósito de Descartes está formulado explícitamente en el subtítulo del *Discurso* e insistentemente en toda su obra. Aquí nos lo repite: se trata de *investigar*, de descubrir lo que no conocemos (ed. lat.: *ad ea quae ignoramus investiganda*). Desde esta posición, la crítica del silogismo, que repite en la conclusión lo que ya está contenido en las premisas, y del *Ars magna* de Lulio (cf. a *Beeckman*, 29 de abril de 1619), porque son métodos formales que no hacen avanzar la ciencia, es evidente.

[11] Un método que no sólo garantice formalmente el rigor lógico, sino que sea, ante todo, un método de invención e investigación; que conserve del análisis geométrico el apoyo de la imaginación, puesto que trabaja con líneas, pero sin someter a ella el intelecto, que simbolice las cantidades mediante signos como el álgebra, pero simplificándolos, como acaba de hacer Descartes, representando mediante las letras del alfabeto todas las cantidades conocidas (*a*, *b*, *c*, etc.) y desconocidas (*x*, *y*), etc. En resumen, un método de investigación, que no obstaculice la razón, sino que facilite su trabajo mediante la simplificación de su lenguaje, Cf. J. Echeverría, «Nota sobre la Geometría de 1637 y el Método cartesiano», en *Descartes* (de V. Gómez Pin), Dopesa, Barcelona, 1979, Apéndice II, pp. 117-127.

como la multiplicidad de leyes a menudo sirve de excusa para los vicios, de tal forma que un Estado está mucho mejor regido cuando no existen más que unas pocas, pero muy estrictamente observadas, así también, en lugar del gran número de preceptos de los que la lógica está repleta, estimé que tendría suficiente con los cuatro siguientes, con tal de que tomase la firme y constante resolución de no dejar de observarlos ni una sola vez.

El primero consistía en no admitir jamás cosa alguna como verdadera sin haber conocido con evidencia que así era[12]; es decir, evitar con sumo cuidado la precipitación y la prevención, y no admitir en mis juicios nada más que lo que se presentase tan clara y distintamente a mi espíritu que no tuviese motivo alguno para ponerlo en duda[13].

El segundo, en dividir cada una de las dificulta-

[12] El nuevo criterio de verdad, la *evidencia*, y sus condiciones (la claridad y distinción, y la indudabilidad) es decisivo en la epistemología cartesiana. No sólo rompe la identidad tradicional expresada en la «adaequatio intellectus et rei», sino que excluye, junto con la *probabilidad* aristotélica y la *verosimilitud* escolástica la *conjetura*. Al contrario de ésta, lo evidente es aquello cuya verdad aparece a la mente de manera inmediata, es decir, en la operación denominada *intuición*, según el enunciado de la Regla III: los que buscan el recto camino de la verdad no deben ocuparse sino de «lo que podamos intuir clara y evidentemente o deducir con certeza». Se añade en este enunciado una operación complementaria de la intuición, la deducción, y la exigencia de certeza.

[13] Al remitirnos típicamente a la «claridad y distinción» como condiciones de la evidencia, esto es, de la verdad, se olvida que la *indudabilidad* es su condición más radical. La exigencia metodológica de la *duda* aplicable a «todo», incluida la certeza de la matemática, no sólo funda la verdad del primer principio, sino que somete a prueba la verdad de cualquier otro axioma. Con ello Descartes abre el debate de la filosofía moderna en torno al problema de la coincidencia o no de la certeza (pensamiento) y la verdad (cosa). La *duda* misma es la problematización de la identidad entre certeza y verdad (cf. E. Severino, *op. cit.*, p. 44).

des a examinar en tantas partes como fuera posible y necesario para su mejor solución[14].

El tercero, en conducir con orden mis pensamientos, empezando por los objetos más simples y más fáciles de conocer, para ascender poco a poco, gradualmente, hasta el conocimiento de los más complejos y suponiendo incluso un orden entre
19 aquellos / que no se preceden naturalmente unos a otros[15].

Y el último, en hacer en todo enumeraciones tan completas y revisiones tan amplias, que llegase a estar seguro de no haber omitido nada[16].

Las largas cadenas de razones muy simples y

[14] Una «dificultad» (término empleado en el *Discurso*) no es sino una «cuestión» o un conjunto de cuestiones (Regla XIII). De ahí la necesidad de simplificar la dificultad o dividir las cuestiones (precepto 2.º), y la consiguiente enumeración (precepto 4.º), para «conducir con orden» el pensamiento desde lo simple a lo complejo (precepto 3.º). E. Gilson (*op. cit.*, pp. 204 ss.) subraya la complementariedad de estos preceptos. Se ha de tener en cuenta, además, la primacía que Descartes concede al *análisis* sobre la síntesis, así como las demás observaciones hechas en las notas 9 y 15.

[15] ¿Complementariedad del análisis y de la síntesis o primacía del análisis? Al primar éste Descartes no ha comprendido que el método sintético es el que mejor corresponde al proceder *more geométrico* de los *Elementos* de Euclides, según Spinoza (*Tratado de la reforma del entendimiento*, par. 30-51, 105-107). Leibniz, en cambio, defiende una posición de equilibrio, si bien critica todo el planteamiento cartesiano (cf. Y. Belaval, *Leibniz critique de Descartes*, Gallimard, Paris, 1960, cap. III, «La critique des quatre préceptes», pp. 133-198). J. Marrades Millet («Descartes, Newton y Hegel sobre el método de análisis y síntesis», *Pensamiento*, 164, 1985, 393-429) define así la tesis de la complementariedad: «Aun cuando el análisis, como método de descubrimiento, tiene para nosotros una prioridad heurística respecto a la síntesis en metafísica y en física, la síntesis posee siempre una primacía epistemológica sobre el análisis, pues demuestra deductivamente lo que en el análisis sólo vale a título de suposición» (p. 400).

[16] Más importante que las condiciones de la *enumeración*, es tener en cuenta la función que Descartes le asigna en la Regla VII: por una parte, la de constituir una *prueba* de la verdad inducida por medio del análisis entendido como momento inductivo de la investigación; en

fáciles, que los geómetras acostumbran a emplear para llegar a sus demostraciones más difíciles, me habían proporcionado la ocasión de imaginar que todas las cosas que pueden ser objeto de conocimiento humano se encadenan de la misma manera; y que, con sólo abstenerse de admitir como verdadera alguna que no lo sea y guardando siempre el orden necesario para deducir unas de otras, no puede haber algunas tan alejadas de nuestro conocimiento a las que, finalmente, no podamos llegar ni tan ocultas que no podamos descubrir. No me originó excesiva preocupación el buscar por cuáles era preciso comenzar, pues ya sabía que debía ser por las más simples y las más fáciles de conocer; y considerando que, entre todos los que hasta ahora han investigado la verdad en las ciencias, sólo los matemáticos han podido encontrar algunas demostraciones, es decir, algunas razones ciertas y evidentes, en modo alguno dudaba que debía comenzar por las mismas que ellos han examinado, aunque no esperase de ellas ninguna otra utilidad, sino habituar mi ingenio a conocer la verdad y a no contentarse con falsas razones. Mas no por ello he tenido el propósito de procurar aprender todas esas ciencias particulares, que comúnmente se llaman matemáti-
20 cas; / viendo también que aunque sus objetos son diferentes todas coinciden, sin embargo, en que no

este sentido Descartes habla de «enumeración o inducción» (AT, X, pp. 388-389); por otra, la de ser *garantía* del camino del pensamiento, que a través de todos los pasos deductivos o eslabones de la cadena, va desde el primer axioma hasta la última conclusión (AT, X, 389-390). La enumeración, en cualquier caso, pone en juego la memoria para «no omitir nada»: ningún dato informativo, ninguna noción simple, ningún eslabón de la cadena deductiva. Pero ¿tendrá razón Leibniz al objetar que sólo se puede sustituir la debilidad de la memoria, no por la incesante enumeración y revisión, sino por un lenguaje simbólico universal? ¿No es ésta la tendencia actual de los códigos simplificados? (Y. Belaval, *op. cit.*, p. 197).

consideran sino las diversas relaciones o proporciones que en ellos se encuentran, pensé que era más interesante que examinara únicamente esas proporciones en general, suponiéndolas sólo en aquellas materias más idóneas para facilitarme su conocimiento o incluso no vinculándolas a ellas de ninguna manera para poder aplicarlas tanto mejor después a todas aquellas a las que pudiera convenir. Al darme cuenta, luego, de que para conocer tales proporciones tendría unas veces necesidad de considerar cada una en particular, y otras tan sólo comprender varias conjuntamente y retenerlas en la memoria, pensé que, para mejor analizarlas en particular, las tenía que suponer con líneas, pues no encontraba nada más simple ni que pudiera representar más distintamente ante mi imaginación y mis sentidos; y que, para retener o comprender varias conjuntamente, era preciso que las explicara mediante algunas cifras lo más cortas que fuera posible; de este modo, pensé también, tomaba todo lo mejor del análisis geométrico y del álgebra, y corregía todos los defectos de uno por medio del otro[17].

Y, en efecto, me atrevo a decir que el cumplimiento exacto de estos pocos preceptos que había elegido me dio tal facilidad para resolver todas las

[17] Del análisis geométrico conserva la ayuda de la imaginación, dado que trabaja con líneas (ed. lat.: *in lineis rectis*), y del álgebra la brevedad que proporciona el simbolismo, tal como Descartes lo acaba de simplificar. La posibilidad de una matemática universal exige reducir a lo «más simple» las dos ramas más generales de la matemática: la geometría, cuyo objeto es la magnitud o cantidad continua, y la aritmética, cuyo objeto es el número o cantidad discontinua. La reducción se efectúa eligiendo la más simple de las dos cantidades como símbolo de la otra (cf. E. Gilson, *op. cit.*, pp. 219-222). ¿No es éste el trabajo llevado a cabo en su ensayo *Geometria?* Cf. Echeverría, «Nota sobre la Geometría de 1637 y el Método cartesiano», en *Descartes*, de V. Gómez Pin, Dopesa, Barcelona, 1979, Apéndice II, pp. 117-127.

cuestiones tratadas por estas dos ciencias, que en dos o tres meses que empleé en examinarlas, habiendo comenzado por las más simples y generales, y siendo cada verdad que descubría una regla que
21 me / servía luego para encontrar otras, no sólo conseguí resolver algunas cuestiones que otras veces había juzgado muy difíciles, sino que me pareció también, hacia el final, que podía determinar por qué medios y hasta qué punto era posible resolver incluso las que yo ignoraba. Lo cual no puede parecer acaso muy pretencioso, si se considera que no habiendo en matemáticas más que una verdad de cada cosa quien la descubre sabe acerca de ella todo cuanto se puede saber; así, por ejemplo, un niño que sabe aritmética, si ha realizado una suma siguiendo las reglas, puede estar seguro de haber alcanzado con relación a la suma que examinaba todo cuanto el ingenio humano es capaz de hallar. Pues, en definitiva, el método que nos enseña a seguir el verdadero orden y a enumerar exactamente todas las circunstancias de lo que se busca, contiene todo lo que confiere certeza a las reglas de la aritmética.

Pero lo que más me satisfacía de este método era que, por su medio, estaba seguro de usar en todo mi razón, si no de modo perfecto, al menos de la mejor forma que me fuera posible; más aún, me daba cuenta de que la práctica del mismo habituaba poco a poco mi ingenio a conocer más clara y distintamente los objetos, y que, no habiendo limitado este método a una determinada materia en particular, me prometía aplicarlo tan útilmente a las dificultades de otras ciencias como lo había hecho a las del álgebra. No quiero decir con ello que pretendiera examinar todas aquellas dificultades que se presentaran en un primer momento, pues hasta hubiera sido contrario al orden que el método prescribe. Pero, al darme cuenta de que todos los principios de las Ciencias

22 debían tomarse de / la filosofía, en donde aún no hallaba ninguno cierto, pensé que era necesario ante todo que me propusiera investigarlos[18]; pero, siendo esto la cosa más importante del mundo y en la que son más de temer la precipitación y la prevención, juzgué que no debía intentar llevar a cabo tal proyecto hasta no haber alcanzado una madurez mucho mayor que la que se posee a los veintitrés años, que yo tenía entonces[19], y hasta que no hubiese empleado antes mucho tiempo en prepararme para él, tanto desarraigando de mi espíritu todas las malas opiniones recibidas con anterioridad a este momento, como haciendo acopio de varias experiencias destinadas a ser luego la materia de mis razonamientos y ejercitándome siempre en el método que me había prescrito con el fin de afianzarme cada vez más en él.

---

[18] Esta consideración constituye una de las claves arquitectónicas del *Discurso del método*. Se acaba de diseñar un proyecto metodológico que, en el marco de la reflexión sobre lo matemático, define un nuevo criterio de verdad que exige someter a duda «todo» saber existente y, al mismo tiempo, levantar el edificio del saber moderno sobre cimientos o «principios» nuevos. De otro modo, simbolizada *la filosofía* en el nuevo árbol del saber, cuya raíz es la metafísica, el tronco la física y las ramas de las demás ciencias: medicina, mecánica, moral, etc. («Lettre-Préface» a la edición francesa de *Principes*, AT, IX), Descartes se encuentra con el problema de que, habiendo «ensayado» el método en algunas ciencias (geometría, óptica, física), no ha fundamentado la unidad sistemática de la ciencia *desde sus cimientos*, no ha meditado sobre sus *principios*, esto es, no ha hecho metafísica o *Meditationes de prima philosophia* (1641). Por eso justifica ahora la necesidad de una meditación sobre los principios, que aparecerá esbozada en la Parte IV. «Los principios que sólo surgen de la razón de acuerdo al rasgo fundamental matemático del pensar, se convierten en los principios del saber auténtico, es decir, de la filosofía en sentido estricto, de la metafísica» (Heidegger, *La pregunta por la cosa*, p. 97). Cf. S. Gaukroger (ed.), *Descartes: Philosophy, Mathematics and Physic*, Harvester Press, Hassocks, 1980.

[19] En 1619, cuando descubre el principio de la unidad del saber y su conexión con el problema del método.

TERCERA PARTE

Así como para empezar a reconstruir la casa en la que se habita no basta haberla derribado y haber hecho la reserva de materiales y de arquitectos, o haberse ejercitado uno mismo en la arquitectura y, además, haber diseñado cuidadosamente el proyecto, sino que también hay que proveerse de alguna otra en la que se pueda estar alojado cómodamente durante el período de su construcción; de igual modo, con el fin de no permanecer irresoluto en mis acciones mientras la razón me obligara a serlo en mis juicios, y no dejar de vivir desde ese momento lo más felizmente que pudiera, elaboré una moral provisional[1] que constaba solamente de tres o cuatro máximas, que tengo a bien haceros partícipes.

[1] Si la tarea de fundamentación del saber comprende también el saber práctico, la justificación de la «provisionalidad» (ed. fr.: *par provision;* ed. lat. *ad tempus)* radica en el hecho de que en la vida no se puede dejar de tomar decisiones, mientras se lleva a cabo la meditación sobre los principios metafísicos o raíces del árbol de la filosofía antes de estudiar una de sus ramas; la moral. Pero la idea misma de provisionalidad es controvertida: defendida por unos (R. Spaemann, «La morale provisoire de Descartes», *Archives de Philosophie,* 35, 1972, pp. 353-369), es interpretada por otros como un anticipo seguro —*provisión* en sentido jurídico— que se puede incrementar (M. le Doeuf, «En torno a la moral de Descartes», en *Descartes,* de V. Gómez Pin, Dopesa, Barcelona, 1979, Apéndice 1) y, si nos atenemos a la *Conversación con Barman* (*cit.*, p. 179) parece que a Descartes «no le gusta escribir sobre ética». Ahora bien, al subrayar la provisionalidad de este esbozo de moral, aun en la forma de anticipo, ¿no se corre el

La primera consistía en obedecer las leyes y
23 costumbres / de mi país, conservando constantemente la religión en la cual Dios me ha concedido la gracia de ser instruido desde la infancia guiándome en cualquier otra cuestión por las opiniones más moderadas y por las más alejadas de todo extremo, que fuesen comúnmente aceptadas en la práctica por los más sensatos de aquéllos con los cuales tuviera que vivir. Pues, habiendo comenzado desde ese momento a no contar para nada con las mías propias, ya que había decidido someterlas todas a examen, estaba seguro de que lo mejor que podía hacer era seguir las de los más sensatos. Y aun cuando tal vez haya hombres tan sensatos entre los persas y entre los chinos como entre nosotros, me parecía que lo más útil era tomar como regla las opiniones de aquellos con quienes tuviera que vivir; y que para saber cuáles eran verdaderamente sus opiniones debía prestar más atención a lo que practicaban que a lo que decían, no sólo porque dada la corrupción de nuestras costumbres hay pocos que quieran decir todo lo que piensan, sino también porque otros lo ignoran; pues siendo distinto el acto del pensamiento por el que se cree una cosa de aquél por el que se la conoce, con frecuencia se

riesgo tanto de infravalorar su significado histórico como de marginar el desarrollo teórico posterior? En efecto, Descartes ha desarrollado su teoría moral en *Las pasiones del alma* (1649) y en las *Cartas* a Elisabeth (1643-49) y a Cristina de Suecia. Cf. *Lettres sur la morale,* ed. J. Chévalier, Paris, 1935 (trad. esp. en Ed. Yerba Buena, Buenos Aires, 1945); M. Néel, *Descartes et la princesse Elisabeth,* Elzévir, Paris, 1946; E. Cassirer, *Descartes, Corneille, Christine de Suède,* J. Vrin, Paris, 1942; E. Boutroux, «Du rapport de la morale a la science dans la philosophie cartésienne», *Rv. Mét. et Mor.,* 1896; L. Verga, «Ragioni ed esperienza nelle morali di Cartesio e de Cartesiani», *Rv. Intern. de Philos.,* 114, 1975, pp. 453- 475; G. Rodis-Lewis, *La morale de Descartes,* Paris, 1957 (3e. éd., 1970); A. Levi, *French moralist. The theory of the passions 1585 to 1649,* Oxford, 1964.

dan el uno sin el otro. Y, entre varias opiniones igualmente aceptadas, no elegía sino las más moderadas, no sólo porque son siempre las más cómodas en la práctica y probablemente las mejores, pues todo exceso es habitualmente pernicioso, sino también con el fin de alejarme menos del verdadero camino en caso de equivocación que si, habiendo elegido una de las opiniones extremas, hubiera sido la otra la que tendría que haber seguido. Consideraba, particularmente, / como un exceso toda promesa por la cual uno se cercena algo de la propia libertad. No quiero decir que desaprobara las leyes que, para remediar la inconstancia de los espíritus débiles o para consolidar la seguridad del comercio, permiten que uno haga votos o contratos que obligan a perseverar en ellos, tanto cuando se tiene un buen propósito como cuando éste no es sino indiferente; pero, como no veía cosa alguna en el mundo que permaneciera siempre en el mismo estado y como, en lo que me concierne, me prometía perfeccionar cada vez más mis juicios y no empeorarlos, hubiera pensado que cometía una gran falta contra el buen sentido si, por el hecho de haber aprobado entonces alguna opinión, me hubiera obligado también a tener que aceptarla posteriormente como buena, cuando tal vez hubiera dejado de serlo o yo hubiera dejado de estimarla como tal[2].

[2] El tono literalmente acomodaticio y conformista con el que comienza esta máxima contrasta con el *sentido crítico* con el que está escrito el parágrafo. Se llama la atención sobre la «corrupción» de las costumbres, sobre determinadas prácticas que limitan la libertad de pensar. Contra tales abusos sugiere cautelosamente «someter a examen», las opiniones recibidas, rehuir la insensatez de la intolerancia y guiarse por las opiniones más sensatas que emergen de la *vida*. El sentido crítico, es cierto, está más explícito en la propuesta moral contenida en las máximas 3.ª y 4.ª. Cf. V. Peña García, «Acerca de la «razón» en Descartes: reglas de la moral y reglas del método», *Arbor*, 112, 1982, pp. 167-183.

Mi segunda máxima consistía en ser lo más firme y lo más decidido que pudiera en mis acciones, y en seguir con no menos firmeza las opiniones más dudosas, una vez determinado a ello, que si hubieran sido muy seguras[3]. Imitaba en esto a los viajeros que, extraviados en algún bosque, no deben vagar dando vueltas, de un lado para otro, ni mucho menos detenerse en un lugar, sino caminar siempre lo más directamente que puedan hacia el mismo punto, sin sustituirlo por razones nimias, aunque en un principio tal vez haya sido el azar solamente lo que les ha determinado a elegirlo; pues, de este modo, si no llegan precisamente allí donde desean, acabarán llegando / al menos a algún lugar en el

[3] Al señalar que esta máxima ha sido criticada, porque se puede llegar a encubrir el error con un gesto de firmeza, se olvida la respuesta de Descartes en carta de marzo de 1638: «He afirmado algo totalmente diferente, a saber, que es preciso ser decidido en las acciones aunque se permanezca indeciso en los juicios [...]. Y no debe temerse que esta firmeza en la acción nos comprometa cada vez más en el error o en el vicio, puesto que el errar no es propio sino del entendimiento, del cual supongo que, no obstante, permanece libre y considera como dudoso lo que es dudoso. Además *relaciono esta regla con las acciones de la vida que no admiten dilación alguna* y, por otra parte, no me sirvo de ella sino provisionalmente y *con el proyecto de cambiar mis opiniones* tan pronto como pueda encontrar otras mejores y *sin perder ocasión alguna de cambiar* [...]. Por todo ello me parece que no he podido usar de mayor circunspección para situar la decisión, en tanto que es una virtud, entre los dos vicios que son contrarios a la misma: la irresolución y la obstinación» (AT, II, pp. 34-36. La cursiva es mía). Basta la «seguridad moral», según E. Gilson (*op. cit.*, p. 243), cuando no es posible un estado de certeza para guiar nuestra voluntad. Tal seguridad moral, observa G. Quintás (*op. cit.*, p. 431), es capaz de liberarnos de la conciencia desdichada que generaba la teología moral tridentina; pues en definitiva, para Descartes basta con que «nuestra conciencia nos testimonie que nunca hemos carecido de decisión [...] para ejecutar cuantas acciones hemos estimado que eran mejores [...], siendo esto suficiente para que vivamos alegres» (AT, IV, p. 266). ¿No es posible leer esta propuesta en el marco de la transformación de una moral de la contemplación hacia una *ética de la acción*, que Weber ha asignado a la ética protestante?

que probablemente estarán mejor que en medio del bosque. Así también, dado que las acciones de la vida frecuentemente no admiten ningún aplazamiento, es una verdad muy cierta que, cuando no está a nuestro alcance discernir las opiniones más verdaderas, debemos seguir las más probables[4]; asimismo, aunque no observemos mayor probabilidad en unas que en otras, debemos, sin embargo, decidirnos por algunas y considerarlas después, en tanto que referidas a la práctica, no ya como dudosas, sino como muy verdaderas y ciertas, porque tal es la razón que nos ha determinado a ello. Y esto fue suficiente para liberarme desde entonces de todos los arrepentimientos y remordimientos que suelen agitar las conciencias de esos espíritus débiles y vacilantes que, sin constancia, se dejan arrastrar a practicar como buenas las mismas acciones que posteriormente juzgan ser malas.

Mi tercera máxima consistía en intentar siempre vencerme a mi mismo antes que a la fortuna y en cambiar mis deseos antes que el orden del mundo; y, en general, en acostumbrarme a que nada excepto nuestros pensamientos está enteramente en nuestro poder[5], de manera que después de haber

[4] Descartes ha eliminado la probabilidad del orden teórico (cf. Parte II, nota 12), pero no del práctico, por razón de la provisionalidad/provisión señalada en la nota 1, y porque con frecuencia «las acciones de la vida no admiten ningún aplazamiento», es decir, por razón del tiempo necesario para liberar al entendimiento de su opinión probable o dudosa (cf. *Les Principes de la philosophie*, I, pars. 2 y 3; AT, IX, pp. 25-26).

[5] El sentido crítico, enmascarado en la primera máxima, aparece en ésta con toda nitidez: en el silencio de la moral católica en estas páginas —teniendo en cuenta la época— es más elocuente la voz *estoica*. Ante la ausencia de San Agustín o de Santo Tomás es más significativa la presencia del estoicismo, que le llega a través de los *Essais* de Montaigne y, sobre todo, de la escuela de filólogos holandeses de finales del XVI y comienzos del XVII, que lleva a cabo la reconstrucción del estoicismo romano: G. Voss, Heinsius, Scioppius (*Elementa philosophiae stoicae moralis*, 1606) y el más célebre filólogo estoico, J. Lipsius (*De constantia*, 1585; *Manuductio ad stoicam philosophiam*, 1604; *Physiologia stoicorum*) (W. Dilthey, *Hombre y mundo en los siglos XVI y XVII*, t. II, FCE, México, 1944, pp. 434-441). Cf. A. Bridoux, *Le stoïcisme et son influence*, Paris, 1966; G. Rodis-Lewis, *La morale de Descartes*, Paris, 1970. Descartes desarrolla estas mismas ideas en las *Cartas* a Elisabeth (v. la del 4 de agosto de 1645) y en el *Tratado de las pasiones*, arts. 144-146. Cf. J. E. d'Angers, «Senèque, Epictète et le stoïcisme dans l'œuvre de R. Descartes», *Rev. de théol. et philos.*, 1954, pp. 169-196.

hecho lo que hemos juzgado mejor, con relación a los asuntos que nos son ajenos, todo aquello que nos sale mal es absolutamente imposible para nosotros. Este solo pensamiento me parecía suficiente para impedirme desear en lo sucesivo nada que no pudiera adquirir, y para volverme contento. Pues, dado que nuestra voluntad no tiende naturalmente a / desear sino las cosas que nuestro entendimiento le presenta de algún modo como posibles, es cierto que, si consideramos todos los bienes que están fuera de nosotros como igualmente alejados de nuestro poder, no sentiremos mayor disgusto por carecer de aquellos que parecen debidos a nuestro nacimiento, cuando nos veamos privados de ellos por nuestra culpa, que el que se siente por no poseer los reinos de China o de México. Y haciendo, como suele decirse, de la necesidad virtud, no desearemos mucho más estar sanos, estando enfermos, o ser libres, hallándonos encarcelados, como no deseamos ahora poseer cuerpos de una materia tan poco corruptible como los diamantes o alas para volar como los pájaros. Confieso, no obstante, que es preciso un prolongado ejercicio y una meditación frecuentemente reiterada, para acostumbrarse a mirar todas las cosas desde esta perspectiva; pienso también que en esto consistía principalmente el secreto de aquellos fi-

26

lósofos[6] que en otro tiempo fueron capaces de sustraerse al imperio de la fortuna y, a pesar de los dolores y la pobreza, competir en felicidad con sus dioses[7]. Pues, ocupándose sin cesar en considerar los límites que les habían sido prescritos por la naturaleza, tanto se persuadían de que nada tenían en su poder sino sus propios pensamientos, que esto solo era suficiente para impedirles sentir afecto alguno hacía otras cosas; y disponían de sus pensamientos tan absolutamente, que tenían por ello cierta razón para estimarse más ricos y poderosos más libres y más dichosos que cualquiera de los demás hombres que, careciendo de esta filoso-
27 fía, / nunca llegan a disponer de todo lo que quieren, por muy favorecidos de la naturaleza y de la fortuna que pudieran ser.

En fin, como conclusión de esta moral, se me ocurrió hacer una revisión de las diferentes ocupaciones que los hombres tienen en esta vida, con el fin de intentar elegir la mejor; y, sin pretender decir nada de las de los demás, pensé que no podía hacer

[6] Cuando Descartes propone a Elisabeth la moral como tema de reflexión y, en consecuencia, le sugiere «examinar lo que sobre ello han escrito los antiguos», le señala una primera lectura: Séneca, *De vita beata* (a Elisabeth, 21 de julio de 1645); habrá que dilucidar desde Séneca, la oposición que separa las tesis de Zenón y las de Epicuro sobre la felicidad (A Elisabeth, 4 de agosto de 1645); en la discusión también interviene Aristóteles (a Elisabeth, 18 de agosto de 1645) y, por supuesto, Epitecto y Marco Aurelio, autores éstos preferidos por Cristina de Suecia (Chanut a Mazarin, 12 de octubre de 1648). Cf. el art. de J. E. d'Angers (nota 5).

[7] Si un determinado concepto de *felicidad* define una determinada moral, el que aquí propone Descartes no es el de la moral católica, sino un concepto terreno de felicidad. Los supuestos: la *autonomía* del propio pensamiento, es decir, de «todas las operaciones del alma» (AT, II, p. 36), escuchar y seguir la *naturaleza*, una nueva definición de *bien* y de la *norma* suprema de moralidad (véase par. sig.: «conclusión») y, en coherencia con su proyecto teórico el «recto uso de la razón» (a Elisabeth, 4 de agosto de 1645).

nada mejor que seguir en aquella misma que ya tenía, es decir, emplear toda mi vida en cultivar mi razón y avanzar tanto como pudiera en el conocimiento de la verdad, siguiendo el método que me había prescrito. Había experimentado tan vivas satisfacciones desde que comencé a poner en práctica este método, que no creía que se pudieran recibir más agradables ni más inofensivas en esta vida; y como todos los días descubría mediante la práctica del mismo algunas verdades que me parecían bastante importantes y comúnmente ignoradas por otros hombres, la satisfacción que esto me producía colmaba de tal manera mi espíritu que todo lo demás no me afectaba en modo alguno. Además las tres máximas precedentes no estaban fundadas sino en el propósito que tenía de continuar instruyéndome[8]; pues, habiendo dado Dios a cada uno alguna luz[9] para distinguir lo verdadero de lo falso, no habría creído que debiera contentarme ni por un momento con opiniones ajenas, si no me hubiera propuesto emplear mi propio juicio en examinarlas, cuando haya ocasión; y no habría podido librarme de escrúpulos, al seguirlas, si no me hubiera prometido aprovechar toda ocasión de encontrar otras
28 mejores, en caso de que / las hubiere. Y, en fin, no habría podido limitar mis deseos ni sentirme dichoso, si no hubiera seguido un camino por el que pensaba estar seguro de adquirir todos los conocimientos de los que fuera capaz y, al mismo tiempo y por el mismo medio, todos los verdaderos bienes a cuya posesión me es posible aspirar[10]; ya que

[8] La relación entre la reflexión moral y el proyecto de investigar la verdad está más explícita en la ed. lat.: ... *nisi in veritate per hanc Methodum investiganda perseverare decrevissem* (AT, VI, p. 555).

[9] La luz de la razón (ed. lat., *rationis lumen*).

[10] ¿De qué bienes se trata? El *soberano bien*, según Descartes, «consiste en el *ejercicio de la virtud* o, lo que es lo mismo, en la pose-

nuestra voluntad no se determina a seguir o a evitar algo, sino porque nuestro entendimiento se lo representa bueno o malo, basta juzgar bien para obrar bien, y juzgar lo mejor posible, para hacer también lo mejor[11], es decir, para adquirir todas las virtudes y juntamente con ellas todos los bienes que puedan conseguirse; y, cuando se está seguro de que ello es así, no se puede sino estar contento.

sión de *todos los bienes cuya adquisición depende de nuestro libre arbitrio*» (a Elisabeth, 6 de octubre de 1645). Cierto, para Descartes —y en esto se distancia del estoicismo— la libertad es la condición de toda moralidad. Pero en el *Discurso*, la consecución de «todos los verdaderos bienes» no se hace depender de la libertad, sino de la *razón*, es decir, del «mismo medio» con el que adquirimos todos nuestros conocimientos. Mediante la razón como capacidad de juzgar se conoce la verdad, pero también es la razón la que guía hacia el bien. En esta tesis no difiere de los estoicos. «Por lo cual, la verdadera función de la razón consiste en examinar el justo valor de todos los bienes cuya adquisición parece depender de alguna manera de nuestra conducta, con el fin de que nunca dejemos de emplear toda nuestra atención en procurarnos los que, en efecto, son más deseables; de modo que, si la fortuna se opone a nuestros propósitos y nos impide lograrlos, tendremos al menos la satisfacción de que no ha sido debido a nuestro descuido y no dejaremos de gozar de toda la *dicha natural* cuya adquisición habrá estado en nuestro poder» (a Elisabeth, 1 de noviembre de 1645). He subrayado «dicha natural» (*béatitude naturelle*) o satisfacción interior profunda y durable, porque presupone el *soberano bien*, esto es, la posesión de todos los bienes cuya adquisición depende de la *libertad* y, en definitiva, de la *razón:* la felicidad, según Descartes, depende ante todo del «recto uso de la razón» (a Elisabeth, 4 de agosto de 1645).

[11] Esta *norma de la moralidad*, «basta juzgar bien, para obrar bien», es válida siempre que el juicio sea evidentemente cierto. De otro modo, la filosofía cartesiana no admite la determinación de la voluntad por el entendimiento, sino en el caso en el que el juicio por el que se regula la voluntad sea evidente y cierto. Pero en el caso frecuente en el que a la voluntad se le presentan varias alternativas al mismo tiempo, Descartes propone la segunda fórmula: «[...] y juzgar lo mejor posible para hacer también lo mejor„ En estos términos se expresa Descartes al responder a la objeción de Mersenne (a Mersenne, 27 de abril de 1637). Tal norma moral está cargada de significado crítico de la moral tradicional, por el silencio elocuente del recurso a la ayuda de la gracia, para obrar bien.

Después de haberme provisto de estas máximas y de haberlas puesto aparte junto con las verdades de la fe, que siempre han sido para mí las más importantes, pensé que podía intentar libremente deshacerme de todas las otras opiniones. Y, como esperaba poder conseguirlo mejor conversando con los hombres que permaneciendo por más tiempo encerrado junto a la estufa donde había tenido todos estos pensamientos, continué mi viaje aunque el invierno no había terminado todavía. Y en los nueve años siguientes[12] no hice otra cosa sino andar por el mundo de acá para allá, intentando ser más bien espectador que actor en todas las comedias que en él se representan; y reflexionando, en cada materia, particularmente sobre aquello que pudiera hacerla dudosa y dar ocasión a equivocarnos, erradicaba entre tanto de mi espíritu cuantos errores se
29 hubieran podido / deslizar en él anteriormente. No es que imitara en esto, sin embargo, a los escépticos[13], que no dudan sino por dudar y fingen ser siempre indecisos; pues mi único deseo, al contrario, sólo consistía en llegar a descubrir algo firme, apartando la tierra movediza y la arena con el fin de encontrar la roca o la arcilla. Lo cual conseguía, a mi parecer, bastante bien, ya que intentando descubrir la falsedad o incertidumbre de las proposicio-

[12] Del 10 de noviembre de 1619, fecha de su gran descubrimiento, hasta el 8 de octubre de 1628, fecha de su segundo encuentro con Beeckman (cf. *Journal* de I. Beeckman, AT, X, p. 331). En el par. siguiente alude de nuevo a «aquellos nueve años».

[13] Se refiere principalmente a F. Sánchez (*Que nada se sabe*, 1581) y a Montaigne (*Ensayos* I y II, 1580; III, 1588). Cf. R. H. Popkin, *Historia del escepticismo desde Erasmo hasta Spinoza*, FCE, México, 1983. Descartes, al distanciarse de la tesis escéptica, delimita el alcance de la actitud de duda que practica como exigencia de evidencia y certeza. Es significativo que reitere este distanciamiento nada más formular el primer principio (v. Parte IV), al que ha llegado tras la exigencia al límite de la actitud de duda.

nes que examinaba, no mediante débiles conjeturas sino siguiendo razonamientos evidentes y seguros, no hallaba ninguna tan dudosa que no pudiera sacar de ella alguna conclusión bastante cierta, aunque sólo fuese la de que no contenía nada cierto. Y así como cuando se derriba una casa vieja se conservan generalmente los materiales para aprovecharlos en la construcción de una nueva, de igual modo cuando destruía todas aquellas opiniones propias que consideraba mal fundadas, hacía diferentes observaciones y adquiría algunas experiencias[14], que me han servido más tarde para establecer otras más ciertas. Y continuaba, además, ejercitándome en el método que me había propuesto; pues, además de preocuparme habitualmente de conducir mis pensamientos según sus reglas, reservaba en ocasiones algunas horas que dedicaba especialmente a ponerlo en práctica en dificultades de matemática o, incluso también, en algunas otras que podía considerar casi semejantes a las de las matemáticas, separándolas de los principios de otras ciencias que no estimaba suficientemente firmes, como veréis que he hecho en algunas cuestiones que han sido tratadas en el presente volumen. De este modo, no
30 viviendo / en apariencia sino como los que no tienen otra ocupación que la de llevar una vida agradable e inocente, aplicándose a separar los placeres de los vicios, y que para disfrutar de su ocio sin aburrirse no se privan de ninguna diversión honesta, no dejaba de perseverar en mi propósito y de avanzar en el conocimiento de la verdad, tal vez más que si me hubiese limitado a leer libros o a frecuentar gentes de letras.

[14] Sobre el papel de la *experiencia* en la ciencia cartesiana, véase: D. M. Clarke, *La filosofía de la ciencia en Descartes*, cap. 2, pp. 30-57. L. Verga, «Ragione ed esperienza nelle morali di Cartesio e de Cartesiani», *Revue Internationale de Philosophie*, 114, 1975, pp. 453-475.

Sin embargo, pasaron esos nueve años sin que hubiera llegado a tomar partido alguno con relación a las dificultades que suelen ser discutidas entre los doctos y sin que hubiera comenzado a buscar los fundamentos de una filosofía más cierta que la vulgar[15]. El ejemplo, por otra parte, de algunos ingenios excelentes que habiendo concebido antes el proyecto me parece que no han llegado a realizarlo, me hacía imaginar tanta dificultad en ello, que tal vez aún no me hubiera atrevido a emprenderlo tan pronto, si no me hubiera enterado de que algunos hacían correr ya el rumor de que lo habían concluido. No me es posible decir en qué fundaban tal opinión, y si en algo he contribuido a ella a través de mis conversaciones, debe haber sido por haber confesado lo que ignoraba más ingenuamente de lo que acostumbran a hacer los que han estudiado un poco o, quizás también, por haber dado a conocer las razones que tenía para dudar de muchas cosas que los demás consideran ciertas, y no por haberme vanagloriado de doctrina alguna. Pero, teniendo valor suficiente como para no querer que se me tome por lo que no soy, pensé que era preciso intentar por
31 todos los medios hacerme digno de la reputación / que se me daba; y hace precisamente ocho años[16] que a causa de este deseo decidí alejarme de todos los lugares de donde podía tener amistades y retirarme a un país, en el que la larga duración de la guerra[17] ha hecho posible instituir tales ordenanzas, que los ejércitos que en él se mantienen pare-

[15] Es decir, más cierta que la filosofía escolástica.

[16] Si terminó de escribir el *Discurso* en 1636, la referencia nos remite a 1628, esto es, al regreso de Descartes a Holanda (cf. nota 12).

[17] Se trata de la guerra, en los Países Bajos, que comienza con la rebelión de Guillermo de Orange en 1568 contra España, y termina con el reconocimiento de la independencia holandesa en la Paz de Westfalia (1648).

cen servir únicamente para que sus habitantes gocen de los frutos de la paz con mucha más seguridad, y en donde, en medio de la multitud de un pueblo muy activo, más atento a los propios negocios que curioso de los ajenos, he podido vivir tan retirado y solitario como en los desiertos más apartados, sin carecer de ninguna de las comodidades que se tienen en las ciudades más frecuentadas.

CUARTA PARTE

No sé si debo entreteneros con las primeras meditaciones que allí he hecho, pues son tan metafísicas y tan fuera de lo común que tal vez no sean del gusto de todos[1]. Sin embargo, con el fin de que se pueda apreciar si los fundamentos que he establecido son bastante firmes, me veo en cierto modo obligado a hablar de ellas. Desde hace mucho tiempo había observado que, en lo que se refiere a las costumbres, es a veces necesario seguir opiniones que tenemos por muy inciertas como si fueran indudables, según se ha dicho anteriormente[2]; pero, dado que en ese momento sólo pensaba dedicarme a la

[1] Cuando Hegel llama a Descartes «héroe» del pensamiento, no creo que lo haga subrayando el término *metafísica* en el sentido en el que lo hacen Gilson y Alquié, esto es, como reflexión abstracta, aun concediendo que «la metafísica es una ciencia que casi nadie entiende» (AT, II, 570, 18-20). Más bien lo haría subrayando con Heidegger el término meditación, es decir, una «determinada interpretación de lo existente» y una «determinada concepción de la verdad», que no sólo convierte en «lo más discutible la verdad de los propios axiomas y el ámbito de los propios fines», sino además «funda una época al darle un fundamento de su figura esencial» (*Sendas perdidas*, p. 68). El título *Meditationes de prima philosophia* (1641) responde a esta versión. Y aunque evoque la interrogación de la πρωτη φιλοσοφια aristotélica, lo innovador de la meditación cartesiana consiste en que somete a *duda* precisamente la tradición que se apoya en Aristóteles —tal vez por esto no sea «del gusto de todos»— y funda la filosofía (metafísica) moderna al establecerla sobre «fundamentos» o principios nuevos. Cf. J. Vuillemin, *Mathématiques et métaphysique chez Descartes*, PUF, Paris, 1960.

[2] Véase Parte III, nota 4.

investigación de la verdad[3], pensé que era preciso que hiciera lo contrario y rechazara como absolutamente falso todo aquello en lo que pudiera imaginar la menor duda, con el fin de comprobar si, hecho esto, no quedaba en mi creencia algo que fuera en-
32 teramente indudable. Así, / puesto que nuestros sentidos nos engañan algunas veces, quise suponer que no había cosa alguna que fuera tal como nos la hacen imaginar. Y como existen hombres que se equivocan al razonar, incluso en las más sencillas cuestiones de geometría, y cometen paralogismos, juzgando que estaba expuesto a equivocarme como cualquier otro, rechacé como falsos todos los razonamientos que había tomado antes por demostraciones. Y, en fin, considerando que los mismos pensamientos que tenemos estando despiertos pueden venirnos también cuando dormimos, sin que en tal estado haya alguno que sea verdadero, decidí fingir que todas las cosas que hasta entonces habían entrado en mi espíritu no eran más verdaderas que las ilusiones de mis sueños[4]. Pero, inmediatamente después, advertí que mientras quería pensar de ese modo que todo es falso, era absolutamente necesario que yo, que lo pensaba, fuera alguna cosa. Y observando que esta verdad: *pienso, luego soy*[5], era

[3] En el momento en el que esbozaba el problema práctico de una moral, Descartes se ocupaba preferentemente del problema teórico de investigar la verdad de las ciencias, sobre la base de un fundamento metafísico.

[4] El proceso de *duda* que conduce al descubrimiento y formulamos del primer principio está mejor desarrollado en *Meditaciones metafísicas*, Alfaguara, Madrid, 1977, 1.ª y 2.ª.

[5] La fórmula de la ed. fr. *je pense, donc je suis*, ha sido traducida al latín de este modo: *ego cogito, ergo sum, sive existo* (AT, VI, p. 558). Según Gilson, la adición *existo* se explica por la dificultad de usar el verbo latino *sum* en el sentido de *existir* que sugiere el verbo *être* en francés (*op. cit.*, p. 292). Esta matización apoya la interpretación de Heidegger expuesta en la nota 1. En el comienzo de la tercera medita-

tan firme y tan segura[6] que todas las más extravagantes suposiciones de los escépticos no eran capaces de socavarla, juzgué que podía admitirla como el primer principio[7] de la filosofía que buscaba.

Al examinar, después, atentamente lo que yo era, y viendo que podía fingir que no tenía cuerpo y que no había mundo ni lugar alguno en el que me encontrase, pero que no podía fingir por ello que yo no existía, sino que, al contrario, del hecho mismo de pensar en dudar de la verdad de otras cosas se seguía muy evidente y ciertamente que yo era;
33 mientras, que, con sólo haber / dejado de pensar, aunque todo lo demás que alguna vez había imagi-

ción puede verse, en cambio, el significado de *pensar-cogitare*. Por otra parte, a la objeción que le hace Gassendi, según la cual el enunciado es la conclusión de un silogismo, Descartes niega que se trate de un razonamiento, porque no es necesario suponer premisa alguna mayor. Sobre la posible conexión del enunciado cartesiano con el de S. Agustín, *si enim fallor, sum* (*De lib. arbitr.*, 2, 3, 7), Descartes rechaza tal conexión (a Mersenne, 25 de mayo de 1637) y, en cualquier caso, en S. Agustín no tiene la función de primer principio o cimiento del edificio del saber moderno. Cf. L. Blanchet, *Les antécédents historiques du* «Je pense, donc je suis», Paris, 1920 (reed. J. Vrin, 1985).

[6] En la ed. lat. se hace referencia más explícita al nuevo modelo de *verdad* definido en términos de *evidencia: adeo certam esse atque evidentem* (AT, VI, p. 558). Fundamentada esta primera verdad, Descartes se distancia de nuevo de la posición escéptica (cf. Parte III, nota 13).

[7] En respuesta a algunas objeciones, Descartes puntualiza lo que debe entenderse por *primer principio* (a Clerselier, jun.-jul. 1646; AT, IV, pp. 443-445). Este *premier principe* (ed. fr.) o *primum fundamentum* (ed. lat.) es interpretado por Heidegger en los términos siguientes: «El principio supremo es el principio-proposición del yo (*Ichsatz*): *cogito-sum*. Es el axioma fundamental de todo saber, pero no el único axioma fundamental», ya que la razón, en tanto que el pensar es su acto fundamental, y el principio de no contradicción, en cuanto pertenece a la esencia misma del pensar, constituyen el orden del *fundamento* de todo saber (*La pregunta por la cosa*, pp. 96-97). Cf. Husserl, *Méditationes cartésiennes*, p. 6. La originalidad de la filosofía cartesiana radica en el modo de reflexión que le lleva a fundamentar el edificio del saber en este *primer principio*.

nado existiera realmente, no tenía ninguna razón para creer que yo existiese, conocí por ello que yo era una sustancia cuya esencia o naturaleza no es sino pensar[8], y que, para existir, no necesita de lugar alguno ni depende de cosa alguna material. De manera que este yo, es decir, el alma por la cual soy lo que soy, es enteramente distinta del cuerpo e incluso más fácil de conocer que él y, aunque el cuerpo no existiese, el alma no dejaría de ser todo lo que es[9].

Después de esto, examiné lo que en general se requiere para que una proposición sea verdadera y cierta, pues, ya que acababa de descubrir una que sabía que lo era, pensé que debía saber también en qué consiste esa certeza. Y habiendo observado que no hay absolutamente nada en *pienso, luego soy* que me asegure que digo la verdad, a no ser que

[8] En la 2.ª de las *Meditaciones metafísicas* se aclara el término *pensar*: «¿Qué soy yo, entonces? Una cosa que piensa. ¿Y qué es una cosa que piensa? Es una cosa que duda, que entiende, que afirma, que niega, que quiere, que no quiere, que imagina también y que siente» (AT, VII, p. 28; trad. de V. Peña, p. 26). Ahora bien, al afirmar que esta «cosa que piensa» (*res cogitaos*) es una sustancia radicalmente distinta de la cosa extensa (*res extensa*), Descartes tiene que plantear los problemas desde un dualismo ontológico, que será criticado bien por el pluralismo, como el de Leibniz, bien por el monismo, como el de Spinoza. Según Heidegger, lo que Descartes «deja indeterminado en este comienzo "radical" es la forma de ser de la *res cogitans*, o más exactamente, el *sentido del ser del "sum"*» [*Ser y tiempo*, FCE, México, 1974 (5.ª), pp. 34-35].

[9] El dualismo antropológico aquí esbozado es la consecuencia lógica del dualismo ontológico. Al exponer la naturaleza del «yo pienso», Descartes necesita distinguirlo y diferenciarlo del cuerpo y al contrario, al explicar el cuerpo desde los principios mecanicistas de su teoría física, necesita igualmente diferenciarlo del yo o del alma (*Discurso*, Parte V; *Tratado del hambre*, trad. de G. Quintás, p. 117). Sobre el problema de la unión del alma y del cuerpo, véase *Las pasiones del alma*, arts. 29-34. J. Broughton y R. Mattern, «Reinterpreting Descartes on the notion of the union of mind and body» *Journal of the History of Philosophy*, XVI:I, 1978. pp. 23-33.

veo muy claramente que para pensar es preciso ser, juzgué que podía admitir esta regla general: las cosas que concebimos muy clara y distintamente son todas verdaderas[10], si bien sólo hay alguna dificultad en identificar exactamente cuáles son las que concebimos distintamente.

Reflexionando, a continuación, sobre el hecho de que yo dudaba y que, por lo tanto, mi ser no era enteramente perfecto, pues veía con claridad que había mayor perfección en conocer que en dudar, se me ocurrió indagar de qué modo había llegado a pensar en algo más perfecto que yo; y conocí con
34 evidencia que debía ser / a partir de alguna naturaleza que, efectivamente, fuese más perfecta. Por lo que se refiere a los pensamientos que tenía de algunas otras cosas exteriores a mí como el cielo, la tierra, la luz, el calor, y otras mil, no me preocupaba tanto por saber de dónde procedían, porque, no ob-

[10] Esta regla de verdad o, mejor, la que formula en los mismos términos al comienzo de la 3.ª *Meditación*, ha suscitado algunas objeciones por ejemplo, las de Gassendi (*Disquisitio... adversus Renati Cartesii Metaphysicam et Responsa*, Amsterdam, 1644) y las de Huet (*Censura philosophiae cartesianae*). En respuesta a Gassendi, esto es, a las Quintas Objeciones, Descartes escribe: «Y, por último, hay mucha verdad en lo que añadís: que no deberíamos trabajar tanto por confirmar la verdad de dicha regla como por conseguir un buen método para saber si nos engañamos o no cuando pensamos concebir claramente una cosa; pero yo mantengo que eso es lo que he hecho en su lugar» (ed. de Vidal Peña, Alfaguara, p. 287). Ahora bien, si en las *Regulae* (III, IV) y en la primera regla del método ha establecido ya las condiciones de un nuevo modelo de verdad, ¿qué significa esta reiteración ahora de la regla de verdad? Cabe pensar que, si la interpretación de la verdad que propone está vinculada a una determinada concepción de lo existente, al hacer explícita esta vinculación del *pensar* al *ser* («no hay absolutamente nada en *pienso, luego existo* que me asegure que digo la verdad, a no ser que veo muy claramente que *para pensar es preciso ser*»), Descartes no hace sino matizar su meditación sobre la verdad. Obsérvese, además, que esta formulación es incompleta con relación a la de la primera regla (Parte II), dado que sólo en ésta aparece la condición más radical de la verdad: la indudabilidad.

servando en tales pensamientos nada que me pareciera hacerlos superiores a mí, podía pensar que, si eran verdaderos era por ser dependientes de mi naturaleza en tanto que dotada de cierta perfección; y si no lo era, que procedían de la nada, es decir, que los tenía p(que había en mi imperfección. Pero no podía suceder lo mismo con la idea[11] de un ser más perfecto que el mío; pues, que procediese de la nada era algo manifiestamente imposible; y puesto que no es menos contradictorio pensar que lo más perfecto sea consecuencia y esté en dependencia de lo menos perfecto, que pensar que de la nada provenga algo, tampoco tal idea podía proceder de mí mismo. De manera que sólo quedaba la posibilidad de que hubiera sido puesta en mí por una naturaleza que fuera realmente más perfecta que la mía y que poseyera, incluso, todas las perfecciones de las que yo pudiera tener alguna idea, esto es, para decirlo en una palabra, que fuera Dios[12]. Añadí a esto que,

[11] La ed. lat. identifica *idea y pensamiento* («de cogitatione sive idea», AT, VI, p. 559). En las Definiciones previas a las «Razones que prueban la existencia de Dios» leemos: «Con el nombre de *pensamiento* comprendo todo lo que está en nosotros de modo tal que somos inmediatamente conscientes de ello» (def. I). «Con la palabra *idea*, entiendo aquella forma de todos nuestros pensamientos, por cuya percepción inmediata tenemos consciencia de ellos» (def. II). Al definir la idea como forma del pensamiento, con el término *forma* —de herencia aristotélica— no se designa el aspecto sensible de una cosa, sino lo que ésta tiene de *representativa*. En efecto, al ser de la cosa representada por la idea denomina Descartes «realidad objetiva de una idea»; el ser de la cosa representada puede ser también una «perfección objetiva» (def. III), como sucede en el texto (*Resp. II Objec., Meditaciones*, pp. 129-130: AT, VII, pp. 160-161).

[12] Para Descartes probar la existencia de un ser perfecto y probar la existencia de Dios es lo mismo. Cf. *Meditaciones*, 3.ª, p. 44; AT, VII, p. 51. También: «Llamamos *Dios* a la substancia que entendemos supremamente perfecta, y en la cual nada concebimos que incluya defecto alguno, o limitación de la perfección» (def. VIII, *Resp. II Obj.*, p. 130: AT. VII, p. 162).

puesto que conocía algunas perfecciones que en modo alguno tenía, no era yo el único ser que existiese (usaré aquí libremente, si me lo permitís, términos de la escuela), sino que era absolutamente necesario que existiera otro ser más perfecto, de quien yo dependiese y del que hubiese adquirido todo lo que tenía. Pues, si yo hubiera existido solo e independientemente de todo otro, de tal manera que
35 de mí mismo / procediese todo lo poco que participaba del ser perfecto, por idéntica razón hubiera podido tener por mí mismo todo lo demás que sabía que me faltaba y, de este modo, ser yo mismo infinito, eterno, inmutable, omnisciente, omnipotente y, en fin, tener todas las perfecciones que según podía comprender existen en Dios. Siguiendo, pues, los razonamientos que acabo de hacer, para conocer la naturaleza de Dios hasta donde sea posible a la mía, sólo tenía que considerar si era perfección o no poseer todas las cosas de las cuales hallaba en mí alguna idea, y estaba seguro de que ninguna de las que indicaban alguna imperfección estaba en él, pero si todas las demás. Veía de este modo que la duda, la inconstancia, la tristeza, y cosas semejantes no podían convenir a Dios, dado que yo mismo hubiera sido muy dichoso si estuviera libre de ellas. Además de esto, tenía ideas de algunas cosas sensibles y corporales; pues, aunque supusiese que soñaba y que era falso todo lo que veía o imaginaba, no podía negar, sin embargo, que tales ideas estuvieran verdaderamente en mi pensamiento. Pero como había conocido en mí muy claramente que la naturaleza inteligente es distinta de la corporal, considerando que toda composición expresa dependencia y que la dependencia es manifiestamente un defecto, juzgaba por ello que en Dios no podía ser una perfección el estar compuesto de estas dos naturalezas y que, por consiguiente, no lo estaba; en

cambio, si en el mundo existían algunos cuerpos, o bien algunas inteligencias, u otras naturalezas que
36 no fueran totalmente / perfectas, su ser debía depender del poder divino de tal forma que éstas no podrían subsistir sin él ni un solo momento[13].

Quise buscar, después, otras verdades y, habiéndome propuesto el objeto de los geómetras, que concebía como un cuerpo continuo o un espacio indefinidamente extenso[14] en longitud anchura y altura o profundidad, divisible en diversas partes, que podían tener diferentes figuras y tamaños, y ser movidas o trasladadas de todas las maneras posibles, pues los geómetras suponen todo esto en su objeto, repasé algunas de sus más simples demostraciones. Y habiendo advertido que la gran certeza que todo el mundo les atribuye sólo está fundada en que se las concibe con evidencia, siguiendo la regla antes formulada[15], advertí también que no había en ellas absolutamente nada que me asegurase la exis-

[13] Alusión a la teoría de la creación continua: el acto creador de Dios no sólo ha tenido lugar en el origen del mundo, sino también en cada instante. Dos supuestos de esta teoría: 1) la relación de dependencia del *ens creatum* respecto del *ens increatum*, 2) la noción de tiempo discontinuo, tal como la expone en *Principios*, I, par. 21 (AT, IX, p. 34) y en el Axioma II de *Resp. II Objec., cit.*, p. 133 (AT, IX, p. 164). Cf. J. Wahl, *Du rôle de l'idée d'instant dans la philosophie de Descartes*, Paris, 1920.

[14] La expresión francesa *indéfiniment étendu* no puede traducirse por «infinitamente extenso», pues Descartes mismo precisa: «Hago aquí distinción entre *indefinido* e *infinito*. Sólo llamo infinito, hablando con propiedad, a aquello en que en modo alguno encuentro límites, y, en este sentido, sólo Dios es infinito. Pero aquellas cosas en las que sólo bajo cierto respecto no veo límite —como la extensión de los espacios imaginarios, la multitud de los números, la divisibilidad de las partes de la cantidad, y cosas por el estilo— las llamo *indefinidas*, y no *infinitas*, pues no en cualquier sentido carecen de límites» *(Resp. I Objec., cit.*, p. 95; AT, VII, p. 113). Cf. A. Koyré (1957), *Du monde clos à l'univers infini*, Gallimard, Paris, 1973; Th. S. Kuhn (1957), *La revolución copernicana*, Ariel, Barcelona, 1978 pp. 298 ss.

[15] Cf. Parte II, primera regia, y nota 12; Parte IV, nota 10.

tencia de su objeto. Porque, por ejemplo, veía bien que, si suponemos un triángulo, sus tres ángulos tienen que ser necesariamente iguales a dos rectos, pero en tal evidencia no apreciaba nada que me asegurase que haya existido triángulo alguno en el mundo. Al contrario, volviendo a examinar la idea que tenía de un ser perfecto, encontraba que la existencia estaba comprendida en ella del mismo modo que en la de un triángulo está comprendido el que sus tres ángulos son iguales a dos rectos, o en la de una esfera, el que todas sus partes equidistan de su centro, e incluso con mayor evidencia; y, en consecuencia, es al menos tan cierto que Dios, que es ese ser perfecto, es o existe, como puede serlo cualquier demostración de la geometría[16].

37 Pero si hay algunos que están persuadidos de que es difícil conocerle y aun de conocer la naturaleza del alma, es porque no elevan nunca su pensamiento más allá de las cosas sensibles y porque están tan habituados a no considerar nada si no es mediante la imaginación —que es su peculiar modo de pensar las cosas materiales— que todo lo que no es imaginable les parece ininteligible. Lo cual está suficientemente patente en lo que hasta los filósofos admiten como máxima en las escuelas: nada hay en el entendimiento que no haya estado previamente en los sentidos, en donde, no obstante, es cierto que las ideas de Dios y del alma nunca han estado. Y me parece que los que quieren hacer uso de la imaginación para comprenderlas, obran del mismo

[16] Sobre el significado de esta prueba, v. E. Gilson, *op. cit.*, pp. 347-353. Después de la demostración *a posteriori*, de tradición tomista, Descartes reproduce la prueba *a priori* de S. Anselmo (*Proslog*, cap. II), esto es, el argumento ontológico, que será criticado por Kant (A 592 y B 620 a A 602 y B 630). F. Duque, «Sentido del argumento ontológico en Descartes y Leibniz», *Pensamiento*, 42. 1986, pp. 159-180.

modo que si para oír los sonidos o sentir los olores quisieran servirse de sus ojos; pero aún hay otra diferencia: que el sentido de la vista no nos asegura menos de la verdad de sus objetos que el olfato y el oído de los suyos, mientras que ni nuestra imaginación ni nuestros sentidos podrían asegurarnos nunca de cosa alguna si no interviene en ello nuestro entendimiento.

En fin, si aún hay hombres a quienes las razones que he presentado no han convencido suficientemente de la existencia de Dios y del alma, quiero que sepan que todas las demás cosas, de las que tal vez piensan estar más seguros, como tener un cuerpo, que hay astros y una tierra, y cosas semejantes, son menos ciertas.

Pues, aunque se tenga una seguridad moral[17] de
38 estas cosas tal que parece / que, a menos de ser un extravagante, no se puede dudar de ellas, sin embargo, cuando se trata de una certeza metafísica, tampoco se puede negar, a menos que uno sea poco razonable, que sea motivo suficiente para no estar totalmente seguros de haber advertido que uno puede imaginar de la misma manera, estando dormido, que se tiene otro cuerpo, que se ven otros astros y otra tierra, sin que ninguna de estas cosas existan. Porque ¿cómo es posible saber que los pensamientos que se nos ocurren en sueños son más falsos que los demás, si frecuentemente no son menos vivos ni menos precisos? Y, por mucho que lo estudien los mejores ingenios, no creo que puedan dar razón alguna que sea suficiente para disipar esa duda, si no presuponen la existencia de Dios. Pues, en primer lugar, incluso lo que anteriormente he

[17] «Seguridad moral» es la certeza suficiente para regular la vida práctica, aunque no sea teóricamente cierta (cf. *Principios*, IV, par. 205-206; AT, IX, pp. 323-325).

adoptado como una regla, a saber, que las cosas que concebimos muy clara y distintamente son todas verdaderas, no es cierto sino porque Dios es o existe, porque es un ser perfecto, y porque todo lo que hay en nosotros procede de él[18]. De donde se sigue que nuestras ideas o nociones, en tanto que son claras y distintas, siendo cosas reales, y procediendo de Dios, no pueden ser por ello sino verdaderas. De modo que, si con bastante frecuencia tenemos ideas que encierran falsedad, es tal vez porque hay en ellas algo confuso y oscuro, ya que en esto participan de la nada, es decir, que no se dan tan confusas en nosotros, sino porque no somos enteramente perfectos. Y es evidente que no hay menor contradicción en pensar que la falsedad o la
39 imperfección, / en tanto que tal, procede de Dios, que en pensar que la verdad o la perfección procede de la nada. Pero si no supiéramos que todo cuanto

[18] En la cuarta serie de *Objeciones* Arnauld acusa a Descartes de «haber cometido círculo vicioso, cuando dice que *sólo estamos seguros de que son verdaderas las cosas que concebimos clara y distintamente, en virtud de que Dios existe*. Pues no podemos estar seguros de que existe Dios, si no concebimos eso con toda claridad y distinción» (*Resp. IV Objec., cit.*, p. 174; AT, VII, p. 214). La respuesta de Descartes en la que puntualiza que la certeza de que Dios existe garantiza el *recuerdo* de lo que ha sido probado con claridad y distinción no convence (*Resp. Objec., cit.*, II, p. 115, IV, p. 197; AT, VII, pp. 140 y 246). No convence al mismo Descartes, quien en la *Conversación con Burman* propone, para activar la memoria, «hacer uso de notas escritas o algo» similar; pues lo que garantiza la veracidad divina, no es el recuerdo correcto, sino el de no engañarme al pensar que son verdaderas aquellas proposiciones que recuerdo haber percibido clara y distintamente (*Entretien avec Burman*, ed. Ch. Adam, Paris, 1937, pp. 8-9; trad. esp. *cit.*, pp. 128-129). Sin embargo, coincidimos con la interpretación no teológica de este pasaje hecha por Vidal Peña: «Postular a Dios significa postular las condiciones que hacen posible la racionalidad» (Introducción a *Meditaciones metafísicas*, p. XXXVI). Sobre el problema del posible círculo vicioso, véase el debate en los trabajos reseñados por G. Rodis-Lewis, *L'Oeuvre de Descartes*. t. II, p. 528, n. 59.

en nosotros es real y verdadero proviene de un ser perfecto e infinito, por claras y distintas que fuesen nuestras ideas, no tendríamos razón alguna que nos asegurase que tienen la perfección de ser verdaderas.

Ahora bien, después de que el conocimiento de Dios y del alma nos ha probado así la certeza de aquella regla, es muy fácil conocer que los sueños que imaginamos cuando dormimos, no deben hacernos dudar de la verdad de los pensamientos que tenemos cuando estamos despiertos. Pues, si ocurriera que, incluso mientras dormimos, tuviéramos alguna idea muy distinta como, por ejemplo, que un geómetra inventase alguna nueva demostración, su sueño no impediría que fuese verdadera. Y en cuanto al error más común de nuestros sueños, que consiste en representarnos diversos objetos del mismo modo que lo hacemos mediante los sentidos externos, importa poco que nos dé ocasión para desconfiar de la verdad de tales ideas, ya que éstas también pueden engañarnos con bastante frecuencia aunque no estemos dormidos: como cuando los que tienen la ictericia lo ven todo de color amarillo, o cuando los astros u otros cuerpos muy alejados nos parecen mucho más pequeños de los que son. Pues, en fin, ya estemos despiertos o ya estemos dormidos, no debemos dejarnos persuadir nunca si no es por la evidencia de nuestra razón[19]. Y se ha de subrayar que digo por nuestra razón, y no por nuestra imagi-

[19] En la ed. lat.: *solam evidentiam rationis judicia nostra sequi debent* (AT, VI, p. 562). Cf. W. Thayer, «Descartes, la vigilancia del sueño», *Revista de Filosofía* (Chile), 22-24, 1984, pp. 99-108; H. G. Frankfurt, *Demons, dreamers and madmen: The defense of reason in Descartes' Meditations*, Bobbs, Merril, Indianapolis, 1970; G. Bachelard, *La formation de l'esprit scientifique*. Contribution a une psychanalyse de la connaissance objective, J. Vrin, Paris, 1972 (8e. éd.); trad. esp. en Siglo XXI.

nación ni por nuestros sentidos. Del mismo modo,
40 aunque veamos el sol / muy claramente, no debemos por ello juzgar que no sea sino del tamaño que lo vemos; como podemos muy bien imaginar distintamente una cabeza de león pegada al cuerpo de una cabra, sin que por ello haya que concluir que exista en el mundo una quimera; pues la razón no nos impone que lo que vemos o imaginamos de este modo sea verdadero. Pues nos ordena que todas nuestras ideas o nociones deben tener algún fundamento de verdad; pues no sería posible que Dios, que es sumamente perfecto y veraz, las hubiera puesto en nosotros sin tal fundamento. Y puesto que nuestros razonamientos no son nunca tan evidentes ni completos cuando soñamos como cuando estamos despiertos, si bien a veces nuestras imaginaciones son en aquel caso tanto o más vivas y precisas, la razón nos ordena también que no pudiendo ser verdaderos todos nuestros pensamientos porque no somos totalmente perfectos debe encontrarse infaliblemente la verdad que haya en ellos más bien en los que tenemos cuando estamos despiertos que en los que tenemos mientras soñamos.

QUINTA PARTE

Me gustaría mucho proseguir y mostrar aquí toda la cadena de las otras verdades que he deducido de estas primeras. Pero, como sería necesario para ello que hablase ahora de algunos problemas actualmente en discusión entre los doctos[1], con quienes no deseo indisponerme, pienso que será mejor que me abstenga y que me limite únicamente a decir de modo general cuáles son, con el fin de permitir juzgar a los más sabios si sería útil que el
41 público estuviera mejor informado. He / permanecido siempre firme en la decisión que había tomado de no suponer ningún otro principio[2] que el que acabo de aplicar en la demostración de la existencia de Dios y del alma, y de no admitir como verdade-

[1] *El Mundo o tratado de la luz* (1633), que Descartes presenta aquí en síntesis, constituye una aportación decisiva a la nueva física —al menos hasta la de Newton—, basada no en las tesis aristotélicas de los «doctos» escolásticos que han promovido la condena de Galileo (1633) —de ahí la extrema cautela que se observa en este preámbulo—, sino en la hipótesis heliocéntrica y en el mecanicismo. Cf. P. Tannery, «Descartes physicien», *Rev. Mét. et Mor.*, 4, 1896, pp. 478-488; S. Gaukroger (ed.), *Descartes: Philosophy, Mathematics and Physic*, Harvester Press, Hassocks, 1980; R. S. Westfall, *The construction of modern science: Mechanism and Mechanics*, J. Wiley & Sons, New York, 1971. Una de las grandes aportaciones de *El Mundo* y de la síntesis aquí expuesta es la explicación mecanicista del cuerpo, derivada de la teoría física, equiparando el organismo a una máquina.

[2] Con el término «principio» se refiere esta vez al *cogito*, calificado en la Parte IV de *primer principio*. Su mención muestra la cohe-

ra cosa alguna que no me pareciera más clara y cierta de lo que antes me habían parecido las demostraciones de los geómetras. Y, sin embargo, me atrevo a decir que no sólo he encontrado un medio, que en poco tiempo me ha satisfecho, con relación a las principales dificultades que suelen tratarse en filosofía, sino que además he observado ciertas leyes[3] que Dios ha establecido de tal modo en la naturaleza y de las que ha impreso en nuestras almas tales nociones[4], que después de haber reflexionado suficientemente sobre ello, no nos fuera posible dudar de que se cumplen exactamente en todo lo que existe o acontece en el mundo. Al considerar luego el conjunto de estas leyes, me parece haber descubierto algunas verdades más útiles y más importantes que cuanto había aprendido anteriormente o incluso esperaba aprender.

Pero como he intentado explicar las principales en un tratado que determinadas razones me impiden publicar[5], no podría hacer nada mejor para darlas a

rencia del proyecto cartesiano, hecho explícito al final de la Parte II (cf. nota 18), de fundar en la filosofía «todos los principios de las ciencias», incluida la física. Sobre la relación Metafísica/Física, véase el excelente estudio de D. M. Clarke, *La filosofía de la ciencia de Descartes*, pp. 90-117.

[3] «Leyes de la naturaleza» llama Descartes a «las reglas siguiendo las cuales tienen lugar los cambios» (*Le Monde*, AT, XI, p. 37). Entre otros, «el logro más brillante del Descartes físico es, sin duda, haber proporcionado una fórmula «clara y distinta» del principio de inercia» (A. Koyré, *Estudios galileanos*, p. 149), principio o ley de suma importancia para comprender las demás leyes del movimiento.

[4] ¿Innatismo platónico-agustiniano? G. Quintás, en desacuerdo con E. Gilson, estima que «expresiones como ésta no deben utilizarse para construir explicación alguna del innatismo cartesiano» dada su imprecisión (*op. cit.*, p. 442, n. 7). Con esta interpretación coincide la de D. M. Clarke (*op. cit.*, pp. 196 ss.).

[5] Las razones por las cuales Descartes decide no publicar *Le Monde* se reducen a una: la condena de Galileo es el síntoma evidente que desaconseja iniciar al menos la discusión, polémica, con los «doctos»

conocer que indicar aquí sumariamente lo que contiene. Mi propósito, antes de redactarlo, era incluir en él todo lo que creía saber sobre la naturaleza de las cosas materiales. Ahora bien, así como los pintores, no pudiendo representar con igual detalle sobre una tabla lisa las diferentes caras de un cuerpo sólido, eligen una de las principales y la iluminan,
42 mientras que sombreando las / demás no las hacen aparecer sino tal como se las puede ver al mirar la principal: así también, temiendo no poder expresar en mi discurso todo lo que tenía en mi pensamiento, me propuse sólo exponer en él muy ampliamente lo que pensaba sobre la luz; luego, con este motivo, añadí algo acerca del sol y de las estrellas fijas, dado que casi toda la luz procede de estos cuerpos; de los cielos, ya que la transmiten; de los planetas, de los cometas y de la tierra, que la reflejan; y en particular de todos los cuerpos que hay sobre la tierra, porque son o coloreados o transparentes o luminosos; y, finalmente, del hombre, porque es el espectador. Más aún, para sombrear un poco todo esto y poder expresar con mayor libertad lo que yo pienso al respecto, sin verme obligado a seguir ni a refutar las opiniones admitidas entre los doctos, decidí abandonar todo este mundo a sus disputas y hablar solamente de lo que sucedería en uno nue-

de la tesis copernicana del movimiento de la tierra. Además de éste, otros principios de la física mecanicista sistematizada en *Le Monde* son: la reducción de la materia a la extensión, su expresión cuantitativa a través del número y la figura, y la conceptualización del movimiento a partir del principio de inercia. Si, según R. Taton, la física de Descartes domina todo el siglo hasta la aparición de los *Principia* (1687) de Newton, ello no se debe, claro está, a la síntesis de esta V Parte. Aunque *Le Monde* no es editado sino en 1664, Descartes publica los fundamentos de su física en la segunda parte de *Principia philosophiae* (1644). J. W. Lynes, «Descartes theory of elements: From "Le Monde", to the "Principies", *Journal of the History of Ideas*, 43, 1982, pp. 55-72.

vo[6], si Dios crease ahora en alguna parte, en los espacios imaginarios, bastante materia para componerlo, y agitase de forma diversa y sin orden las diferentes partes de esta materia de manera tal que formase un caos tan confuso como pudieran representarlo los poetas[7], y que no hiciese otra cosa, luego, sino prestar su concurso ordinario a la naturaleza y dejarla obrar según las leyes por él establecidas[8]. De este modo describí, en primer lugar, esta materia y traté de representarla de tal forma que, a mi parecer, no hay nada en el mundo más claro ni más inteligible, excepto lo que se ha dicho antes acerca de Dios y del alma; pues hasta supuse
43 expresamente que / no había en ella ninguna de esas formas o cualidades de las que se discute en las escuelas, ni generalmente cosa alguna cuyo conocimiento no fuera tan natural a nuestras almas, que ni siquiera se pudiera fingir que se ignora. Además, hice ver cuáles eran las leyes de la naturaleza; y, sin fundar mis razones sobre ningún otro principio que en las infinitas perfecciones de Dios, traté de demostrar todas aquéllas sobre las que pudiera haber alguna duda y de probar que son tales que, aunque Dios hubiera creado varios mundos, no podría haber ni siquiera uno en el que no se cumplieran. Mostré, después, cómo la mayor parte de la materia de ese caos debía disponerse y ordenarse, de acuerdo con estas leyes, de una forma tal que lle-

[6] Habiendo roto en la Parte I del *Discurso* con los doctos escolásticos, lo importante es poder hablar de una concepción nueva del mundo. Aunque sea retóricamente, con el fin de rehuir la discusión polémica, Descartes menciona el mundo *nuevo* del que trata en el cap. VI de *El Mundo*.

[7] Alusión a la fábula poética del caos originario y, concretamente, a la obra de Lucrecio, *De rerum natura*.

[8] Se trata de las leyes según las cuales acontecen los cambios. Véase su enumeración y definición en el cap. VII de *El Mundo*.

gaba a ser semejante a nuestros cielos; cómo, mientras tanto, algunas de sus partes debían componer una tierra, otras, planetas y cometas, y algunas otras un sol y estrellas fijas. Y, en este punto, extendiéndome en el estudio de la luz, expliqué detenidamente cuál era la que se debía hallar en el sol y las estrellas, y cómo desde allí atravesaba en un instante los espacios inmensos[9] de los cielos, y cómo se reflejaba desde los planetas y los cometas hacia la tierra. Añadí también algunas observaciones acerca de la sustancia, la situación, los movimientos y otras diversas cualidades de estos cielos y de estos astros; de tal modo que pensaba haber dicho bastante para hacer comprender que nada se observa en los de este mundo que no deba, o al menos no pue-
44 da, parecer totalmente similar a los del mundo / que describía. De ahí pasé a tratar de la tierra en particular: expliqué cómo, aun habiendo supuesto expresamente que Dios no había concedido peso alguno a la materia de la que estaba compuesta, todas sus partes tendían exactamente hacia su centro; cómo, habiendo agua y aire sobre su superficie, la disposición de los cielos y de los astros, de la luna principalmente, debía producir en ella un flujo y reflujo que fuera semejante en todas sus circunstancias al que se observa en nuestros mares; y, además, cierta corriente tanto de agua como de aire, que va de levante a poniente tal como se la observa también entre los trópicos; cómo pueden formarse naturalmente las montañas, los mares, las fuentes y los ríos, y cómo pueden aparecer los metales en las minas, las plantas crecer en los campos y, en general, producirse todos los cuerpos llamados mezclas o compuestos. Y entre otras cosas, dado que, además de los astros, no conozco nada en el mundo que produzca luz sino el fuego, me esforcé en hacer comprender con mucha claridad todo lo que atañe a su naturaleza, cómo se produce, cómo se alimenta y cómo a veces da calor sin luz y otras luz sin calor; cómo puede dar lugar en diversos cuerpos a diferentes colores y a otras varias cualidades; cómo funde unos cuerpos y endurece otros; cómo puede consumir casi todos o convertirlos en humo y cenizas; y, en fin, cómo de esas cenizas forma vidrio sólo por la fuerza de su acción; y pareciéndome
45 esta transformación / de cenizas en vidrio tan admirable como ninguna otra que tenga lugar en la naturaleza, tuve especial agrado en describirla.

Sin embargo, de todo ello no quería inferir que este mundo haya sido creado de la manera que yo exponía; pues es mucho más verosímil[10] que desde el principio Dios lo ha hecho tal como debía ser. Pero es cierto, y esta opinión es comúnmente admitida entre los teólogos, que la acción por la cual lo conserva ahora es exactamente la misma que aquélla por la cual lo ha creado; de manera que, aunque no le hubiese dado al principio otra forma que la del caos, con tal que le prestara su concurso, habiendo establecido las leyes de la naturaleza, para obrar tal como ella acostumbra, es posible pensar,

[9] El espacio infinito es una de las consecuencias más innovadoras de la tesis copernicana, tal como ya lo intuyó G. Bruno. Véase el excelente estudio de A. Koyré, *Du monde clos à l'univers infini*, Gallimard, Paris, 1973.

[10] Consciente Descartes de que su teoría difiere del relato del *Génesis*, prefiere dejar la exégesis de éste a los teólogos sin entrar en abierta discusión con ellos. Así, el término *verosímil* forma parte de la retórica en la que envuelve su teoría, pero esta vez necesita aclararla en los siguientes términos: «[...] al decir que es verosímil (a saber, según la razón humana), que el mundo ha sido creado tal como debía ser, no pretendo en modo alguno negar que sea cierto para la fe que es perfecto» (a Mersenne, 27 de abril de 1637). Cf. *Conversación con Burman*, cit., p. 164.

sin menoscabo del milagro de la creación, que de este único modo todas las cosas que son puramente materiales habrían podido llegar a ser, con el tiempo, tal y como ahora las vemos. Y su naturaleza es mucho más fácil de comprender, cuando se ven nacer poco a poco de ese modo, que cuando se las considera ya hechas del todo.

De la descripción de los cuerpos inanimados y de las plantas pasé a la de los animales y particularmente a la de los hombres. Pero como aún no tenía conocimiento suficiente para tratar este tema de la misma manera que los demás, es decir, demostrando los efectos por las causas y haciendo ver de qué semillas y de qué modo debe producirlos la naturaleza, me contenté con suponer que Dios formó el cuerpo de un hombre enteramente seme-
46 jante a / uno de los nuestros tanto en la figura exterior de sus miembros como en la conformación interior de sus órganos, sin componerlo de otra materia que aquella que ya había descrito y sin darle al principio alma racional alguna ni siquiera otra cosa que le sirviera de alma vegetativa o sensitiva[11], sino que hizo nacer en su corazón uno de esos fuegos sin luz, que ya había explicado, y que concebía de una naturaleza igual a la del que calienta el heno, cuando se lo encierra antes de que esté seco, o como el que lleva a ebullición los vinos nuevos cuando se los deja fermentar con su hollejo. Así pues, exa-

[11] En síntesis, ésta es la conclusión a la que llega el *Tratado del hombre*: «No debemos concebir en esta máquina alma vegetativa o sensitiva alguna, ni otro principio de movimiento y de vida» (ed. de G. Quintás, Ed. Nacional, p. 177; AT, XI, p. 202). Como explicación del cuerpo-máquina, es decir, de las funciones vegetativas, locomotrices, sensitivas, Descartes no sigue los términos aristotélico-tomistas, sino los principios mecanicistas de la física. Si recordamos que el *Tratado del hombre* constituye el cap. XVIII de *El Mundo*, comprenderemos también el orden temático de esta Parte V del *Discurso*.

minando las funciones que como consecuencia de todo esto podían darse en ese cuerpo, encontraba exactamente las mismas que las que puede haber en nosotros sin que pensemos en ellas y, por consiguiente, sin que en nada contribuya nuestra alma, es decir, aquella parte distinta del cuerpo de la que se ha dicho anteriormente que su naturaleza consiste sólo en pensar; y que son aquellas mismas funciones en virtud de las cuales se puede decir que los animales irracionales se parecen a nosotros; pero, en cambio, no pude encontrar en ese cuerpo función alguna de las que, siendo dependientes del pensamiento, son las únicas que nos pertenecen en tanto que hombres; mientras que enseguida encontraba todas éstas, si suponía que Dios creó un alma racional y la unió a ese cuerpo en la forma concreta que describía[12].

Pero, con el fin de que se pueda ver de qué modo trataba esta materia[13], deseo dar aquí la explicación del movimiento del corazón y de las arte-

[12] Este fragmento constituye una modificación con relación al proyecto abandonado en 1633. Al comienzo del *Tratado del hombre* leemos: «Es necesario que, en primer lugar, describa su cuerpo y, en segundo lugar, su alma; finalmente, debo mostrar cómo estas dos naturalezas deben estar unidas para dar lugar a la formación de hombres que sean semejantes a nosotros» (p. 49; AT, XI, p. 119). Sólo la primera parte del proyecto se convirtió en realidad. ¿A qué conclusión hubiera llegado Descartes en su reexamen de la noción de «alma»? Cf. G. Ryle, *El concepto de lo mental*, Paidós, Buenos Aires, 1967; J. Searle, *Mentes, cerebros y ciencia*, trad. de L. Valdés, Cátedra, Madrid, 1985.

[13] Como muestra de la explicación mecanicista del cuerpo elige la función de la circulación de la sangre. Cf. Introducción de G. Quintás al *Tratado del hombre*, pp. 35- 45; B. de Saint-Germain, Descartes consideré comme physiologiste et comme médecin, Paris, 1869; A. C. Crombie, «Descartes on method and physiology», *Cambridge Journal*, 5, 1951, pp. 178-186; T. S. Hall, «Descartes' physiological method: position, principles, exemples», *Journal of the History of Biology*, 3, 1970, pp. 53-79; L. Chauvois, *Descartes. Sa méthode et ses erreures en physiologie*, Les Ed. du Cèdre, Paris, 1966.

rias, pues siendo el primero y el que más comúnmente se observa en los animales, a partir de él
47 será más fácil juzgar lo que se debe / pensar de los demás. Y con el fin de que sea menos difícil de comprender lo que voy a exponer, desearía que los que están menos versados en anatomía se tomasen el trabajo, antes de leer estas páginas, de hacer cortar en su presencia[14] el corazón de algún animal grande, que tenga pulmones, pues es muy semejante en todo al del hombre, con el fin de que les hagan ver las dos cámaras o cavidades que hay en él[15]. En primer lugar, la que está en el lado derecho, a la que van a parar dos tubos muy anchos, a saber: la vena cava, que es el principal receptáculo de la sangre, y como el tronco del árbol cuyas ramas son las restantes venas del cuerpo, y la vena arteriosa, mal llamada así porque es en realidad una arteria[16] que, teniendo su origen en el corazón, se ramifica a la salida del mismo en varias ramas que van a extenderse por todas partes en los pulmones. Después, la que está en el lado izquierdo, a la que corresponden igualmente dos tubos, que son tanto o más anchos que los anteriores, a saber: la arteria venosa, que también ha sido impropiamente denomi-

[14] El método de observación directa, que Descartes practicó no sólo en su estudio anatómico del cuerpo en general, constituye uno de los hechos que exige un estudio más atento de la función de la experiencia en la filosofía cartesiana. Por lo que se refiere a la práctica anatómica, L. Belloni escribe: «Esta nueva anatomía, inspirada en la iatromecánica pretende descomponer en sus partes más diminutas a la máquina de nuestro organismo, es una anatomía destinada a conseguir su plena realización mediante el acoplamiento del artificio anatómico con el empleo del microscopio» (*cit.* por Laín Entralgo, *Historia universal de la medicina*, IV, Barcelona, 1973, p. 220). Cf. Cf. P. Gallois, «La méthode de Descartes et la médecine». *Hippocrate*, 6, 1938, pp. 65-77.

[15] Los dos ventrículos.

[16] La arteria pulmonar, que lleva la sangre venosa del ventrículo derecho al pulmón. En la época de Harvey, las opiniones acerca del papel de esta arteria no siempre eran compartidas.

nada porque no es sino una vena[17] que viene de los pulmones, donde está dividida en múltiples ramificaciones entrelazadas con las de la vena arteriosa y con las del conducto llamado silbato[18], por donde penetra el aire de la respiración; y la gran arteria[19] que, al salir del corazón, se ramifica por todo el cuerpo. Desearía también que se les mostrara con todo cuidado las once películas[20] que, como otras tantas puertecillas, abren y cierran los cuatro orifi-
48 cios que hay en estas dos cavidades, a saber: / tres a la entrada de la vena cava[21], en donde están de tal forma dispuestas que en modo alguno pueden impedir que la sangre que contiene entre en la concavidad derecha del corazón, pero evitando con precisión, sin embargo, que pueda salir; tres a la entrada de la vena arteriosa[22] que, dispuestas de forma totalmente contraria, permiten que la sangre que hay en esta cavidad pase con facilidad a los pulmones, pero no que la sangre alojada en los pulmones retorne al mismo lugar; y otras dos a la entrada de la arteria venosa[23], que dejan que la sangre de los pulmones pase a la concavidad izquierda del corazón, pero obstruyen su retorno; y tres a la entrada de la gran arteria[24] que, si bien permiten que

[17] Las venas pulmonares, que llevan al corazón la sangre oxigenada en los pulmones.

[18] *Sifflet* en la ed. fr. Es obvio que «silbato» no es sino el sentido figurado de lo que se conoce como tráquea.

[19] La arteria aorta.

[20] En el sentido de piel fina o delicada (ed. fr.: *petites peaux*). Se trata de las válvulas.

[21] La válvula tricúspide, que es en realidad una válvula aurículo-ventricular.

[22] Las tres válvulas sigmoideas, situadas en el orificio de la arteria pulmonar.

[23] La válvula mitral o bicúspide, que es, como la tricúspide, una válvula aurículo-ventricular.

[24] Las tres válvulas sigmoideas situadas en el orificio de la arteria aorta.

la sangre salga del corazón, le impiden volver a él. Y no hay que buscar más razón del número de estas películas que la siguiente: siendo ovalado el orificio de la arteria venosa a causa de su ubicación, puede cerrarse cómodamente con dos, mientras que al ser circulares los otros, pueden cerrarse mejor con tres. Desearía, además, que se les hiciera considerar que la gran arteria y la vena arteriosa son de una composición mucho más dura y consistente que la arteria venosa y la vena cava; que estas dos últimas se ensanchan antes de entrar en el corazón y que allí forman como dos bolsas, llamadas orejas del corazón, que están compuestas de una carne muy parecida a la de éste; que hay siempre más calor en el corazón que en ninguna otra parte del cuerpo; y, en fin, que la acción del calor es tal que, si alguna gota de sangre entra en sus concavidades, ésta se infla inmediatamente y se
49 / dilata, como ocurre generalmente con todos los líquidos cuando se los deja caer gota a gota en un vaso muy caliente.

Dicho esto, basta añadir, para explicar el movimiento del corazón, que, cuando sus concavidades no están llenas de sangre, ésta corre necesariamente de la vena cava a la concavidad derecha y de la arteria venosa a la izquierda; tanto más cuanto que estos dos vasos están siempre llenos, y sus orificios, que miran hacia el corazón, no pueden en tal caso estar cerrados; pero tan pronto como han entrado de este modo dos gotas de sangre, una en cada concavidad, estas gotas, que no pueden ser sino muy gruesas ya que los orificios de entrada son muy anchos y los vasos por donde circulan están muy llenos de sangre, se rarifican y se dilatan a causa del calor que allí encuentran, por medio del cual, haciendo dilatar todo el corazón, empujan y cierran las cinco puertecillas que están en las entradas de

los dos vasos de donde provienen, impidiendo así que se vierta más sangre en el corazón; y, al continuar dilatándose cada vez más, presionan y abren las otras seis puertecillas que están en las entradas de los otros dos vasos por donde salen, haciendo dilatar de este modo todas las ramificaciones de la vena arteriosa y de la gran arteria casi al mismo tiempo que el corazón; éste se contrae inmediatamente después, como hacen también las arterias, porque se enfría la sangre que en ellos ha entrado, y se cierran sus seis puertecillas, mientras se abren de nuevo las cinco de la vena cava y de la arteria ve-
50 nosa dando paso a / otras dos gotas de sangre, que hacen que se dilate nuevamente el corazón y las arterias de igual manera que las precedentes. Y como la sangre que así entra en el corazón pasa por las dos bolsas que se llaman sus orejas, de ahí viene que el movimiento de éstas sea contrario al de aquél, contrayéndose éstas cuando aquél se dilata. Por lo demás, con el fin de que los que no conocen la fuerza de las demostraciones matemáticas, y no están habituados a distinguir las verdaderas razones de las verosímiles, no se aventuren a negar todo esto sin examinarlo, deseo advertirles que este movimiento, que acabo de explicar, se sigue también necesariamente de la sola disposición de los órganos que se puede ver a simple vista en el corazón, del calor que en él se puede sentir con los dedos y de la naturaleza de la sangre que se puede conocer por experiencia, así como el movimiento de un reloj es consecuencia de la fuerza, de la situación y de la figura de sus contrapesos y de sus ruedas[25].

[25] Al comparar el movimiento de la circulación de la sangre al movimiento de un reloj, esto es, a la mayor o menor fuerza, a la disposición y figura de sus órganos (y contrapesos), etc., el principio de explicación mecanicista se hace patente. Con todo, a diferencia de

Pero si se pregunta por qué la sangre de las venas no se agota, circulando de modo continuo a través del corazón, y por qué las arterias no están excesivamente repletas de sangre, ya que toda la que pasa por el corazón a ellas se dirige, no necesito contestar otra cosa que lo que ha escrito un médico inglés [Heruaeus, *de motu cordis*][26], al cual hay que conceder el honor de haber roto el hielo en esta materia y de ser el primero que ha defendido la existencia de pequeñas comunicaciones en las extremidades de las arterias, por donde pasa la sangre que reciben del corazón a las pequeñas ramificaciones de las venas, desde donde se dirige de nuevo al corazón, de manera que su curso no es sino una
51 circulación / incesante. Y esto lo prueba muy bien por medio de la experiencia ordinaria de los cirujanos, quienes habiendo atado el brazo con mediana fuerza por la parte superior del punto en donde abren la vena, hacen que la sangre salga más abundantemente que si no lo hubiesen atado; y sucedería todo lo contrario si lo atasen por la parte inferior, entre la mano y la abertura, o si lo atasen con mucha fuerza por la parte superior. Pues es evidente

Harvey, Descartes considera que la razón última de los movimientos del corazón no es de carácter muscular (AT, XI, pp. 169-170), es decir, no se debe a contracciones.

[26] Obra citada por Descartes al margen. Se trata de *Exercitatio anatomica de motu cordis et sanguinis in animalibus*, Frankfurt, 1628, de W. Harvey (1578-1657), profesor de anatomía y cirugía en el Colegio de Medicina de Londres. Tras sugerirle Mersenne la conveniencia de leer esta obra, Descartes le manifiesta, a finales de 1632, que tiene el propósito de leerla (AT, I, p. 263). La pregunta con la que comienza este parágrafo es la misma que condujo a Harvey al descubrimiento de la circulación de la sangre. Un excelente estudio de *De matar cordis* ha sido realizado por Woodger. *Biología y lenguaje*, trad. de M. Garrido, Madrid, 1978. Sobre la relación Descartes/Harvey, véase el de E. Gilson, «Descartes, Harvey et la scolastique», en *Études de philosophie médiévale*, 1921, pp. 220-222. G. Witteridge, *W. Harvey and the circulation of the blood*, MacDonald, London; Elsevier, New York, 1971.

que la atadura hecha con mediana fuerza, si bien puede impedir que la sangre que ya está en el brazo vuelva al corazón por las venas, no por eso impide que nueva sangre venga sin cesar por las arterias, porque éstas van por debajo de las venas y porque, siendo sus pieles más duras, son más difíciles de apretar y, además, porque la sangre que procede del corazón tiende con más fuerza a pasar por las arterias hacia la mano que a volver al corazón por las venas. Y puesto que la sangre sale del brazo por la abertura hecha en una de sus venas, debe haber necesariamente algunos pasos por debajo de la atadura, es decir, hacia las extremidades del brazo, por donde la sangre pueda venir de las arterias. Prueba asimismo muy bien lo que afirma sobre el curso de la sangre, por la existencia de ciertas películas[27] que de tal modo están dispuestas en diversos lugares, a lo largo de las venas, que no permiten que la sangre vaya desde el centro del cuerpo a las extremidades, pero sí que retorne de las extremidades al corazón; y, además, sostiene que la experiencia demuestra que toda la sangre que hay en el cuerpo puede salir en poco tiempo por una sola arteria que se haya seccionado, incluso aunque se atase fuertemente muy cerca del corazón y se cor-
52 tase entre éste y la atadura, de tal manera que / no haya motivo para imaginar que la sangre vertida pueda venir de otra parte.

Pero hay muchas otras razones que prueban que la verdadera causa de este movimiento de la sangre es la que he indicado[28]. En primer lugar, la diferencia que se advierte entre la sangre de las venas y la que sale de las arterias no puede proceder sino de

[27] Véase la nota 20 de esta Parte.

[28] Hasta aquí Descartes ha elogiado a Harvey por el descubrimiento de la circulación de la sangre. Pero ahora procede a su crítica. El texto muestra los argumentos que pretenden probar que la verdade-

que, habiéndose rarificado y como destilado al pasar por el corazón, es más sutil, más viva y más caliente nada más haber salido de él, es decir, cuando está en las arterias, que poco antes de entrar, esto es, cuando está en las venas; y si se observa atentamente, se verá que esta diferencia sólo aparece claramente cerca del corazón, pero no tanto en los lugares más alejados de él[29]. Además, la dureza de la piel de que están hechas la vena arteriosa y la gran arteria demuestra suficientemente que la sangre golpea contra ellas con mayor fuerza que contra las venas[30]. y ¿por qué la concavidad izquierda del corazón y la gran arteria han de ser más amplias y más anchas que la concavidad derecha y la vena arteriosa, sino porque la sangre de la arteria venosa, que ha estado solamente en los pulmones desde que ha pasado por el corazón, es más sutil y se expande con mayor fuerza y más fácilmente que la que viene inmediatamente de la vena cava?[31]. ¿Y qué es lo que los médicos pueden averiguar al tomar el pulso, si no saben que, cuando la sangre cambia de naturaleza, puede rarificarse por el calor del corazón con mayor o menor intensidad y con

ra causa del movimiento de la sangre no es la contracción del corazón, como sostiene Harvey, sino la dilatación de la sangre provocada por el calor del corazón.

[29] En el primer argumento Descartes señala un error de la explicación de Harvey. Pero incurre, a su vez, en otro error al ignorar —como ignoraban todos hasta Lavoisier, 1777— que la transformación de la sangre venosa en arterial es el resultado de la respiración pulmonar, y que constituye una verdadera combustión.

[30] El segundo argumento no es válido, porque no es cierta la hipótesis de Descartes, según la cual la sangre que corre por la arteria pulmonar es sangre venosa.

[31] El tercer argumento refuerza la teoría cartesiana contra la opinión de Harvey. Si la misma cantidad de sangre pasa por los dos ventrículos y el izquierdo es más grande que el derecho, es preciso, para llenarlo, que la sangre se dilate. Tal dilatación sólo proviene del calor del corazón.

mayor o menor velocidad que anteriormente?[32]. Y si se examina cómo se comunica este calor a los de-
53 más miembros, habrá que admitir que es / por medio de la sangre, la cual al pasar por el corazón vuelve a calentarse y a extenderse desde él a todo el cuerpo; por ello acontece que si se extrae sangre de alguna parte, se priva a ésta por el mismo hecho del calor; y aun cuando el corazón estuviera ardiendo como un hierro candente no bastaría para mantener el calor en los pies y en las manos como lo hace, si no les enviase continuamente sangre nueva. Por otra parte, también se llega a conocer a partir de ahí[33] que la verdadera función de la respiración consiste en aportar suficiente aire fresco al pulmón, para hacer que la sangre que a él llega desde la concavidad derecha del corazón, donde se ha rarificado y como vaporizado, se espese y se convierta de nuevo en sangre antes de retornar a la concavidad izquierda, sin lo cual no estaría en condición de alimentar el fuego que allí hay. Lo cual se confirma al observar que los animales que carecen de pulmones sólo tienen una concavidad en el corazón, y que los niños, que no pueden utilizarlos mientras están en el seno materno, tienen un orificio por donde pasa sangre de la vena cava a la concavidad izquierda del corazón y un conducto por donde va de la vena arteriosa a la gran arteria sin pasar por el pulmón. Además, ¿cómo se hará la digestión en el

[32] Cuarto argumento: las fiebres serían alteraciones de la sangre, que provocarían las variaciones en el pulso.

[33] Una nueva serie de argumentos muestran la diferencia de puntos de vista entre Harvey y Descartes. El primero se interesa por el problema de la circulación de la sangre y su causa. El segundo pretende encontrar en el problema de la circulación el principio fundamental de toda la fisiología. Cf. *Tratado del hombre*, Editora Nacional, Madrid, 1980. T. S. Hall, «Descartes' physiological method: position, principles, exemples», *Journal of the History of Biology*, 3, 1970, pp. 53-79.

estómago, si el corazón no le enviase calor por las arterias y, con él, algunas de las partes más fluidas de la sangre que ayudan a disolver los alimentos que allí se han introducido? Y la acción que convierte en sangre el jugo de tales alimentos, ¿no es fácil de conocer si se considera que, al pasar una y otra vez por el corazón, se destila quizá más de
54 cien o doscientas veces al día? / Y para explicar la nutrición y la producción de los diversos humores que hay en el cuerpo, ¿qué necesidad hay de otro supuesto que no sea la afirmación de que la fuerza con que la sangre, al dilatarse, pasa del corazón a las extremidades de las arterias, hace que algunas de sus partes se detengan entre las de los miembros donde se hallan y tomen allí el lugar de otras a las que expulsan, y que, según la situación o la figura o la pequeñez de los poros que encuentran, unas afluyen a ciertos lugares y otras a otros, de la misma manera que, como todos han podido ver, las diferentes cribas, por estar agujereadas de distinto modo, sirven para clasificar los granos según su tamaño? Y, en fin, lo más relevante en todo esto es la generación de los espiritas animales, que son como un viento muy sutil o más bien como una llama muy pura y muy viva que, ascendiendo incesante y profusamente desde el corazón al cerebro, de ahí se comunica a través de los nervios con los músculos y produce el movimiento de todos los miembros; de tal manera que no es preciso imaginar otra causa, para explicar que las partes de la sangre más agitadas y más penetrantes son, a su vez, las más adecuadas para formar estos espíritus que se dirigen al cerebro antes que a otras partes, sino la siguiente: las arterias que allí las llevan son las que, saliendo del corazón, van más en linea recta que todas las demás y, según las reglas de la mecánica, que son las mismas que las de la naturaleza, cuando

varias cosas tienden conjuntamente a moverse hacia el mismo lado, en donde no hay espacio suficiente para todas, como sucede con las partes de la sangre que salen de la concavidad izquierda del corazón y
55 tienden hacia el cerebro, / las más débiles y menos agitadas deben ser apartadas por las más fuertes, que por estas razones sólo ellas llegan.

Había explicado detalladamente todas estas cosas en el tratado que hace algún tiempo tuve el propósito de publicar. Y en él había mostrado, a continuación[34], cuál debe ser la contextura de los nervios y de los músculos del cuerpo humano, para que los espíritus animales, alojados en su interior, tengan la fuerza necesaria para mover sus miembros, como se observa en las cabezas que, poco después de ser cortadas, aún se mueven y muerden la tierra, aunque ya no estén animadas; qué cambios deben darse en el cerebro, para que se produzca la vigilia, el sueño y los sueños; cómo la luz, los sonidos, los olores, los sabores, el calor y demás cualidades de los objetos exteriores pueden imprimir en el cerebro distintas ideas por medio de los sentidos; cómo pueden también suscitar allí las suyas el hambre, la sed y demás pasiones interiores; qué debe entenderse por el sentido común, donde tales ideas son recibidas, qué por la memoria, que las conserva, y qué por la fantasía, que puede combinarlas de diferentes modos y componer otras nuevas, y del mismo modo también puede, al distribuir los espíritus animales en los músculos, hacer que se muevan los miembros del cuerpo de tan diversas maneras y tan de acuerdo con los objetos que se presentan a los sentidos externos y con las pasiones internas, como puedan moverse sin que la voluntad los guíe. Lo

[34] Se refiere a *El Mundo*, y al cap. XVIII de esta obra, es decir, el *Tratado del hombre* (AT, XI, pp. 119-202).

cual no parecerá en modo alguno extraño a quienes, sabiendo cuántos diferentes *autómatas* o máquinas
56 que se mueven puede hacer la industria humana, / empleando en ello muy pocas piezas en comparación con el gran número de huesos músculos, nervios, arterias, venas y todas las demás partes que hay en el cuerpo de cada animal, consideren este cuerpo como una máquina que, habiendo sido hecha por la mano de Dios está incomparablemente mejor ordenada y es capaz de movimientos más admirables que ninguna de las que pueden ser inventadas por los hombres[35].

Y en este punto me detuve particularmente para mostrar que, si hubiera máquinas tales que tuvieran los órganos y la figura de un mono o de cualquier otro animal irracional, no tendríamos medio alguno para reconocer que las máquinas no eran exactamente de la misma naturaleza que estos animales; mientras que, si las hubiese semejantes a nuestros cuerpos e imitasen nuestras acciones tanto como moralmente fuese posible tendríamos siempre dos

[35] La tesis del *cuerpo-máquina* con la que Descartes sintetiza la explicación mecanicista del cuerpo tiene, al menos, dos consecuencias importantes: 1) explica las funciones fisiológicas y sensibles del cuerpo sin recurrir a la terminología aristotélico- tomista (cf. nota 11 de esta parte); 2) sienta las bases no sólo para un estudio científico del cuerpo humano, sino también para un reexamen del papel del cuerpo en la constitución del hombre. Todavía no sabemos lo que puede el cuerpo —dirá Spinoza— (*Ética*, III, prop. 2, Esc.). Que la tesis del animal-máquina haya sido enunciada casi un siglo antes (E. Gilson cita el libro del médico español, Gómez Pereira, *Antoniana Margarita*, Medina del Campo, 1554) tiene sin duda un interés histórico. Pero la importancia teórica antes señalada sólo se mide por el eco despertado por una teoría que se divulga en el *Discurso* (1637) y en el *Tratado del hombre* (1662, 1664, 1667). La polémica se aprecia en las conferencias, luego editadas, de Darmanson, *La bête transformée en machina* (Amsterdam, 1684), de las que P. Bayle toma nota en su *République des Lettres* (D, 889). Pero el eco resuena sobre todo en pensadores del XVIII como Helvetius o La Mettrie (*El hombre máquina*, 1748).

medios muy seguros para reconocer que no por eso eran verdaderos hombres[36]. El primero es que no podrían hacer uso de palabras ni de otros signos compuestos de ellas, como hacemos nosotros para manifestar a los demás nuestros pensamientos. Pues es posible concebir que una máquina esté hecha de tal forma que profiera palabras, e incluso que profiera algunas a propósito de acciones corporales que causaran cierto cambio en sus órganos: como si, tocándole en una parte, preguntase lo que se le quiere decir, y si, en otra, gritase que se le hace

[36] La teoría del animal-máquina, válida para la explicación mecanicista del cuerpo, ¿permite explicar otros fenómenos de la conducta humana como el *lenguaje* o el pensamiento? Descartes cree que estos fenómenos muestran el límite del principio mecanicista. Por eso postula otro principio explicativo, un «principio creador» —opina N. Chomsky (*Lingüística cartesiana*, Gredos, Madrid, 1978, p. 25)—, para estos fenómenos. Es importante anotar esta línea de investigación, que ni siquiera inició en el *Tratado del hombre*, y que aquí aparece apenas esbozada. Me refiero a los «dos medios muy seguros» que, según Descartes, distinguen al hombre del animal (y la máquina), esto es, que son específicos de lo humano: el *lenguaje* y la *conciencia*. Sin duda alguna, al plantear lo específicamente humano en términos de lo que diferencia al hombre de la máquina, el autor controvertido en la época de la contrarreforma abrió una de las polémicas más vivas de la actualidad, es decir, del momento presente en el que el cerebro programador y la inteligencia artificial suscita preguntas sin fácil respuesta. ¿Puede la máquina *pensar?* ¿Se logrará una máquina que pueda *hablar* como el hombre? A pesar de teorías funcionalistas que no dudarían en responder positivamente, J. Searle contestaría con un «no» rotundo, en base al nivel semántico del lenguaje —una máquina no supera el nivel sintáctico por muchas combinaciones que haga— y a la conciencia (otro nombre cartesiano de la razón) que sólo el hombre tiene de lo que hace, incluso cuando piensa y habla (*Mentes, cerebros y ciencia*, trad. de M. L. Valdés, Cátedra, Madrid, 1985). Más aún, M. Garrido no duda en calificar de «provocativa» la tesis del modelo computacional de la mente, de A. M. Turing, según la cual un computador puede hacer todo lo que hace el hombre, incluso la función de pensar (*Mentes y máquinas*, de A. M. Turing, H. Putnam, D. Davidson, Tecnos, Madrid, 1985, p. 10). No cabe duda de que la inspiración cartesiana al plantear el problema proporciona a la vez el principio de solución, subrayado tanto por Chomsky como por Searle.

daño, y cosas semejantes; pero no que sea capaz de
57 ordenar de forma diversa las palabras, para / responder con sentido a todo lo que se diga en su presencia, como pueden hacerlo los hombres más estúpidos. Y el segundo consiste en que, aun cuando estas máquinas pudieran hacer algunas cosas tan bien o quizá mejor que cualquiera de nosotros, fallarían infaliblemente en otras, a través de las cuales se descubriría que no actuaban por conocimiento, sino sólo por la disposición de sus órganos. Pues, mientras que la razón es un instrumento universal, capaz de servir en cualquier circunstancia, estos órganos necesitan una determinada disposición particular para cada acción concreta; de donde resulta que es moralmente imposible que haya en una máquina los suficientes resortes como para hacerla actuar en todas las circunstancias de la vida, tal y como nos hace actuar nuestra razón.

Ahora bien, por estos dos medios también es posible conocer la diferencia que hay entre los hombres y los animales.

Pues es algo bien patente que no hay hombres tan embrutecidos ni tan estúpidos, sin exceptuar siquiera a los locos, que no sean capaces de coordinar diversas palabras y de componer un discurso mediante el cual den a conocer sus pensamientos; y, al contrario, no hay animal alguno, por perfecto y afortunadamente dotado que sea, que pueda hacer algo semejante. Y esto no sucede por carecer de órganos, ya que es posible observar que las urracas y los loros pueden emitir palabras como nosotros y, sin embargo, no pueden hablar como nosotros, es decir, mostrar que piensan lo que dicen, en cambio, los hombres que, habiendo nacido sordos y mudos, están tan privados o más que los animales de los ór-
58 ganos que a los otros sirven / para hablar, suelen inventar por sí mismos algunos signos, mediante los

cuales se hacen entender por aquellos que, viviendo habitualmente con ellos, tienen ocasión de aprender su lenguaje. Y esto no sólo prueba que los animales tienen menos razón que los hombres, sino que carecen totalmente de ella, pues se puede ver que basta muy poca para saber hablar; y aunque se observe tanta desigualdad entre los animales de una misma especie como entre los hombres, y que unos son más fáciles de adiestrar que otros, no es creíble que un loro o un papagayo, aun siendo los más perfectos de su especie, igualen en esto a un niño de los más estúpidos o, al menos, a un niño que tuviera perturbado el cerebro, a no ser que su alma fuera de naturaleza totalmente distinta a la nuestra. Y no deben confundirse las palabras con los movimientos naturales, que manifiestan las pasiones y pueden ser imitados por las máquinas tan bien como por los animales; ni debe pensarse, como algunos antiguos, que los animales hablan, aunque no entendamos su lenguaje; pues si fuera verdad, dado que tienen algunos órganos que se parecen a los nuestros, podrían hacerse entender por nosotros tan bien como por sus semejantes. Es asimismo algo muy destacable que, aunque haya algunos animales que dan muestras de mayor habilidad que nosotros en algunas de sus acciones, se observa, sin embargo, que estos mismos no muestran ninguna en muchas otras; de manera que lo que hacen mejor que nosotros no prueba que tengan ingenio, porque, en ese caso,
59 tendrían más que ninguno de nosotros y / lo harían mejor en todo; por el contrario, carecen totalmente de él, y no es sino la naturaleza quien obra en ellos, según la disposición de sus órganos, tal como se observa que un reloj, compuesto únicamente de ruedas y resortes, puede marcar las horas y medir el tiempo con mayor precisión que nosotros a pesar de nuestra prudencia.

Traté, a continuación, el problema del alma racional[37]. Mostraba que en modo alguno puede ser sacada de la potencia de la materia, como las otras cosas de las que había hablado, sino que debe ser expresamente creada; y no basta que esté alojada en el cuerpo humano, como un piloto en su navío[38], a no ser, tal vez, para mover sus miembros, sino que es necesario que esté junta y unida más estrechamente con él para tener, además, sentimientos y apetencias semejantes a los nuestros, y constituir así un verdadero hombre. Por lo demás, me he extendido aquí un poco sobre el tema del alma, porque es de los más importantes; ya que, dejando aparte el error de los que niegan a Dios, que pienso haber refutado anteriormente de modo conveniente, no hay nada que aleje más a los espíritus débiles del recto camino de la virtud que el imaginar que el alma de los animales sea de la misma naturaleza que la nuestra, y que, en consecuencia, no tenemos nada que temer ni que esperar después de esta vida como nada tienen las moscas y las hormigas mientras que, cuando se conoce la diferencia que existe entre el alma de los animales y la nuestra se comprenden mucho mejor las razones que prueban que la nuestra es de una naturaleza enteramente independiente del cuerpo y, por lo tanto, que no está
60 condicionada a morir con él; y, en fin, dado que / no se ven otras causas que puedan destruirla, es natural inclinarse a pensar que es inmortal.

---

[37] No poseemos la parte del *Tratado del hombre*, que Descartes anuncia en su comienzo (cf. nota 12 de esta Parte) y menciona aquí.

[38] Se atribuye a Platón la opinión según la cual la relación entre el alma y el cuerpo se concibe como la que existe entre el piloto y su navío (Aristóteles, *De anima*. II, 1, 413a).

SEXTA PARTE

Hace ya tres años[1] que había llegado a concluir el tratado que contiene todas estas cuestiones y comenzaba a revisarlo con el fin de entregarlo a un impresor, cuando tuve noticia de que determinadas personas, a quienes profeso deferencia y cuya autoridad sobre mis acciones no es mucho menor que la de mi razón sobre mis pensamientos, habían condenado una opinión de física publicada poco antes por otro[2]; no quiero decir que yo la compartiera, pero sí que no había observado en ella, antes de su censura, nada que pudiera imaginar como perjudicial para la religión ni para el estado ni, por consiguiente, que me hubiese impedido escribir sobre ella si la razón me hubiera persuadido. Esto me hizo temer que se pudiera encontrar, del mismo modo, alguna opinión entre las mías en la cual me

[1] Se refiere a 1633, fecha en la que decide no publicar *El Mundo*. Esta parte VI se divide en tres secciones que reflejan las oscilaciones de Descartes a partir de ese momento. Sección A (AT, VI, 60-65): es esencialmente un resumen del trabajo preparatorio del manuscrito de *El Mundo*, para su publicación. Sección B (AT, VI, 65-74): exposición de las causas que le llevan a aplazar la publicación e, incluso, a decidir no publicar nada en toda su vida; empieza así: «Pero desde entonces me han asaltado otras razones que me han hecho cambiar de opinión [...]». Sección C (AT, VI, 74-78): muestra un nuevo cambio de opinión, hacia 1635, por el que decide publicar los ensayos *Dióptrica* y *Meteoros:* la referencia a ellos es más explícita, sobre todo porque la redacción de esta sección fue pensada como introducción a los dos ensayos científicos, a los que luego añade un tercero, la *Geometría*.

[2] Galileo.

hubiera equivocado, a pesar del gran cuidado que he tenido siempre de no aceptar novedades en mis creencias de las cuales no tuviera demostraciones muy ciertas y de no escribir sobre algo que pudiera volverse en perjuicio de alguien. Estas circunstancias han sido suficientes para obligarme a cambiar la resolución que había tomado de publicar dicho tratado. Pues, aunque las razones en virtud de las cuales había tomado anteriormente tal decisión fueran muy fuertes, mi inclinación, que me había llevado siempre a aborrecer el oficio de hacer libros, me ha facilitado inmediatamente bastantes otras para excusarme. Y tales son esas razones, de una y
61 otra parte, que no / sólo tengo cierto interés en exponerlas aquí, sino también quizá el público lo tenga a su vez en conocerlas.

Nunca he atribuido excesiva importancia a aquellas cosas que procedían de mi ingenio y, mientras no he recogido del método que uso otros frutos sino la solución de algunas dificultades pertenecientes a las ciencias especulativas, o mientras he intentado regular mis costumbres de acuerdo con las razones que ese método me proporcionaba, no me he creído obligado a escribir sobre tales cosas. Pues, por lo que atañe a las costumbres, cada uno abunda tanto en su sentido que podrían encontrarse tantos reformadores como cabezas, si se permitiera emprender la tarea de realizar algún cambio a personas distintas de las que Dios ha establecido como soberanos de sus pueblos o bien ha dado suficiente gracia y celo para ser profetas; y en cuanto a mis especulaciones, si bien me complacían en alto grado, pensé que también los demás tendrían otras que les gustarían acaso más. Pero, tan pronto como hube adquirido algunas nociones generales relacionadas con la física, y, comenzando a ponerlas a prueba en diversas dificultades concretas, me di cuenta hasta

dónde pueden conducir y cuánto difieren de los principios comúnmente admitidos hasta el presente, pensé que no podía tenerlas ocultas sin infringir gravemente la ley que nos obliga a procurar el bien general de todos los hombres en tanto nos sea factible. Pues tales nociones me han hecho ver que es posible lograr conocimientos muy útiles para la vida y que, en lugar de la filosofía especulativa que
62 se enseña en las escuelas, se pueda encontrar / una filosofía práctica en virtud de la cual, conociendo la fuerza y las acciónes del fuego, del agua, del aire, de los astros, de los cielos y de todos los demás cuerpos que nos rodean con tanta precisión como conocemos los diferentes oficios de nuestros artesanos, podríamos aprovecharlos de la misma manera en todos los usos para los cuales son apropiados y convertirnos, de este modo, en dueños y poseedores de la naturaleza[3]. Lo cual no sólo es deseable con vistas a la invención de una infinidad de artificios, que nos permitirían disfrutar sin esfuerzo alguno de los frutos de la tierra y de todas las comodidades que hay en ella, sino principalmente también para la conservación de la salud, que es sin duda el primer bien y el fundamento de todos los demás bienes de esta vida; pues hasta el ingenio depende tanto del temperamento y de la disposición de los órganos del cuerpo que, si es posible encontrar algún medio que haga a los hombres más sabios y más hábiles de lo que hasta aquí lo han sido, pienso que hay que buscarlo en la medicina[4].

[3] Descartes retama el ideal de la ciencia que tema F. Bacon: «La ciencia y el poder humanos vienen a ser lo mismo» (*Novum organum*, 1620, Lib. I, afor. III).

[4] La medicina es, juntamente con la moral y la mecánica, una de las tres ramas del árbol de la ciencia («Lettre-Préface», a la ed. fr. de *Principes de philosophie*, AT, IX-II, p. 14). Cf. P. Gallois, «La méthode de Descartes et la médicine», *Hippocrate*, 6, 1938, pp. 65-77.

Sin duda es cierto que la que se practica en nuestros días ofrece pocas cosas cuya utilidad sea muy destacable; pero, aun sin propósito alguno de despreciarla, estoy convencido de que no hay nadie, incluso entre aquéllos que la ejercen como profesión, que no reconozca que lo que en ella se sabe es casi nada si se compara con lo que todavía queda por saber y que podríamos vernos libres de una infinidad de enfermedades, tanto del cuerpo como de la mente[5] y quizá también hasta de la debilidad que acompaña a la vejez, si se tuviera conocimiento suficiente de sus causas y de todos los remedios de los cuales nos ha provisto la naturaleza[6]. Ahora bien,
63 teniendo el propósito / de emplear toda mi vida en la búsqueda de una ciencia tan necesaria y habiendo encontrado un camino que, si alguien lo sigue, me parece tal que infaliblemente debe encontrarla, a no ser que se lo impida la brevedad de la vida o la falta de experiencias, estimaba que no había mejor remedio contra estos dos obstáculos que comunicar fielmente al público todo lo poco que hubiera encontrado e invitar a los ingenios capaces a intentar avanzar más allá, contribuyendo cada uno según su inclinación y poder a realizar las experiencias que fueran necesarias, y comunicando también al público cuanto hayan aprendido, con el fin de que comenzando los últimos por donde sus predecesores hubieran acabado y uniendo, de este modo, las vidas y los trabajos de muchos, lleguemos mediante el trabajo conjunto mucho más lejos de lo que cada uno en particular podría conseguir.

[5] *Esprit* (ed. fr.), *animi* (ed. lat.: AT, VI, p. 575).

[6] Aconseja siempre los remedios naturales frente a los artificiales. Así, sugiere a la princesa Elisabeth «una buena dieta», «pues, en cuanto a las drogas, ya sean de los farmacéuticos ya de los empíricos, las tengo en tan mala estima que no me atrevería nunca a recomendar a nadie su uso» (a Elisabeth, marzo de 1647).

Asimismo advertía, con relación a las experiencias[7], que son tanto más necesarias cuanto más se ha avanzado en el conocimiento. Pues al comienzo es más conveniente utilizar sólo las que se presentan por sí mismas a nuestros sentidos, y que no podríamos ignorar aunque no hagamos sino una mínima reflexión, que buscar otras más raras y preparadas; la razón es que estas últimas nos inducen con frecuencia a error, cuando no se conocen todavía las causas más comunes y cuando las circunstancias de que dependen son casi siempre tan singulares y tan insignificantes que es muy difícil darse cuenta de ellas. Pero el orden que he seguido en todo esto ha sido el siguiente: primero he procu-
64 rado formular / los principios o primeras causas[8] de todo lo que es o puede existir en el mundo, sin considerar, a este efecto, nada más que a Dios que lo ha creado, ni obtener tales principios sino a partir de ciertas semillas de verdades que están naturalmente en nuestras almas[9]. Examiné después cuáles eran los primeros y más comunes efectos que podían ser deducidos de estas causas: me parece que de este modo he encontrado cielos, astros, una tierra y también, sobre la tierra, agua, aire, fuego, minerales y algunas otras cosas que son las más comunes y las más simples y, por consiguiente, las más fáciles de conocer. A continuación, cuando me propuse

[7] Sobre el papel de la experiencia en la ciencia cartesiana véase: D. M. Clarke, *La filosofía de la ciencia de Descartes*, cap. 2, pp. 30-57. G. Quintás considera que se inicia aquí uno de los parágrafos más representativos de la concepción metodológica de Descartes en relación con las ciencias de la naturaleza. Cf. E. Denissoff, *Descartes, premier théoricien de la physique mathématique*, Nauwelaerts, Louvain, 1970, pp. 61-74.

[8] Se refiere a los principios y leyes generales que ha establecido en el tratado *El Mundo*.

[9] Sobre el significado concepto «semillas de verdad», véase el excelente comentario de D. M. Clarke (*op. cit.*, pp. 196-204).

llegar a conocer las más particulares, acudieron a mi pensamiento tantas y tan diversas que no creía que fuera posible al ingenio humano distinguir las formas o especies de cuerpos que hay en la tierra de una infinidad de otras que podría haber si Dios hubiera querido ponerlas en ella, ni, por lo tanto, someterlas a nuestro uso, si no se investigan las causas por los efectos y si no se utilizan muchas experiencias particulares. Después de esto, volviendo a pensar en todos los objetos que alguna vez se habían presentado a mis sentidos, me atrevo a decir que no he observado en ellos nada que no pudiera explicarse cómodamente mediante los principios que había formulado. Pero también he de reconocer que el poder de la naturaleza es tan amplio y tan grande, y que estos principios son tan simples y tan generales, que ya casi no observo efecto particular alguno que desde el principio no conozca
65 que se puede / ser deducido de muchas y diversas formas; mi mayor dificultad, por lo tanto, consiste generalmente en encontrar en qué forma concreta depende de tales principios. Para este problema[10] no veo otra solución que la de buscar de nuevo experiencias tales, que varíe su resultado según que se tenga que explicar por una u otra de esas formas posibles. Por lo demás, he llegado ya a tal punto que creo conocer bastante bien qué perspectiva se debe adoptar para llevar a cabo la mayor parte de las experiencias que pueden servir para aquella finalidad;

[10] La deducción a partir de principios evidentes es de limitada utilidad en la ciencia. Puede dar lugar sólo a leyes generales. De ahí el *problema* que Descartes resuelve mediante el recurso a la experiencia (cf. nota 7 de esta Parte). Un papel importante en la teoría cartesiana del método científico desempeñan, pues, la observación y la experimentación al proporcionar la información concreta que se corresponde o no con las leyes generales (J. Losee, *Introducción histórica a la filosofía de la ciencia*, Alianza, Madrid, 1979, pp. 85 ss.).

pero me doy cuenta, también, de que son de tales características y tan numerosas, que ni mis manos ni mis rentas, annque tuviera mil veces más de lo que tengo, serían suficientes para todas ellas; de modo que, según tenga en adelante la posibilidad de llevar a cabo más o menos experiencias, así también avanzaré más o menos en el conocimiento de la naturaleza. Todo esto pensaba darlo a conocer en el tratado que había preparado, mostrando tan claramente la utilidad que el público podría obtener que obligaría a todos aquellos que en general desean el bien de los hombres, es decir, a todos los que son realmente virtuosos y no son guiados por falsa apariencia o mera opinión, no sólo a comunicarme las experiencias realizadas por ellos sino también a cooperar conmigo en la investigación de las que quedan por hacer.

Pero desde entonces me han asaltado otras razones que me han hecho cambiar de opinión[11]: pensé que verdaderamente debía continuar escribiendo todo lo que juzgara de alguna importancia, a medida que descubriera la verdad, poniendo en ello el
66 mismo cuidado que si deseara hacerlo imprimir, / no sólo para disponer de una nueva ocasión de examinarlo detenidamente (pues, sin duda, se pone siempre mayor atención en lo que se estima que debe ser conocido por muchos que en lo que se hace únicamente para sí mismo, y, con frecuencia, lo que ha parecido verdadero cuando he empezado a concebirlo, luego lo he juzgado falso al intentar redactarlo), sino también para no perder ocasión alguna de ser útil al público, si soy capaz de ello, y para que, si mis escritos tienen algún valor, aquellos a cuyas manos lleguen después de mi muerte pue-

[11] Comienza la sección B, en la que Descartes expone los motivos que le llevan a decidir no publicar El *Mundo*.

dan utilizarlos como consideren más conveniente; pero pensé también que de ningún modo debía consentir que mis escritos fuesen publicados durante mi vida, con el fin de que ni las réplicas y controversias, que tal vez podrían suscitar, ni cualquiera que fuese la reputación que pudieran depararme, constituyesen una ocasión de perder el tiempo que tengo el propósito de dedicar a instruirme. Pues, si bien es cierto que cada hombre está obligado, en cuanto le sea posible, a procurar el bien de los demás ya que verdaderamente quien a nadie es útil nada vale, sin embargo también lo es que nuestras preocupaciones deben sobrepasar el tiempo presente y que es conveniente prescindir de cosas que tal vez podrían aportar algún provecho a los que hoy viven, cuando se tiene la intención de hacer otras que proporcionen mayor utilidad a la posteridad. Y, en efecto, quiero que se sepa que lo poco que hasta ahora he aprendido no es casi nada en comparación con lo que ignoro y que no desespero de poder conocer; pues ocurre casi lo mismo con aquellos que descu-
67 bren poco a poco la verdad en las / ciencias que con aquellos que, habiendo comenzado a enriquecerse, tienen menos dificultad para hacer grandes adquisiciones que la que tuvieron anteriormente, cuando eran más pobres, para hacer otras menores. También se los puede comparar con los jefes del ejército, cuyas fuerzas suelen crecer en proporción a sus victorias, y a quienes es necesaria mayor sagacidad para mantenerse como tales después de la pérdida de una batalla que para ocupar villas y provincias después de haberla ganado. Porque es, en realidad, como librar batallas el intentar vencer todas las dificultades y errores que nos impiden llegar al conocimiento de la verdad, pero es como perder una el admitir alguna opinión acerca de algo un tanto general e importante; se necesita, después,

mucha más destreza para volver al estado inicial que para hacer grandes progresos cuando se poseen ya principios ciertos. En cuanto a mí, si he logrado alcanzar anteriormente algunas verdades en las ciencias (y espero que las cosas contenidas en este volumen[12] mostrarán que he encontrado algunas), puedo decir que no son sino consecuencia de cinco o seis dificultades principales que he logrado superar y que considero como otras tantas batallas en las que he tenido la fortuna de mi parte. No temo incluso afirmar que sólo necesito ganar dos o tres más como éstas para alcanzar totalmente el fin que me he propuesto; además, mi edad no es tan avanzada como para que, de acuerdo con el transcurso ordinario de la naturaleza, no pueda disponer aún
68 del tiempo necesario a tal efecto. / Así pues, creo estar tanto más obligado a no malgastar el tiempo que me queda cuanto que tengo mayores esperanzas de poder emplearlo bien; y sin duda no me faltarían muchas ocasiones de perderlo, si publicara los fundamentos[13] de mi física. Pues, aunque sean casi todos tan evidentes que basta comprenderlos para asentir a ellos y que no hay uno solo del que no pueda dar demostraciones, sin embargo, puesto que es imposible que sean acordes con todas las diferentes opiniones de los hombres, presumo que suscitarían objeciones que me distraerían a menudo de mi propósito.

[12] Se refiere a los tres ensayos, con relación a los cuales el *Discurso del método* no es sino una introducción.

[13] Los fundamentos de la física cartesiana —la tesis copernicana del movimiento de la tierra, la teoría de los tres elementos (extensión, figura y movimiento) y las tres leyes del movimiento— no aparecen publicadas en la síntesis de *El Mundo*, que constituye la Parte V. Descartes los omite en 1637 para evitar polémicas comprometidas. Los publicará, luego, en los *Principios* (1644). J. W. Lynes, «Descartes theory of Elements: From "Le Monde" to the "Principes"», *Journal of the History of Ideas*, 43, 1982, 57-72.

Se podría decir que tales objeciones serían útiles, no sólo para hacerme conocer mis errores, sino también para que, si aportara algo de provecho, los demás tuvieran por ese medio una mejor comprensión de todo ello; y como muchos ojos pueden ver más que uno solo, comenzando desde ahora a hacer uso de las mías, me ayudaran a su vez con sus invenciones. Pero, aunque reconozca que estoy muy expuesto a equivocarme y aunque no me fíe nunca de los primeros pensamientos que se me ocurren, sin embargo, la experiencia que tengo de las objeciones que se me pueden hacer me impiden esperar de ellas provecho alguno; pues ya he comprobado con frecuencia los juicios, tanto de aquellos que he tenido como amigos, como de aquéllos otros para quienes me consideraba indiferente y hasta de otros cuya malicia y envidia sabía que les haría intentar más de una vez descubrir lo que el afecto hacía ocultar a mis amigos; con todo, rara vez me ha sucedido que se me haya objetado algo que no hubiera previsto en cierto modo, a no ser que la objeción
69 / se alejara mucho de mi tema; de modo que casi nunca he encontrado un censor de mis opiniones que no me pareciese menos riguroso o menos justo que yo mismo. Tampoco he observado nunca que por medio de las disputas que se practican en las escuelas se haya descubierto alguna verdad antes ignorada[14]; pues, luchando cada uno por vencer a su contrario, se dedica mucho más tiempo a hacer valer la verosimilitud que a sopesar las razones de

[14] Tales disputas tienen su base en el silogismo, sobre cuyo valor escribe Descartes en la Parte II de este *Discurso* que «sirven más para explicar a otro cuestiones ya sabidas o, incluso como el arte de Lulio, para hablar sin juicio de las que se ignoran, que para investigar las que desconocemos» (AT, VI, p. 17). El método que propone para «investigar la verdad en las ciencias» es la cara positiva de ese distanciamiento crítico. Cf. *Conversación con Burman*, *cit.*, pp. 175-176.

uno y del otro; y los que durante largo tiempo han sido buenos abogados, no por ello son después mejores jueces.

En cuanto a la utilidad que otros pudieran recibir de la comunicación de mis pensamientos, tampoco podría ser muy grande, puesto que aún no los he desarrollado hasta tal punto que no sea preciso añadir muchas cosas antes de poder llevarlos a la práctica. Y creo poder afirmar sin vanidad que, si hay alguien capaz de hacerlo, he de ser yo mejor que cualquier otro; y no porque no pueda haber en el mundo otros ingenios incomparablemente mejores que el mío, sino porque no sería tan fácil concebir una cosa y admitirla como propia, cuando se la ha aprendido de otro que cuando la ha inventado uno mismo. Lo cual es tan cierto en esta materia que, aunque frecuentemente he explicado algunas de mis opiniones a personas de gran talento que parecían comprenderlas muy claramente cuando les hablaba, sin embargo, he observado que, al repetirlas, las han alterado casi siempre de tal manera que no podía reconocerlas como propias. Por este motivo / me complace rogar a la posteridad que jamás 70
considere como mío pensamiento alguno, aunque alguien lo diga, si yo mismo no lo hubiera publicado. Más aún, no me extraño en modo alguno de las extravagancias que se atribuyen a todos los antiguos filósofos, cuyos escritos no poseemos, ni juzgo por ello que sus pensamientos hayan sido desatinados, ya que eran los mejores ingenios de su época, sino únicamente que de ellos hemos tenido una información inadecuada. También se observa que casi nunca ha sucedido que algunos de sus discípulos les hayan superado; y estoy seguro de que los más apasionados de quienes en la actualidad siguen a Aristóteles se considerarían dichosos si tuvieran tanto conocimiento de la naturaleza como

él tuvo[15], aunque fuera incluso con la condición de que nunca sobrepasarían tales conocimientos. Son como la yedra, que no puede subir a mayor altura que la de los árboles que la sostienen y que incluso muy a menudo desciende después de haber llegado hasta la copa; me parece que aquellos también descienden, es decir, se vuelven en cierto modo menos sabios que si se abstuvieran de estudiar, pues no contentos con saber todo lo que está explicado de forma clara e inteligente en las obras de su autor, desean además encontrar en él la solución de muchas dificultades sobre las que nada dice y en las que tal vez nunca pensó. Tal manera de filosofar es, sin embargo, muy cómoda para quienes no poseen sino un ingenio muy mediocre; puesto que la oscuridad de las distinciones y de los principios de que se sirven les permite opinar sobre cualquier materia tan audazmente, como si la conocieran, y sostener
71 todo lo que / afirman contra los más hábiles y sutiles sin que haya forma de convencerlos. En lo cual me parecen semejantes a un ciego que, para luchar en igualdad de condiciones contra otro que ve, lo hubieran conducido hasta el fondo de una cueva muy oscura; y puedo afirmar que éstos tienen interés en que me abstenga de publicar los principios de la filosofía que utilizo, pues siendo tan simples y tan evidentes como son, al publicarlos, ocurrirá como si abriera algunas ventanas e hiciera entrar la luz del día en la cueva a donde han descendido para batirse. Pero ni siquiera los mejores ingenios tienen motivos para desear conocerlos, pues, si lo que pre-

[15] El elogio que aquí se hace de Aristóteles, criticando en cambio a sus seguidores los escolásticos, se enmarca en la línea de las posiciones críticas que frente a éstos adoptaron pensadores renacentistas como Petrarca: a estos «empachados de erudición», escribe, «no lograría sosegarles ningún amigo, ni aun el propio Aristóteles tirando de las riendas» (Petrarca, *Obras*, I, Alfaguara, Madrid, 1978, p. 194).

tenden es saber hablar de todas las cosas y adquirir la reputación de doctos, lo conseguirán más fácilmente contentándose con lo verosímil, que puede ser alcanzado sin gran dificultad en toda clase de materias, que investigando la verdad, que no se descubre sino poco a poco en algunas y que, cuando se trata de hablar de otras, obliga a reconocer con franqueza que se ignoran. Pero si profieren el conocimiento de algunas pocas verdades à la vanidad de parecer saberlo todo, como sin duda es preferible, y desean seguir un proyecto similar al mío, no es preciso para ello que les diga nada más que lo que ya he dicho en este discurso. Pues, si son capaces de llegar más allá de lo que yo he podido, con mayor razón lo serán también de encontrar por ellos mismos todo lo que creo haber encontrado. Porque, si siempre he examinado los problemas siguiendo un orden, ciertamente lo que aún me queda por descu-
72 brir es / en sí más difícil y problemático que lo que hasta ahora he descubierto; por lo tanto, tendrían mucha menos satisfacción en llegar a conocerlo por mí que por ellos mismos; además, el hábito que podrían adquirir investigando al comienzo cuestiones fáciles y pasando poco a poco y gradualmente a otras más difíciles será más útil que todas mis instrucciones. Porque, en lo que a mí me atañe, estoy persuadido de que si desde mi juventud se me hubiesen enseñado todas las verdades cuyas demostraciones he buscado después, y que no hubiera tenido ninguna dificultad en alcanzarlas, tal vez nunca hubiera llegado a descubrir ninguna otra ni hubiera adquirido, al menos, el hábito y la facilidad que creo poseer para encontrar siempre otras nuevas, a medida que me dedique a buscarlas. En una palabra, si hay en el mundo alguna obra que no pueda ser bien acabada por ningún otro, excepto por el mismo que la empezó, ésa es aquella en la que trabajo.

Pero es cierto que, en relación con las experiencias que puedan contribuir a tal fin, un hombre solo no podría realizarlas todas; pero tampoco podría emplear útilmente otras manos que las suyas, a excepción de los artesanos u otras gentes a las que pudiera pagar, y a quienes la esperanza del beneficio, que es un medio muy eficaz, llevaría a hacer exactamente todo cuanto les prescribiera. Pues, en cuanto a los voluntarios que, por curiosidad o deseo de aprender, se ofrecieran tal vez a ayudarle, aparte de que prometen generalmente más de lo que cumplen y que no hacen sino hermosas propuestas de
73 las que jamás alguna tiene éxito, / desearían inequívocamente ser compensados mediante la explicación de algunas dificultades o, al menos, por medio de halagos y conversaciones inútiles, en las que no podría emplear parte de su tiempo sin gran detrimento. Y en cuanto a las experiencias que otros han realizado ya, aunque se las quisieran comunicar, cosa que no harían nunca aquellos que las califican de secretos, están constituidas en su mayor parte de tantas circunstancias o elementos superfluos que le sería muy difícil descifrar la verdad en ellas; además, las encontraría tan mal explicadas, o incluso tan falseadas, porque quienes las han realizado han tratado de hacerlas parecer conformes a sus principios que, si hubiera algunas útiles, no le compensarían el tiempo que tendría que dedicar a seleccionarlas. De modo que, si existiera alguien en el mundo del que se supiera con seguridad que es capaz de descubrir las cosas más grandes y más útiles para el público y que, por esta razón, otros hombres se esforzasen por todos los medios en colaborar con él para lograr el objetivo de sus propósitos, no veo que pudieran hacer otra cosa por él sino contribuir a sufragar los gastos de las experiencias que necesitara y, además, a impedir que

alguien le hiciera perder el tiempo con inoportunidades. Pero, aparte de que no presumo tanto como para querer prometer algo extraordinario, ni me alimento de pensamientos tan vanos como el de creer que el público tenga que estar muy interesado en mis proyectos, tampoco soy de ánimo tan
74 abatido como para desear aceptar de quienquiera / algún favor que pudiera pensarse que no he merecido.

Todas estas consideraciones conjuntamente valoradas fueron la causa de que, hace tres años[16], no quisiera publicar el tratado que tenía entre manos e, incluso, de que tomara la resolución de no publicar, durante mi vida, ningún otro que fuera tan general ni a partir del cual se pudieran conocer los fundamentos de mi física. Sin embargo, posteriormente dos nuevas razones me han obligado a insertar aquí algunos ensayos concretos[17] y a dar al público cuenta de mis acciones y proyectos. La primera es que, si dejaba de hacerlo, muchos que han conocido la intención que anteriormente había tenido de hacer imprimir algunos escritos, podrían imaginar que las causas por las cuales me he abstenido serían para mí más desfavorables de lo que realmente son. Pues, aunque no amo excesivamente la gloria, y hasta me atrevo a decir que la odio, en tanto que la considero contraria a la tranquilidad, que estimo por encima de todo, sin embargo, tampoco he intentado nunca ocultar mis acciones

[16] Se refiere a 1633, fecha en la que decide no publicar *El Mundo*. La reiteración de lo ya dicho al comienzo de esta parte VI se explica por la compleja estructura de la misma, detallada en la nota 1. Comienza aquí la sección C, escrita inicialmente como breve introducción a los dos ensayos —*Dióptrica y Meteoros*— que decide publicar hacia 1635. Por ello están ausentes en estas páginas las referencias al tercer ensayo, *Geometría*, escrito a finales de 1636.

[17] Ver nota anterior.

como si fueran delitos ni he tomado muchas precauciones para permanecer desconocido, tanto porque hubiera creído perjudicarme como porque ello me hubiera producido cierta especie de inquietud, que de nuevo hubiera sido contraria a la perfecta tranquilidad de ánimo que busco. Y como, mientras permanecía siempre indiferente ante la preocupación de ser o no ser conocido, no he podido impedir que adquiriese una determinada reputación, he pensado que debía hacer todo lo posible para evitar al menos tenerla mala. La segunda ra-
75 zón, que me ha obligado a escribir / esto, es que, observando cada día el retraso cada vez mayor que sufre el proyecto que tengo de instruirme, debido a una infinidad de experiencias que me son necesarias y que sin la ayuda de otro me es imposible realizar, aunque no me jacto hasta tal punto que espere del público una gran participación en mis planes, sin embargo, tampoco quiero ser tan descuidado que llegue a dar motivo, a los que me sobrevivan, para reprocharme algún día el haber podido dejarles algunas cosas mucho mejores, si no hubiera descuidado excesivamente el hacerles comprender de qué forma podrían contribuir a mis proyectos.

He pensado también que sería fácil escoger algunas materias que, sin estar sujetas a muchas controversias ni obligarme a declarar mis principios más de lo que quiero, me permitiesen mostrar con suficiente claridad lo que puedo o no puedo probar en las ciencias. En lo cual no podría decir si he triunfado, pues no quiero anticipar las opiniones de nadie hablando yo mismo de mis propios escritos; pero me agradaría mucho que fueran examinados y, con el fin de que haya una mayor oportunidad, ruego a todos aquéllos que tengan que hacerme alguna objeción que se tomen la molestia de enviarlas a mi

librero[18] e, informado por éste, trataré al mismo tiempo de juntar a ellas mi respuesta; y de este modo, disponiendo los lectores de ambas opiniones juntas, podrán juzgar mucho más fácilmente acerca de la verdad. Pues prometo no dar jamás respuestas largas, limitándome a reconocer con toda franqueza
76 mis equivocaciones, si me doy cuenta de ellas, / o bien, si no me es posible percibirlas, decir simplemente lo que crea necesario para la defensa de las tesis que he formulado, sin añadir explicación de ninguna materia nueva, con el fin de no comprometerme sin fin en pasar de una a otra.

Pero si alguna de las que he hablado al comienzo de la *Dióptrica* y de los *Meteoros* sorprende a primera vista, porque las denomino hipótesis[19] y no parece que tenga intención de probarlas, ruego que se tenga la paciencia de leer todo el tratado con atención y confío en que se encontrará satisfecho. Pues considero que en él las razones se siguen unas de otras de tal forma que, así como las últimas se demuestran por las primeras, que son sus causas, estas primeras lo son recíprocamente por las últimas, que son sus efectos. Y no hay que pensar que por ello cometo el defecto que los lógicos llaman círculo; porque al mostrar la experiencia que la mayor parte de estos efectos son muy ciertos, las causas de donde los deduzco no sirven tanto para probarlos como para explicarlos; y, al contrario, son

[18] Jean Maire, en Leyden (Holanda), donde apareció publicado el *Discurso* el 8 de mayo de 1637.

[19] Las denomina *hipótesis* en la ed. lat. (*quia hypothesis voco*, AT, VI, p. 582) y *supuestos* en la ed. fr. (*que je les nomme des suppositions*, AT, VI, p. 76). Se trata, en definitiva, de los principios físicos, a partir de los cuales se deduce una serie de consecuencias o fenómenos, que serán probados mediante la experiencia o método experimental. Cf. M. Martinet, «Science et hypothèse chez Descartes», *Archives Intern. d'hist. des sciences*, 24, 1974, pp. 319-339; G. Milhaud, «Descartes expérimentateur», en *Descartes savant*, J. Vrin, Paris, 1921.

éstas las que son probadas por aquéllos[20]. Y no las he denominado hipótesis sino con el único fin de que se sepa que creo poder deducirlas de estas primeras verdades que he explicado anteriormente, pero que expresamente no he querido hacerlo para impedir que ciertos ingenios, que se imaginan poder conocer en un día todo lo que otro ha pensado en veinte años tan pronto como se les diga dos o tres palabras del tema, y que son tanto más propensos al error y menos capaces de saborear la verdad cuanto
77 más penetrantes y vivos son, pudieran / aprovechar de ahí la ocasión para construir alguna filosofía extravagante sobre lo que creyeran ser mis principios, y luego se me echara la culpa. Pues, en cuanto a las opiniones que son enteramente mías, no pretendo justificarlas como nuevas, dado que si se consideran atentamente las razones en que se fundan, estoy seguro de que se las encontrará tan simples y tan conformes con el sentido común que parecerán menos extraordinarias y menos extrañas que algunas otras

[20] Esta reflexión metodológica contrasta, sin duda alguna, con la que propone el mismo Descartes tanto en las *Reglas* como en la Parte II del *Discurso*. El modelo matemático, deductivo, que allí se configura, sólo pudo ser aplicado en el ensayo *Geometría*, en la investigación de los principios de la nueva filosofía, la metafísica, así como en la física especulativa. Consciente de los límites de dicho modelo (DM, VI, p. 88, n. 10), Descartes nos expone aquí el método practicado en los demás ensayos, concretamente en *Dióptrica* y *Meteoros*, que no es sino el método hipotético-deductivo, muy eficaz en la ciencia de Galileo. Al comienzo de dichos ensayos científicos nos habla de *hipótesis* que han de ser *probadas*. Y si la hipótesis no es un enunciado que exprese una verdad exacta, sino provisional o conjetural (Popper), ello significa que el racionalista Descartes se ha visto en la necesidad de recurrir a la observación empírica y al experimento (cfr. *Meteoros*, Discurso Octavo), bien para formular la hipótesis, bien para probar o refutar la verdad de la misma. Ello significa un compromiso con la hipótesis, en la nueva ciencia, pero sin renunciar al deductivismo (cfr. D. M. Clarke, *La filosofía de la ciencia de Descartes*, Alianza, Madrid, 1986, p. 28).

que se pudieran tener acerca de los mismos temas. Tampoco presumo de ser el primer inventor de algunas, sino sólo de no haberlas admitido jamás como mías, porque hayan sido dichas por otros o porque no lo hayan sido, sino únicamente porque la razón me ha persuadido de su verdad[21].

Si los artesanos no pueden ejecutar rápidamente la invención explicada en la *Dióptrica*[22], no creo que por ello se pueda decir que carece de valor; pues, dado que se necesita destreza y práctica para construir y para ajustar las máquinas que he descrito sin que falte detalle alguno, si tuvieran éxito al primer intento, me sorprendería tanto como si alguien pudiera aprender en un día a tocar el laúd a la perfección, sólo porque se le hubiera entregado una buena partitura. Y si escribo en francés, que es la lengua de mi país, y no en latín[23], que es la de mis preceptores, es porque confío en que aquellos que únicamente se sirven de su pura razón natural juzgarán mejor mis opiniones que aquellos que no creen sino en los libros antiguos[24]. Y en cuanto a los

[21] Insiste Descartes en el rechazo del argumento de autoridad y en constituir, en cambio, a la razón en tribunal de la verdad.

[22] Se refiere al discurso X de la *Dióptrica*, en donde explica la forma de tallar las lentes (cf. la correspondencia con Ferrier en 1629).

[23] ¿Por qué escribe esta primera edición en francés? ¿Para garantizar una mayor difusión? ¿Para mostrar mayor distanciamiento respecto de los «doctos»? Lo cierto es que, si bien casi todas sus obras posteriores fueron escritas originalmente en latín —y hasta una trad. lat. del *Discurso* aparece en Amsterdam en 1644—, en general utiliza alternativamente el francés y el latín. Por otra parte, la ed. lat. no reproduce desde «Y si escribo en francés...» hasta la conclusión de este parágrafo. Ésta es la variante más notoria (cf. AT, VI, p. 583). Cf. J. Derrida, «Descartes: lengua e institución filosófica», en *La filosofía como institución*, Granica, Barcelona, 1984. pp. 145-186.

[24] Descartes concluye el *Discurso* como lo empezó, esto es, reafirmando el principio de la racionalidad. La luz de la razón como fundamento del saber en general y, en concreto, de la ciencia, se contrapone aquí a los libros antiguos. No otro es el tema central de *La*

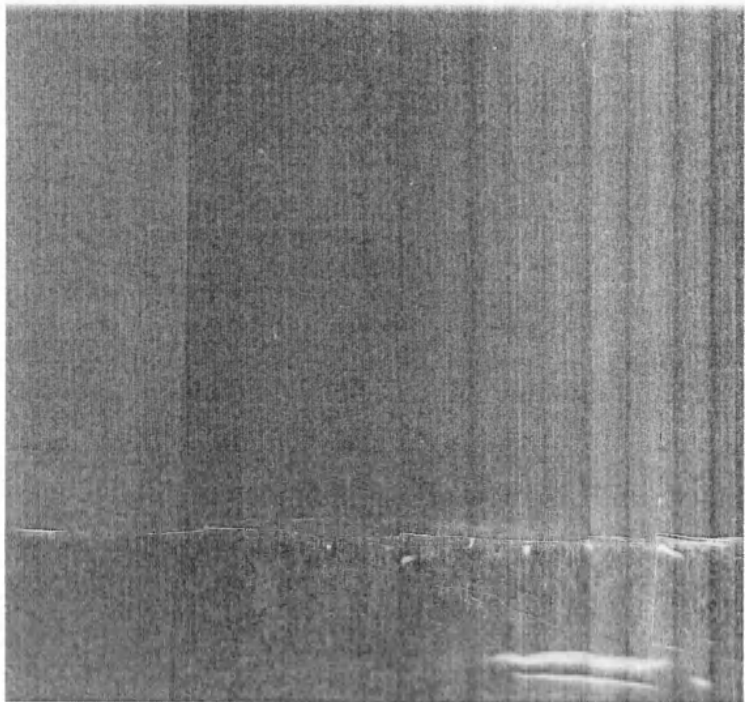
que unen al buen sentido el estudio, los únicos que
78 deseo tener / por mis jueces, estoy seguro de que no serán tan partidarios del latín que rehúsen conocer mis razonamientos porque los exponga en lengua vulgar.

Por lo demás, no quiero hablar aquí particularmente de los progresos que sucesivamente espero hacer en las ciencias, ni comprometerme con el público mediante alguna promesa que no esté seguro de poder cumplir; solamente diré que he decidido emplear el tiempo que me queda de vida exclusivamente en tratar de adquirir algún conocimiento de la naturaleza, que sea tal que puedan deducirse de él normas aplicables en la medicina, más seguras que las que ha habido hasta el presente, y que mi inclinación me aleja hasta tal punto de todo otro tipo de proyectos, principalmente de aquellos que no podrían ser útiles a unos sino perjudicando a otros, que, si determinadas circunstancias me forzaran a dedicarme a ellos, creo que no sería capaz de llevarlos a buen término. Sé que la declaración que aquí hago no puede contribuir a hacerme importante en el mundo, pero tampoco deseo serlo; pues siempre me consideraré más obligado hacia aquéllos por cuyo favor gozaré sin obstáculo de mi tiempo de ocio, que con aquéllos que me ofrecieran los empleos más honorables de la tierra.

recherche de la vérité par la lumière naturelle (AT, X, pp. 495 ss.). No otra es la tesis de la razón moderna. Cf. P. A. Schouls, «Descartes and the autonomy of Reason», *Journal of the History of Philosophy*, X:3, 1972, pp. 307-323; J. Morris, «Descartes' natural light», *Journal of the Hist of Philosophy*, XI:2, 1973, pp. 169-187; P. A. Schouls, «Cartesian certainty and "natural light"», *Australian Journal of Philosophy*, 48, 1970, pp. 116-119; «Reason, method, and science in the philosophy of Descartes», *ibid.*, 50, 1972, pp. 30-39; G. Buchdahl, «Descartes' anticipation of a "logic scientific discovery"», en *Scientific Change*, ed. A. C. Crombie, Heinemann, London, 1963, pp. 399-417.

Eduardo Bello, catedrático de Filosofía de la Universidad de Murcia, ha cursado estudios de postgrado en Leuven como becario del CSIC. Ha publicado, sobre la modernidad, trabajos como «Descartes, lo matemático y la constitución del saber moderno» (1997); sobre el pensamiento ilustrado, *La aventura de la razón* (1997) y, en coedición, *La actitud ilustrada* (2002), así como un monográfico en la revista *Daimon* (1993), de la que es su actual director. Entre sus publicaciones sobre filosofía de la existencia hay que mencionar *De Sartre a Merleau-Ponty* (1979).

Signo expresivo del pensamiento de su autor, el *Discurso del método* (1637) es también huella fehaciente de las tensiones y problemas de una época. Confluencia de diferentes proyectos, la articulación del texto se observa, más que en el discurso del método como tal, en la tarea de fundamentar el nuevo saber —teórico y práctico— moderno. El estilo autobiográfico, más vivo en la Parte I, opera como máscara que acentúa un determinado gesto: destruir críticamente el viejo edificio del saber y alzar sobre otros cimientos el saber moderno. La Parte II especifica el cimiento epistemológico (metodológico), iniciado en las Reglas, y formula la exigencia de nuevo fundamento (ontológico). El nuevo saber es también práctico; de ahí el esbozo original de la moral —Parte III—, que desarrollará en *Cartas* y en las *Pasiones del alma*. La meditación metafísica de la Parte IV —continuada en *Meditaciones metafísicas*— constituye uno de los signos de la época moderna, al darle un fundamento de su figura mediante una determinada interpretación de lo existente y de la verdad. Nuevo signo es la ciencia física de la Parte V, donde resuenan *El Mundo* y la voz polémica de Galileo. En la Parte VI aparece un tercer signo, la máquina o la técnica, aplicación práctica del saber.



«En lo cual me parecen semejantes a un ciego que, para luchar en igualdad de condiciones contra otro que ve, le hubieran conducido hasta el fondo de una cueva muy oscura; y puedo afirmar que éstos tienen interés en que me abstenga de publicar los principios de la filosofía que utilizo, pues siendo tan simples y tan evidentes como son, al publicarlos, ocurrirá como si abriera algunas ventanas e hiciera entrar la luz del día en la cueva adonde han descendido para batirse.»

Tercer milenio

Importador por:
Librerias Crisol S.A.C.
Av. Tudela y Varela 445 - Lima 18
RUC 20501457869 - Telef. 440 8060
importaciones@crisol.com.pe

ISBN 978-84-309-4371-5

tecnos
CLÁSICOS DEL PENSAMIENTO

1229620
www.tecnos.es

LA GEOMETRIA

Advertencia:

Hasta aquí he intentado que cualquier persona pudiera entender mis escritos; sin embargo, temo que este tratado no podrá ser leído sino por aquellos que ya tienen conocimiento de lo que se expone en los estudios de Geometría, pues, considerando que incluyen verdades muy correctamente demostradas que me han sido de gran utilidad, he considerado superfluo repetirlas[1].

LA GEOMETRIA

LIBRO PRIMERO

A-T, VI, 369 SOBRE LOS PROBLEMAS QUE PUEDEN CONSTRUIRSE EMPLEANDO SOLAMENTE CÍRCULOS Y LÍNEAS RECTAS

Todos los problemas de la Geometría pueden ser reducidos fácilmente a términos tales que no sea necesario posteriormente para construirlos, sino conocer la longitud de algunas líneas.

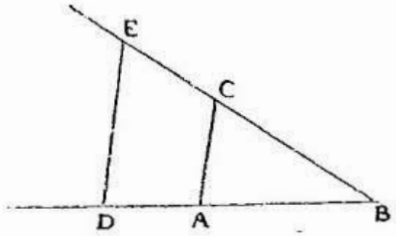
Cómo el cálculo de la aritmética se relaciona con las operaciones geométricas

Así como la Aritmética se basa en cuatro o cinco operaciones, a saber, la adición, la sustracción, la multiplicación, la división y la extracción de raíces (que puede ser considerada como una especie de división), de igual forma no es necesario en Geometría para llegar a conocer las líneas que se buscan y para disponerlas a ser conocidas, sino añadir o sustraer otras, o bien tomando una línea que consideraré como la unidad, para relacio-
A-T, VI, 370 narla tanto más fácilmente con los números (pudiendo ser tomada generalmente a discreción), y teniendo otras dos líneas, encontrar una cuarta línea que sea a cada una de las líneas dadas como la otra es a la unidad (lo cual es lo mismo que la multiplicación); o, en segundo

lugar, encontrar una cuarta línea que es a una de estas dos como la unidad es a la otra (lo que equivale a la división); o, finalmente, hallar una, dos o varias medias proporcionales entre la unidad y alguna otra línea (lo cual es equivalente a la obtención de la raíz cuadrada o cúbica, etc...). No temeré introducir estos términos en la Geometría con el fin de hacerme más inteligible[2].

La multiplicación

Sea, por ejemplo, *AB* la unidad y que sea preciso multiplicar *BD* por *BC*; solamente debo unir los puntos



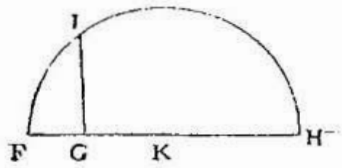
A y *C*, trazando *DE* paralela a *CA*, siendo *BE* el resultado de esta multiplicación[3].

La división

O bien, si es preciso dividir *BE* por *BD*, habiendo unido los puntos *E* y *D*, trazo *AC* paralela a *DE*, siendo, pues, *BC* el resultado de tal división[4].

La extracción de la raíz cuadrada

Si se desea calcular la raíz cuadrada de *GH*, se agrega *FG* que es la unidad y dividiendo *FH* en dos partes



iguales, tomando como centro el punto *K*, trazo el círculo *FIH*. Seguidamente trazo sobre el punto *G* hasta

el punto I la perpendicular a FH, siendo GI la raíz
A-T, VI, 371 buscada. No me refiero a la raíz cúbica ni a otras, puesto que las trataré más pormenorizadamente en otro lugar[5].

Sobre el uso de letras en Geometría

Pero frecuentemente no es necesario trazar de esta forma tales líneas sobre el papel, siendo suficiente designar cada una de ellas por una letra[6]. Así, para sumar la línea BD y GH, llamo a la una a y a la otra b y escribo $a+b$. Por tanto, escribiré $a-b$ para indicar la resta de b respecto de a. Y escribiré ab para indicar la multiplicación de la una por la otra. Asimismo, $\frac{a}{b}$ para dividir a por b. Y aa o a^2 para multiplicar a por sí misma. Y a^3 para multiplicar este resultado una vez más por a, y así hasta el infinito; y $\sqrt{a^2+b^2}$ para obtener la raíz cuadrada de a^2+b^2; finalmente $\sqrt{C\cdot a^3-b^3+abb}$ para obtener la raíz cúbica de a^3-b^3+abb, y de forma similar para otras[7].

Debe observarse que con a^2, b^3, y expresiones similares no concibo ordinariamente sino líneas simples, aunque las nombre cuadrados o cubos porque me sirvo de las expresiones utilizadas en el álgebra[8].

Igualmente debe tenerse en cuenta que todas las partes de cada línea deben, ordinariamente, expresarse por el mismo número de dimensiones, cuando la unidad no ha sido establecida en la formulación del problema. Así, a^3, contiene las mismas dimensiones que abb o b^3, siendo éstas las componentes de la línea que he denominado $\sqrt{C\cdot a^3-b^3+abb}$. No sucede lo mismo, sin embargo, cuando la unidad está determinada, pues ésta puede suponerse siempre cualquiera que sean las dimensiones. Así, si es preciso extraer la raíz cúbica de $aabb-b$, debemos considerar que la cantidad $aabb$ está dividida una vez por la unidad y la cantidad b
A-T, VI, 372 está multiplicada dos veces por la misma unidad.

Finalmente, con el fin de no olvidar los nombres de estas líneas, debe ser confeccionada una lista tantas veces cuantas introduzcamos variaciones. Por ejemplo, escribiremos:

$AB=1$, es decir, AB igual a 1
$GH=a$,
$BD=b$, etc...[9]

Sobre el procedimiento para acceder a las ecuaciones que sirven para resolver los problemas

Si, pues, deseamos resolver un problema, inicialmente debe suponerse efectuada la resolución, dando nombre a todas las líneas que se estimen necesarias para su construcción, tanto a las que son desconocidas[10] como a las que son conocidas. A continuación, sin establecer distinción entre las líneas conocidas y las desconocidas, debemos descifrar el problema siguiendo el orden que muestre de modo más natural las relaciones entre estas líneas, hasta que se indentifique un medio de expresar una misma cantidad de dos formas: esto es lo que se entiende por una ecuación, pues los términos de una de estas expresiones son iguales a los de la otra. Deben hallarse tantas ecuaciones como líneas desconocidas se han supuesto. Pero si no se logra esto y, no obstante, no se ha omitido consideración alguna de lo especificado en el problema, esto testimonia que el problema no está completamente determinado. En tal caso podemos elegir arbitrariamente líneas de longitud conocida para A-T, VI, 373
cada línea a la que no corresponda una ecuación. Si después de esto aún existen varias, es preciso servirse por orden de cada una de las ecuaciones restantes, bien sea considerándolas aisladamente, bien cada una en comparación con las otras, para obtener un valor para cada una de las líneas desconocidas; debe procederse de este modo hasta que no exista sino una sola línea desconocida que sea igual a alguna línea conocida o cuyo cuadrado, cubo, cuadrado del cuadrado, supersólido, cuadrado del cubo, etc.., sea igual a la suma o diferencia de dos o más cantidades, una de las cuales sea conocida y las otras estén compuestas de algunas medias proporcionales entre la unidad y ese cuadrado, cubo, cuadrado del cuadrado, etc..., multiplicado por otras conocidas[11]. Esto lo expreso del modo siguiente:

$z=b$
o $z^2=-az+bb$,
o $z^3=+az^2+bbz-c^3$,
o $z^4=az^3-c^3z+d^4$, etc...

Es decir, z, tomada como la cantidad desconocida es igual a b; o el cuadrado de z es igual al cuadrado de b menos a multiplicado por z; o el cubo de z es igual a a multiplicado por el cuadrado de z más el cuadrado de b

multiplicado por z, menos el cubo de c, y así en otros casos.

A-T, VI, 374 De esta manera pueden reducirse todas las cantidades desconocidas a una sola, siempre que el problema pueda ser construido mediante círculos y líneas rectas o por medio de secciones cónicas, o también por medio de cualquier otra línea curva de grado no superior al tercero o cuarto. Pero no me detengo en la explicación detallada de esto, pues os privaría del placer de aprender por vosotros mismos y de la utilidad de cultivar vuestro espíritu al ejercitarse en estas cuestiones, que es, según mi opinión, el principal resultado que se puede obtener de esta ciencia. Pues no creo exista entre estas cuestiones alguna tan difícil que no puedan solucionar aquellos que sean un poco más versados en Geometría común y en el Algebra si prestan atención a cuanto se diga en este tratado.

Por esta razón me limitaré a advertiros que, si resolviendo estas ecuaciones no se descuida realizar todas las divisiones posibles, infaliblemente se alcanzarán los términos más simples a los cuales puede ser reducido el problema.

Cuáles son los problemas planos

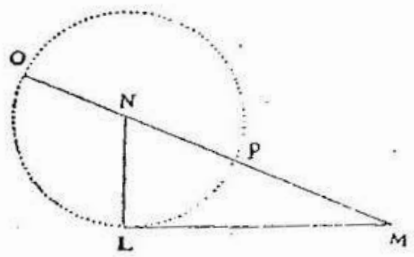
Si el problema puede ser solucionado mediante la Geometría ordinaria, esto es, mediante el uso exclusivo de líneas rectas y círculos trazados sobre una superficie plana, cuando la última ecuación haya sido resuelta no nos encontraremos sino un cuadrado desconocido igual al resultado de multiplicar su raíz por alguna cantidad conocida y sumar o restar alguna otra cantidad conocida[12].

Cómo se resuelven

Entonces esta raíz o línea desconocida es fácilmente calculable[13], pues si tengo, por ejemplo:

$$z^2 = az + bb,$$

A-T, VI, 375 construyo el triángulo rectángulo NLM, cuyo lado LM es igual a b (raíz cuadrada de la cantidad conocida bb) y el otro lado, LN, es igual a $\frac{1}{2}a$, esto es: la mitad de la



otra cantidad conocida que estaba multiplicada por z, la cual he supuesto que es la línea desconocida[14]. Seguidamente, prolongando MN, que es la base[15] de este triángulo, hasta O, de suerte que NO sea igual a NL, la línea OM es z, la línea buscada. La cual se expresa de la siguiente forma:

$$z = \frac{1}{2}a + \sqrt{\frac{1}{4}aa + bb}$$

Pero si,

$$yy = -ay + bb,$$

siendo y la cantidad que deseamos calcular, entonces construyo el mismo triángulo NLM y de la hipotenusa MN resto NP que es igual a NL, obteniendo que PM es y, que era la raíz que deseaba calcular. De forma que:

$$y = -\frac{1}{2}a + \sqrt{\frac{1}{2}aa + bb}.$$

Y de la misma forma si tuviera

$$x^4 = -ax^2 + b^2,$$

PM sería x^2, teniendo entonces que:

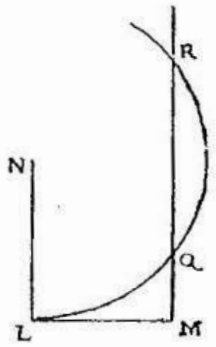
$$x = \sqrt{-\frac{1}{2}a + \sqrt{\frac{1}{4}aa + bb}};$$

y así en otros casos.

Finalmente, si tengo A-T, VI, 376

$$z^2 = az - bb,$$

supongo NL igual a $\frac{1}{2}a$ y LM igual a b como en el caso anterior. Después, en lugar de unir los puntos M y N, trazo MQR paralela a LN, y tomando N como centro



trazo por L un círculo que corta a MQR en los puntos Q y R. Por tanto, la línea buscada, z, es MQ o bien MR, pues en este caso se expresa de dos formas[16], a saber:

$$z = \frac{1}{2}a + \sqrt{\frac{1}{4}aa - bb} \quad ,$$

$$z = \frac{1}{2}a - \sqrt{\frac{1}{4}aa - bb} \quad .$$

Y si el círculo que, teniendo su centro en el punto N y pasando por el punto L, no corta ni toca la línea recta MQR, entonces no existe raíz alguna de la ecuación. En consecuencia, puede asegurarse que es imposible la construcción del problema propuesto.

Finalmente, estas mismas raíces pueden encontrarse por una infinidad de medios. Solamente he deseado presentar éstos por ser muy simples con el fin de mostrar que pueden construirse todos los problemas de la geometría ordinaria sin hacer otra cosa que lo expuesto en las cuatro figuras explicadas. Esto no creo que haya llegado a ser conocido por los antiguos, pues en caso contrario no se hubiesen tomado la molestia de escribir tan gruesos volúmenes, ya que el solo orden de sus proposiciones nos permite concluir que no han conocido el verdadero método para calcular todas, sino que solamente han compilado lo que ocasionalmente llegaron a conocer.

Ejemplo tomado de Pappus

A-T, VI, 377 Esto mismo puede observarse con gran claridad en lo que Pappus[17] ha afirmado al inicio de su séptimo libro,

dada, el punto se encontrará sobre una cierta línea. Si las rectas dadas son seis y se da la proporción del sólido comprendido bajo tres de las trazadas con el sólido bajo las otras tres, el punto se encontrará igualmente sobre cierta línea. Pero si hay más de seis rectas, no podemos conocer si se guarda la proporción de algo comprendido por cuatro líneas con algo comprendido bajo el resto, pues nada hay que esté comprendido bajo más de tres dimensiones.

Os ruego que observéis de paso que el escrúpulo que tenían los antiguos en usar notaciones propias de Aritmética en Geometría, que no podía proceder sino de que no veían con bastante claridad su relación, era la causa de la gran oscuridad y dificultad en el modo en que se expresaban, pues Pappus sigue expresándose del modo siguiente:

A-T, VI, 379 *Están de acuerdo en esto[20] quienes en épocas pasadas han analizado tales cuestiones en cuanto defienden que la figura comprendida por tales líneas no es comprensible en modo alguno. Sin embargo, es posible enunciar y demostrar en general por medio de relaciones compuestas las anteriores proposiciones del modo siguiente: Si desde un punto se trazan sobre rectas dadas otras rectas que formen ángulos dados y, a la vez, se posee la relación compuesta de una de las trazadas a otra de ellas, de la segunda con la segunda, de la tercera con la tercera y, si son siete, de la última con una línea dada o, si son ocho, la de la cuarta con su correspondiente, el punto se encontrará sobre una de las líneas dadas en posición. De modo semejante, cualquiera que sea el número de rectas —par o impar—, puede afirmarse lo mismo. Pero, como he dicho, para cualquiera de los lugares que siguen al que corresponde a cuatro rectas, ninguna síntesis ha sido propuesta de modo que la línea sea conocida, etc...*

Así pues, la cuestión cuya solución había sido iniciada por Euclides y continuada por Apolonio sin haber llegado a ser solventada por nadie, es la siguiente: Teniendo tres, cuatro o un mayor número de rectas dadas en posición[21], se intenta hallar en primer lugar un punto desde el cual se pudiesen trazar tantas líneas rectas, una sobre cada una de las dadas, formando ángulos dados, de modo que el rectángulo formado por dos de las trazadas desde el mismo punto[22], guarde una proporción dada con el cuadrado de la tercera, en

donde después de haberse detenido algún tiempo en exponer cuanto se había escrito en Geometría por quienes le habían precedido, finalmente expone una cuestión de la que afirma que ni Euclides ni Apolonio ni ningún otro habían logrado responderla. Estas son sus palabras.

Apolonio afirma en el libro III que el problema del lugar con tres o cuatro líneas no fue enteramente solucionado por Euclides, así como tampoco lo fue por él mismo ni por ningún otro; es más, tampoco se completó lo escrito por Euclides, ateniéndose a los Elementos de las cónicas que ya habían sido demostrados en tiempos de Euclides..:

Cito la versión latina y no la griega con el fin de que pueda ser entendido con más facilidad.

Un poco más adelante expone este problema del modo siguiente:

El problema del lugar relacionado con tres o cuatro líneas, el cual sirve de motivo a Apolonio para atribuirse grandes elogios por sus oposiciones sobre él mismo sin conceder mérito alguno a quien le había precedido en sus análisis es el siguiente[18]*: Si son dadas tres líneas rectas en posición, y si son trazadas otras tres líneas rectas desde un mismo punto formándose ángulos conocidos con las tres líneas dadas, y si, a su vez, es conocida la proporción del rectángulo formado por dos de las líneas trazadas con el cuadrado de la otra, entonces el punto se encuentra en un lugar sólido, dado en posición, es decir, sobre una de las tres secciones cónicas*[19]*. Y si, de nuevo, se trazan líneas sobre cuatro rectas dadas en una* A-T, VI, 378 *determinada posición, en ángulos dados, y se da la proporción del rectángulo formado por dos de las trazadas con el formado por las otras dos, entonces y de modo semejante el punto se encuentra en una sección cónica. Por otra parte, se ha demostrado que únicamente a dos líneas corresponde un lugar plano; pero, si hay más de cuatro líneas, el lugar del punto no es de los que son conocidos; es de los llamados simplemente «líneas», sin conocerse nada más sobre su naturaleza o propiedades. Una de ellas, no la primera, pero sí la más clara, ha sido examinada, siendo de utilidad. Las proposiciones relacionadas con las mismas son éstas:*

Si desde un punto son trazadas cinco líneas rectas, dadas en posición, sobre otras rectas bajo ángulos dados, y se da la proporción entre el paralelepípedo rectángulo comprendido bajo tres de las trazadas y el paralelepípedo rectángulo comprendido bajo las otras dos y otra línea

el caso de que no haya sino tres; o bien con el rectángulo de las otras dos si no hay más que cuatro. O bien, si hay cinco, que el paralelepípedo formado por tres[23] guarde la proporción dada con el paralelepípedo construido sobre las dos restantes y otra línea dada; si hay seis, que el paralelepípedo construido sobre tres guarde una proporción dada con el paralelepípedo formado por otras tres; si hay siete, que el resultado A-T, VI, 380 obtenido cuando se multipliquen cuatro de ellas entre sí, guarde la proporción dada con el resultado de la multiplicación de las otras tres y también de una línea dada; si hay ocho, que el resultado obtenido de la multiplicación de cuatro guarde la proporción dada con el resultado obtenido de las otras cuatro. De este modo, tal cuestión puede hacerse extensiva a cualquier otro número de líneas[24]. A continuación y puesto que existen una infinidad de puntos que pueden cumplir lo que se trata de hallar, es necesario conocer y trazar la línea en la que se encuentran todos; Pappus afirma que, cuando no existen sino tres o cuatro líneas rectas dadas, se encuentra en una de las tres secciones cónicas, pero no intenta determinarla ni describirla, al igual que tampoco intenta explicar aquellas en que deben encontrarse todos estos puntos cuando la cuestión es planteada para un número mayor de líneas. Solamente indica que los antiguos habían imaginado una de aquellas que habían mostrado ser muy útil, que parecía la más simple, pero, sin embargo, no era la más importante. Esto me ha dado la oportunidad para intentar si puedo llegar tan lejos como lo que ellos han alcanzado mediante el método que utilizo.

Respuesta al problema de Pappus

En primer lugar, he llegado a conocer que si este problema es propuesto únicamente para tres, cuatro o cinco líneas, siempre podemos conocer los puntos buscados mediante la Geometría Simple, es decir, sirviéndonos exclusivamente de la regla y el compás y no haciendo sino lo que se ha dicho; únicamente debe exceptuarse cuando hay cinco líneas dadas y todas ellas son paralelas. En este caso, al igual que cuando hay seis, A-T, VI, 381 siete, ocho o nueve líneas, pueden siempre ser hallados los puntos deseados mediante la Geometría de sólidos,

esto es, empleando alguna de las tres secciones cónicas; debemos exceptuar solamente aquel caso en el que hay nueve líneas dadas y todas ellas son paralelas. Para tal caso, como para cuando hay 10, 11, 12 ó 13 líneas, pueden hallarse los puntos buscados por medio de una línea curva que sea de un grado mayor que las secciones cónicas, exceptuando el caso en que fueren trece las líneas dadas si son todas ellas paralelas. En este caso, como cuando son dadas 14, 15, 16, 17, será preciso emplear una línea curva de un grado aún mayor que la precedente y así indefinidamente.

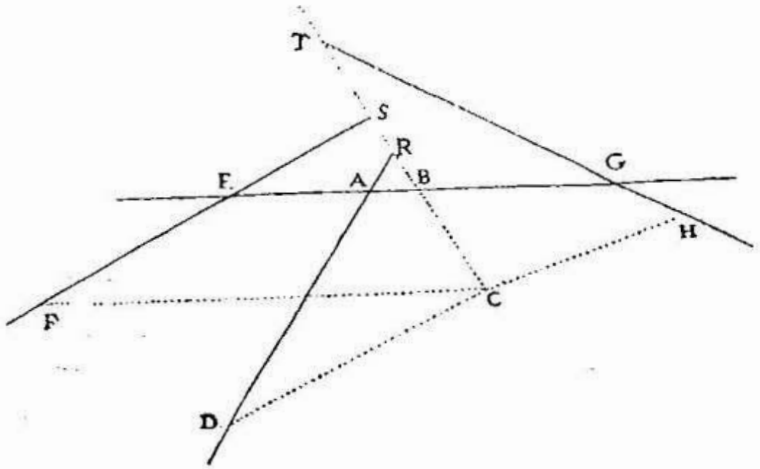
En segundo lugar, he hallado que cuando no hay sino tres o cuatro líneas dadas, los puntos buscados se encuentran todos no solamente en una de las tres secciones cónicas, sino también y en algunas ocasiones en la circunferencia de un círculo o en una línea recta. Igualmente, cuando hay 5, 6, 7 u 8, todos estos puntos se encuentran en alguna de las líneas que son de un grado más que las secciones cónicas, siendo imposible imaginar alguna que no sea útil para tal problema; pero también puede encontrarse excepcionalmente en una sección cónica, en un círculo o en una línea recta; si hay 9, 10, 11 ó 12, estos puntos se encuentran en una línea que no puede ser sino de un grado superior a las precedentes; pero todas las que son de un grado más pueden ser usadas para el problema y así al infinito.

Finalmente, la primera y la más simple de todas las
A-T, VI, 382 líneas después de las secciones cónicas, es la que puede describirse por la intersección de una parábola y de una línea recta en la forma que explicaremos. De suerte que yo pienso haber respondido enteramente a lo que Pappus dice haber sido investigado en relación con esto durante la antigüedad; por ello, trataré de dar la demostración en pocas palabras, pues ya me hastía extenderme de tal modo.

Sean AB, AD, EF, GH, etc..., varias líneas dadas, debiendo hallarse un punto como C, desde el que trazando varias líneas rectas sobre las líneas dadas, como CB, CD, CF y CH, de modo que los ángulos CBA, CDA, CFE, CHG sean dados, y de modo tal que el resultado de la multiplicación de una parte de estas líneas sea igual al resultado obtenido por la multiplicación de las otras, o bien que guarden alguna otra proporción dada, lo cual en nada dificulta el problema.

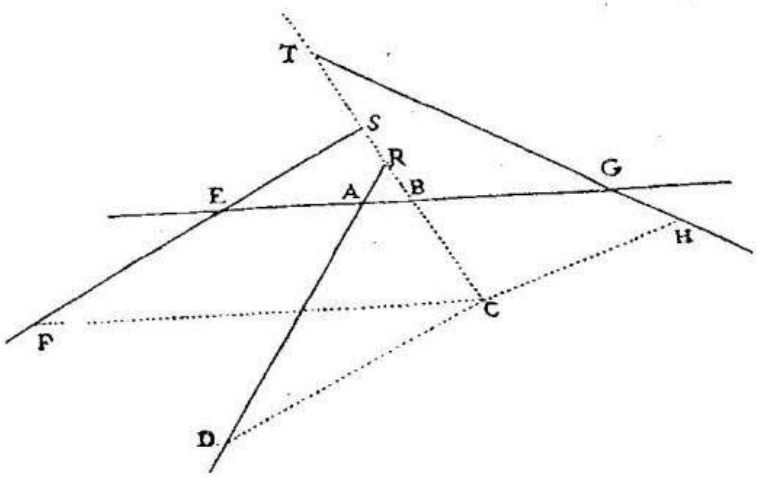
CR y CD, que establezco[30] como z es a c; de modo que si CR es $y+\frac{bx}{z}$, entonces[31] CD será $\frac{cy}{z}+\frac{bcx}{zz}$.

Después de esto, puesto que las líneas AB, AD y EF son dadas en posición, la distancia que hay entre A y E también es conocida; si tal distancia la llamo k, entonces[32] EB será igual a $k+x$; pero sería $k-x$, si el punto B se encontrase entre E y A, y $-k+x$, si E



estuviese entre A y B. Y puesto que los ángulos del triángulo ESB son dados, la proporción entre BE y BS también es conocida; la establezco como z es a d. Por tanto, $BS=\frac{dk+dx}{z}$ y la línea completa[33] $CS=\frac{zy+dk+dx}{z}$; pero sería $\frac{zy-dk-dx}{z}$ si el punto
A-T, VI, 384 S se encontrase entre B y C; y sería $\frac{-zy+dk+dx}{z}$ si C se encontrase entre B y S. Además los tres ángulos del triángulo FSC son dados y, por tanto, también lo es la proporción de CS con CF, que es la misma[34] de z a e; por tanto, la línea[35] CF será $\frac{ezy+dek+dex}{zz}$.

Del mismo modo siendo AG, a la que llamo l, dada, BG será igual a $l-x$; porque los ángulos del triángulo BGT son también dados, tenemos dada la proporción de BG



Cómo deben establecerse los términos para alcanzar la ecuación en este ejemplo

En primer lugar, supongo resuelto el problema y, para librarme de la confusión[25] de todas estas líneas, A-T, VI, 383
considero solamente una de las dadas y una de las que es preciso calcular, por ejemplo AB y CB, como las principales y con las que intento relacionar todas las otras. Concedamos que el segmento de la línea AB, que está entre los puntos A y B, sea llamado x, y que BC sea y; asimismo, supongamos que todas las otras líneas dadas sean prolongadas hasta que corten estas dos[26], también prolongadas, si es necesario y si no son paralelas: como veis aquí cortan la línea AB en los puntos A, E, G; asimismo cortan BC en los puntos R, S, T. Seguidamente como los ángulos del triángulo ARB son dados, la proporción que se da entre los lados AB y BR también es conocida, estableciéndola[27] como z es a b; de modo[28] que AB siendo x, RB será $\frac{bx}{z}$, pues el punto B se encuentra entre C y R; CR tenemos[29] que será $y+\frac{bx}{z}$; si R se encontrase entre C y B, CR sería $y-\frac{bx}{z}$; si C se encontrase entre B y R, entonces CR sería $-y+\frac{bx}{z}$. De igual modo los tres ángulos del triángulo DCR son dados y, en consecuencia, también lo es la proporción que existe entre los lados

con BT, que es como $z:f$, y, por tanto, BT será $\frac{fl-fx}{z}$ y CT será igual a $\frac{zy+fl-fx}{z}$. Ahora bien, puesto que la proporción de TC con CH es dada en virtud del triángulo TCH, estableciéndola como $z:g$, se obtendrá[36] que:

$$CH=\frac{gzy+fgl-fgx}{zz}$$

De este modo podéis ver que cualquiera que sea el número de líneas dadas en posición, cuantas líneas sean trazadas desde el punto C formando ángulos dados, siguiendo el enunciado del problema, pueden siempre ser expresadas por tres términos: uno estará compuesto por una cantidad desconocida y, multiplicada o dividida por alguna otra cantidad conocida; el otro estará integrado por la cantidad desconocida x, también multiplicada o dividida por alguna otra conocida; el tercero, por una cantidad conocida. La única excepción será A-T, VI, 385
cuando las líneas dadas sean paralelas a la línea AB, en cuyo caso el término compuesto de la cantidad x será cero; o bien a la línea CB, en cuyo caso el compuesto de la cantidad y será cero, siendo esto muy claro y, por tanto, no siendo necesario que me detenga en su explicación. En relación con los signos «+» y «−» que se unen a estos términos pueden ser combinados de todos los modos imaginables.

Así mismo podéis ver que multiplicando varias de estas líneas entre sí, las cantidades x e y que se encuentran en el producto pueden tener solamente tantas dimensiones como líneas haya, a cuya explicación sirven las que han sido multiplicadas. De suerte que nunca tendrán más de dos dimensiones, cuando han resultado de la multiplicación de dos líneas; ni más de tres, cuando son el resultado de una multiplicación de tres líneas y así hasta el infinito.

Modo de hallar que el problema es plano cuando no se ha propuesto para más de cinco líneas

Además, para determinar el punto C solamente es requerida una condición: que el producto de la multiplicación de un cierto número de líneas sea igual al

obtenido por la de otras, o (lo que no es más difícil) que guarde una proporción dada con el resultado obtenido por la multiplicación de otras; podemos tomar a discreción una de las dos cantidades desconocidas x o y, buscando la otra por esta ecuación, en la que es evidente que, cuando el problema no es propuesto para un número superior a cinco líneas, la cantidad x que no es utilizada para la expresión de la primera de las líneas nunca puede tener más de dos dimensiones. De modo
A-T, VI, 386 que, tomando y como una cantidad conocida, tendremos:

$$x^2 = \pm ax \pm b^2.$$

Por tanto y de la forma indicada, se podrá encontrar la cantidad x mediante la regla y el compás. Asimismo, tomando sucesivamente un infinito número de diferentes valores para la línea y, podremos hallar otros tantos para x; de este modo, podremos también encontrar un número infinito de diversos puntos tales como el C, pudiendo describirse la línea curva exigida.

También puede procederse de igual modo cuando el problema está planteado para seis o más líneas, si, entre las que han sido dadas, hay algunas que son paralelas a BA o BC; en este caso, una de las dos cantidades x o y no debe tener sino dos dimensiones en la ecuación[37]; de este modo es posible encontrar el punto C mediante la utilización de la regla y el compás. Pero, por el contrario, si todas son paralelas, aunque el problema sea propuesto para cinco líneas, este punto C no podrá ser hallado de este modo, puesto que no encontrándose la cantidad x en la ecuación, no será permitido tomar una cantidad conocida para la que llamamos y, sino que el valor de ésta es el que es preciso calcular. Y, puesto que y, tendrá tres dimensiones, no se la podrá calcular sino obteniendo la raíz de una ecuación cúbica, lo que usualmente no se puede lograr sin emplear por lo menos una de las secciones cónicas. Y aunque tengamos hasta nueve líneas dadas, con tal de que no todas sean
A-T, VI, 387 paralelas, puede siempre lograrse que la ecuación no sobrepase el cuadrado del cuadrado[38]; tales ecuaciones, pueden siempre resolverse por medio de secciones cónicas, del modo que más adelante explicaré. Y aunque sean dadas 13 líneas siempre puede expresarse la ecuación de modo que no alcance sino el cuadrado del cubo[39]; consecuentemente, puede resolverse del modo que explicaré por medio de una línea que es solamente

con estas líneas como con las que son más simples[40]. No afirmaré tampoco que tal exclusión sea debida a que no desearon aumentar el número de sus postulados y se consideraron satisfechos con admitir que pueden unirse dos puntos dados por una línea recta y describir un círculo desde el centro que pasase por un punto dado, pues no dudaron en modo alguno al realizar su estudio de las secciones cónicas en que debía suponerse la posibilidad de cortar un cono dado por un plano dado. Y, no es necesario suponer nada para trazar todas las líneas curvas que pretendo introducir, sino que dos o más líneas puedan moverse entre sí y que sus intersecciones engendren otras curvas; esto en nada me parece más difícil. Verdad es que no han introducido de modo perfecto las secciones cónicas en su geometría, pero yo no deseo introducir modificación alguna en los nombres normalmente aceptados. Pero me parece totalmente claro que si entendemos, como generalmente se hace, por «geométrico» lo que es preciso y exacto y, en segundo lugar, por «mecánica» lo que no lo es; y, asimismo, si consideramos la Geometría como una ciencia que enseña en general a conocer las medidas de todos los cuerpos, no existe razón alguna para excluir de la misma el estudio de las líneas más complejas y no
A-T, VI, 390 el de las más simples, con tal de que puedan imaginarse descritas por un movimiento continuo o por varios movimientos sucesivos y en los que los últimos vienen determinados por los anteriores[41]; pues, por este medio, puede siempre tenerse conocimiento exacto de su medida. Pero lo que puede quizá haber impedido la admisión por los antiguos geómetras de aquellas curvas que eran más complejas que las secciones cónicas, es que las primeras que analizaron, por casualidad, fueron la Espiral, la Cuadratriz y semejantes[42], que no pertenecen propiamente sino a las Mecánicas y no son del número de las que pienso que deban ser incluidas aquí; pienso así, puesto que podemos imaginarlas descritas por dos movimientos independientes, no guardando el uno con el otro relación alguna que pueda ser exactamente medida; aunque posteriormente hayan examinado la Concoide, la Cisoide, y algunas otras que nosotros aceptamos, sin embargo, es posible que al no haber analizado suficientemente sus propiedades, no se hayan cuestionado su importancia más de lo que cuestionaron la de las primeras. También podría haber sucedido que pensasen que no debían iniciar el análisis de una materia más difícil, ya que no conocían sino algunas

un grado superior a las secciones cónicas. Esto constituye la primera parte de lo que aquí deseaba demostrar; pero antes de iniciar la segunda, es necesario que facilite algunas observaciones relacionadas con la naturaleza de las líneas curvas.

LIBRO SEGUNDO

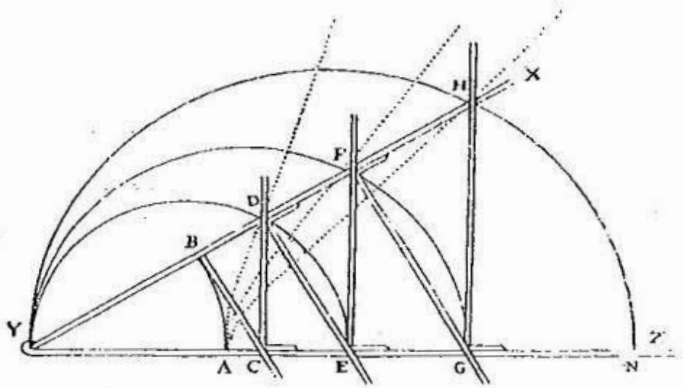
SOBRE LA NATURALEZA DE LAS LÍNEAS CURVAS A-T, VI, 388

Sobre las líneas curvas que pueden admitirse en Geometría

Los antiguos distinguieron con toda perfección la existencia de tres clases de problemas en Geometría: planos, sólidos y lineales; es decir, unos pueden ser construidos con sólo trazar líneas rectas y círculos; los segundos, por el contrario, no pueden serlo sin realizar la introducción de alguna sección cónica; finalmente, los terceros requieren el empleo de una línea más compleja. Pero no logro evitar la extrañeza que me produce el que, además de esto, no llegasen a distinguir diversos grados entre estas líneas más complejas, al igual que no logro comprender por qué las han denominado Mecánicas y no Geométricas. Pues si pensamos que las han denominado de tal modo porque es necesario utilizar algún instrumento para trazarlas, entonces deberíamos rechazar por la misma razón los círculos y las líneas rectas, puesto que no se trazan sobre el papel, sino utilizando la regla y el compás que también pueden ser considerados como máquinas. Tampoco puede explicarse porqué los instrumentos utilizados para trazarlas, A-T, VI, 389
siendo más complejos que la regla y el compás, no pueden llegar a alcanzar la misma precisión; en tal caso, deberían considerarse fuera de la Mecánica, pues la precisión de las obras realizadas por la mano es aún más buscada en ella que en la Geometría, donde exclusivamente se desea lograr precisión en el razonamiento, pudiendo obtenerse tanta perfección en relación

cosas sobre las secciones cónicas e ignoraban aún mucho de cuanto puede construirse mediante la regla y el compás. Pero, puesto que tengo la esperanza de que aquellos que tengan destreza para servirse del cálculo geométrico aquí propuesto, no encontrarán tema alguno como para hacerlos vacilar en relación con los problemas planos o sólidos, creo pues oportuno invitarles a realizar otras investigaciones en las que nunca les faltará oportunidad para ejercitarse.

Considérense las líneas *AB*, *AD*, *AF* y semejantes, que A-T, VI, 391
supongo han sido descritas con la ayuda del instrumento *XYZ*, compuesto de varias reglas. Estas están unidas de tal forma que cuando la conocida como *YZ* está sobre la línea *AN*, puede abrirse y cerrarse el ángulo



XYZ; cuando está totalmente cerrada, los puntos *B*, *C*, *D*, *F*, *G*, *H*, se encuentran todos en *A*; pero a medida que se abre, la regla *BC*, unida a *XY* de modo que forma ángulos rectos en el punto *B*, empuja hacia *Z* la regla *CD*; *CD* resbala sobre *YZ* formando siempre ángulos rectos con ella y empuja a *DE*, que se desliza igualmente sobre *YX*, permaneciendo paralela a *BC*. *DE* empuja a *EF*; *EF* a *FG* y *FG* a *GH*; asimismo, podemos concebir una infinidad de otras, que se impulsan consecutivamente de igual modo: unas forman los mismos ángulos con *YX*, y los otros con *YZ*. Así cuando se abre el ángulo *XYZ*, el punto *B* describe la A-T, VI, 392
línea *AB*, que es un círculo; los otros puntos, *D*, *F* y *H*, en donde se encuentran las intersecciones de las otras reglas, describen otras curvas, como *AD*, *AF* y *AH*, siendo cada una de ellas progresivamente más compleja que el círculo[43]. Pero no veo qué puede impedir que se

conciba tan clara y distintamente la descripción de la primera como la del círculo o, al menos, como la de las secciones cónicas; nada veo que nos impida concebir la segunda, la tercera y cuantas se puedan describir, tan adecuadamente como la primera. En consecuencia, no veo por qué todas ellas no son igualmente aceptadas con el fin de contribuir a las especulaciones de la Geometría[44].

Sobre la clasificación de todas las curvas en géneros y sobre la relación de sus puntos con los de las rectas

Podría exponer en este lugar otros medios para trazar y concebir líneas curvas que fuesen cada vez más complejas hasta el infinito. Pero, para comprender a la vez todas aquellas que se dan en la naturaleza y para clasificarlas por orden según ciertos tipos, no conozco nada más apropiado que afirmar que todos los puntos de las que pueden llamarse geométricas, es decir, de aquellas que caen bajo alguna medida precisa y exacta, tienen necesariamente alguna relación con todos los puntos de una línea recta y esta relación puede ser expresada por alguna ecuación válida para todos los puntos[45]. Así mismo la línea curva será del primero y más simple tipo, en el que no están comprendidas sino el círculo, la parábola, la hipérbola y la elipse, cuando esta ecuación no alcanza sino al rectángulo[46] de dos cantidades indeterminadas o bien al cuadrado de una misma cantidad. Pero la curva pertenece al segundo
A-T, VI, 393 tipo cuando una o ambas cantidades desconocidas de la ecuación (deben ser dos para explicar la relación entre dos puntos), alcanza el tercer o cuarto grado[47]; cuando la ecuación alcanza el quinto o sexto grado[48], la curva es del tercer tipo y así para otras hasta el infinito.

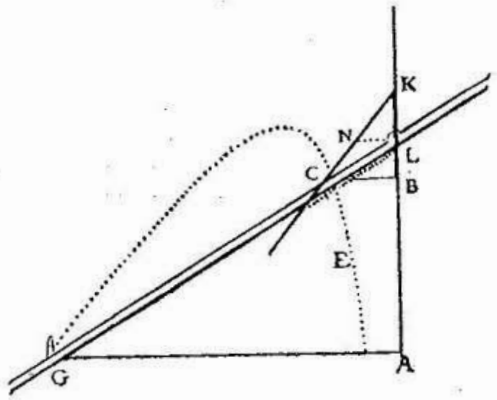
Así, supongamos que deseo saber a qué clase pertenece la línea *EC*, que imagino descrita por la intersección de la regla *GL* y del plano rectilíneo *CNKL* cuyo lado *KN* es prolongado indefinidamente hacia *C*, y que moviéndose sobre el plano en línea recta (de tal forma que su diámetro *KL* se encuentre siempre aplicado sobre algún lugar de la línea *BA*, prolongada por ambos extremos) da lugar al movimiento circular de esta regla *GL*, alrededor del punto *G*, puesto que tal regla está unida de tal modo a *CNKL* que pasa siempre por el punto *L*. En tal caso, escojo una línea recta como *AB* para relacionar con tales puntos

car la primera por la última. Por tanto, la ecuación buscada es:

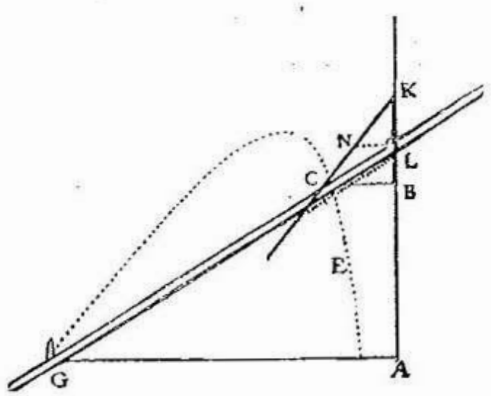
$$y^2 = cy - \frac{c}{b}xy + ay - ac.$$

En virtud de la misma conocemos que la línea *EC* es de las del primer tipo; en efecto, no es sino una hipérbola.

Si en el instrumento utilizado para describir la curva sustituimos la línea recta *CNK*, de modo que limitemos el plano *CNKL* mediante la hipérbola o cualquier otra línea curva de la primera clase, entonces la intersección de este línea y de la regla *GL* en vez de describir la hipérbola *EC*, describirá otra línea curva que pertenece-
A-T, VI, 395 rá al segundo tipo. Así, si *CNK* es un círculo, siendo *L* su centro, se describirá la primera concoide de los antiguos; pero si es una parábola, siendo su diámetro —eje— *KB*, entonces se describirá la línea curva que, como ya he dicho, era la primera y más simple de las curvas buscadas para el problema de Pappus, cuando



no son sino cinco las líneas rectas dadas en posición. Pero si en lugar de una de estas líneas curvas del primer género, es una de las del segundo género la que delimita el plano *CNKL*, entonces se describirá una del tercer tipo; mientras que si utilizamos una del tercer tipo, entonces se describirá una del cuarto tipo y así hasta el infinito. Estas afirmaciones es fácil probarlas mediante el cálculo. Y de cualquier otra forma que se imagine la descripción de una curva, siendo del número de las que son llamadas geométricas, siempre se podrá hallar una



todos los de la línea curva *EC*; asimismo, en esta línea, *AB*, tomo un punto tal como *A* para iniciar por él el cálculo. Digo que escojo tanto uno como la otra, porque pueden tomarse los que se deseen ya que aunque existan muchas alternativas que abrevien y simplifiquen la ecuación, sin embargo, de cualquier modo que se seleccionen, la curva pertenecerá siempre a la misma clase tal como es fácilmente demostrable. Después de esto, tomando un punto a discreción en la curva tal como *C*, sobre el que supongo que aplico el instrumento que sirve para describirla, trazo desde el punto *C* la línea *CB* paralela a *GA*; puesto que *CB* y *BA* son dos cantidades indeterminadas y desconocidas, las denomino *y* e *x*, respectivamente. Pero con el fin de encontrar la relación de una con la otra[48bis], considero también las cantidades conocidas que determinan la descripción o trazado de esta curva: así *GA* (llamada *a*), *KL* (llamada *b*) y *NL*, paralela a *GA* (llamada *c*). Seguidamente afirmo: que como *NL* es a *LK* (o bien *c* es a *b*), así *CB* o *y* es a *BK* que es, por tanto, igual a $\frac{b}{c}y$. Por tanto, *BL* es igual a $\frac{b}{c}y-b$, y *AL* es igual a $x+\frac{b}{c}y-b$. Además, como *CB* es a *LB*, así *GA* es a *LA*, o bien como *y* es a $\frac{b}{c}y-b$, así *a* es a $x+\frac{b}{c}y-b$. Multiplicando la segunda por la tercera tenemos $\frac{-ab}{c}y-ab$, que es igual a $xy+\frac{b}{c}y^2-by$, que es el resultado obtenido al multipli-

A-T, VI, 394

ecuación para determinar de este modo todos sus puntos.

Así mismo, considero las líneas curvas, cuya ecuación alcanza el cuarto grado, en el mismo género que aquellas que solamente alcanzan el tercer grado; aquéllas cuyas ecuaciones alcanzan hasta el cuadrado del cubo las considero en el mismo que aquellas que no alcanzan sino al supersólido; de modo semejante para otras. La razón para establecer tal clasificación se fundamenta en que existe una regla general para reducir al cubo cuantas dificultades aparezcan en el cuadrado del cuadrado y para reducir al supersólido cualquiera del cuadrado del cubo, de modo tal que en cualquier caso la última no es necesario que deba ser considerada más compleja que la precedente[49].

A-T, VI, 396

Pero es preciso señalar que entre las líneas de cada uno de estos géneros, aunque la mayor parte de ellas tengan la misma complejidad, de modo tal que puedan servir para determinar los mismos puntos y construir los mismos problemas, sin embargo, hay siempre algunas más simples, cuya utilidad es más limitada. Así entre las del primer tipo, además de la Elipse, la Hipérbola, y la Parábola, cuya complejidad es la misma, también está incluido el círculo que obviamente es una curva más simple. Entre las de la segunda clase se encuentra la Concoide vulgar que toma su origen del círculo y algunas otras que, aunque no tengan una utilidad tan amplia como la mayor parte de las de su clase, sin embargo, no pueden ser consideradas como pertenecientes al primer tipo[49 bis].

Prosigue la explicación del problema de Pappus expuesta en el libro anterior

Ahora, después de haber clasificado todas las líneas curvas según ciertos tipos o clases, me es fácil continuar con la demostración de la respuesta que anteriormente he facilitado en relación con el problema de Pappus. En primer lugar, he mostrado que cuando no hay sino tres o cuatro líneas rectas dadas, la ecuación que permite determinar los puntos buscados es de segundo grado[50]. Se infiere de ello que la curva en que tales puntos se encuentran necesariamente ha de ser alguna de las del primer tipo, puesto que esta misma ecuación expresa la relación existente entre todos los puntos de las líneas

del primer género con los de una línea recta. Cuando no
A-T. VI. 397 han sido dadas más de ocho líneas rectas, esta ecuación será, a lo sumo, una bicuadrática y, en consecuencia, la curva resultante pertenecerá a la segunda clase o a la primera. Cuando no hay más de doce líneas dadas, la ecuación no excede del cuadrado del cubo y, en consecuencia, la curva buscada no pertenecerá sino a la tercera clase o a una inferior, y así sucesivamente para otros casos. Pero, puesto que la posición de las líneas dadas puede variar de todas las formas imaginables y, en consecuencia, se introducen modificaciones en los valores de las cantidades conocidas así como en los signos «+» y «−» de la ecuación, de todas las formas imaginables, es evidente que no hay alguna línea curva del primer tipo o clase que no sea útil para este problema cuando se plantea respecto de cuatro líneas rectas; ni alguna del segundo tipo que no ofrezca solución cuando el problema ha sido planteado para ocho líneas; ni una de la tercera clase, cuando se ha propuesto para doce, y así para otros casos. De modo que no hay una línea curva, cuya ecuación pueda ser obtenida y que pueda ser aceptada en Geometría, que no sea útil para algún número de líneas.

Solución de esta cuestión cuando se propone para tres o cuatro líneas

Es necesario proceder a indicar con mayor precisión el método para descubrir la línea curva buscada en cada caso, cuando no son dadas sino tres o cuatro líneas rectas; tal análisis mostrará que la clase primera no contiene otras líneas que las tres secciones cónicas y el círculo.

Considérense nuevamente las cuatro líneas AB, AD, EF y GH, anteriormente mencionadas, y que sea preciso hallar otra línea en la que se encuentren una infinidad de puntos tales como C, de tal manera que, trazando
A-T. VI. 398 desde él mismo las cuatro líneas CB, CD, CF, CH, con ángulos dados sobre las líneas dadas, CB multiplicada por CF sea igual a CD multiplicada por CH. Esto sería equivalente a afirmar que si

$$CB = y,$$

$$CD = \frac{czy + bcx}{z^2},$$

De este modo tenemos:

$$y^2 = 2my - \frac{2n}{z}xy + \frac{bcfglx - bcfgx^2}{ez^3 - cgz^2},$$

cuya raíz es:

$$y = m - \frac{nx}{z} + \sqrt{m^2 - \frac{2mnx}{z} + \frac{n^2x^2}{z^2} + \frac{bcfglx - bcfgx^2}{ez^3 - cgz^2}}.$$

De nuevo, por razones de brevedad, en lugar de

$$-\frac{2mn}{z} + \frac{bcfgl}{ez^3 - cgz^2},$$

anotaremos o; y en lugar de

$$\frac{n^2}{z^2} - \frac{bcfg}{ez^3 - cgz^2},$$

escribiremos $-\frac{p}{m}$. Hacemos esto, pues podemos designarlas como nos plazca, considerando que todas son cantidades dadas. Tendremos entonces que:

$$y = m - \frac{n}{z}x + \sqrt{m^2 + ox - \frac{p}{m}x^2}.$$

Tal debe ser la longitud de la línea BC, quedando AB, esto es x indeterminada[51]. También es evidente que, no siendo propuesto el problema sino para tres o cuatro líneas, es obvio que siempre tendremos tales términos, aunque alguno de ellos puedan ser nulos y los signos + y − puedan variar.

A-T. VI. 400 Después de esto, trazo KI igual y paralela a BA, de modo que de BC obtengamos el segmento BK igual a m (puesto que la expresión para BC es $+m$; si hubiese sido $-m$, esta línea IK debería haberla trazado del otro lado de AB y, finalmente, no la hubiera trazado si el valor de m hubiera sido *cero*). Seguidamente trazo IL de modo que $IK : KL = z : n$; esto es, de modo que si IK es igual a x, KL es igual a $\frac{n}{z}x$. De esta misma forma conozco la proporción existente entre KL e IL que puedo establecer como $n : a$, de modo que si KL es igual a $\frac{n}{z}x$, entonces IL es igual a $\frac{a}{z}x$. De este modo establezco el punto K entre L y C, pues la ecuación

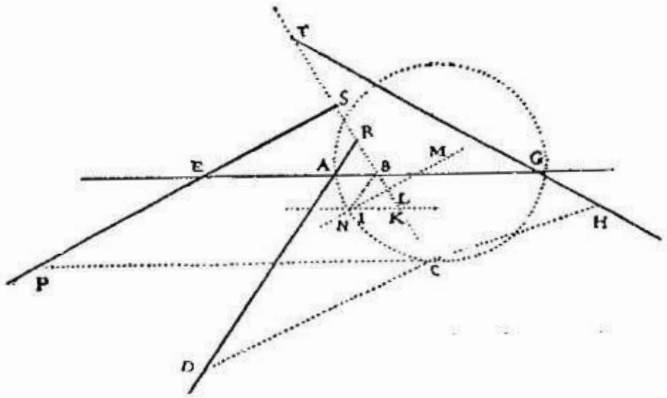
$$CF = \frac{ezy + dek + dex}{z^2},$$

$$CH = \frac{gzy + fgl - fgx}{z^2},$$

entonces la ecuación es:

$$y^2 = \frac{(cfglz - dekz^2)y - (dez^2 + cfgz - bcgz)xy + bcfglx - bcfgx^2}{ez^3 - cgz^2}$$

Se ha supuesto que *ez* es mayor que *cg*; si fuese menor sería preciso cambiar los signos «+» y «−». Si el valor de *y* fuese *cero* o menor que *cero* en esta ecuación,



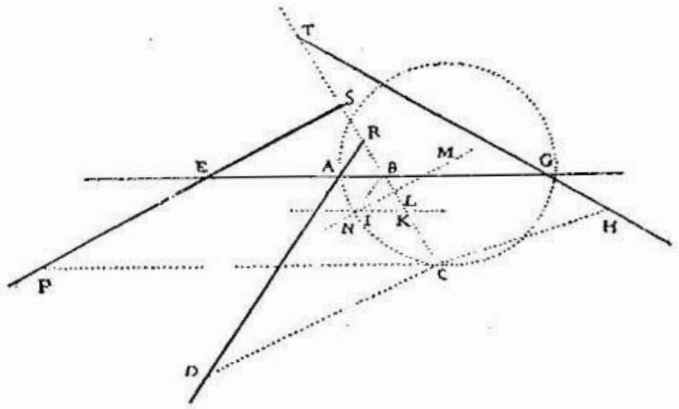
habiendo supuesto que el punto *C* se encuentra en el ángulo *DAG*, será preciso suponerlo también en el ángulo *DAE*, o *EAR*, o *RAG*, debiendo ser alterados los signos «+» y «−» según sea requerido. Si para cada una de estas cuatro posiciones el valor de *y* fuese igual a *cero*, entonces el problema no tendría solución para el caso propuesto. Pero supongamos que es posible y, para abreviar las expresiones, en lugar de las cantidades A-T, VI, 399

$$\frac{cflgz - dekz^2}{ez^3 - cgz^2},$$

escribiremos $2m$; y $\frac{2n}{z}$ en lugar de:

$$\frac{dez^2 + cfgz - bcgz}{ez^3 - cgz^2}.$$

indica $-\frac{n}{z}x$; hubiera establecido *L* entre *K* y *C*, en el caso de que $+\frac{n}{z}x$; en el caso de que $\frac{n}{z}x$ fuera igual a *cero*, no hubiera sido trazada esta línea, *IL*.



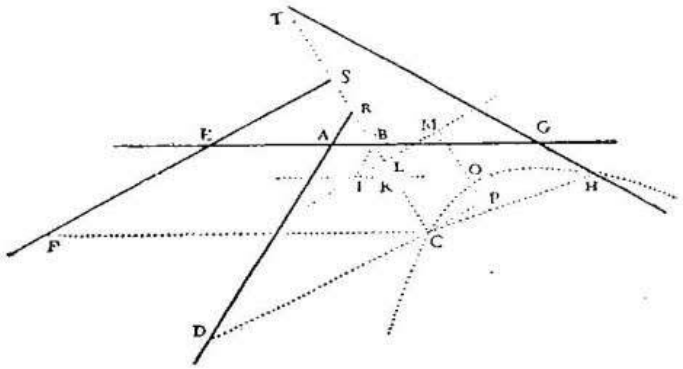
Hecho esto, obtenemos la expresión:

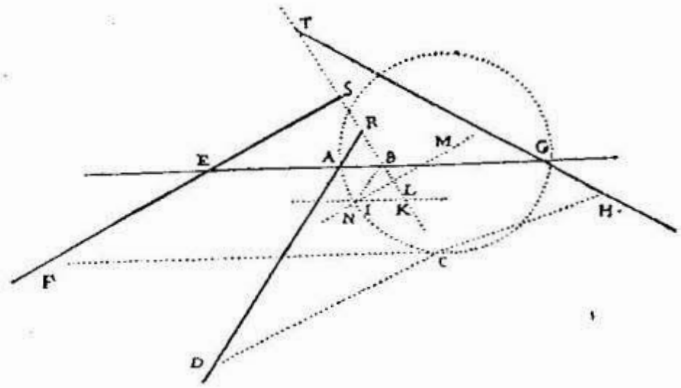
$$LC = \sqrt{m^2 + ox - \frac{p}{m}x^2}.$$

Si el valor de estos términos fuese *cero*, el punto *C* debería localizarse en la línea recta *IL*; si fuesen sus valores tales que fuese posible extraer su raíz, esto es, si m^2 y $\frac{p}{m}x^2$ tuviesen el mismo signo, bien positivo o bien negativo, y o^2 fuese igual a $4pm$, o también si m^2 y ox, o bien ox y $\frac{p}{m}x^2$ tuviesen como valor *cero*, este punto *C* debería localizarse en otra línea recta cuya posición podría ser tan fácilmente calculable como la de *IL*. Pero cuando no se da ninguna de estas posibilidades, el punto *C* se encuentra siempre en una de las tres secciones cónicas o en un círculo, uno de cuyos diámetros se encuentra en la línea *IL*, y la línea *LC* es una de las que se aplican por orden a este diámetro; también, por el contrario, puede ser *LC* paralela al diámetro al que aquélla, que está en la línea *IL*, se aplica por orden. A saber si el término $\frac{p}{m}x^2$ tiene valor *cero*, entonces tal sección cónica será una parábola; si tiene A-T, VI, 401

valor positivo (+), entonces es una hipérbola y si, por el contrario, tiene valor negativo (−), entonces es una elipse. La única excepción sería cuando la cantidad a^2m es igual a pz^2, cuando ILC es un ángulo recto, en cuyo caso tenemos un círculo en vez de una elipse. Y si esta sección es una parábola, su lado recto es igual a $\frac{oz}{a}$ y su eje se encuentra siempre sobre la línea IL; para hallar el punto N, que es el vertex, es necesario construir IN igual a $\frac{am^2}{oz}$, y el punto I debe encontrarse entre L y N, si m^2 es positivo y ox también lo es[52]; también el punto L debe hallarse entre I y N si m^2 es positivo y ox tiene valor negativo[53]; igualmente debería ser necesario para N encontrarse entre I y L si m^2 tiene valor negativo y ox es positivo[54]. Pero nunca puede darse el valor $-m^2$
A-T, VI, 402 cuando los términos han sido establecidos en la forma que hemos indicado. Finalmente, el punto N debe coincidir con el punto I si m^2 es igual a *cero*. Por este medio es fácil encontrar esta parábola de acuerdo con el primer problema del primer libro de Apolonio[50bis].

Si la línea pedida es un círculo, una elipse o una hipérbola, es preciso en primer lugar hallar el punto M, que es el centro de la figura. Tal punto siempre se encuentra en la línea recta IL y puede ser hallado tomando IM igual a $\frac{aom}{2pz}$. De este modo, si o es igual a *cero*, entonces M coincidirá con I. Si la línea que deseamos hallar es un círculo o una elipse, debemos tomar el punto M en el mismo lado que el punto L respecto de I, cuando ox tiene valor positivo; pero en lado opuesto cuando su valor es negativo. Por el





LM, así como tomar MO igual a $\sqrt{m^2 - \frac{o^2m}{4p}}$; pero si ox es igual a *cero*, MO debe ser tomada igual a m. Seguidamente considerando O como el vértex de esta hipérbola, siendo OP el diámetro y CP la línea aplicada;
A-T, VI, 404 en tal caso, el lado recto es:

$$\sqrt{\frac{4a^4m^4}{p^2z^4} - \frac{a^4o^2m^3}{p^3z^4}}$$

y su diámetro será:

$$\sqrt{4m^2 - \frac{o^2m}{p}}.$$

No obstante, debemos considerar una excepción: cuando ox es igual a *cero*; en tal caso el lado recto es $\frac{2a^2m^2}{pz^2}$ y el valor del diámetro $2m$. A partir de estos datos la curva puede ser determinada de acuerdo con el problema tercero del primer libro de Apolonio.

Demostración de lo que ha sido explicado

Las demostraciones de las afirmaciones anteriores son evidentes, pues componiendo un espacio con las cantidades que he asignado para el lado recto, el transverso y para el segmento del diámetro, NL u OP, siguiendo los teoremas 11, 12 y 13 del primer libro de Apolonio, podremos hallar los mismos términos que componen el cuadrado de la línea CP o CL, aplicada

contrario, en el caso de una hipérbola, si se tuviese $-ox$, el centro M debería estar hacia L; si tuviésemos $+ox$, debería ser tomado hacia la otra parte. Según todo esto, tenemos que el lado recto de la figura debe ser

$$\sqrt{\frac{o^2z^2}{a^2}+\frac{4mpz^2}{a^2}},$$ A-T, VI, 403

si m^2 es positivo y la línea es un círculo o una elipse; o si m^2 es negativo y el lugar es una hipérbola. Deberá ser:

$$\sqrt{\frac{o^2z^2}{a^2}-\frac{4mpz^2}{a^2}};$$

si la línea buscada es un círculo o una elipse y m^2 es negativo, o si es una hipérbola y o^2 es mayor que $4mp$, siendo positivo el valor de m^2. Pero si el valor de m^2 es cero, entonces el lado recto es $\frac{oz}{a}$; y si oz es igual a *cero*, entonces será:

$$\sqrt{\frac{4mpz^2}{a^2}}.$$

Así mismo, en relación con el correspondiente diámetro debe ser hallada una línea que sea al lado recto como a^2m es a pz^2; esto es, si el lado recto es

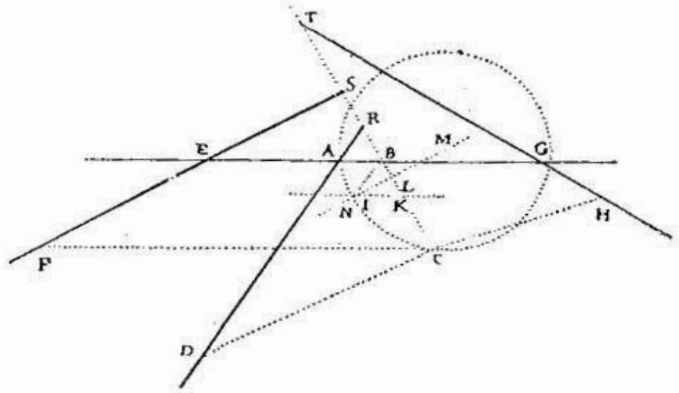
$$\sqrt{\frac{o^2z^2}{a^2}+\frac{4mpz^2}{a^2}},$$

entonces el diámetro será:

$$\sqrt{\frac{a^2o^2m^2}{p^2z^2}+\frac{4a^2m^3}{pz^2}}.$$

En cada uno de estos casos el diámetro de la sección se encuentra en la línea IM, siendo LC de las aplicadas por orden. De tal modo esto es evidente que siendo MN igual a la mitad del diámetro y tomando N y L en el mismo lado de M, el punto N será el vertex de este diámetro. Siguiendo lo expuesto es fácil hallar la sección (determinar la curva) de acuerdo con los problemas segundo y tercero del primer libro de Apolonio.

Pero, cuando esta sección es una hipérbola y tenemos que m^2 es positivo y que o^2 es igual a *cero* o menor que $4pm$, entonces debemos trazar la línea MOP por M de modo que sea paralela a LC y de modo que CP lo sea a



por orden a este diámetro. Así, en este ejemplo, tomando IM, que es $\frac{aom}{2pz}$, de NM cuyo valor es:

$$\frac{a}{2pz}\sqrt{o^2+4mp},$$

se obtiene IN; sumando a ésta IL, que es $\frac{a}{z}x$, tenemos NL que es igual a:

$$\frac{a}{z}x-\frac{aom}{2pz}+\frac{am}{2pz}\sqrt{o^2+4mp}.$$

Y siendo esto multiplicado por

$$\frac{z}{a}\sqrt{o^2+4mp},$$

que es el lado recto de la figura, obtenemos:

$$x\sqrt{o^2+4mp}-\frac{om}{2p}\sqrt{o^2+4mp}+\frac{mo^2}{2p}+2m^2$$

para el rectángulo, del cual debemos restar un espacio que debe ser al cuadrado de NL como el lado recto es al diámetro. Este cuadrado de NL es: A-T, VI, 405

$$\frac{a^2}{z^2}x^2-\frac{a^2om}{pz^2}x+\frac{a^2m}{pz^2}x\sqrt{o^2+4mp}+$$
$$+\frac{a^2o^2m^2}{2p^2z^2}+\frac{a^2m^3}{pz^2}-\frac{a^2om^2}{2p^2z^2}\sqrt{o^2+4mp}.$$

A su vez, debe ser dividido por a^2m y debe ser multiplicado el cociente por pz^2, puesto que éstos

expresan la razón o proporción existente entre el diámetro y el lado recto. El resultado será:

$$\frac{p}{m}x^2 - ox + x\sqrt{o^2+4mp} +$$

$$+\frac{o^2m}{2p} - \frac{om}{2p}\sqrt{o^2+4mp} + m^2.$$

Restando esta cantidad del rectángulo previamente obtenido, tenemos:

$$CL^2 = m^2 + ox - \frac{p}{m}x^2.$$

En consecuencia, CL es una línea aplicada por orden en una elipse o en un círculo al segmento del diámetro NL.

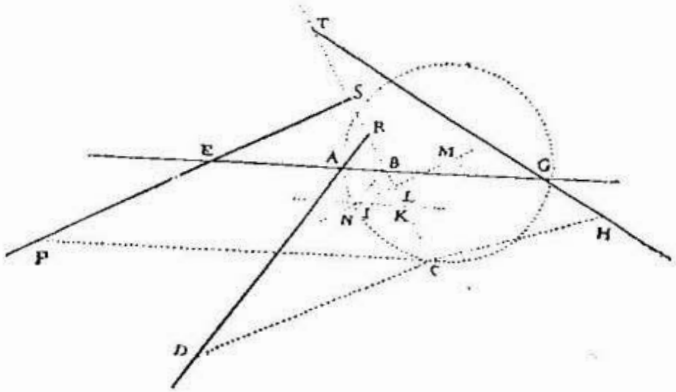
Si deseamos explicar todas las cantidades dadas mediante números, suponemos, por ejemplo, que $EA=3$, $AG=5$, $AB=BR$, $BS=\frac{1}{2}BE$, $GB=BT$, $CD=\frac{1}{2}CR$, $CF=2CS$, $CH=\frac{2}{3}CT$, el ángulo $ABR=60°$ y, finalmente, suponemos que $CB\cdot CF = CD\cdot CH$; todas estas cantidades deben ser conocidas si el problema ha de ser completamente determinado. Así mismo, suponiendo que $AB=x$ y $CB=y$, obtendremos mediante el método explicado que:

$$y^2 = 2y - xy + 5x - x^2,$$

y que

$$y = 1 - \tfrac{1}{2}x + \sqrt{1+4x-\tfrac{3}{4}x^2};$$

Por consiguiente, BK debe ser igual a 1, y KL lo es a la mitad de KI; y puesto que el ángulo IKL es igual al
A-T, VI, 406 ángulo ABR, que es igual a 60° y el ángulo KIL, que es



deseo de solventar los problemas relacionados con la composición de tales lugares.

Así mismo, se ve que lo que he considerado como el primer género de líneas curvas no puede incluir otras que el círculo, la parábola, la hipérbola y la elipse, que es cuanto había intentado probar.

Sobre la primera y más simple de todas las líneas curvas utilizable en el problema de los antiguos cuando está propuesto para cinco líneas

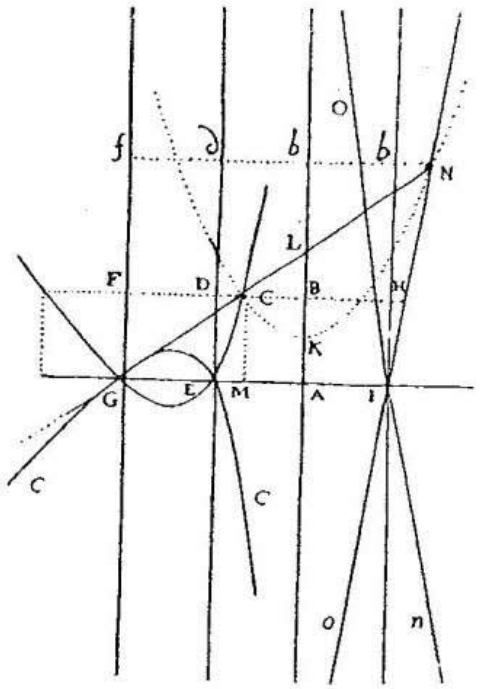
Si el problema de los antiguos fuese propuesto para cinco líneas, siendo todas paralelas, es evidente que el punto buscado se encontrará siempre en una línea recta.
A-T, VI, 408 Pero si es propuesto para cinco líneas de modo que cuatro de ellas sean paralelas y la quinta sea perpendicular a las otras cuatro; y si todas las líneas trazadas desde el punto buscado deben alcanzar las mencionadas líneas paralelas de modo que formen ángulos rectos; y si, finalmente, sucede que el paralelepípedo formado por tres líneas así trazadas sobre tres paralelas, debe ser igual al paralelepípedo formado por dos líneas que son trazadas, una sobre la cuarta de las que son paralelas y la otra sobre la que las corta formando ángulo recto, y por una tercera línea dada (siendo éste el caso más simple que podemos imaginar después de el precedente), entonces el punto que buscamos debe encontrarse en una línea curva descrita por el movimiento de una parábola del modo explicado anteriormente.

Sean, por ejemplo, las líneas AB, IH, ED, GF y GA, y que sea preciso hallar el punto C de modo que si CB, CF, CD, CH y CM son perpendicularmente trazadas sobre cada una de las líneas dadas, el paralelepípedo formado por las tres líneas CF, CD y CH sea igual al formado por CB, y CM y por una tercera línea, AI. Establezco que $CB=y$. $CM=x$, AI o AE o $GE=a$; por consiguiente, si C se encuentra entre AB y DE, tenemos que $CF=2a-y$, $CD=a-y$, $CH=y+a$. Multiplicando estas tres, una por otra, se obtiene $y^3-2ay^2-a^2y+2a^3$ igual al producto de las otras tres, esto es axy. Seguidamente considero la curva CEG, que imagino descrita
A-T, VI, 409 por la intersección de la parábola CKN (moviéndose de modo tal que su eje KL se encuentra sobre la línea recta AB) y de la regla GL (que gira sobre el punto G de modo que siempre pasa por el punto L). Tomamos $KL=a$, y el lado principal, es decir, aquel que corres-

la mitad del ángulo KIB o del IKL, es de 30°, el ángulo ILK es un ángulo recto. Así mismo, puesto que IK o $AB=x$, KL es un $\frac{1}{2}x$; IL es igual a $x\sqrt{\frac{3}{4}}$, y la cantidad representada por z es igual a 1, tenemos que $a=\sqrt{\frac{3}{4}}$, $m=1$, $o=4$, $p=\frac{3}{4}$; por consiguiente, $IM=\sqrt{\frac{16}{3}}$, $NM=\sqrt{\frac{19}{3}}$; en consecuencia puesto que a^2m (que es $\frac{3}{4}$) es igual a pz^2, y el ángulo ILC es un ángulo recto, se sigue de todo ello que la curva NC es un círculo. Un tratamiento similar de los otros casos puede realizarse con facilidad.

Cuáles son los lugares planos, sólidos y sobre el modo de hallarlos

Puesto que todas las ecuaciones que no son superiores a las de segundo grado caen bajo las explicaciones dadas, no solamente hemos solventado de modo definitivo el problema que los antiguos plantearon para tres y cuatro líneas, sino también cuanto se relaciona con lo que ha sido llamado composición de los lugares sólidos y, en consecuencia, con la de los lugares planos puesto que éstos están comprendidos en los sólidos; digo tal, pues estos lugares no son otra cosa que esto: la solución de uno de estos problemas no consiste más que A-T, VI, 407
en hallar algún punto para cuya completa determinación es necesaria una condición, siendo tales todas las otras condiciones (como en este ejemplo) que cualquier punto de una línea recta puede ser tomado por el que es pedido. Si esta línea es recta o circular decimos que se trata de un lugar plano; pero si es una parábola, una hipérbola o una elipse, entonces se dice que se trata de un lugar sólido. En cualquiera de los casos puede llegarse a una ecuación con dos cantidades desconocidas, que es semejante a alguna de las que acabo de resolver. Si la curva en la que se encuentra el punto buscado es de un grado superior al de las secciones cónicas, puede ser llamada, siguiendo el mismo procedimiento, un lugar supersólido, y así sucesivamente para otros casos. Si para la determinación del punto son necesarias dos condiciones, el lugar sobre el que se encuentra es una superficie que puede ser completamente plana, esférica o más compleja. Los antiguos llegaron a alcanzar como nivel superior de análisis la composición de los lugares sólidos y cuanto Apolonio expuso acerca de las secciones cónicas podría responder al



ponde al eje de esta parábola, igual a a; asimismo, $GA=2a$, CB o $MA=y$, CM o $AB=x$. Puesto que los triángulos GMC y CBL son semejantes, tenemos que $GM(2a-y)$ es a $MC(x)$ como $CB(y)$ es a BL, que es igual a

$$\frac{xy}{2a-y}.$$

Y puesto que LK es a, BK es

$$a-\frac{xy}{2a-y},$$

o también

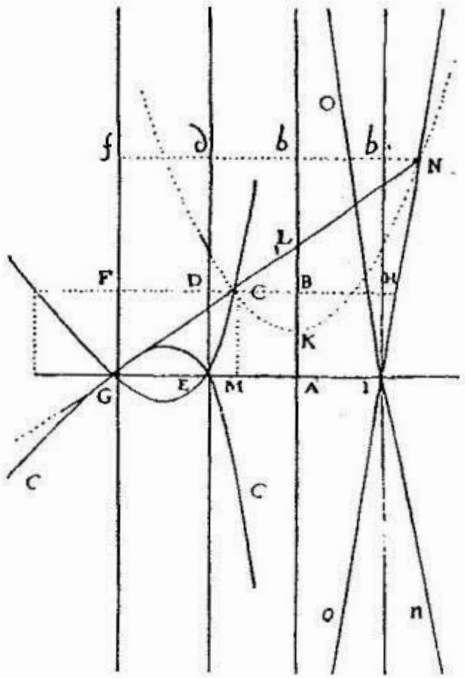
$$\frac{2a^2-ay-xy}{2a-y}.$$

Finalmente, puesto que BK es un segmento del eje de la parábola, BK es a BC (su ordenada) como BC es a a (lado recto) y, en consecuencia, en virtud del cálculo ya podemos establecer que:

$$y^3-2ay^2-a^2y+2a^3=axy,$$

y, por tanto, que el punto C es el que era exigido. Puede ser tomado en cualquier lugar de la línea CEG o
A-T. VI, 410 también en $cEGc$ que es descrita del mismo modo con excepción de que su vértex se encuentra en dirección opuesta; finalmente, C puede hallarse en NIo, nIO, descritas por la intersección que la línea GL produce en el otro lado de la parábola KN.

Aunque las líneas paralelas AB, IH, ED y GF no fuesen equidistantes y ni GA ni las líneas trazadas sobre ellas desde el punto C las cortasen formando ángulos



rectos, este punto C no dejaría de encontrarse en una línea curva que fuese de la misma naturaleza; es más, podría encontrarse aunque ninguna de las líneas dadas fuese paralela. Pero, supóngase que tenemos cuatro líneas paralelas y una quinta que las corta, y que el paralelepípedo formado por tres líneas trazadas por el punto buscado (una de ellas sobre la quinta línea y

Sobre las que pueden trazarse utilizando una cuerda y que son admisibles

De igual modo, no debe rechazarse el procedimiento o forma según la cual nos servimos de un hilo para determinar la igualdad o diferencia existente entre dos o más líneas rectas que pueden trazarse desde cada uno de los puntos de la curva buscada a otros, o bien sobre ciertas líneas con ciertos ángulos. Yo mismo he utilizado este procedimiento en *La Dióptrica* cuando explicaba la elipse o la hipérbola. Pues aunque no podamos aceptar en Geometría líneas que sean como cuerdas, es decir, que tanto puedan ser rectas como curvas, ya que las proporciones existentes entre las líneas rectas y curvas no son conocidas y no creo que puedan llegar a serlo y, en consecuencia, nada exacto y seguro podría concluirse; sin embargo, puesto que en tales construcciones nos servimos de cuerdas solamente para determinar las líneas rectas, cuya longitud nos es perfectamente conocida, no estimo que deba ser éste un motivo para rechazarlas[55].

Que para hallar las propiedades de las líneas curvas, basta con conocer la relación que guardan sus puntos con los de las rectas, y la forma de trazar otras líneas formando ángulos rectos

Cuando conocemos la relación que guardan todos los puntos de una línea curva con todos los de una recta, según el modo que he explicado, también es fácil conocer la relación que guardan con todos los otros puntos y líneas dadas; en consecuencia, será fácil conocer los diámetros, ejes, centros y otras líneas o puntos con los que pueda tener cada línea curva alguna
A-T. VI, 413 relación más específica o más simple; de esta forma, podremos imaginar diversos modos para describirlas y hacer una selección. Así mismo podremos hallar cuanto se encuentre relacionado con la magnitud de sus áreas, no siendo necesario pasar a ofrecer una explicación más detallada. Finalmente, y en relación con todas las otras propiedades atribuibles a las líneas curvas, no dependen sino de los ángulos que estas curvas formen con otras líneas. Pero, cuando puedan trazarse líneas rectas que las corten formando ángulos rectos, en los puntos en que hay intersección formada por aquéllas

otras dos sobre dos de las líneas paralelas) fuese igual al paralelepípedo formado por dos líneas trazadas sobre las otras dos paralelas y una tercera línea dada; el punto buscado se encuentra en tal caso en una curva de otra naturaleza, es decir, en una curva tal que todas las líneas rectas aplicadas por orden a su diámetro siendo iguales a las de una sección cónica, los segmentos de este diámetro que se encuentren entre el vertex y estas líneas guardan la misma proporción con una línea dada que la que ésta guarda con los segmentos del diámetro de la sección cónica a los cuales tales líneas están aplicadas por orden. No afirmo que esta línea sea menos simple que la precedente, la cual he pensado que debía considerarla en primer lugar puesto que tanto su descripción como su cálculo son en cierto modo más fáciles. A-T, VI, 411

No me detendré realizando un análisis y distinción por especies de las líneas relacionadas con otros casos, pues no he deseado acometer una exposición completa y, habiendo explicado el procedimiento para hallar una infinidad de puntos por donde pasan, estimo haber facilitado el medio que nos permitirá describirlas.

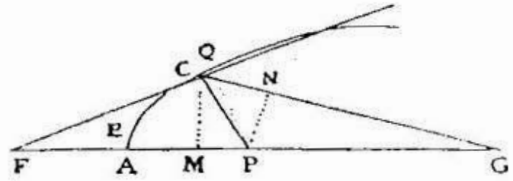
Sobre las líneas que se trazan hallando varios de sus puntos y que pueden ser admitidas en Geometría

De igual modo estimo oportuno señalar que existe una gran diferencia entre la forma expuesta para llegar a conocer varios puntos que permitan trazar una curva y la que se utiliza para trazar la espiral y sus semejantes. En este último caso no se calculan indiferentemente todos los puntos de la línea buscada, sino sólo aquellos que pueden ser determinados por algún procedimiento más simple que aquel que es requerido para formarla; propiamente hablando, no hallamos uno de sus puntos, esto es, alguno de aquellos que son puntos propios de esta línea de modo tal que no pudiesen ser hallados sino por su mediación. Por el contrario, no hay un punto en estas líneas que sirven para el problema propuesto, que no pueda ser determinado mediante el procedimiento indicado. Debe considerarse que si bien tal forma de determinar una línea curva, hallando indiferentemente diversos puntos de la misma, no es aplicable sino a las que puedan ser también descritas por un movimiento regular y continuo, no por ello se la debe excluir de la Geometría. A-T, VI, 412

con las que forman ángulos que se desean medir, o, lo que estimo como lo mismo, en que cortan sus contingentes[56], la dimensión de los ángulos no es más difícil de conocer que si se formasen entre o por dos líneas rectas. Por ello estimo haber expuesto cuanto se requiere en un estudio introductorio para realizar el análisis de las curvas, cuando haya desarrollado el procedimiento para trazar líneas rectas que formen ángulos rectos sobre cualesquiera de los puntos de aquellas que se elijan. Me atrevo a afirmar que éste es el problema cuyo conocimiento es más útil y no sólo el más general que yo conozco, sino también el que más he deseado llegar a conocer.

Procedimiento general para hallar líneas rectas que corten las curvas dadas o sus tangentes, formando ángulos rectos

Sea CE la línea curva y que sea preciso trazar por el punto C una línea recta que forme ángulos rectos. Supongo solucionado el problema y que tal línea es CP, la cual prolongo hasta el punto P en el cual alcanza la línea recta GA con la que supongo que se relacionan A-T, VI, 414



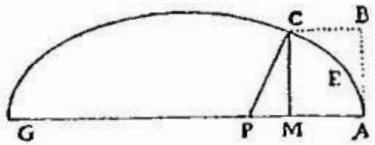
todos los puntos de la línea CE. De modo que siendo MA o $CB=y$, CM o $BA=x$, puedo establecer una ecuación que indica la relación entre x e y. Se establece que $PC=s$, $PA=v$, por lo que[57] $PM=v-y$; pero en virtud de que el triángulo PMC es un triángulo rectángulo, tenemos que s^2, el cuadrado de la hipotenusa, es igual a $x^2+v^2-2vy+y^2$, la suma de los cuadrados de los dos lados[58]. Es decir, que:

$$x=\sqrt{s^2-v^2+2vy-y^2},$$

o también que:

$$y=v+\sqrt{s^2-x^2}\ ^{59}.$$

Mediante esta ecuación puedo eliminar una de las dos cantidades indeterminadas, x o y, de la otra ecuación, de aquella que expresa la relación de los puntos de la línea curva CE con los de la recta GA. Si x debe ser eliminada puede realizarse fácilmente con sólo sustituir x por $\sqrt{s^2-v^2+2vy-y^2}$, y x^2 por el cuadrado de esta misma expresión, x^3 por el cubo de la misma expresión, etc...; pero si deseásemos eliminar y, entonces y debería sustituirse por $v+\sqrt{s^2-x^2}$, así como y^2, y^3 por el cuadrado o cubo de esta misma expresión, y así sucesivamente. El resultado que llegaremos a obtener sería el de una ecuación con una sola cantidad desconocida, bien x o bien y.



Por ejemplo, si CE es una elipse, MA el segmento de su eje, correspondiéndole CM, r su lado recto y q su eje transversal, entonces mediante el teorema 13 del libro I
A-T, VI, 415 de Apolonio tenemos que:

$$x^2 = ry - \frac{r}{q}y^2.$$

Eliminando x^2 obtendríamos que:

$$s^2 - v^2 + 2vy - y^2 = ry - \frac{r}{q}y^2,$$

o bien que:

$$y^2 + \frac{qry - 2qvy + qv^2 - qs^2}{q - r} = 0,$$

Pues en este caso es preferible considerar conjuntamente la expresión que establecer una igualdad entre sus partes.

De igual modo, si CE es una curva generada por el movimiento de una parábola, tal como anteriormente expuse[60], y si a GA la llamamos b, a KL c, y el lado recto del eje KL de la parábola le llamamos d, entonces la ecuación que expresaría la relación entre x e y sería:

$$y^3 - by^2 - cdy + bcd + dxy = 0.$$

CF sobrepasa a FA sea a la que GA sobrepasa a GC como d es a e; de modo que si la cantidad indeterminada la llamamos z, entonces tendremos que FC es igual a $c+z$, y GC lo será a $b-\frac{e}{d}z$. Seguidamente estableciendo $MA=y$, $GM=b-y$, $FM=c+y$; puesto que CMG es un triángulo rectángulo, restando el cuadrado de GM del cuadrado de GC, obtenemos el cuadrado de CM que es:

$$\frac{e^2}{d^2}z^2 - \frac{2be}{d}z + 2by - y^2.$$

Seguidamente, restando el cuadrado de FM del cuadrado de FC, obtenemos de nuevo el cuadrado de CM, pero expresado de modo distinto, es decir:

$$z^2 + 2cz - 2cy - y^2.$$

Siendo iguales estas dos expresiones, nos permitirán conocer el valor de y o MA, que es

$$\frac{d^2z^2 + 2cd^2z - e^2z^2 + 2bdez}{2bd^2 + 2cd^2};$$

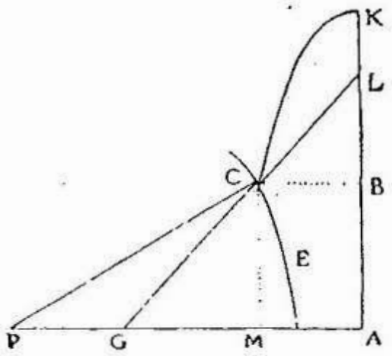
sustituyendo este valor por y en la expresión equivalente al cuadrado de CM, tenemos que:

A-T, VI, 417
$$CM^2 = \frac{bd^2z^2 + ce^2z^2 + 2bcd^2z - 2bcdez}{bd^2 + cd^2} - y^2.$$

A continuación, si suponemos que la línea PC alcanza la curva formando ángulos rectos en C, y estableciendo $PC=s$ y $PA=v$ como anteriormente, entonces $PM=v-y$; puesto que PCM es un triángulo rectángulo, tenemos $s^2-v^2+2vy-y^2$ para el cuadrado de CM. Sustituyendo y por su valor, tenemos la ecuación buscada:

$$z^2 + \frac{2bcd^2z - 2bcdez - 2cd^2vz - 2bdevz - bd^2s^2 + bd^2v^2 - cd^2s^2 + cd^2v^2}{bd^2 + ce^2 + e^2v - d^2v} = 0$$

Hallada tal ecuación no debe utilizarse para determinar x, y o z, que son dadas, puesto que el punto C ha sido dado, sino que debe utilizarse para hallar v o s, que determinan el punto exigido, P. A tal efecto, es preciso considerar que si este punto P cumple las condiciones exigidas, el círculo con centro en P y que pasa por C, tocará pero no cortará la curva CE; pero si el punto P se encuentra a mayor o menor distancia de A de lo que



Si eliminamos x, tenemos:

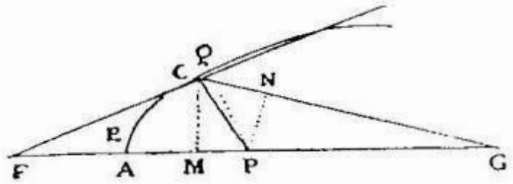
$$y^3 - by^2 - cdy + bcd + dy\sqrt{s^2 - v^2 + 2vy - y^2}.$$

Ordenando tales términos, tenemos:

$$y^6 - 2by^5 + (b^2 - 2cd + d^2)y^4 + (4bcd - 2d^2v)y^3 + \\ + (c^2d^2 - d^2s^2 + d^2v^2 - 2b^2cd)y^2 - 2bc^2d^2y + \\ + b^2c^2d^2 = 0.$$

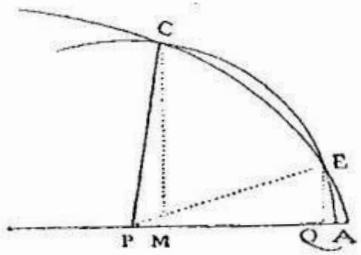
Tal procedimiento es válido para otros casos.

Pero aún cuando los puntos de la línea curva no se relacionen en la forma indicada con los de una recta, sino con cualquiera otra imaginable, podríamos establecer una ecuación similar. Así, si CE es una línea relacionada con los puntos F, G y A de modo tal que las A-T, VI, 416



líneas rectas trazadas desde cada uno de sus puntos, tal como C, hasta el punto F, excedan el segmento FA por una cantidad que guarda una cierta proporción dada con otra cantidad por la que GA sobrepasa a las líneas trazadas desde los mismos puntos a G. En tal caso, estableciendo $GA = b$, $AF = c$ y, tomando a discreción el punto C en la curva, supongo que la cantidad en que

debe estar, este círculo cortará la curva no sólo en C, sino también y necesariamente en algún otro punto[61]. Asimismo, es preciso considerar que si este círculo corta la línea curva CE, la ecuación por la que se halla el valor de x e y o cualquiera otra, suponiendo que conozcamos PA y PC, debe tener necesariamente dos raíces diferentes. Supóngase, por ejemplo, que el círculo A-T, VI, 418



corta la curva en los puntos C y E; trácese EQ paralela a CM; por otra parte vemos que x e y convienen por igual a EQ y a QA del mismo modo que a CM y MA, ya que PE es igual a PC, pues son dos radios. Si buscamos EQ y QA, suponiendo dadas PE y PA, tendremos la misma ecuación que si hubiésemos buscado CM y MA, suponiendo que PC y PA sean dados. Se deduce de ello que el valor de x, de y o de cualquier otra cantidad supuesta debe ser doble, es decir, la ecuación tendrá dos raíces distintas entre sí; si deseamos conocer el valor de x, una de estas raíces será CM y la otra EQ; si se desea conocer y, una raíz será MA y la otra QA. Es verdad que si E no se encuentra en el mismo lado de la curva que C, entonces solamente una de ellas será la raíz verdadera y la otra deberá ser opuesta o menor que cero; asimismo, cuanto más próximos se encuentren estos dos puntos, C y E, menor será la diferencia entre estas dos raíces; cuando ambos puntos coincidan serán iguales. Este sería el caso en que el círculo trazado por C toca la curva CE sin llegar a cortarla[62].

Además debe considerarse que cuando una ecuación tiene dos raíces iguales, necesariamente tiene la misma forma que si se multiplica, por sí misma, la cantidad que se supone desconocida menos la cantidad conocida igual a ella; si la expresión o suma resultante no es de grado semejante a la ecuación original, multiplicando ésta por otra expresión con tantas dimensiones como las que le faltan, se logrará que sea del mismo grado. De A-T, VI, 419

este modo cada uno de los términos de las dos expresiones se corresponderán término a término.

Así, por ejemplo, establezco que la primera ecuación hallada en este análisis, a saber:

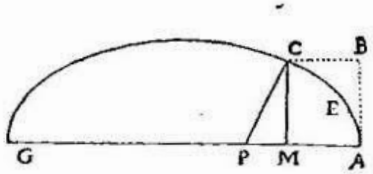
$$y^2 + \frac{qry - 2qvy + qv^2 - qs^2}{q-r},$$

debe tener la misma forma que la expresión obtenida estableciendo $e = y$, y multiplicando $y - e$ por sí misma. De este modo obtenemos $y^2 - 2ey + e^2$. Podemos comparar las dos expresiones término a término y afirmar que el primer término, y^2, es el mismo en cada una; el segundo término de la primera

$$\frac{qry - 2qvy}{q-r},$$

es igual a $-2ey$, el segundo de la otra; a partir de esto, buscando la cantidad v, que es PA, tenemos:

$$v = e - \frac{r}{q}e + \frac{1}{2}r;$$



o bien, puesto que hemos supuesto que $e = y$, tenemos:

$$v = y - \frac{r}{q}y + \frac{1}{2}r.$$

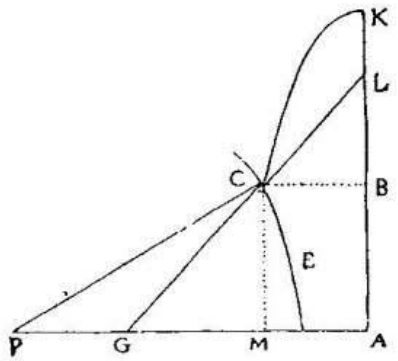
Del mismo modo, podemos hallar s a partir del tercer término,

$$e^2 = \frac{qv^2 - qs^2}{q-r};$$

pero puesto que v determina completamente P, el único punto que tratamos de hallar, no es necesario proseguir.

A-T, VI, 420 Del mismo modo, la segunda ecuación hallada, a saber,

$$y^6 - 2by^5 + (b^2 - 2cd + d^2)y^4 + (4bcd - 2d^2v)y^3 + \\ + (c^2d^2 - 2b^2cd + d^2v^2 - d^2s^2)y^2 - 2bc^2d^2y + b^2c^2d^2,$$



Así mismo, la tercera ecuación, es decir,

$$z^2 + \frac{2bcd^2z - 2bcdez - 2cd^2vz - 2bdevz - bd^2s^2 + bd^2v^2 - cd^2s^2 + cd^2v^2}{bd^2 + ce^2 + e^2v - d^2v}$$

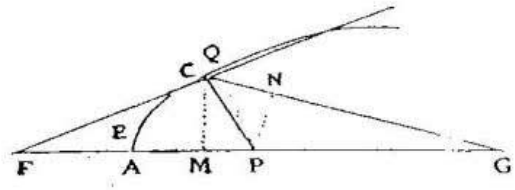
A-T, VI, 422 es de la misma forma que $z^2 + 2fz + f^2$, suponiendo que $f = z$, de modo que $-2f$ o $-2z$ deba ser igual a

$$\frac{2bcd^2 - 2bcde - 2cd^2v - 2bdev}{bd^2 + ce^2 + e^2v - d^2v};$$

a partir de esto podemos conocer que:

$$v = \frac{bcd^2 - bcde + bd^2z + ce^2z}{cd^2 + bde - e^2z + d^2z}.$$

Si establecemos la línea AP igual a v, cuyas cantidades son todas conocidas, y unimos el punto P con el punto C, esta línea cortará a la curva CE formando



ángulos rectos, siendo esto lo que era preciso. No veo que exista razón alguna para que esta solución no sea aplicada a cualquier curva a la que sea aplicable algún cálculo geométrico.

debe tener la misma forma que la expresión obtenida cuando se multiplica $y^2-2ey+e^2$ por $y^4+fy^3+g^2y^2+$ $+h^3y+k^4$, que es:

$$y^6+(f-2e)y^5+(g^2-2ef+e^2)y^4+(h^3-2eg^2+e^2f)y^3+$$
$$+(k^4-2eh^3+e^2g^2)y^2+(e^2h^3-2ek^4)y+e^2k^4$$

A partir de estas dos ecuaciones pueden obtenerse otras seis que permiten determinar f, g, h, k, v, s. A partir de esto es fácil entender que cualquiera que sea la clase a la que la línea pertenezca, este procedimiento permite establecer siempre tantas ecuaciones como cantidades desconocidas supuestas. Pero para desarrollar por orden estas ecuaciones y, finalmente, hallar v, que realmente es lo que es necesario calcular, ya que las otras pueden ser halladas mediante v, primeramente por el segundo término debe determinarse f, que es la primera de las cantidades desconocidas en la última expresión; así hallamos que $f=2e-2b$. Seguidamente debemos hallar k, la última de las cantidades desconocidas de la misma expresión; así tenemos

$$k^4=\frac{b^2c^2d^2}{e^2};$$

a partir del tercer término es preciso hallar g, la segunda cantidad, obteniendo $g^2=3e^2-4be-2cd+b^2+d^2$; a partir del penúltimo término es preciso hallar h, la penúltima cantidad, que es: A-T, VI, 421

$$h^3=\frac{2b^2c^2d^2}{e^3}-\frac{2bc^2d^2}{e^2}.$$

Mediante el mismo procedimiento podemos continuar siguiendo este mismo orden hasta que hayamos calculado la última cantidad.

Así, a partir del término correspondiente, que en este caso es el cuarto, podemos hallar v, y tenemos que:

$$v=\frac{2e^3}{d^2}-\frac{3be^2}{d^2}+\frac{b^2e}{d^2}-\frac{2ce}{d}+e+\frac{2bc}{d}+\frac{bc^2}{e^2}-\frac{b^2c^2}{e^3};$$

pero si introducimos y en lugar de e, puesto que $y=e$, tenemos que:

$$v=\frac{2y^3}{d^2}-\frac{3by^2}{d^2}+\frac{b^2y}{d^2}-\frac{2cy}{d}+y+\frac{2bc}{d}+\frac{bc^2}{y^2}-\frac{b^2c^2}{y^3}$$

para la línea AP.

Así mismo, debe observarse al considerar la expresión tomada a discreción para elevar el resultado inicial al grado necesario, como antes hemos tomado:

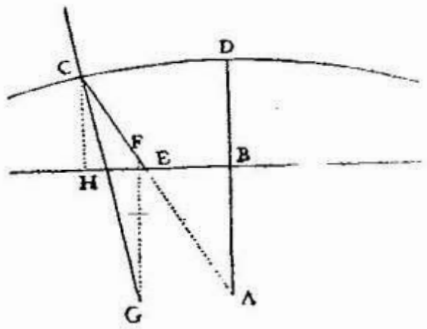
$$y^4+fy^3+g^2y^2+h^3y+k^4,$$

que los signos «+» y «−» pueden introducirse como se desee sin que la línea v o AP tenga diferentes valores, tal como fácilmente se puede comprobar, pues si fuera preciso que me detuviese realizando la demostración de todos los teoremas a los que aludo me vería obligado a redactar un volumen cuya amplitud sería mucho mayor de lo que deseo. Más bien deseo advertiros que la invención que consiste en suponer dos ecuaciones de la misma forma con el fin de comparar todos los términos de una con los de la otra, obteniéndose de este modo una sola, según se mostró en el ejemplo anterior, puede servir para una infinidad de problemas y no es de las menos importantes que forman parte del método que utilizo. A-T, VI, 423

No facilitaré las construcciones en virtud de las que pueden describirse las tangentes o las perpendiculares buscadas, siguiendo el cálculo que acabo de exponer, puesto que son fácilmente calculables, aunque frecuentemente sea necesaria una destreza significada para que se realice de modo simple y breve.

Ejemplo de la construcción de este problema en la concoide

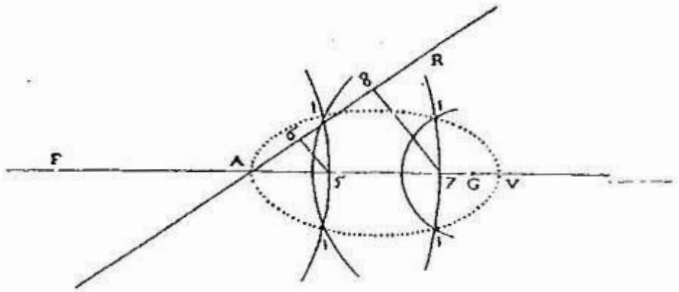
Así, por ejemplo, si DC es la primera concoide de los antiguos, A es su polo y BH la regla, de tal modo que los segmentos de todas las líneas rectas, como CE y



DB, que convergen en *A* y están comprendidas entre la curva *CD* y la línea recta *BH*, son iguales; pensemos que sea requerido hallar la línea *CG* que la corta en el punto *C*, formando ángulos rectos; se podría, calculando el punto por donde esta línea *CD* debe pasar en *BH*, según el método explicado, aventurarse en un cálculo tanto o
A-T, VI, 424 más extenso que ninguno de los precedentes. Sin embargo, la construcción que debe ser deducida es muy simple. Sólo es preciso tomar *CF* en la recta *CA*, siendo igual a *CH*, que cae perpendicular sobre *HB*; seguidamente, desde el punto *F* debe trazarse *FG* paralela a *BA* e igual a *EA*. Por este medio se halla el punto *G*, por el que debe pasar *CG*, la línea pedida.

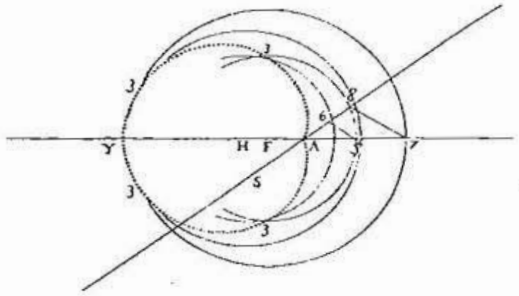
Explicación de cuatro nuevos géneros de óvalos de utilidad en la Optica

Finalmente, con el fin de que conozcáis que la consideración de las líneas curvas propuestas no carece de uso y que tienen diversas propiedades cuya importancia no es menor que las de las secciones cónicas, deseo aún realizar la explicación de algunos óvalos que, como se comprobará, son de gran utilidad tanto en la teoría de la Catóptrica como de la Dióptrica. Pueden describirse del siguiente modo.

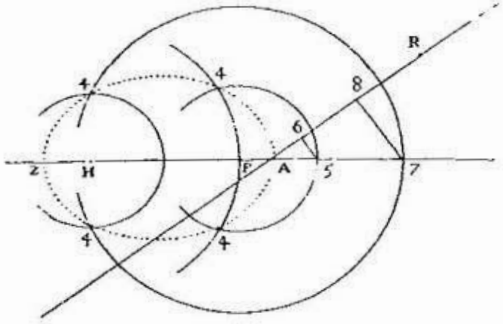


Primeramente, habiendo trazado las líneas rectas *FA* y *AR* que se cortan en el punto *A*, siendo indiferente los ángulos formados, tomo a discreción en una el punto *F*, más o menos alejado del punto *A*, según desee construir
A-T, VI, 425 óvalos de mayores o menores dimensiones. Tomando el punto *F* como centro describo un círculo que pasa un poco más allá del punto *A*, como por el punto que llamamos 5. En segundo lugar, desde este punto 5 trazo la línea recta 56 que corta a la anterior en el punto 6, de

Para la construcción del tercero y cuarto en vez de tomar la línea *AG* es preciso tomar *AH* del otro lado del punto *A*, a saber, del mismo en el que está el punto *F*. Debemos observar, además, en relación con *AH* que debe ser mayor que *AF*, cuya dimensión puede ser equivalente a cero, de modo que los puntos *F* y *A* coincidan en la construcción de estos óvalos. Seguidamente, siendo iguales las líneas *AR* y *AS* a *AH*, para describir el tercer oval, *A*3*Y*, con centro en *H* trazo un círculo cuyo radio sea igual a *S*6 que corte en el punto 3
A-T, VI, 427 al trazado con centro en *F* que pasa por el punto 5; el



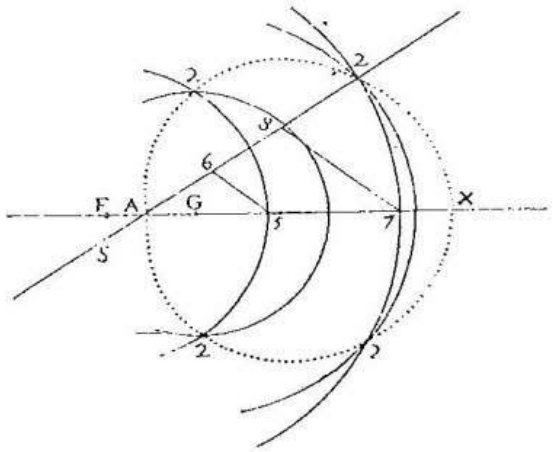
segundo, cuyo radio será igual a *S*8 que corta al que pasa por el punto 7 en el punto 3, y así sucesivamente para otros puntos. Finalmente, y en relación con el último óvalo, trazo dos círculos con centro en *H*, cuyos radios han de ser iguales a las líneas *R*6, *R*8 y semejantes que cortan los otros círculos en los puntos que llamamos 4.



Aún podríamos referirnos a una infinidad de procedimientos existentes para describir estos mismos óvalos como, por ejemplo, puede trazarse *AV* cuando se supone que las líneas *FA* y *AG* son iguales si se divide la

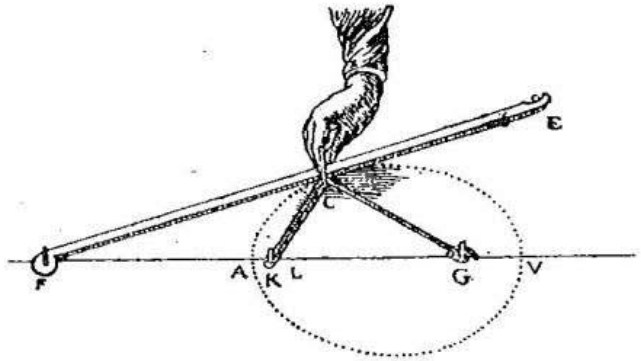
modo que *A*6 sea menor que *A*5 según una proporción dada, a saber, según la que mide las refracciones, si es que deseamos tenga alguna aplicación para la Dióptrica. En tercer lugar se toma a discreción el punto *G* en la línea *FA* y del mismo lado que se encuentra el punto 5; es decir, haciendo que entre las líneas *AF* y *GA* se guarde la proporción deseada. A continuación, en la línea *A*6 establezco *RA* igual a *GA* y desde el punto *G*, describiendo un círculo cuyo radio sea igual a *R*6 corto el otro círculo de una a otra parte, punto que será uno de aquéllos por donde debe pasar uno de los óvalos deseados. Por último, desde *F* describo un círculo que corta a *FA* un poco más allá del punto 5, por ejemplo, hacia el punto 7 y, trazando la línea 78 paralela a 56, desde el punto *G* trazo otro círculo cuyo radio sea igual a la línea *R*8. Este círculo cortará al que pasa por el punto 7 en el punto 1 que es otro de los del mismo óvalo. De este modo pueden hallarse cuantos se deseen, trazando otras líneas paralelas a 78 y otros círculos desde los centros *F* y *G*. A-T. VI. 426

La única diferencia existente en la construcción del segundo óvalo reside en que en vez de *AR* deberemos tomar *AS* en el otro lado de *A*, igual a *AG*, y, en segundo lugar, en que el radio del círculo descrito desde el punto *G*, cortando el círculo que se describe con centro en *F* y que pasa por el punto 5, sea igual a la línea *S*6, o bien que sea igual a la línea *S*8 si ha de cortar el que pasa por el punto 7 y así sucesivamente. De este modo, tales círculos se cortan entre sí en los puntos 2, 2, que son los del segundo óvalo, *A*2*X*.



FG en el punto *L*, de modo que *FL* sea a *LG* como *A*5 es a *A*6(*FL*:*LG*=*A*5:*A*6), es decir, según la razón que corresponde al índice de refracción. A continuación dividiendo *AL* en dos partes iguales, siendo *K* el punto equidistante, se procede a hacer girar una regla, presionando con el dedo *C* la cuerda *EC*, que estando unida al extremo de esta regla *E*, se repliega desde *C* hacia *K* y, seguidamente, desde *K* hacia *C* y desde *C* hacia *G*, punto en el que se une al otro extremo. De modo que la longitud de esta cuerda es igual a *GA*+*AL*+*FE*−*AF*. Será, pues, el movimiento del punto *C* el que describirá este óvalo de modo semejante al expuesto en la Dióptrica para la elipse y la hipérbola. Pero no deseo detenerme más en este tema. A-T. VI. 428

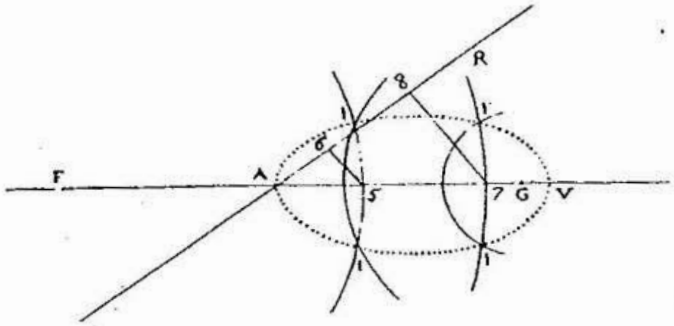
Aunque todos estos óvalos parezcan ser de la misma naturaleza, sin embargo, pertenecen a cuatro clases diferentes conteniendo cada una de ellas una infinidad de subclases, cada una de las cuales a su vez comprende tantas especies diferentes como la clase de las elipses o



las hipérbolas; las subclases dependen del valor de la razón de *A*5 respecto de *A*6, ya que según sea semejante o diferente, será diferente el género subalterno de estos óvalos. Así mismo, según que la proporción que exista entre las líneas *AF* y *AG* o *AH*, haya variado, los óvalos de cada subclase cambian de especie; según que la longitud de *AG* o *AH* sea mayor o menor, también variarán sus diversos tamaños. Si las líneas *A*5 y *A*6 son iguales, en vez de los óvalos del primer género o del tercero no se describirán sino líneas rectas; pero entre los de la segunda clase tenemos todas las posibles hipérbolas; y entre los de la cuarta, todas las posibles elipses. A-T. VI. 429

Propiedades de estos óvalos en relación con las reflexiones y refracciones

Además, en cada uno de estos óvalos es preciso considerar dos partes que tienen diferentes propiedades. En el óvalo primero la parte que se encuentra hacia A da lugar a que los rayos procedentes del punto F a través del aire converjan hacia G después de haber alcanzado la superficie convexa $1A1$ de una lente cuyo índice de refracción, de acuerdo con lo expuesto en la Dióptrica, sea tal que todas las refracciones puedan ser medidas considerando la proporción que existe entre las líneas $A5$ y $A6$ o semejantes, mediante las que ha sido descrito este óvalo.



A-T, VI, 430 Pero la parte que se encuentra hacia V da lugar a que todos los rayos procedentes de G converjan hacia F en caso de alcanzar una superficie cóncava de un espejo de la forma $1V1$, cuyo material fuese tal que disminuyese la fuerza de los rayos según la proporción que existe entre las líneas $A5$ respecto de $A6$. Pues, a partir de lo que ha sido demostrado en la Dióptrica es evidente que en tales circuntancias tanto los ángulos de reflexión como de refracción no serán iguales, pudiendo ser medidos de igual forma.

Considérese el segundo oval. También en este caso la parte $2A2$ sirve para las reflexiones de las que se supone que los ángulos son desiguales. Afirmo tal, pues, siendo la superficie del espejo del mismo material que lo era la del óvalo precedente, los rayos procedentes del punto G se reflejarían de tal modo que parecería que proceden del punto F. Obsérvese asimismo que habiendo trazado la línea AG de dimensiones mucho mayores que la línea AF, este espejo sería convexo en el centro, hacia A, pero cóncavo en sus extremidades, pues tal es la figura de

tómese el punto C a discreción en la parte primera del primer óvalo y trácese la línea recta CP de modo que
A-T, VI, 432 corte a la curva en el punto C formando ángulos rectos. Esto puede construirse siguiendo el procedimiento expuesto anteriormente del modo siguiente:

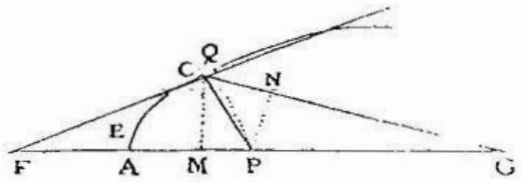
Sea $AG=b$, $AF=c$, $FC=c+z$ y supongo que la proporción existente entre d y e, que tomo como aquella que mide el poder refractor de la lente en consideración, designa la que existe también entre las líneas $A5$ y $A6$ o semejantes que han servido para trazar este óvalo. En tal caso tenemos que:

$$GC=b-\frac{e}{d}z.$$

De donde hallamos que la línea:

$$AP=\frac{bcd^2-bcde+bd^2z+ce^2z}{bde+cd^2+d^2z-e^2z}.$$

Seguidamente y desde P se traza la línea PQ perpendicular a FC, y PN perpendicular a GC. Consideramos que si PQ es a PN como d es a e, es decir, como son entre sí las líneas que miden las refracciones del vidrio convexo AC, entonces el rayo que procediendo del punto F llega a C debe curvarse de tal modo en este punto al penetrar en el vidrio que llegue a alcanzar el punto G, tal como es evidente si consideramos lo expuesto en *La Dióptrica*. Pero expongamos ahora mediante el cálculo si es verdadero que PQ sea a PN como d es a e ($PQ:PN=d:e$). Los triángulos rectángu-



los PQF y CMF son semejantes; de ello deducimos que CF es a CM como FP es a PQ. En consecuencia, si multiplicamos FP por CM y dividimos por CF, esto es igual a PQ. Así mismo, puesto que los triángulos rectángulos PNG y CMG son semejantes, se sigue que GP multiplicado por CM y dividido por CG es igual a PN. Pero puesto que las multiplicaciones y divisiones

esta línea que más bien representa la forma del corazón que de un óvalo.

Pero la otra parte, 2*X*2, es útil para las refracciones, sirviendo para que los rayos que atraviesan el aire y tienden a *F*, se desvíen hacia *G* al atravesar la superficie de un vidrio que tuviera tal figura.

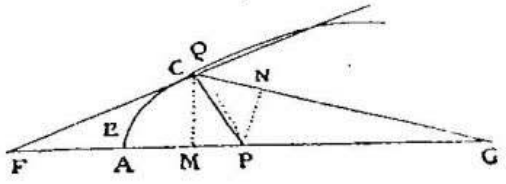
El tercer óvalo solamente es usado para la refracción y causa el que los rayos desplazándose a través del aire hacia *F* se dirijan a través del cristal hacia *H* después de haber atravesado la superficie cuya forma sería la de *A*3*Y*3, que en todas sus partes es convexa con excepción de *A* donde es un poco cóncava, teniendo al igual que A-T, VI, 431
en el caso precedente la forma de un corazón. La diferencia existente entre las dos partes de este óvalo reside en que una de ellas está más próxima a *F* y más alejada de *H*, mientras que en la otra está más próxima a *H* y más alejada de *F*.

De igual forma el último de los óvalos es útil solamente en el caso de las reflexiones. Su efecto consiste en dar lugar a que todos los rayos procedentes de *H* y que encuentran la superficie cóncava de un espejo cuyo material sea el mismo que en los casos precedentes, y que tenga la forma *A*4*Z*4, converjan hacia *F* después de producirse la reflexión.

Por ello, los puntos *F* y *G* pueden ser llamados focos de estos óvalos siguiendo en esto lo expuesto sobre las elipses y las hipérbolas que han sido llamados de igual modo en la Dióptrica.

Demostración de las propiedades en relación con las reflexiones y refracciones

No menciono otras diversas clases de reflexión y refracción que pueden ser producidas por estos óvalos puesto que siendo nuevos u opuestos los efectos, pueden ser fácilmente conocidos. Sin embargo, no debo silenciar la demostración de lo que he expuesto. Con tal fin,



de dos cantidades por una misma no modifica la A-T, VI, 433
proporción existente entre ellas, si

$$\frac{FP\cdot CM}{CF}:\frac{GP\cdot CM}{CG}=d:e,$$

entonces, dividiendo cada término de la primera por *CM* y multiplicando cada uno por *CF* y *CG*, tenemos que $FP\cdot CG:GP\cdot CF=d:e$. En virtud de la construcción:

$$FP=c+\frac{bcd^2-bcde+bd^2z+ce^2z}{cd^2+bde-e^2z+d^2z},$$

o que

$$FP=\frac{bcd^2+c^2d^2+bd^2z+cd^2z}{cd^2+bde-e^2z+d^2x}$$

y que $CG=b-\frac{e}{d}z$.

Entonces tenemos que:

$$FP\cdot CG=\frac{b^2cd^2+bc^2d^2+b^2d^2z+bcd^2z-bcdez-c^2dez-bdez^2-cdez^2}{cd^2+bde-e^2z+d^2z}$$

Tenemos también que:

$$GP=b-\frac{-bcd^2bcde-bd^2z-ce^2z^{62\,\text{bis}}}{cd^2+bde-e^2z+d^2z};$$

o bien

$$GP=\frac{b^2de+bcde-be^2z-ce^2z}{cd^2+bde-e^2z+d^2z};$$

y $CF=c+z$. De modo que

$$GP\cdot CF=\frac{b^2cde+bc^2de+b^2dez+bcdez-bce^2z-c^2e^2z-be^2z^2-ce^2z^2}{cd^2+bde-e^2z+d^2z}$$

Puesto que el primero de estos productos ($FP\cdot CG$) dividido por *d* es igual al segundo $GP\cdot CF$ dividido por *e*, se sigue que:

$$PQ:PN=FP:CG:GP\cdot CF=d:e,$$

que era lo que se debía demostrar.

Tal demostración es extensiva a cuanto se ha dicho de las otras refracciones o reflexiones que se daban en
A-T, VI. 434 los óvalos propuestos sin otra modificación que las convenientes alteraciones de los signos «+» y «−». Por ello cada uno puede examinarlos por sí mismo sin que sea necesario que me detenga en ello.

Pero sí que es necesario en este lugar que considere una cuestión silenciada en *La Dióptrica* cuando después de haber notado que pueden darse vidrios de diversas figuras que, tanto unos como otros, diesen lugar a que los rayos procedentes de un mismo punto del objeto se reúnan en otros puntos después de haberlos atravesado; que, entre éstos, los que son convexos por uno de los lados y cóncavos por el otro tienen un mayor poder para quemar que aquéllos cuya convexidad es la misma por ambas caras o lados; por el contrario, estos últimos son los más adecuados para la construcción de las lentes. Por ello he creído suficiente explicar aquellos que he estimado eran los más adecuados en razón de la práctica, puesto que supongo la dificultad que los artesanos deberán encontrar en la talla de los mismos. Considerando todo esto y con el fin de que no omitamos punto alguno que pueda estimarse necesario en relación con la teoría de esta ciencia, debo aún explicar la figura de vidrios que teniendo una de sus superficies tan convexa o cóncava como se desee, no por ello dejan de dar lugar a que todos los rayos que se dirigen hacia ellos procedentes de un mismo punto o bien paralelos se reúnan posteriormente en un mismo punto; así mismo, expondré la explicación relacionada con aquellos que dan lugar a un fenómeno semejante siendo convexos por ambos lados o bien guardando proporción la convexidad de uno de sus lados con la del otro.

Cómo se puede construir una lente, tan cóncava o convexa en una de sus caras como se deseee y que sea capaz de reunir en un punto dado los rayos procedentes de un punto dado

En primer lugar sean G, Y, C y F puntos dados tales que los rayos procedentes de G o paralelos a GA
A-T, VI. 435 converjan en F después de haber atravesado una lente cóncava que sea tal que Y sea el centro de su superficie interior y su extremidad se encuentre en el punto C de modo que tanto la cuerda CMC como la altura YM del arco CYC sean dadas. Previamente debemos determinar

la diferencia existente entre las bases, CH, y el lado pedido, MH, que será fácilmente calculable a partir de estos datos del modo siguiente: sea $k=CH-MH$ y $n=CM$; en consecuencia,

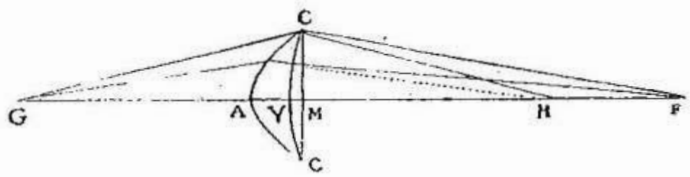
$$\frac{n^2}{2k}-\frac{1}{2}k=MH,$$

determinándose de este modo el punto H. Así calculado, si HY es mayor que HF, la curva CY debe ser la primera parte de un óvalo de la tercera clase, anteriormente llamada $3A3$. Pero supongamos que HY es menor que FY, entonces pueden darse dos posibilidades; en la primera, HY excede a HF de modo tal que la razón de la misma respecto de la línea FY es mayor que la razón de e, la menor de las dos líneas que miden las refracciones, respecto de la línea d, la mayor de ellas; esto es, si $HF=c$, $HY=c+h$, entonces dh es mayor que $2ce+eh$.
A-T, VI. 437 En este caso CY debe ser la segunda parte de $3Y3$ del mismo óvalo perteneciente a la tercera clase y al que anteriormente nos hemos referido. En la segunda posibilidad dh es igual o menor que $2ce+eh$; CY será la segunda parte, $2X2$, del óvalo de la segunda clase. Finalmente, si los puntos H y F coinciden, lo cual se da cuando FY y FC son iguales, entonces la línea YC es un círculo.

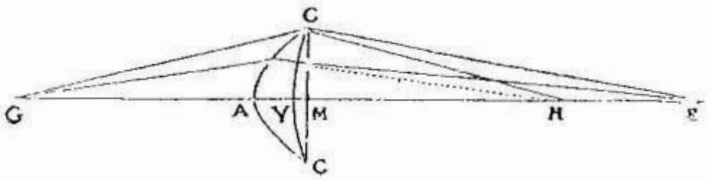
Finalmente, es preciso determinar CAC, la otra superficie. Si suponemos que los rayos que inciden son paralelos, será una elipse que tiene H como uno de sus focos, siendo fácilmente determinable su forma. Pero si suponemos que estos rayos proceden del punto G, la lente debe tener la forma correspondiente a la primera parte de un óvalo de la primera clase, debiendo ser los focos los puntos G y H, y pasando por el punto C; a partir de aquí se concluye que el punto A es su vertex, puesto que GC debe ser mayor que GA en una cantidad tal que sea a la que HA sobrepasa a HC como d lo es a e. Es tal, pues, si k representa la diferencia entre CH y HM, y x representa AM, entonces $x-k$ representará la diferencia entre AH y CH; si g representa la diferencia entre GC y GM, que son dadas, $g+x$ representará la diferencia entre CG y GA; puesto que $g+x:x-k=d:e$, tenemos que $ge+ex=dx-dk$, o que

$$AM=x=\frac{ge+dk}{d-e},$$

siendo posible determinar el punto A.



cuál de los dos óvalos explicados debe tener la figura de la superficie YC para que todos los rayos que al atravesarlo tienden hacia un punto como H, que aún no es conocido, sin embargo, finalicen reuniéndose en otro, hacia F, después de haberlo atravesado. Debe ser así, pues no se produce cambio alguno de dirección de los rayos por medio de la reflexión o refracción que no sea efectuado por uno de estos óvalos; fácilmente, pues, se ve que éste puede ser producido por la parte del tercer óvalo, que conocemos como $3A3$ o por la del mismo óvalo conocido como $3Y3$ o, finalmente, por la parte del segundo óvalo llamado $2X2$. Puesto que todos ellos son sometibles a un mismo cálculo, debemos tanto para uno como para otro tomar Y como su vertex, C como un punto de su curva y F como uno de sus focos; por tanto, sólo deberemos hallar H que debe ser otro de los focos. Puede ser hallado al considerar que la diferencia que existe entre las líneas FY y FC debe ser a la que existe entre las líneas HY y HC como d es a e, es decir, como la mayor de las líneas que mide las refracciones del vidrio propuesto es a la menor, tal como claramente puede verse a partir de la descripción de estos óvalos. Pero puesto que las líneas FY y FC son dadas, también lo es su diferencia y, en consecuencia, la que existe entre HY y HC, puesto que es conocida la proporción que A-T, VI, 436
existe entre estas dos diferencias. Así mismo, puesto que YM es dada, también lo es la diferencia que existe entre

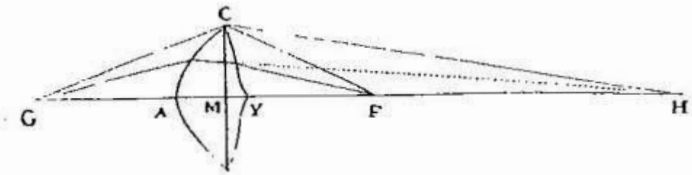


MH y HC; finalmente, puesto que CM es dada, solamente deberemos calcular MH, el lado del triángulo CMH, cuyo otro lado CM es conocido al igual que lo es

Cómo se puede elaborar una lente que produzca el mismo efecto que la precedente guardando la convexidad de una de sus caras una proporción dada con la de la otra

En relación con el otro caso, supongamos que solamente son dados los puntos G, C y F, junto con la razón de AM respecto de YM, siendo preciso hallar la figura de la lente ACY, que dará lugar a que todos los rayos procedentes del punto G se reúnan en el punto F. A-T, VI, 438

En este caso podemos utilizar dos óvalos, AC y CY, cuyos focos sean G y H, o F y H, respectivamente. Para determinarlos supongamos en primer lugar que H, foco común a ambos sea conocido; determino AM por los tres puntos G, C y H del modo anteriormente explicado:



Si k representa la diferencia entre CH y HM, y g la diferencia entre GC y GM, siendo a su vez AC la primera parte del óvalo de la primera clase, entonces tendremos que:

$$AM = \frac{ge + dk}{d - e}.$$

Seguidamente debe hallarse MY por medio de los tres puntos F, C y H. Si CY es la primera parte de un óvalo de tercera clase y tomamos y por MY, así como f por la diferencia entre CF y FM, tendremos que la diferencia entre CF y FY es igual a $f + y$; así mismo, si la diferencia entre CH y HM es igual a k, tenemos que la diferencia entre CH y HY es igual a $k + y$. De donde se deduce que $k + y : f + y = e : d$, y, por tanto, que el óvalo es de la tercera clase y que

$$y = MY = \frac{fe - dk}{d - e}.$$

Por esto,

$$AM + MY = AY = \frac{ge + fe}{d - e}.$$

de donde se sigue que en cualquier lado que sea situado

A-T, VI, 439 el punto H, la razón de la línea AY al exceso de $GC + CF$ sobre GF es siempre igual a la razón de e respecto de $d-e$, la diferencia que existe entre estas dos líneas, lo cual es un interesante teorema. Hallada la línea AY será preciso dividirla según la proporción que sus partes deben guardar, AM y MY; puesto que el punto M es conocido, los puntos A e Y así como, a continuación, el punto H serán hallados en virtud del problema precedente. Pero previamente es preciso considerar si la línea AM, hallada de este modo, es mayor que $\frac{ge}{d-e}$, o bien si es igual o menor. Pues si es mayor, se deduce de ello que la curva AC debe ser la primera parte de un óvalo de la primera clase, y CY la primera de un óvalo de la tercera, tal como se han considerado; pero si es menor, CY debe ser la primera parte de uno de la primera clase y AC la primera de uno de la tercera. Finalmente, si AM es igual a $\frac{ge}{d-e}$, las curvas AC y CY deben ser hipérbolas.

Estos problemas podrían hacerse extensivos a una infinidad de casos que no deseo explicar, puesto que carecen de uso en Dióptrica.

Aún se podría avanzar y mostrar cómo en el caso de que una superficie de una lente fuera dada y no fuera totalmente plana o compuesta de secciones cónicas o
A-T, VI, 440 círculos, cómo debería trazarse la otra superficie con el fin de que todos los rayos procedentes de un punto dado incidiesen en otro que también fuera dado. Lo más difícil acabo de explicarlo; abierto el camino todo resultará más fácil. Prefiero, sin embargo, que tal tarea sea cumplida por otros con el fin de que si aún encuentran dificultades, esto les suscite estima por cuanto se ha demostrado.

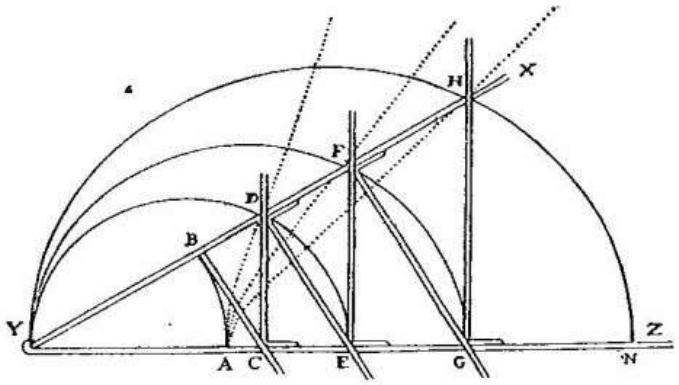
Sobre la aplicación de lo expuesto sobre las líneas curvas trazadas sobre una superficie plana, a las que se describen en un espacio con tres dimensiones

En toda esta exposición no me he referido sino a las líneas curvas que pueden ser descritas sobre una superficie plana, pero fácilmente podría ser relacionada con cuantas pudiéramos imaginar y que han sido formadas por el movimiento regular de los puntos de algún cuerpo en un espacio con tres dimensiones. Podría realizarse trazando dos perpendiculares desde cada uno

por las más simples no sólo deben entenderse aquellas que más fácilmente pueden ser descritas, ni aquellas que permiten una construcción o demostración con más facilidad del problema, sino que debemos entender aquellas que son de la clase más simple que puede utilizarse para determinar la cantidad exigida.

Ejemplo relacionado con la determinación de diversas medias proporcionales

Así, por ejemplo, no creo que exista modo más fácil para hallar cuantas medias proporcionales se deseen y
A-T, VI, 443 cuya demostración sea más evidente, que el empleo de las líneas curvas que se describen mediante la utilización del instrumento XYZ, anteriormente descrito. Pues si se desea hallar dos medias proporcionales entre YA e YE no es preciso sino describir un círculo cuyo diámetro sea YE; puesto que este círculo corta la curva AD en el punto D, YD será una de las medias proporcionales



buscadas. La demostración aparece como evidente tan pronto como este instrumento es aplicado sobre la línea YD, pues como $YA = YB$ es a YC, así YC es a YD, e YD es a YE.

De igual modo para hallar cuatro medias proporcionales entre YA e YG, o para hallar seis entre YA e YN, no será preciso sino trazar el círculo YFG que, cortando en el punto F a AF, determina la línea recta YF que es una de las cuatro; igualmente si se traza YHN de modo que el punto H corta a AH, se determina YH, una de las seis y así para otros casos.

de los puntos de la línea curva que se desea considerar sobre dos planos que se corten formando ángulos rectos. Afirmo tal, pues, los extremos de estas perpendiculares describen otras dos líneas curvas, una sobre cada uno de estos planos a partir de los cuales y según la forma explicada podríamos determinar todos los puntos y relacionarlos con los de la línea recta que es común a estos dos planos; por este medio los de la curva que tienen tres dimensiones son enteramente determinados. Igualmente, si se desea trazar una línea recta que corte esta curva en el punto dado formando ángulos rectos, solamente es preciso trazar otras dos líneas rectas en los dos planos, una sobre cada uno, de modo que formen ángulos rectos las dos líneas curvas en los dos puntos en que caen las dos perpendiculares trazadas desde el punto dado. Pues, habiendo trazado otros dos planos, uno sobre cada una de las rectas dadas, cortando en ángulo recto el plano en que está, tendremos la intersección de estos dos planos como la recta buscada. De este modo espero no haber omitido A-T, VI, 441
elemento alguno decisivo para alcanzar el conocimiento de las líneas curvas.

LIBRO TERCERO

SOBRE LA CONSTRUCCIÓN DE PROBLEMAS SÓLIDOS Y SUPERSÓLIDOS[63] A-T, VI, 442

Sobre las líneas curvas de las que podemos servirnos en la construcción de cada problema

Aunque todas las líneas curvas que pueden ser descritas en virtud de algún movimiento regular deben considerarse como objeto de estudio en la Geometría, esto no quiere decir que deba utilizarse indiferentemente cualquiera de ellas para la construcción de cada uno de los problemas; por el contrario debe seleccionarse siempre la más simple, siempre y cuando haga posible la solución del problema. De igual modo debe notarse que

Pero puesto que la curva AD es de la segunda clase, es posible hallar dos medias proporcionales mediante A-T, VI, 444
las secciones cónicas, que pertenecen a la primera clase; puesto que también pueden hallarse cuatro o seis medias proporcionales mediante líneas curvas de clase más simple de lo que son AF y AH, sería un error en Geometría utilizarlas en este caso. Por otra parte sería igualmente un error el trabajar inútilmente para construir un problema por medio de una clase de líneas más simples de lo que permite su naturaleza.

Sobre la naturaleza de las ecuaciones

Así pues, con el fin de facilitar algunas reglas que permitan evitar tanto el primero como el segundo de estos errores, es preciso realizar algunas observaciones generales sobre la naturaleza de las ecuaciones. Una ecuación está integrada por varios términos, alguno de ellos conocido y alguno de ellos desconocido, siendo unos iguales a los otros o, más bien, considerados todos conjuntamente son iguales a cero; digo tal, pues lo más conveniente será considerarlos de este modo.

Sobre el número de raíces en cada ecuación

Cada ecuación, pues, puede tener tantas raíces (valores de la cantidad desconocida), cuantas dimensiones tiene la cantidad desconocida: así, por ejemplo, supongamos que x es igual a $2(x=2)$, o bien que $x-2$ es igual a *cero* $(x-2=0)$ y, a su vez, que x es igual a $3(x=3)$ o que $x-3$ es igual a *cero* $(x-3=0)$; multiplicando estas dos ecuaciones, $x-2=0$ y $x-3=0$, tenemos $x^2-5x+6=0$, o bien que $x^2=5x-6$. Es ésta una ecuación en la que x tiene el valor 2 y, a la vez, el valor 3. Si $x-4=0$ y multiplicamos por $x^2-5x+6=0$, tenemos otra ecua- A-T, VI, 445
ción: $x^3-9x^2+26x-24=0$; en esta ecuación x, teniendo tres dimensiones, tiene también tres valores, que son 2, 3 y 4.

Sobre las raíces falsas

Pero frecuentemente se da el caso de que alguna de estas raíces es falsa[64] o menor que cero. Así, si suponemos que x representa la carencia de una cantidad $(x=$

-5) que fuese 5, tenemos que $x+5=0$; cuando se multiplica por $x^3-9x^2+26x-24=0$, resulta:

$$x^4-4x^3-19x^2+106x-120=0.$$

Estamos ante una ecuación que tiene cuatro raíces, a saber, tres que son verdaderas 2, 3 y 4, y una que es falsa, 5.

Sobre la reducción del número de dimensiones de una ecuación cuando se conoce alguna de sus raíces

Es evidente a partir de lo expuesto que la ecuación que tiene varias raíces es siempre divisible por un binomio formado por la cantidad desconocida menos el valor de una de las raíces verdaderas, o más el valor de una de las raíces falsas. De este modo pueden ser reducidas las dimensiones (el grado) de una ecuación.

Cómo puede examinarse si una cantidad dada es el valor de una raíz

De modo recíproco si la ecuación no puede ser dividida por un binomio compuesto de la cantidad desconocida más («+») o menos («−») alguna otra cantidad, esto muestra que esta última cantidad no es valor de alguna de sus raíces. Así, esta última ecuación,
A-T. VI. 446 $x^4-4x^3-19x^2+106x-120=0$; puede ser dividida por $x-2$, $x-3$, $x-4$, $x+5$, pero no por x más («+») o menos («−») alguna otra cantidad, lo cual muestra que solamente tiene cuatro raíces, 2, 3, 4, 5.

Cuántas raíces verdaderas pueden darse en cada ecuación

Igualmente podemos determinar el número de raíces verdaderas y falsas que pueda tener alguna ecuación del modo siguiente: pueden darse tantas verdaderas como cambios de signos de «+» (más) a «−» (menos) y de «−» (menos) a «+» (más); igualmente, pueden darse tantas falsas como veces se encuentren dos signos «+», o dos signos «−» que se encuentren en sucesión. Así, en la última ecuación, puesto que $+x^4$ es seguido de $-4x^3$,

$$\begin{array}{r}
y^4-12y^3+54y^2-108y+\ 81\\
+\ 4y^3-36y^2+108y-108\\
-19y^2+114y-171\\
-106y+318\\
-120\\
\hline
y^4-\ 8y^3-\ \ y^2+\ \ 8y\ \ \ \ \ \ =0
\end{array}$$

o

$$y^3-8y^2-y+8=0,$$

cuya verdadera raíz no es ahora 5 sino 8, puesto que ha sufrido un incremento de 3.

A-T. VI. 448 Por el contrario, si lo que deseamos es disminuir en 3 la raíz de esta misma ecuación, será preciso que $y+3=x$, y que $y^2+6y+9=x^2$, y así para otros casos. De modo que en lugar de:

$$x^4+4x^3-19x^2-106x-120=0,$$

tendremos:

$$\begin{array}{r}
y^4+12y^3+54y^2+108y+\ 81\\
+\ 4y^3+36y^2+108y+108\\
-19y^2-114y-171\\
-106y-318\\
-120\\
\hline
y^4+16y^3+71y^2-\ \ 4y-420=0
\end{array}$$

Aumentando las raíces verdaderas disminuyen las falsas, y a la inversa

Debe tenerse en cuenta que al incrementar las raíces verdaderas de una ecuación, las falsas disminuyen en la misma cantidad y, por el contrario, si se disminuyen las verdaderas, aumentan las falsas; pero, sean unas u otras, si disminuyen en una cantidad igual a ellas, entonces la raíz será cero y si disminuimos en una cantidad mayor que la raíz, entonces la raíz verdadera pasa a ser falsa o la falsa pasa a ser verdadera. Así, en nuestro caso, incrementando en 3 la raíz verdadera, 5, disminuyen en 3 cada una de las falsas de modo que la que era 4 no será sino 1, la que era 3 será 0, y la raíz 2 es ahora una raíz verdadera, pues $-2+3=+1$. Por ello en esta ecuación:

$$y^3-8y^2-y+8=0,$$

existiendo un cambio de signo de «+» a «−», y puesto que $-19x^2$ es seguido de $+106x$, y $+106x$ de -120, se deduce que hay tres raíces verdaderas y una falsa puesto que $-4x^3$ es seguido de $-19x^2$.

Cómo en una ecuación pueden convertirse las raíces falsas en verdaderas y a la inversa

Así mismo, es fácil lograr en una misma ecuación que todas las raíces falsas pasen a ser verdaderas y, por el mismo medio, que todas las verdaderas pasen a ser falsas. Esto se logra modificando los signos del segundo, cuarto, sexto y de todos los términos pares sin modificar los del primero, tercero, quinto y términos que ocupen un puesto impar. Así si en vez de

$$+x^4-4x^3-19x^2+106x-120=0,$$

ponemos

$$+x^4+4x^3-19x^2-106-120=0,$$

obtendremos una ecuación en la que no hay sino una raíz verdadera, que es 5, y tres falsas, que son 2, 3 y 4. A-T, VI, 447

Cómo pueden aumentarse o disminuirse las raíces de una ecuación sin conocerlas

Pero si los valores de las raíces de una ecuación son desconocidos y deseamos aumentar o disminuir cada una de las raíces en una cantidad conocida, solamente es necesario que en el lugar del término desconocido se suponga otro que sea mayor o menor que esta misma cantidad, sustituyéndolo en toda la ecuación. Así, si deseamos aumentar en 3 la raíz de esta ecuación,

$$x^4+4x^3-19x^2-106x-120=0,$$

debemos introducir y en lugar de x, y suponer que y es mayor que x en 3, de modo que $y-3=x$; así mismo, en vez de x^2 debemos introducir el cuadrado de $y-3$, que es y^2-6y+9; en vez de x^3, debe introducirse $y^3-9y^2+27y-27$; y en lugar de x^4, debemos introducir $y^4-12y^3+54y^2-108y+81$. De este modo, sustituyendo todos estos valores en la anterior ecuación, tenemos:

solamente hay tres raíces, de las cuales solamente dos son verdaderas, 1 y 8, y una falsa, que también es 1. A-T, VI, 449
Pero en esta otra ecuación:

$$y^4+16y^3+71y^2-4y-420=0,$$

solamente tenemos una raíz verdadera, 2, puesto que $+5-3=+2$, y tres raíces falsas, 5, 6 y 7.

Cómo se puede sacar el segundo término de una ecuación

Este procedimiento para transformar las raíces de una ecuación sin determinar sus valores permite realizar dos observaciones cuya utilidad posteriormente comprobaremos. En primer lugar, siempre podemos suprimir el segundo término de la ecuación que examinamos restando sus raíces verdaderas de la cantidad conocida de este segundo término, dividida por el número de dimensiones del primero si uno de estos términos es positivo y el otro negativo (si tienen signos opuestos); si tienen signos iguales, incrementando las raíces con la misma cantidad. Así para suprimir el segundo término de la ecuación:

$$y^4+16y^3+71y^2-4y-420=0,$$

dividido 16 por 4 (el exponente de y en y^4), siendo el cociente 4. Seguidamente establezco $z-4=y$. Entonces

$$\begin{array}{r}
z^4-16z^3+\ 96z^2-256z+\ 256\\
+16z^3-192z^2+768z-1.024\\
+\ 71z^2-568z+1.136\\
4z+\ 16\\
-\ 420\\
\hline
z^4\quad -\ 25z^2-\ 60z-\ 36=0
\end{array}$$

La verdadera raíz de esta ecuación que era 2, ha pasado a ser 6, puesto que ha sido incrementada en 4; las falsas que eran 5, 6 y 7 son en el caso presente 1, 2 y 3, pues A-T, VI, 450
cada una de ellas ha sido reducida en 4.

De modo similar, si deseamos suprimir el segundo término de:

$$x^4-2ax^3+(2a^2-c^2)x^2-2a^3x+a^4=0.$$

puesto que $2a$ dividido por $4 = \frac{1}{2}a$, debemos introducir $z + \frac{1}{2}a = x$ y tenemos que:

$$\begin{aligned} & z^4 + 2az^3 + \tfrac{3}{2}a^2z^2 + \tfrac{1}{2}a^3z + \tfrac{1}{16}a^4 \\ & -2az^3 - 3a^2z^2 - \tfrac{3}{2}a^3z - \tfrac{1}{4}a^4 \\ & +2a^2z^2 + 2a^3z + \tfrac{1}{2}a^4 \\ & - c^2z^2 - ac^2z - \tfrac{1}{4}a^2c^2 \\ & - 2a^3z - a^4 \\ & + a^4 \end{aligned}$$

$$z^4 + \left(\frac{1}{2}a^2 - c^2\right)z^2 - (a^3 + ac^2)z + \frac{5}{16}a^4 - \frac{1}{4}a^2c^2 = 0.$$

Y si posteriormente hallamos el valor de z, sumándole $\frac{1}{2}a$, obtendremos el valor de x.

Sobre el procedimiento para lograr que las raíces falsas pasen a ser verdaderas, sin que suceda lo inverso

En segundo lugar será igualmente de utilidad para el desarrollo de este tratado el considerar que siempre que incrementemos las raíces verdaderas en una cantidad mayor de lo que es cualquiera de las raíces falsas, todas ellas pasarán a ser verdaderas, de modo que no se registrarán dos signos, «+» o «−», consecutivamente; además de esto, que la cantidad conocida del tercer término sea mayor que el cuadrado de la mitad del segundo. Pues, aunque esto pueda realizarse cuando las
A-T. VI. 451 raíces falsas son desconocidas es fácil estimar aproximadamente sus valores y adoptar una cantidad que sea tal como se requiere o mayor aún para lograr tal efecto. Así, si tenemos:

$$x^6 + nx^5 - 6n^2x^4 + 36n^3x^3 - 216n^4x^2 + \\ + 1.296n^5x - 7.776n^6 = 0,$$

Sobre la multiplicación o división de raíces desconocidas

Así mismo, sin conocer el valor de las raíces de una ecuación podemos multiplicarlas o dividirlas por cualquier cantidad conocida que deseemos. Esto se hace suponiendo que la cantidad desconocida, siendo multiplicada o dividida por la cantidad que debe multiplicar o dividir las raíces, es igual a otra cantidad; seguidamente, multiplicando o dividiendo la cantidad conocida del segundo término por la cantidad dada, y la cantidad conocida del tercer término por el cuadrado de la cantidad dada, y la del cuarto por el cubo, y así sucesivamente hasta el último.

Sobre la reducción de los quebrados en una ecuación

Esto puede ser útil para reducir a números enteros y racionales tanto las fracciones como frecuentemente los números irracionales que se encuentran en los términos de las ecuaciones. Así, si tenemos

$$x^3 - \sqrt{3}x^2 + \frac{26}{27}x - \frac{8}{27\sqrt{3}} = 0,$$

y deseamos establecer otra en la que todos sus términos
A-T. VI. 453 sean expresados mediante números racionales, es preciso suponer:

$$y = x\sqrt{3},$$

y multiplicar por $\sqrt{3}$ la cantidad conocida del segundo término, que es también $\sqrt{3}$; y por su cuadrado, que es 3, la del tercer término, que es $\frac{26}{27}$; por su cubo, que es 3 $\sqrt{3}$, la del último, que es $\frac{8}{27\sqrt{3}}$.

Obtendremos como ecuación resultante:

$$y^3 - 3y^2 + \frac{26}{9}x - \frac{8}{9} = 0.$$

Posteriormente si se desea reemplazar esta ecuación por otra cuyas cantidades conocidas sean solamente expresadas en números enteros, debemos suponer que $z = 3y$, y multiplicando 3 por 3, $\frac{26}{9}$ por 9, y $\frac{8}{9}$ por

estableciendo $y-6n=x$, tendremos:

$$\begin{array}{rrrrrrr}
y^6-36ny^5+540n^2y^4-4.320n^3y^3+19.440n^4y^2-46.656n^5y+46.656n^6 \\
+\ ny^5-\ 30n^2y^4+\ 360n^3y^3-\ 2.160n^4y^2+\ 6.480n^5y-\ 7.776n^6 \\
-\ 6n^2y^4+\ 144n^3y^3-\ 1.296n^4y^2+\ 5.184n^5y-\ 7.776n^6 \\
+\ 36n^3y^3-\ 648n^4y^2+\ 3.888n^5y-\ 7.776n^6 \\
-\ 216n^4y^2+\ 2.592n^5y-\ 7.776n^6 \\
+\ 1.296n^5y-\ 7.776n^6 \\
-\ 7.776n^6 \\
\hline
y^6+35ny^5+504n^2y^4-3.780n^3y^3+15.120n^4y^2-27.216n^5y \quad =0.
\end{array}$$

Es claro en este caso que $504n^2$, la cantidad conocida del tercer término, es mayor que $\left(\frac{35}{2}n\right)^2$, el cuadrado de la mitad de la del segundo; no hay caso en el que la cantidad, en que se incrementen las verdaderas raíces, deba ser necesariamente mayor, en proporción a las que son dadas, de lo que ha sido para éste.

Procedimiento para completar los diversos lugares de una ecuación

Pero si no se desea que el último término sea nulo (cero), como en este caso, debe ser necesariamente incrementado el valor de las raíces aunque sólo sea en una pequeña cantidad, no debiendo ser tan reducida que no sea suficiente a tal efecto; no más que si se desea aumentar el número de dimensiones (el grado) de una ecuación y también asegurar que todos los lugares de sus términos estén completos, como si en vez de

$$x^5-b=0$$

deseamos una ecuación de sexto grado y en la que ninguno de sus términos fuere nulo; en este caso, debemos primeramente para

$x^5-b=0$, escribir A-T. VI. 452

$$x^6-bx=0;$$

en segundo lugar, habiendo establecido $y-a=x$, tendremos:

$$y^6-6ay^5+15a^2y^4-20a^3y^3+$$
$$+15a^4y^2=(6a^5+b)y+a^6+ab=0.$$

Es evidente que por muy pequeña que se haya supuesto la cantidad a, cada término de esta ecuación ha sido completado.

27, hallamos que $z^3-9z^2+26z-24=0$. Las raíces son 2, 3 y 4. A partir de esto deducimos que las raíces de la ecuación precedente eran $\frac{2}{3}$, 1 y $\frac{4}{3}$, siendo las de la primera

$$\frac{2}{9}\sqrt{3},\ \frac{1}{3}\sqrt{3}\ \text{y}\ \frac{4}{9}\sqrt{3}.$$

Procedimiento para establecer una igualdad entre la cantidad conocida de uno de los términos de la ecuación y otra cantidad dada

Esta operación también puede utilizarse para hacer que la cantidad conocida de algún término de la ecuación sea igual a alguna otra dada. Así, si teniendo la ecuación:

$$x^3-b^2x+c^3=0,$$

deseamos establecer otra ecuación en la que la cantidad conocida del término que ocupa el tercer lugar, esto es b^2, sea $3a^2$, es preciso suponer que:

$$y=x\sqrt{\frac{3a^2}{b^2}}$$

y tenemos

$$y^3-3a^2y+\frac{3a^3c^3}{b^3}\sqrt{3}=0.$$

Las raíces, verdaderas o falsas, pueden ser reales o imaginarias

Nótese, finalmente, que tanto las raíces verdaderas como las falsas no siempre son reales; es decir, podemos imaginar cuanto he dicho en cada ecuación, pero algunas veces no hay cantidad alguna que corresponda a las raíces imaginadas. Así, aunque pudiésemos imaginar tres raíces en esta ecuación: A-T. VI. 454

$$x^3-6x^2+13x-10=0,$$

no hay sino una real, 2; en relación con las otras dos, aunque se las incremente, disminuya o multiplique de

acuerdo con las reglas cuya exposición he realizado, siempre serán imaginarias.

La reducción de ecuaciones cúbicas cuando el problema es plano

Cuando la construcción de algún problema implica la solución de una ecuación en la que la cantidad desconocida tiene tres dimensiones, deben adoptarse los siguientes pasos: en primer lugar si la ecuación incluye algunas fracciones, debemos reducir a números enteros, siguiendo el procedimiento indicado; si contiene algún número irracional, en cuanto sea posible mediante multiplicación o por otros fáciles procedimientos, debemos reducirlos a racionales; en segundo lugar examinando por orden todas las cantidades que pueden dividir el último término sin dar lugar a fracción, es preciso comprobar si alguna de ellas es divisible por un binomio integrado por la cantidad desconocida más o menos alguna de ellas[65]. Si tal es el caso, entonces el problema es plano, esto es, puede ser construido por medio de la regla y el compás; pues, o bien la cantidad conocida de este binomio es la raíz exigida, o bien dividida la ecuación por el binomio, el cociente es de segundo grado (se reduce a dos dimensiones), de modo que podamos a partir de este cociente hallar la raíz, como se explicó en el primer libro de este tratado.

Por ejemplo, si tenemos:

$$y^6 - 8y^4 - 124y^2 - 64 = 0,$$

A-T. VI. 455 el último término, que es 64, es divisible por 1, 2, 4, 8, 16, 32 y 64; por tanto, debemos comprobar si la ecuación es divisible por y^2-1, y^2+1, y^2-2, y^2+2, y^2-4 y así sucesivamente. Hallaremos que es divisible por y^2-16 del modo siguiente:

$$\begin{array}{r} +y^6 - 8y^4 - 124y^2 - 64 = 0 \\ -y^6 - 8y^4 \dot{-} \quad 4y^2 - 16 \\ \hline -16y^4 - 128y^2 \\ \hline -16 \quad - \quad 16 \\ \hline +y^4 + \overline{8y^2} + 4 = 0 \end{array}$$

Esto muestra que la raíz buscada es a^2+c^2, siendo fácil de establecer la prueba mediante una multiplicación.

Qué problemas son sólidos cuando la ecuación es cúbica

Pero cuando no se encuentra algún binomio que pueda dividir la ecuación propuesta, es cierto que el
A-T. VI. 457 problema que de ella depende es sólido. En tal caso no es un error menor intentar su construcción utilizando solamente círculos y líneas rectas de lo que sería emplear secciones cónicas en la construcción de aquellos que no exigen sino círculos, ya que todo aquello que testimonia una cierta ignorancia suele ser llamado error.

Sobre la reducción de las ecuaciones con cuatro dimensiones cuando el problema es plano y sobre los que son sólidos

De igual modo, si tenemos una ecuación cuya cantidad desconocida tiene cuatro dimensiones, es preciso después de haber suprimido los números irracionales y fracciones, caso de haberlos, comprobar si es posible establecer algún binomio que divida la suma, estando integrado por una de las cantidades que dividiesen sin fracción el último término. Si se halla tal binomio, o bien la cantidad conocida de este binomio es la raíz buscada o al menos después de realizada tal división, la ecuación resultante solamente es de tres dimensiones, debe ser examinada de igual modo. Pero cuando tal binomio no se halla, es preciso, aumentando o disminuyendo el valor de la raíz, suprimir el segundo término del modo explicado; posteriormente, debe reducirse a otra que no tenga sino tres dimensiones. Esto se hace del modo siguiente: en vez de

$$x^4 \pm px^2 \pm qx \pm r = 0,$$

es preciso afirmar

$$y^6 \pm 2py^4 + (p^2 \pm 4r)y^2 - q^2 = 0.$$

En relación con los signos «+» o «−» (que he omitido[66]), si en la primera ecuación tenemos $+p$, es preciso que en la segunda introduzcamos $+2p$, o si en la primera se da $-p$, en la segunda debemos introducir

Modo de dividir una ecuación por un binomio que contiene su raíz

Comienzo por el último término, divido -64 por -16, que da $+4$, que anoto en el cociente; seguidamente se multiplica $+4$ por y^2, cuyo resultado es $+4y^2$ y consigno en el dividendo $-4y^2$, pues debemos introducir el signo «+» o «−» de modo opuesto al del resultado de la multiplicación. Sumando $-124y^2$ con $-4y^2$, obtenemos $-128y^2$. Dividiendo esto por -16, se obtiene $+8y^2$ en el cociente, y multiplicando por y^2 se obtiene $-8y^4$ que debe sumarse al término que corresponde, $-8y^4$, en el dividendo. El resultado es $-16y^4$ que dividido por -16 da como resultado y^4 para el cociente; por otra parte, $-y^6$ unido con $+y^6$, será cero el resultado, mostrándose que la división ha sido concluida. Si aún resta alguna cantidad o bien si alguno de los términos no fuese divisible sin obtenerse fracción, entonces estaría claro que no podría realizarse.

De modo semejante, si tenemos: A-T. VI. 456

$$\left.\begin{matrix} y^6+a^2 \\ -2c^2 \end{matrix}\right\} y^4 \quad \left.\begin{matrix} -a^4 \\ +c^4 \end{matrix}\right\} y^2 \quad \left.\begin{matrix} -a^6 \\ -2a^4c^2 \\ -a^2c^4 \end{matrix}\right\} = 0,$$

el último término es divisible sin fracción por a, a^2, a^2+c^2, a^3+ac^2 y semejantes. Pero solamente deben ser considerados dos, esto es, a^2 y a^2+c^2. Es tal, pues los otros darían un término en el cociente de mayor o menor grado que la cantidad conocida del penúltimo término, haciendo esto imposible la división. Nótese que estoy considerando y^6 como de tercer grado, puesto que no hay términos tales como y^5, ni y^3, ni y. Así, examinando el binomio $y^2-a^2-c^2=0$, hallamos que la división puede establecerse como sigue:

$$\begin{array}{l} \left.\begin{matrix} +y^6+a^2 \\ -y^6-2c^2 \\ 0\ -2a^2 \\ +\ c^2 \end{matrix}\right\} y^4 \quad \left.\begin{matrix} -a^4 \\ +c^4 \\ -a^4 \\ -a^2c^2 \end{matrix}\right\} y^2 \quad \left.\begin{matrix} -\ a^6 \\ -2a^4c^2 \\ -\ a^2c^4 \end{matrix}\right\} = 0 \\ \hline \begin{array}{ll} -a^2-c^2 & -a^2-c^2 \ \ -\ a^2-c^2 \\ \hline +y^4 & +2a^2 \\ & -\ c^2 \end{array} \left.\begin{matrix} \\ \\ +\ a^4 \\ +\ a^2c^2 \end{matrix}\right\} y^2 \left.\begin{matrix} \\ \\ \\ \end{matrix}\right\} = 0; \end{array}$$

$-2p$; por el contrario, si tenemos $+r$, debe introducirse $-4r$, y $+4r$ si tenemos $-r$; pero si tenemos $+q$ o $-q$ siempre deberemos introducir $-q^2$ y $+p^2$ en la segunda, al menos si se supone que x^4 e y^6 tienen signo positivo (+); en caso contrario debería ser $+q^2$ y $-p^2$. A-T. VI. 458

Por ejemplo, si tenemos:

$$x^4-4x^2-8x+35=0,$$

debemos introducir en su lugar

$$y^6-8y^4-124y^2-64=0,$$

pues la cantidad llamada p siendo igual a -4, debemos reemplazar $2py^4$ por $-8y^4$; puesto que r es igual a 35, deberemos sustituir $(p^2-4r)y^2$ por $(16-140)y^2$ o $-124y^2$; y puesto que $q=8$, deberemos reemplazar $-q^2$ por -64.

De modo semejante, en vez de

$$x^4-17x^2-20x-6=0,$$

debemos afirmar

$$y^6-34y^4+313y^2-400=0,$$

pues 34 es el doble de 17, 313 es el cuadrado de 17 sumado al cuádruplo de 6, y 400 es el cuadrado de 20.

Del mismo modo, en vez de

$$+z^4+\left(\frac{1}{2}a^2-c^2\right)z^2-(a^3+ac^2)z-\frac{5}{16}a^4-\frac{1}{4}a^2c^2=0$$

debemos escribir

$$y^6+(a^2-2c^2)y^4+(c^4-a^4)y^2-a^6-2a^4c^2-a^2c^4=0;$$

pues,

$$p=\frac{1}{2}a^2-c^2,\ p^2=\frac{1}{4}a^4-a^2c^2+c^4,\ 4r=-\frac{5}{4}a^4+a^2c^2.$$

Finalmente, tenemos:

$$-q^2=-a^6-2a^4c^4-a^2c^4.$$

Cuando ya hemos reducido la ecuación a tres dimensiones, es preciso calcular el valor de y^2, utilizando el método que ya hemos explicado[66 bis]; si no pudiera ser hallado, no existe necesidad alguna de proseguir, pues de ello se deduce necesariamente que el problema es sólido. Pero si es hallado el valor de y^2, ello hará posible que podamos dividir la ecuación precedente en otras dos, no teniendo la cantidad desconocida en cada A-T. VI. 459

una de ellas, sino dos dimensiones, y siendo sus raíces las mismas que las de la ecuación original. Así en vez de

$$\pm x^4 \pm px^2 \pm qx \pm r = 0,$$

deberemos establecer estas otras dos ecuaciones

$$+x^2 - yx + \frac{1}{2}y^2 \pm \frac{1}{2}p \pm \frac{q}{2y} = 0$$

y

$$+x^2 + yx + \frac{1}{2}y^2 \pm \frac{1}{2}p \pm \frac{q}{2y} = 0.$$

En relación con los signos + y − que he omitido[66], si tenemos $+p$ en la ecuación precedente, entonces debemos anotar $+\frac{1}{2}p$ en cada una de las otras dos ecuaciones; caso de que tengamos $-p$, entonces $-\frac{1}{2}p$. Pero será necesario anotar $+\frac{q}{2y}$ cuando tenemos $-yx$, y $-\frac{q}{2y}$ cuando tenemos $+yx$, siempre que q vaya precedida de signo positivo. Pero, por el contrario, si tuviéramos $-q$ deberíamos poner $-\frac{q}{2y}$ en la que se diera $-yx$; debería anotarse $+\frac{q}{2y}$ si encontramos $+yx$. De este modo es fácil conocer todas las raíces de la ecuación propuesta y, en consecuencia, construir el problema cuya solución expresa, no siendo necesario utilizar sino círculos y líneas rectas.

Por ejemplo, tomando

$$+y^6 - 34y^4 + 313y^2 - 400 = 0$$

en vez de

$$+x^4 - 17x^2 - 20x - 6 = 0,$$

A. T. VI. 460 hallamos que $y^2 = 16$; entonces, en lugar de la ecuación

$$+x^4 - 17x^2 - 20x - 6 = 0,$$

deberemos anotar

$$+x^2 - 4x - 3 = 0$$

puesto que $y^2 = a^2 + c^2$, tendremos que

$$z^2 - \sqrt{a^2+c^2}\,z + \frac{3}{4}a^2 - \frac{1}{2}a\sqrt{a^2+c^2} = 0,$$

y

$$z^2 + \sqrt{a^2+c^2}\,z + \frac{3}{4}a^2 + \frac{1}{2}a\sqrt{a^2+c^2} = 0;$$

puesto que

$$y = \sqrt{a^2+c^2},$$

y que

$$+\frac{1}{2}y^2 + \frac{1}{2}p = \frac{3}{4}a^2;$$

y, finalmente, que

$$\frac{q}{2y} = \frac{1}{2}a\sqrt{a^2+c^2},$$

por tanto, llegamos a establecer que el valor de

$$z = \frac{1}{2}\sqrt{a^2+c^2} + \sqrt{-\frac{1}{2}a^2 + \frac{1}{4}c^2 + \frac{1}{2}a\sqrt{a^2+c^2}}$$

o bien que

$$z = \frac{1}{2}\sqrt{a^2+c^2} - \sqrt{-\frac{1}{2}a^2 + \frac{1}{4}c^2 + \frac{1}{2}\,a\sqrt{a^2+c^2}}.$$

Como se había establecido que $z + \frac{1}{2}a = x$, tenemos que la cantidad x, para cuyo conocimiento hemos realizado estas operaciones, es igual a

$$+\frac{1}{2}a + \sqrt{\frac{1}{2}a^2 + \frac{1}{4}c^2} - \sqrt{\frac{1}{4}c^2 - \frac{1}{2}a^2 + \frac{1}{2}a\sqrt{a^2+c^2}}.$$

Ejemplo del uso de estas reducciones

A. T. VI. 462 Con el fin de que pueda ser más fácilmente reconocida la utilidad de esta regla, es preciso que la aplique a algún problema[67].

Si el cuadrado AD y la línea BN son dados, debemos prolongar el lado AC hasta E, de modo que EF, trazada desde E hacia B sea igual a NB. Conocemos por Pappus

y

$$+x^2+4x+2=0;$$

pues $y=4$, $\frac{1}{2}y^2=8$, $p=17$ y $q=20$, tenemos que

$$+\frac{1}{2}y^2-\frac{1}{2}p-\frac{q}{2y}=-3$$

y

$$+\frac{1}{2}y^2-\frac{1}{2}p+\frac{q}{2y}=+2.$$

Obteniendo las raíces de estas ecuaciones, llegaremos al mismo resultado que en el caso de que se obtuvieran en la que contiene x^4; es decir, tendríamos una verdadera $\sqrt{7+2}$, y tres falsas: $2+\sqrt{2}$, $\sqrt{7-2}$ y $2-\sqrt{2}$. Así, teniendo

$$x^4-4x^2-8x+35=0,$$

puesto que la raíz de

$$y^6-8y^4-124y^2-64=0$$

es 16, debemos anotar

$$x^2-4x+5=0$$

y

$$x^2+4x+7=0.$$

En este caso tenemos que

$$+\frac{1}{2}y^2-\frac{1}{2}p-\frac{q}{2y}=5$$

y

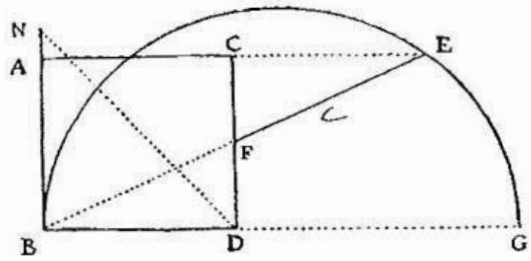
$$+\frac{1}{2}y^2-\frac{1}{2}p+\frac{q}{2y}=7.$$

Como no se encuentra ninguna raíz, verdadera o falsa, en estas dos últimas, deducimos que las cuatro raíces de la ecuación original son imaginarias; asimismo que el problema al cual corresponde es plano, pero que no podrá ser construido puesto que las cantidades dadas no pueden ser unidas. A-T, VI, 461

De igual modo, si tenemos

$$z^4+\left(\frac{1}{2}a^2-c^2\right)z^2-(a^3+ac^2)z+\frac{5}{16}a^4-\frac{1}{4}a^2c^2=0,$$

que si BD se prolonga hasta G de modo que DG sea igual a DN y se describe un círculo cuyo diámetro sea BG, entonces si prolongamos la recta AC, encontrará a la circunferencia de este círculo en el punto E que era el buscado. Pero supondría una gran dificultad para quienes no están familiarizados con tal construcción su hallazgo; sin embargo, si aplican el método aquí propuesto, nunca tomarían DG como la cantidad desconocida, sino más bien CF o FD puesto que conducen con mayor facilidad al establecimiento de una ecuación. En tal caso accederíamos a una ecuación cuya solución no



sería fácil si se prescinde de la regla que acabo de explicar. Pues si establecemos a por BD o CD, c por EF y x por DF, tenemos que $CF=a-x$, y puesto que CF es a FE como FD es a BF, tenemos que

$$a-x:c=x:BF,$$

de donde obtenemos que

$$BF=\frac{cx}{a-x}.$$

Seguidamente, puesto que el triángulo BDF es rectángulo, sus lados son x y a; sus cuadrados, a^2+x^2, son igual al cuadrado de la hipotenusa que es

$$\frac{c^2x^2}{x^2-2ax+a^2}.$$

Si, pues, multiplicamos el todo por $x^2-2ax+a^2$, obtenemos la ecuación

$$x^4-2ax^3+2a^2x^2-2a^3x+a^4=c^2x^2$$

o bien A-T, VI, 463

$$x^4-2ax^3+(2a^2-c^2)x^2-2a^3x+a^4=0;$$

mediante el procedimiento expuesto conocemos que su raíz, que es la longitud de DF, es

$$\frac{1}{2}a+\sqrt{\frac{1}{4}a^2+\frac{1}{4}c^2}-\sqrt{\frac{1}{4}c^2-\frac{1}{2}a^2+\frac{1}{2}a\sqrt{a^2+c^2}}.$$

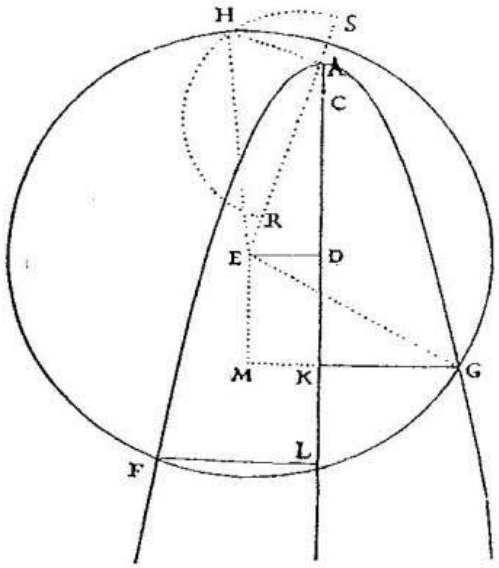
Si considerásemos BF o CE o BE como la cantidad desconocida obtendríamos una ecuación con cuatro dimensiones pero cuyo desarrollo sería más fácil y, a la vez, se accedería a la misma con gran facilidad. Por el contrario, si tomásemos DG se obtendría con mucha mayor dificultad la ecuación, aunque también sería muy simple. Comunico esto para advertir que cuando se propone un problema que no es sólido si al intentarlo resolver por un método accedemos a una ecuación muy compleja, podemos acceder a otra más simple tratando de hallarla por otro camino.

Aún podría comunicar otras reglas para resolver ecuaciones cúbicas o bicuadráticas, pero estimo que son innecesarias, pues si los problemas son planos puede accederse a su solución mediante los procedimientos que hemos propuesto.

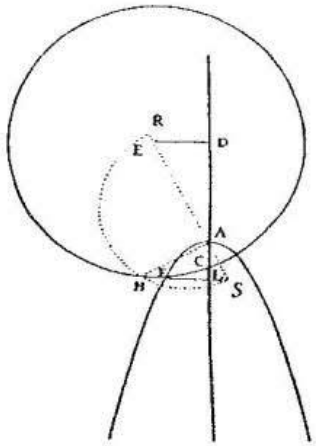
Regla general para reducir las ecuaciones que pasan el cuadrado de cuadrado

Así mismo, podría agregar otras reglas relacionadas con las ecuaciones de quinto, sexto y superior grado. Sin embargo, prefiero reducirlas todas a una sola regla,
A-T, VI, 464 afirmando en general que cuando hemos intentado reducir las ecuaciones dadas a la misma forma de aquellas que con otras tantas dimensiones, provienen de la multiplicación de otras dos con menos dimensiones, y que si habiendo enumerado todos los medios por los cuales esta multiplicación es posible, no ha sido posible obtenerse, entonces podemos asegurar que las ecuaciones dadas no pueden ser reducidas a otras más simples. De modo que si la cantidad desconocida tuviese tres o cuatro dimensiones, el problema correspondiente es sólido; si tiene cinco o seis dimensiones, entonces posee un grado mayor de complejidad, y así sucesivamente.

Finalmente, he omitido las demostraciones de la mayor parte de cuanto he expuesto porque me han parecido tan fáciles que si se toma la molestia de examinar metódicamente si me he equivocado, llegarán



(en el caso de que p sea igual a 0 desde el punto C) se traza la perpendicular DE sobre CD de modo que DE
A-T, VI, 466 sea igual a $\frac{1}{2}q$. Finalmente, desde el punto E y con el radio AE se traza el círculo FG en el caso de que la ecuación dada sea cúbica, esto es, de modo que r sea igual a cero. Pero cuando nos encontramos con que en la ecuación se halla $+r$, debemos tomar sobre la prolongación de AE, por un lado, $AR=ry$, por otro, AS



a hacerse presentes. Es mucho más útil conocerlas de este modo que mediante una simple lectura.

Forma general de reducir todos los problemas sólidos en una ecuación de tercer o cuarto grado

Asegurados de que el problema propuesto es sólido, sea que la ecuación de la cual depende su solución sea de cuarto grado o bien porque solamente sea de tercer grado, siempre podremos hallar su raíz por medio de una cualquiera de las tres secciones cónicas o incluso por alguna de sus partes, por pequeña que fuese, empleando simplemente líneas rectas y circulares. Me sentiré satisfecho en este lugar con facilitar solamente una regla para encontrarlas todas por medio de una parábola puesto que, en cierto modo, es la más simple de estas curvas.

En primer lugar, debe eliminarse el segundo término de la ecuación, caso de que ya no sea nulo y reducir la ecuación a la fórmula

$$z^3 = \pm aqz \pm a^2 q,$$

si la cantidad desconocida no tiene más que tres dimensiones, en el caso de que tuviera cuatro deberíamos reducirla a la fórmula A-T, VI, 465

$$z^4 = \pm apz^2 \pm a^2 qz \pm a^3 r.$$

Si tomamos a como la unidad, entonces la reducción para el primer caso sería

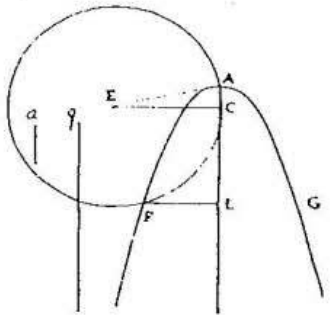
$$z^3 = \pm pz \pm q$$

y para el segundo

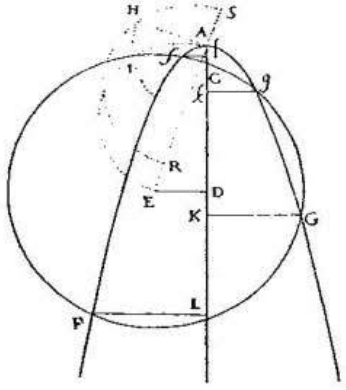
$$z^4 = \pm pz^2 + qz \pm r.$$

A continuación, suponiendo que la parábola FAG ha sido trazada y que su eje es $ACDKL$, que su lado recto es $a = 1$, y que AC es su mitad ($\frac{1}{2}$); suponiendo que el punto C es interior a esta parábola y que A es el vertex. Debemos establecer que $CD = \frac{1}{2}p$ y tomarla de modo que los puntos D y A caigan sobre el mismo lado que C en el caso de que la ecuación contenga $+p$; en el caso de que en la ecuación tengamos $-p$, es preciso que caiga del lado opuesto. Seguidamente desde el punto D

que debe ser igual al lado recto de la parábola, esto es, a 1. Cuando hayamos descrito un círculo cuyo diámetro sea RS, debemos trazar una perpendicular, AH, sobre AE, encontrando tal perpendicular al círculo RHS en el punto H, por el que debe pasar también el otro círculo FHG. En el caso de que tengamos $-r$, es preciso



después de haber hallado de este modo la línea AH, inscribir AI, siendo igual a AH en otro círculo, siendo AE el diámetro. En tal caso pasará por I el primer círculo buscado FIG. Este círculo, FG, puede cortar o ser tangente a la parábola en 1 ó 2 ó 3 ó 4 puntos; si desde ellos trazamos perpendiculares sobre el eje, tendremos todas las raíces de la ecuación, tanto las verdaderas como las falsas. Si la cantidad q está precedida del signo $+$ las raíces verdaderas serán aquéllas, como FL que se encuentran del mismo lado de la parábola que E, centro del círculo; las otras, tal como es el caso de GK, serán las raíces falsas. Pero, si por el A-T, VI, 467



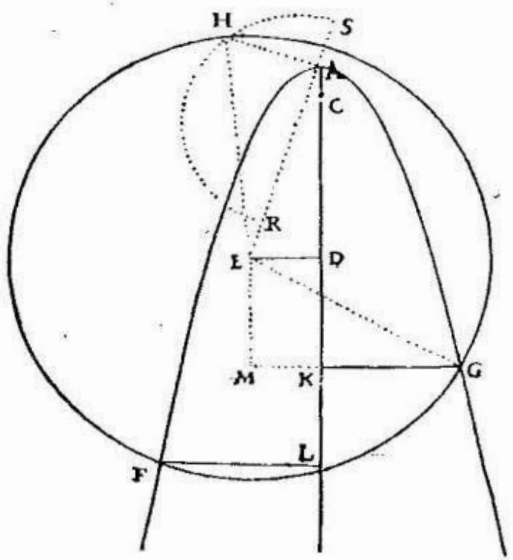
contrario, q está precedida del signo «$-$», entonces las raíces verdaderas se encontrarán al otro lado y las falsas o menores que cero serán las que se encuentran en el lado en que está E, el centro del círculo. Finalmente, si el círculo no cortara ni fuera tangente a la parábola en ningún punto, esto sería indicativo de que la ecuación no tiene ninguna raíz verdadera o falsa, sino que todas son imaginarias. De modo que esta regla es la más general y perfecta que se puede desear.

Por otra parte, su demostración es muy fácil. Pues si la línea GK, hallada mediante esta construcción, la llamamos z, entonces AK será z^2, en virtud de la naturaleza de esta parábola: CK debe ser la media proporcional entre AK y el lado recto cuyo valor es 1. Si, pues, de AK resto AC o $\frac{1}{2}$, y CD o $\frac{1}{2}p$, entonces tendremos que DK o EM, que es igual a

$$z^2 - \frac{1}{2}p - \frac{1}{2},$$

cuyo cuadrado sería

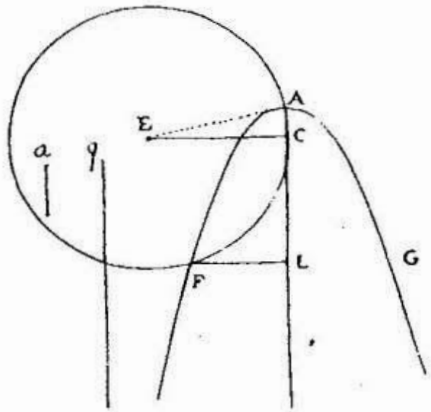
$$z^4 - pz^2 - z^2 + \frac{1}{4}p^2 + \frac{1}{2}p + \frac{1}{4};$$



Sobre la determinación de dos medias proporcionales

Así mismo, si lo deseamos, siguiendo esta regla podemos hallar dos medias proporcionales entre las líneas a y q. Es claro que si a una llamamos z, entonces tendremos que

$$a : z = z : \frac{z^2}{a} = \frac{z^2}{a} : \frac{z^3}{a^2}.$$



De este modo tenemos una ecuación entre q y $\frac{z^3}{a^2}$, esto es que $z^3 = a^2 q$.

La parábola FAG siendo trazada con la parte de su eje AC, siendo que AC es igual a $\frac{1}{2}a$, la mitad del lado recto, debe trazarse por C una perpendicular $CE = \frac{1}{2}q$, describiendo el círculo AF de modo que el centro del mismo sea E y que pase por A. De este modo tendremos
A-T. VI. 470 que FL y LA serán las medias proporcionales buscadas.

Modo de dividir un ángulo en tres partes

De igual modo si deseamos dividir el ángulo NOP, o bien el arco o porción del círculo $NQTP$ en tres partes iguales, tomando $NO = 1$ como el radio del círculo, $NP = q$ como la cuerda del arco dado y $NQ = z$ para la

pero puesto que $DE = KM = \frac{1}{2}q$, tendremos que la línea A-T. VI, 468

$GM = z + \frac{1}{2}q$, y el cuadrado de GM será igual a $z^2 + qz + \frac{1}{4}q^2$. La adición de estos dos cuadrados nos da que

$$z^4 - pz^2 + qz + \frac{1}{4}q^2 + \frac{1}{4}p^2 + \frac{1}{2}p + \frac{1}{4}$$

es el valor para el cuadrado de la línea GE, puesto que es la hipotenusa del triángulo EMG.

Pero puesto que esta misma línea GE es el radio del círculo FG, puede ser expresada en otros términos; esto es, si

$$ED = \frac{1}{2}q,$$

y

$$AD = \frac{1}{2}p + \frac{1}{2},$$

entonces tendremos que

$$EA = \sqrt{\frac{1}{4}q^2 + \frac{1}{4}p^2 + \frac{1}{2}p + \frac{1}{4}},$$

puesto que ADE es un ángulo recto. Luego, siendo AH A-T. VI, 469
media proporcional entre $AS = 1$ y AR o r, tendremos que AH es igual a $\sqrt{r}$; y puesto que EAH es un ángulo recto, el cuadrado de HE o de EG es

$$\frac{1}{4}q^2 + \frac{1}{4}p^2 + \frac{1}{2}p + \frac{1}{4} + r.$$

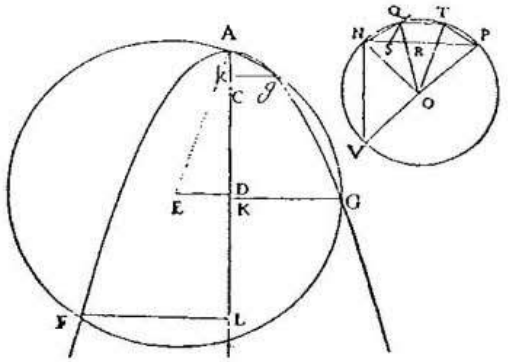
De este modo tenemos una ecuación a partir de esta suma y de la anteriormente obtenida.

Tal ecuación será de la forma $z^4 = pz^2 - qz + r$ y, en consecuencia, tendremos que la línea GK, llamada z, es la raíz de esta ecuación. Tal era el objeto de esta demostración. Si se aplica el mismo cálculo a todos los otros casos de esta regla, adoptando los oportunos cambios de los signos (+ y −), podrán lograrse resultados semejantes, no siendo necesario que yo me detenga en tal tipo de consideraciones[68].

cuerda que corresponde al tercio del arco dado, tenemos que la ecuación es

$$z^3 = 3z - q.$$

Es tal, pues trazando las líneas NQ, OQ y OT, y trazando a su vez QS paralela a TO, se ve que NO es a NQ como NQ es a QR y como QR es a RS. De modo que siendo $NO = 1$, y $NQ = z$, entonces $QR = z^2$ y RS



$= z^3$. Puesto que es preciso solamente RS o z^3 para que $NP = q$ sea triple de $NQ = z$, tenemos que $q = 3z - z^3$ o bien que $z^3 = 3z - q$.

Si posteriormente describimos la parábola FAG, y si CA es la mitad de su lado recto, y si $CD = \frac{3}{2}$ y la perpendicular $DE = \frac{1}{2}q$, y, finalmente, si desde el centro E describimos el círculo $FAgG$ por A, tendremos que cortará a esta parábola en los tres puntos F, g y G sin contar el punto A que es el vértice. Esto muestra que A-T. VI, 471
hay tres raíces en esta ecuación, a saber: GK y gk que son verdaderas, y la tercera que es falsa, FL. De las dos raíces verdaderas, la menor de ellas gk es la que debemos tomar para la línea buscada NQ; la otra, GK, es igual a NV la cuerda correspondiene a la tercera parte del arco NVP que junto con el otro arco NQP constituye el círculo. La falsa, FL, es igual a la suma de las otras dos, QN y NV, como fácilmente puede comprobarse mediante el cálculo[69].

Sobre la reducción de los problemas sólidos a estas dos construcciones

Sería superfluo que me detuviese explicando otros ejemplos ya que todos los problemas sólidos admiten reducciones tales que no necesitamos aplicar esta regla para construirlos con excepción de aquellos casos en que es necesario hallar dos medias proporcionales o dividir un ángulo en tres partes iguales. Esto es evidente si se considera que las dificultades propias de estos problemas pueden estar siempre comprendidas en ecuaciones no superiores a las de tercer o cuarto grado; que todas las ecuaciones de cuarto grado pueden reducirse a las de segundo por medio de otras no superiores a las de tercer grado y, finalmente, que puede eliminarse el segundo término de estas ecuaciones. En consecuencia, no hay ninguna que no pueda reducirse a alguna de estas tres formas

$$z^3 = -pz + q$$
$$z^3 = +pz + q$$
$$z^3 = +pz - q.$$

Así, si tenemos $z^3 = -pz + q$, la regla que atribuye
A-T, VI, 472 Cardan a Escipión Ferreus, nos enseña que la raíz es

$$\sqrt[3]{\frac{1}{2}q + \sqrt{\frac{1}{4}q^2 + \frac{1}{27}p^3}} - \sqrt[3]{-\frac{1}{2}q + \sqrt{\frac{1}{4}q^2 + \frac{1}{27}p^3}}.$$

De modo semejante cuando tenemos $z^3 = +pz + q$, y el cuadrado de la mitad del último término es mayor que el cubo del tercio de la cantidad conocida del penúltimo término, una regla semejante nos muestra que la raíz es

$$\sqrt[3]{\frac{1}{2}q + \sqrt{\frac{1}{4}q^2 - \frac{1}{27}p^3}} + \sqrt[3]{\frac{1}{2}q - \sqrt{\frac{1}{4}q^2 - \frac{1}{27}p^3}}.$$

Por ello parece que pueden construirse todos los problemas cuyas ecuaciones pueden reducirse a una de estas dos formas sin tener necesidad de las secciones cónicas sino para extraer las raíces cúbicas de ciertas cantidades conocidas, es decir, para calcular medias proporcionales entre estas cantidades y la unidad.

De nuevo, si tenemos $z^3 = +pz + q$ cuando el cuadrado de la mitad del último término no es mayor que el cubo del tercio de la cantidad conocida del penúltimo término, describiendo el círculo $NQPV$ con radio NO

Forma de expresar el valor de todas las raíces de las ecuaciones cúbicas y de todas las bicuadráticas

A-T, VI, 474 Por otra parte, debe señalarse que la forma de expresar el valor de las raíces en base a la relación que guardan con los lados de ciertos cubos de los que no se conocen sino sus áreas, no es ni más inteligible ni más simple que la de expresarlas en virtud de su relación con las cuerdas de ciertos arcos o porciones de círculos cuyos triplos son dados. Por ello todas las ecuaciones cúbicas que no pueden ser expresadas mediante las reglas de Cardan, pueden serlo con tanta claridad o aún mayor del modo indicado.

Pues si, por ejemplo, se piensa conocer la raíz de la ecuación

$$z^3 = +pz + q,$$

puesto que se sabe que está compuesta de dos líneas, una de las cuales es el lado de un cubo cuyo contenido es $+\frac{1}{2}q$, agregado al lado de un cuadrado cuya área es

$$\frac{1}{4}q^2 - \frac{1}{27}p^3,$$

y la otra es el lado de otro cubo cuyo volumen es la diferencia entre $\frac{1}{2}q$ y el lado del cuadrado cuyo área es

$$\frac{1}{4}q^2 - \frac{1}{27}p^3.$$

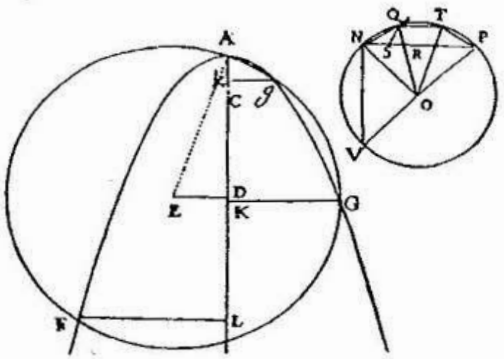
Todo esto es lo que podemos conocer mediante la regla de Cardan. No cabe duda de que podemos conocer el valor de la raíz de

$$z^3 = pz - q$$

considerándola inscrita en un círculo cuyo radio es $\sqrt{\frac{1}{3}p}$ y sabiendo que es la cuerda de un arco cuyo triple tiene por cuerda el valor $\frac{3q}{p}$. Estos mismos términos son menos engorrosos que los otros y, además, más breves si se desea usar algún signo particular para
A-T, VI, 475 expresar estas cuerdas, tal como se hace con $\sqrt{C}$ para expresar el lado de los cubos.

igual a $\sqrt{\frac{1}{3}p}$, es decir, la media proporcional entre el tercio de la cantidad conocida, p, y la unidad; seguidamente tómese $NP=\frac{3q}{p}$, inscrita en este círculo, de modo que NP sea a q, la otra cantidad conocida, como 1 es a $\frac{1}{3}p$. Divídase cada uno de los arcos NQP y NVP en tres partes iguales, y la raíz buscada será la suma de NQ y NV, cuerdas correspondientes a los tercios de los respectivos arcos, que unidas nos dan la raíz buscada. A-T. VI, 473

Finalmente, si tenemos $z^3=pz-q$, suponiendo de nuevo el círculo $NQPV$ cuyo radio $NO=\sqrt{\frac{1}{3}p}$ y que



NP, igual a $\frac{3q}{p}$, esté inscrito en este círculo; entonces NQ la cuerda del tercio del arco NQP, será la primera de las raíces buscadas y NV, la cuerda correspondiente al tercio del otro arco, será la segunda. Esto será así si el cuadrado de la mitad del último término no es mayor que el cubo del tercio de la cantidad conocida del penúltimo. En el caso de que fuera mayor, la línea NP no estaría inscrita en el círculo, puesto que su dimensión sería superior a la del diámetro. En tal caso, las dos raíces que eran verdaderas deberían ser imaginarias y la única raíz verdadera sería la que previamente era falsa que, según la regla de Cardan, sería:

$$\sqrt[3]{\frac{1}{2}q+\sqrt{\frac{1}{4}q^2-\frac{1}{27}p^3}}+\sqrt[3]{\frac{1}{2}q-\sqrt{\frac{1}{4}q^2-\frac{1}{27}p^3}}.$$

Así mismo, es posible expresar las raíces de todas las ecuaciones bicuadráticas mediante las reglas expresadas. De modo que no conozco nada más que sea deseable en esta cuestión. La naturaleza de estas raíces no nos permite que las expresemos en términos más simples ni que se las determine por alguna construcción que sea no sólo más general, sino también más fácil.

Razones por las que los problemas sólidos no pueden ser construidos sin las secciones cónicas y aquéllos cuya complejidad es mayor sin otras de complejidad creciente

Verdad es que aún no he dicho nada sobre las razones por las que me atrevo a afirmar que algo es o no posible. Pero, si nos percatamos de cómo mediante el método que utilizo todo cuanto es considerado por los geómetras se reduce a un mismo género de problemas, que consiste en buscar el valor de las raíces de alguna ecuación, se podrá juzgar correctamente que no es equivocado realizar una enumeración de todos los medios por los que pueden ser halladas de modo que sea suficiente para mostrar que se ha escogido el más general y el más simple. En particular y en relación con los problemas sólidos que he indicado que no pueden ser construidos sin que se emplee alguna línea más compleja que la circular, es algo que también puede ser fácilmente cognoscible a partir de que éstos se reducen a dos tipos de construcciones: uno de ellos supone tener los dos puntos que determinan dos medias proporcionales entre dos líneas dadas; el otro, los dos puntos que dividen en tres partes iguales un arco dado. Pues, en tanto que la curvatura del círculo no depende sino de una simple relación de todas sus partes con el punto A-T, VI, 476
que ocupa el centro, no puede sernos de utilidad sino para determinar un solo punto entre dos extremos, o bien para hallar una media proporcional entre dos líneas rectas dadas, o para dividir en dos un arco dado. Por el contrario la curvatura de las secciones cónicas, dependiendo de dos factores diferentes, puede servir para determinar dos puntos diferentes.

Pero por esta misma razón es imposible que cualquiera de los problemas que son de grado superior a los sólidos y que presuponen la determinación de cuatro medias proporcionales o la división de un ángulo en cinco partes iguales, puedan ser construidos por medio de alguna de las secciones cónicas. Por ello creo realizar

lo mejor si facilito una regla general para construirlos, empleando la línea curva que se describe por la intersección de una parábola y una línea recta de la forma ya explicada. Me atrevo a afirmar que no hay otra más simple en la naturaleza que pueda contribuir a tal efecto y se ha visto como sigue a las secciones cónicas en esta cuestión tan estudiada por los antiguos y cuya solución muestra por orden todas las líneas curvas que deben ser recibidas en Geometría.

Procedimiento general para construir todos los problemas reducidos a una ecuación no superior al grado sexto

Ya se conoce cómo cuando se buscan las cantidades que son requeridas para la construcción de estos problemas pueden siempre reducirse a alguna ecuación que no sea superior a las de quinto o sexto grado. También es conocido cómo aumentando el valor de las raíces en esta ecuación puede lograrse que todas lleguen a ser
A-T, VI. 477 verdaderas; de este modo, que la cantidad conocida del tercer término sea mayor que el cuadrado de la mitad de la del segundo. Finalmente, si no es superior al quinto grado, puede ser modificada de modo que alcance el sexto grado, logrando que cada uno de sus términos estén consignados. Así, con el fin de que todas las dificultades de las que pueden surgir en esto, puedan ser resueltas recurriendo a una misma regla, deseo que todo esto sea realizado y que, por tal medio, se les reduzca siempre a una ecuación del tipo

$$y^6 - py^5 + qy^4 - ry^3 + sy^2 - ty + u = 0$$

en la que q es mayor que el cuadrado de la mitad de la designada p.

Seguidamente prolónguese indefinidamente por ambos extremos la línea BK y trácese AB por B perpendicular a BK, siendo igual a $\frac{1}{2}p$; en un plano separado trácese la parábola CDF cuyo lado recto principal es

$$\sqrt{\frac{t}{\sqrt{u}} + q - \frac{1}{4}p^2},$$

que será representado por n por razones de comodidad. Finalmente, es preciso situar el plano en que se encuentra la parábola sobre el que se han trazado las líneas AB

$$LH = \frac{t}{2n\sqrt{u}};$$

a continuación desde el punto H, así hallado, se traza formando ángulos rectos la línea HI del mismo lado en que se encuentra la curva ACN. Tomando

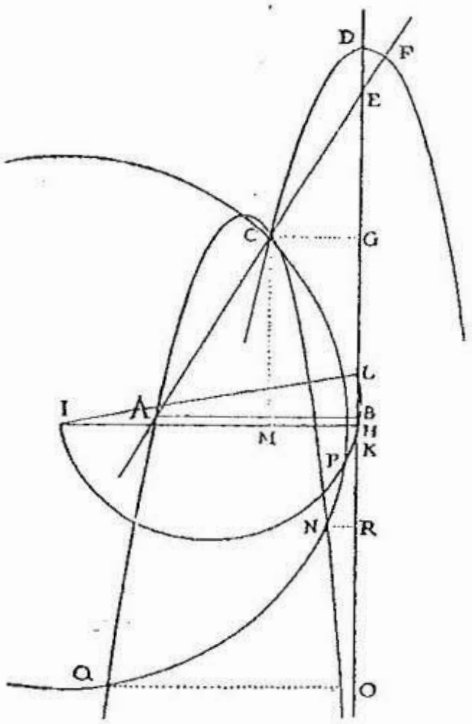
$$HI = \frac{r}{2n^2} + \frac{\sqrt{u}}{n^2} + \frac{pt}{4n^2\sqrt{u}},$$

que, para abreviar, podemos llamar $\frac{m}{n^2}$; uniendo los puntos L e I y describiendo el círculo LPI se tiene que LI es su diámetro e inscribiéndose la línea LP en este círculo, tenemos que es igual a

$$\sqrt{\frac{s + p\sqrt{u}}{n^2}}.$$

Finalmente, descríbase el círculo PCN con centro en I y pasando por el punto P. Este círculo deberá cortar o ser tangente a la curva ACN en tantos puntos como raíces tenga la ecuación. De modo que las perpendiculares trazadas desde estos puntos sobre la línea BK tal como
A-T, VI. 479 CG, NR, QO y semejantes, serán las raíces buscadas sin que exista excepción o fallo alguno dada la aplicación de esta regla. Pues si la cantidad s fuera tan grande, en proporción con las otras, p, q, r, t y u, que la línea LP fuera mayor que el diámetro del círculo IL, de modo que no pudiera inscribirse, en tal caso no habría ninguna raíz en esta ecuación que no fuese imaginaria; de igual modo sucedería si el círculo IP fuese tan pequeño que no cortase a ACN en punto alguno. Pero puede cortarla en seis puntos diferentes, pudiendo, en consecuencia, existir seis raíces diferentes en la ecuación. Cuando la corta en un número menor, esto testimonia que alguna de estas raíces son iguales entre sí o bien que no son sino imaginarias.

En el caso de que el procedimiento utilizado para trazar la línea ACN, mediante el movimiento de una parábola, pudiera parecer incómodo, es fácil hallar otros diversos medios para describirla. Así, teniendo las mismas cantidades que antes para AB y BL y también para BK, que había sido establecida como el lado recto de la parábola, debe describirse el semicírculo KST tomando el centro del mismo a discreción en la línea
A-T, VI. 480 BK, de modo que corte en alguna parte la línea AB, tal

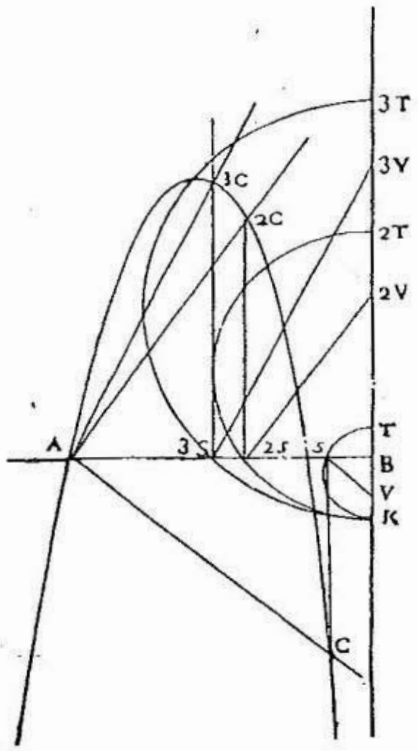


y BK, de modo que su eje, DE, se encuentre sobre la recta BK. Tomando la parte de este eje que está entre los puntos E y $D = \frac{2\sqrt{u}}{pn}$, es preciso situar sobre E una regla de modo tal que conecte el punto E y el punto A del plano inferior. Tal regla debe permanecer conectando ambos puntos mientras asciende o desciende la parábola sobre la línea BK, sobre la que se ha aplicado el eje. Por este medio, la intersección de la parábola y de esta regla, que se dará en el punto C, describirá una línea curva como ACN que es de la que hemos de servirnos para la construcción del problema propuesto. Descrita de este modo, si se toma el punto L en la línea BK hacia el lado en que se encuentra el vértice de la parábola, y si

A-T, VI, 478

$$BL = DE = \frac{2\sqrt{u}}{pn};$$

seguidamente, si desde el punto L y hacia B se toma en la misma línea BK, la línea



como en el punto S; si desde el punto T en que termina, se toma hacia K la línea $TV = BL$ y, seguidamente, habiendo trazado la línea SV, se traza otra que sea paralela por el punto A, como AC; y si se traza otra por el punto S que sea paralela a BK, tal como la línea SC; en tal caso, el punto C, en el que se produce la intersección de AC y SC, tendremos uno de los puntos de la línea buscada. Podemos, siguiendo el mismo procedimiento, hallar cuantos deseemos.

La demostración de todo esto es bastante fácil. Pues aplicando la regla AE y la parábola FD de tal modo que pasen por el punto C, lo cual puede ser ya que el punto C se encuentra en la línea curva ACN que es descrita por la intersección de la parábola y la regla. Si $CG = y$, entonces tendremos que

$$GD = \frac{y^2}{n},$$

puesto que el lado recto, esto es n, es a CG como CG es a GD. Por tanto,

$$D\dot{E} = \frac{2\sqrt{u}}{pn},$$

y restando DE de GD tendremos que

$$GE = \frac{y^2}{n} - \frac{2\sqrt{u}}{pn}.$$

Seguidamente y puesto que AB es a BE como CG es a GE, y puesto que $AB = \frac{1}{2}p$, tenemos que

$$BE = \frac{py}{2n} - \frac{\sqrt{u}}{ny}.$$

De igual modo, suponiendo que el punto C de la curva ha sido hallado mediante la intersección de las líneas rectas SC, paralela a BK, y AC, paralela a SV; sea, por otra parte, que $SB = CG = y$, y $BK = n$, el lado recto de la parábola. Entonces tendremos que

$$BT = \frac{y^2}{n}$$

puesto que KB es a BS como BS es a BT; pero puesto que

$$TV = BL = \frac{2\sqrt{u}}{pn},$$

tenemos que

$$BV = \frac{y^2}{n} - \frac{2\sqrt{u}}{pn}.$$

Y como SB es a BV, así AB es a BE que, en consecuencia, es igual a

$$\frac{py}{2n} - \frac{\sqrt{u}}{ny}.$$

En consecuencia, obtenemos el mismo resultado que en el caso anterior, lo cual nos muestra que se trata de una misma curva la que es descrita de estas dos formas. A-T. VI. 481

Después de esto, puesto que BL y DE son iguales, podemos concluir que también lo son DL y BE. En consecuencia, puesto que

$$LH = \frac{t}{2n\sqrt{n}}$$

$$GH = \frac{-y^3 + \frac{1}{2}py^2 + \frac{ty}{2\sqrt{u}} - \sqrt{u}}{ny}.$$

En consecuencia, el cuadrado de GH es igual a

$$\frac{y^6 - py^5 + \left(\frac{1}{4}p^2 - \frac{t}{\sqrt{u}}\right)y^4 + \left(2\sqrt{u} + \frac{pt}{2\sqrt{4}}\right)y^3 + \left(\frac{t^2}{4u} - p\sqrt{u}\right)y^2 - ty + u}{n^2y^2}$$

Si imaginamos que el punto C se encuentra en cualquier otro lugar de esta línea curva, tal como hacia N o hacia Q, tendremos siempre que el cuadrado de la línea recta situada entre el punto H y aquél en que indica la perpendicular trazada por C sobre BH, puede ser expresada en estos mismos términos y con los mismos signos «+» y «−».

Así, siendo

$$IH = \frac{m}{n^2} \quad \text{y} \quad LH = \frac{t}{2n\sqrt{u}},$$

tendremos que

$$IL = \sqrt{\frac{m^2}{n^4} + \frac{t^2}{4n^2u}},$$

puesto que el ángulo IHL es recto; por otra parte, puesto que

$$LP = \sqrt{\frac{s}{n^2} + \frac{p\sqrt{n}}{n^2}}$$

puede establecerse

$$IC = IP = \sqrt{\frac{m^2}{n^4} + \frac{t^2}{4n^2u} - \frac{s}{n^2} - \frac{p\sqrt{u}}{n^2}}$$

puesto que IPL es un ángulo recto.

Trazando CM perpendicular a IH, tendremos que IM es igual a $HI - HM$ y, por tanto, $IM = HI - CG$. Aplicando los valores establecidos tendríamos que

$$IM = \frac{m}{n^2} - y.$$

Por ello el cuadrado de IM sería igual a

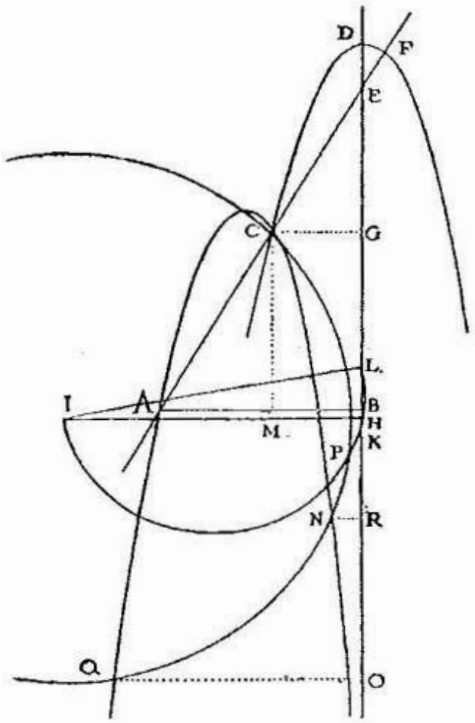
$$\frac{m^2}{n^4} - \frac{2my}{n^2} + y^2.$$

y puesto que

$$DL=\frac{py}{2n}-\frac{\sqrt{u}}{ny},$$

y como tenemos que $DH=LH+DL$, tendremos que

$$DH=\frac{t}{2n\sqrt{u}}+\frac{py}{2n}-\frac{\sqrt{u}}{ny}.$$



Así mismo, puesto que

$$GD=\frac{y^2}{n},$$

tendremos que $GH=DH-GD$ y sustituyendo valores

$$GH=\frac{py}{2n}-\frac{\sqrt{u}}{ny}+\frac{t}{2n\sqrt{u}}-\frac{y^2}{n},$$

A-T. VI, 482 por lo cual podemos establecer que

Si sustraemos tal cantidad del cuadrado de IC, tendremos el cuadrado de CM, o

$$\frac{t^2}{4n^2u}-\frac{s^2}{n^2}-\frac{p\sqrt{u}}{n^2}+\frac{2my}{n^2}-y^2,$$

el cual es igual al cuadrado de GH, previamente A-T, VI, 483 hallado.

O bien, dividiendo esta expresión, como en el otro caso, por n^2y^2, tenemos

$$\frac{-n^2y^4+2my^3-p\sqrt{u}\,y^2-sy^2+\frac{t^2}{4u}y^2}{n^2y^2}.$$

Seguidamente, estableciendo que

$$\frac{t}{\sqrt{u}}y^4+qy^4-\frac{1}{4}p^2y^4$$

para n^2y^4, y

$$ry^3+2\sqrt{u}y^3+\frac{pt}{2\sqrt{u}}y^3$$

para $2my^3$, y multiplicando ambos miembros por n^2y^2, tenemos que

$$y^6-py^5+\left(\frac{1}{4}p^2-\frac{t}{\sqrt{u}}\right)y^4+\left(2\sqrt{u}+\frac{pt}{2\sqrt{u}}\right)y^3+$$
$$+\left(\frac{t^2}{4u}-p\sqrt{u}\right)y^2-ty+u$$

es igual a

$$\left(\frac{1}{4}p^2-q-\frac{t}{\sqrt{u}}\right)y^4+\left(r+2\sqrt{u}+\frac{pt}{2\sqrt{u}}\right)y^3+$$
$$+\left(\frac{t^2}{4u}-s-p\sqrt{u}\right)y^2;$$

es decir, tenemos que

$$y^6-py^5+qy^4-ry^3+sy^2-ty+u=0.$$

De lo que resulta que las líneas CG, NR, QO y otras semejantes son las raíces de esta ecuación, lo cual constituía el objeto e interés de nuestra demostración.

Así, pues, si se desean calcular cuatro medias propor-

cionales entre las líneas a y b, habiendo establecido x para la primera, la ecuación es

$$x^5 - a^4 b = 0$$

o

$$x^6 - a^4 bx = 0.$$

A-T, VI, 484 Siendo $y - a = x$, tendríamos

$$y^6 - 6ay^5 + 15a^2 y^4 - 20a^3 y^3 + 15a^4 y^2 - \\ -(6a^5 + a^4 b)y + a^6 + a^5 b = 0.$$

Por esto debemos tomar $AB = 3a$, y BK, el lado recto de la parábola debe ser

$$\sqrt{\frac{6a^3 + a^2 b}{\sqrt{a^2 + ab}} + 6a^2}$$

al cual he llamado n. Así mismo, DE o BL será:

$$\frac{2a}{3n}\sqrt{a^2 + ab}.$$

Describiendo la curva ACN de acuerdo con estos tres valores, tendremos que

$$LH = \frac{6a^3 + a^2 b}{2n\sqrt{a^2 + ab}}$$

y

$$HI = \frac{10a^3}{n^2} + \frac{a^2}{n^2}\sqrt{a^2 + ab} + \frac{18a^4 + 3a^3 b}{2n^2\sqrt{a^2 + ab}},$$

y finalmente que

$$LP = \sqrt{\frac{15a^4 + 6a^3\sqrt{a^2 + ab}}{n^2}}.$$

Así el círculo que con centro en I pase por el punto P, hallado de esta forma, cortará a la curva en los puntos C y N. Trazando perpendiculares desde estos puntos, NR y CG, tendremos que si restamos la menor, NR, de la mayor CG, obtendremos x, que será la primera de las cuatro medias proporcionales buscadas.

De igual modo y con facilidad puede dividirse un ángulo en cinco partes iguales, e inscribir una figura dentro de un círculo que posea 11 ó 13 lados iguales, o bien tratar otra infinidad de problemas.

Sin embargo, debe hacerse notar que en diversos casos, puede suceder que el círculo corte tan oblicuamente a la parábola de segundo género que el punto de intersección sea difícilmente determinable; en consecuencia, tal construcción no sería cómoda desde el punto de vista práctico. La dificultad puede ser fácil- A-T, VI, 485
mente superable estableciendo otras reglas semejantes a ésta, lo cual puede realizarse de mil modos diferentes.

Pero no es mi propósito confeccionar un gran tratado, sino el analizar muchas cuestiones en pocas palabras. Así espero que seré juzgado si se piensa que he reducido todos los problemas de un mismo género a un mismo tipo de construcción y, a la vez, he facilitado el procedimiento para reducirlos a una infinidad de otras; de este modo cada uno de ellos puede solucionarse de infinidad de formas; además, habiendo construido todos los que son planos, cortando una recta con un círculo, y todos los que son sólidos, cortando con un círculo una parábola, y, en fin, todos aquellos que son de un grado superior cortando un círculo por una curva de un grado superior a la parábola. De esta forma no es preciso sino seguir el mismo procedimiento para construir todos aquéllos cuya complejidad sea creciente. Pues en materia de progresiones matemáticas cuando se tienen los dos o tres primeros términos no es difícil hallar los otros. Espero que nuestros descendientes sabrán reconocerme no sólo las cuestiones que he explicado, sino también aquellas que he omitido voluntariamente con el fin de permitirles disfrutar del placer que produce su descubrimiento.

FIN

INDICE

DE LAS MATERIAS QUE CONTIENE LA GEOMETRIA

Libro Primero

SOBRE LOS PROBLEMAS QUE PUEDEN CONSTRUIRSE EMPLEANDO SOLAMENTE CIRCULOS Y LINEAS RECTAS

Cómo el cálculo de la aritmética se relaciona con las operaciones geométricas (369).
Cómo se realizan en Geometría la multiplicación, la división, la extracción de la raíz cuadrada (369-370).
Cómo se usan las letras o cifras en Geometría (371).
Cómo se accede a las ecuaciones que sirven para resolver los problemas (372-373).
Cuáles son los problemas planos y cómo se resuelven (374-375).
Ejemplo tomado de Pappus (377).
Respuesta al problema de Pappus (380-381).
Cómo deben establecerse los términos para acceder a la ecuación de este ejemplo (382-383).
Cómo se decide que este problema es plano cuando no se ha propuesto para más de cinco líneas (385-386).

Libro Segundo

SOBRE LA NATURALEZA DE LAS LINEAS CURVAS

Cuáles son las líneas curvas que pueden admitirse en Geometría (388).
Sobre la clasificación de todas las curvas en géneros y sobre la relación de sus puntos con los de las rectas (392).
Prosigue la explicación del problema de Pappus expuesta en el libro anterior (396).
Solución de este problema cuando se propone respecto de tres o cuatro líneas (397).
Demostración de lo que ha sido explicado (404).
Cuáles son los lugares planos y sólidos; forma de hallarlos (406).
Cuál es la primera y más simple de todas las líneas curvas utilizable en el problema de los antiguos cuando está propuesto para cinco líneas (407).
Sobre las líneas que se trazan hallando varios de sus puntos y que pueden ser admitidas en Geometría (411).
Cuáles son las que pueden trazarse utilizando una cuerda y que también pueden ser admitidas (412).
Que para hallar todas las propiedades de las líneas curvas, basta con conocer la relación que guardan sus puntos con

Cómo se puede sacar el segundo término de una ecuación (449).
Sobre el procedimiento para lograr que las raíces falsas pasen a ser verdaderas sin que suceda lo inverso (450).
Procedimiento para completar los diversos lugares de una ecuación (451).
Cómo se pueden multiplicar o dividir las raíces de una ecuación (452).
Sobre la reducción de los quebrados en una ecuación (452).
Procedimiento para establecer una igualdad entre la cantidad conocida de uno de los términos de la ecuación y otra cantidad dada (453).
Las raíces, verdaderas o falsas, pueden ser reales o imaginarias (453-454).
La reducción de las ecuaciones cúbicas cuando el problema es plano (454).
Sobre el modo de dividir una ecuación por un binomio que contiene su raíz (455).
Qué problemas son sólidos cuando la ecuación es cúbica (456-457).
Sobre la reducción de las ecuaciones con cuatro dimensiones cuando el problema es plano y cuáles son aquellos que son sólidos (457).
Ejemplo del uso de estas reducciones (461-462).
Regla general para reducir las ecuaciones que pasan el cuadrado del cuadrado (463-464).
Forma general de reducir todos los problemas sólidos en una ecuación de tercer o cuarto grado (464).
Sobre la determinación de las medias proporcionales (469).
Modo de dividir un ángulo en tres partes (470).
Sobre la reducción de los problemas sólidos a estas dos construcciones (471).
Forma de expresar el valor de todas las ecuaciones cúbicas y de todas las bicuadráticas (474).
Razones por las que los problemas sólidos no pueden ser construidos sin las secciones cónicas y aquellos cuya complejidad es mayor sin otras de complejidad creciente (475).
Procedimiento general para construir todos los problemas reducidos a una ecuación no superior al grado sexto (476).
Sobre la invención de las cuatro medias proporcionales (483).

los de las rectas, y la forma de trazar otras líneas formando ángulos rectos (412).
Procedimiento general para hallar líneas rectas que corten las curvas dadas o sus contingentes formando ángulos rectos (413).
Ejemplo de esta operación en una elipse o en una parábola de segundo género (414-415).
Ejemplos respecto de un óvalo de segundo género (415-416).
Ejemplo de la construcción de este problema en la concoide (423).
Explicación de cuatro nuevos géneros de óvalos de utilidad en la óptica (424).
Sus propiedades en relación con las reflexiones y refracciones (429).
Su demostración (431).
Cómo se puede construir una lente tan cóncava o convexa en una de sus caras como se desee, y que sea capaz de reunir en un punto dado los rayos procedentes de un punto dado (435).
Cómo se puede elaborar una lente que produzca el mismo efecto, que la precedente, guardando la convexidad de una de sus caras una proporción dada con la otra (437).
Aplicación de lo expuesto sobre las líneas curvas trazadas sobre una superficie plana a las que se describen en un espacio que tiene tres dimensiones o bien sobre una superficie curva (440).

Libro Tercero

SOBRE LA CONSTRUCCION DE PROBLEMAS SOLIDOS O MAS QUE SOLIDOS

Cuáles son las líneas curvas de que podemos servirnos en la construcción de cada problema (441-442).
Ejemplo relacionado con la determinación de diversas medias proporcionales (442-443).
Sobre la naturaleza de las ecuaciones (444).
Sobre el número de raíces en cada ecuación (444-445).
Cuáles son las raíces falsas (445).
Sobre la reducción del número de dimensiones de una ecuación cuando se conoce alguna de sus raíces (445).
Cómo puede examinarse si una cantidad dada es el valor de una raíz (445).
Cuántas raíces verdaderas pueden darse en cada ecuación (446).
Cómo en una ecuación pueden convertirse las raíces falsas en verdaderas y a la inversa (446).
Cómo pueden aumentarse o disminuirse las raíces de una ecuación sin conocerlas (447).
Aumentando las raíces verdaderas disminuyen las falsas, y a la inversa (448).

NOTAS A LA GEOMETRIA

[1] Esta observación no aparece generalmente recogida en ediciones actuales, v.gr. en las de Olscamp o la de Smith. No obstante, estimamos que es importante, pues, evidencia que Descartes era consciente de las dificultades que podía plantear su lectura. Tales dificultades se veían aumentadas al no facilitar los pasos intermedios de sus demostraciones bien por evitar «pérdida de tiempo», bien por evitar críticas (A-T, II, 30, 18/22). Ciertamente la labor de reproducir las razones de todos y cada uno de los pasos de este tratado supondría multiplicar por cuatro su volumen. Tal fue, por ejemplo, la tarea realizada por Schooten. En su edición de la Geometría reproduce al margen diversas letras que vienen a ser otras tantas notas. En los correspondientes lugares realiza el comentario al texto de Descartes y a la vez introduce una serie de problemas con sus resoluciones y planteamientos. Tales comentarios se realizan desde la página 107 hasta la página 520. Ejemplares de esta edición se encuentran en la Biblioteca Nacional de Madrid y en la Biblioteca Universitaria de Valencia, si bien el estado de conservación de este último ejemplar es lastimoso por cuanto no existe servicio de restauración en la referida biblioteca. La

propia estructura de la edición Schooten hace inviable su inclusión dentro del presente volumen.

Para el lector de nuestros días que se haya formado dentro del tradicional olvido de la Geometría siguen siendo válidas estas apreciaciones y las que Descartes comunicó en carta a Pemplius: sus lectores deben ser «valde laboriosos, ingeniosos et attentos» (A-T, I, 411, 12/20). Siempre y cuando no sea necesario un gran desarrollo algebraico, trataremos de facilitar en notas el proceso deductivo que sigue Descartes.

[2] El lector de *La Geometría* puede recordar algunas reglas (R, XVIII) en las que se refiere a alguna de estas cuestiones. Sin embargo, es claro que no estaba en posesión del tratamiento de la geometría analítica, pues, por ejemplo, en la mencionada R, XVIII aún figura mediante la superficie de un triángulo el producto de dos factores. Sobre la progresión de *La Geometría* en relación con algunas partes de *Las Reglas para la dirección de la mente*, véase el estudio de P. Bourtroux (*L'imagination et les mathématiques selon Descartes*. París, 1900, pp. 43 y ss.).

[3] En éste como en la generalidad de los casos se razona a partir de la semejanza de triángulos, estableciendo las proporciones correspondientes. Así, trazada tal paralela, cabe afirmar:

$$\frac{BA}{BC} = \frac{BD}{BE}.$$

Considerando que BA es la unidad, tenemos que $BE = BC \cdot BD$.

[4] De igual modo tendríamos:

$$\frac{BC}{BE} = \frac{BA}{BD}; \quad \frac{BC}{BA} = \frac{BE}{BD}.$$

Por tanto,

$$BC = \frac{BE}{BD}.$$

[5] El triángulo GIH es semejante a FGI. Por tanto,

$$\frac{GI}{GH} = \frac{GF}{GI}.$$

Dado que $GI \cdot GI = GF \cdot GH$, y que $GF = 1$, tenemos que $GI \cdot GI = GH$. En consecuencia, $GI = \sqrt{GH}$.

[6] Las anotaciones y símbolos fueron objeto de una importante simplificación por parte de Descartes. Como se deducirá de la lectura de la geometria adopta x, y, z, para las cantidades desconocidas; a, b, c, para las cantidades conocidas.

[7] Se trata de raíz cúbica. El equivalente en notación de nuestros días sería $\sqrt[3]{a^3 - b^3 + ab^2}$.

Al igual que Descartes en la traducción respetaremos expresiones del tipo *aa*, *bb*; en otros casos a^2, b^2. Optamos por esta última forma cuando son más complejas las expresiones.

[8] Las magnitudes a, $a+b$, $a \cdot b$, etc., pueden representarse mediante una longitud. De igual modo acontece con expresiones del tipo a^2, a^3, a^n. Para Descartes, al generalizar la noción de dimensión, al prescindir del obstáculo que para los ántiguos representaba («nada hay que no esté comprendido bajo más de tres dimensiones...»), es posible superar su tratamiento.

[9] Descartes no utiliza el signo «=». Por razones de adaptación a la asignatura actual no respetaremos tal signo «∞».

[10] Entiéndase «incógnitas».

[11] Las dos condiciones fundamentales del método analítico, considerar efectuada la solución y considerar las relaciones entre las lineas, han sido dadas. La segunda hace posible el establecimiento de ecuaciones. El estadio analítico del método se complementa con la síntesis algebraica. La construcción sintética se funda en la analítica.

El análisis y el álgebra están ordenados al conocimiento de figuras. En la geometría analítica, pues, deben traducirse los datos geométricos de modo que sean tratables por medio del cálculo algebraico; asimismo debe concluirse el problema de álgebra y, finalmente, los resultados obtenidos mediante el cálculo algebraico deben ser traducidos al lenguaje geométrico.

[12] Como se deduce del libro II de *La Geometría*, existe relación entre los problemas de construcción y los de clasificación. En el presente lugar relaciona claramente un tipo determinado de expresiones algebraicas con los instrumentos que permiten trazar determinadas construcciones: línea (regla) y círculo (compás).

[13] Va a iniciar el estudio de los diferentes tipos de ecuaciones de segundo grado. Estudia tres casos:

1.° Aquél en el que los coeficientes tienen valor positivo (son magnitudes positivas).

2.°. Aquel en el que uno de los coeficientes tiene valor negativo y el segundo positivo.

3.° Aquél en el que el segundo de los coeficientes tiene valor negativo y el primero positivo.

[14] De acuerdo con la figura trazada se establece inicialmente que $ON = NL$ (ambos son radios del círculo con centro en N).

Indica Descartes que el valor de $Z = OM$; en virtud del teorema de Pitágoras, tenemos que

$$OM = \frac{1}{2}a + \sqrt{\frac{1}{4}a^2 + b^2}.$$

Por tanto, OM^2 será igual a

$$\frac{1}{4}a^2 + \frac{1}{4}a^2 + b^2 + a\sqrt{\frac{1}{4}a^2 + b^2}.$$

Simplificando el valor de az será igual a:

$$\frac{1}{2}a^2 + a\sqrt{\frac{1}{4}a^2 + b^2}.$$

Por tanto,

$$az+b^2=\frac{1}{2}a^2+a\sqrt{\frac{1}{4}a^2+b^2}.$$

[15] Es decir, hipotenusa.

[16] El desarrollo total sería el siguiente:

$$z^2=az-b^2.$$

$$z^2-az+b^2=0.$$

Si, pues,

$$\left(z-\frac{1}{2}a\right)^2=z^2+\frac{a^2}{4}-az,$$

entonces podemos obtener que

$$\left(z-\frac{1}{2}a\right)^2-\frac{a^2}{4}+b^2=0.$$

Y,

$$\left(z-\frac{1}{2}a\right)^2=\frac{a^2}{4}-b^2$$

Cuando $b^2>\frac{a^2}{4}$, entonces carece de solución. En consecuencia, puede estableserse que

$$z-\frac{1}{2}a=\pm\sqrt{\frac{a^2}{4}-b^2};$$

por tanto:

$$z=\frac{1}{2}a+\sqrt{\frac{a^2}{4}-b^2}$$

$$z=\frac{1}{2}a-\sqrt{\frac{a^2}{4}-b^2}.$$

Por lo que se refiere a la construcción trazada en el correspondiente gráfico, es claro que $NL=\frac{1}{2}a$, tal como se indica. Por tanto, para hallar el valor de $z=MQ$, sólo tendríamos que trazar por N una perpendicular sobre MR que, sería paralela a LM y alcanzaría a MR en el punto O. En consecuencia, tendríamos que $z=MO-QO$.

Por otra parte, si unimos los puntos Q y N formamos el triángulo rectángulo QNO. Tendríamos que como NQ es igual a NL, pues ambos son radios, entonces $NQ=\frac{1}{2}a$. Por el teorema de Pitágoras, tendríamos que

$$\left(\frac{1}{2}a\right)^2=b^2+m^2.$$

Por ello, podemos establecer que

$$\frac{1}{4}a^2-b^2=m^2.$$

Y, por tanto, puede establecerse que

$$\sqrt{\frac{1}{4}a^2-b^2}=m.$$

Como $MQ=z=(MO-QO)$, y tenemos que $MO=\frac{1}{2}a$, entonces, sustituyendo valores, tenemos que

$$z=\frac{1}{2}a-\sqrt{\frac{1}{4}a^2-b^2}.$$

Conocido esto tenemos que OR, en virtud del teorema de Pitágoras, es igual a

$$\sqrt{\frac{1}{4}a^2-b^2}.$$

Como $MR=MQ+QO+OR$, con sólo sustituir los valores tendríamos que

$$MR=\sqrt{\frac{1}{4}a^2-b^2}+\sqrt{\frac{1}{4}a^2-b^2}+\frac{1}{2}a-\sqrt{\frac{1}{4}a^2-b^2};$$

Por tanto,

$$MR=\frac{1}{2}a+\sqrt{\frac{1}{4}a^2-b^2}.$$

[17] Pappus vivió hacia el 300 d.C., destacando por la gran compilación que realizó y que fue dada a conocer por Commandinus a los escolásticos. En carta a Golius (Amsterdam, enero de 1632) Descartes ya se refiere al problema de Pappus y, a su vez, facilita la razón por la que excluye de la geometría algunas líneas. Esta cuestión la analizaría en el libro II (A-T, I, 232).

[18] Se inicia la exposición de los datos del problema.

[19] Elipse, parábola e hipérbola.

[20] Se refiere a que nada está comprendido bajo más de tres dimensiones.

[21] Segmentos de líneas rectas.

[22] Esto es, el producto de estos dos segmentos.

[23] Esto es, el producto de los tres segmentos.

[24] La geometría antigua solamente permitía explicar el caso en que el problema se refería a tres o cuatro líneas. La geometría cartesiana al adoptar el método analítico de las proporciones hace posible la generalización de tal problema.

Tal método es al que se refería cuando indicaba a Mersenne que en su Geometría «...estimaba dar una demostración de que

mi método es mejor que el ordinario» (A-T, I, 340, 10/11). Según hemos indicado en la introducción a esta edición tal método equivale a un procedimiento de análisis cuya claridad, simplicidad y alcance explicativo-resolutivo le hacen sobresalir sobre el procedimiento seguido desde la antigüedad (establecer ecuaciones que relacionen las áreas dadas con las buscadas o deseadas).

[25] Evitar la confusión y favorecer la claridad es una de las cualidades que Descartes destaca como propias de su método. Desde la comprensión de tal procedimiento debe valorarse la crítica que Descartes dirige a la geometría de los antiguos, cuya confusión y oscuridad le llevaron a pensar en la necesidad de buscar «...otro método que estuviese libre de tales defectos» (A-T, VI, 18, 1/8). El método cuya práctica ilustrará a partir de este momento supone que las cantidades geométricas constituyen ilustraciones de las algebraicas que forman las ecuaciones.

[26] Se trata de los ejes.

[27] Para comprender el razonamiento cuyo inicio se está dando debe considerarse que:

ARB, dado por suposición;

CBA, dado por suposición;

RBA, conocido por ser suplementario;

$RAB = 180 - (ABR + ARB)$.

Por otra parte, tenemos que

$$\frac{AB}{BR} = \frac{z}{b},$$

y de acuerdo con el supuesto establecido que

$$\frac{x}{BR} = \frac{z}{b}.$$

Aplicando que el producto de medios es igual al producto de extremos, tenemos que $xb = BR \cdot z$. En consecuencia, se podrá establecer que

$$BR = \frac{bx}{z}.$$

[28]
$$\frac{x}{BR} = \frac{z}{b}; \quad BR = \frac{bx}{z}.$$

[29] Considerando que $CR = BR + BC$, tenemos que

$$CR = y + \frac{bx}{z}.$$

[30] Esto es: los ángulos del triángulo DRC son dos, y por tanto,

$$\frac{CR}{CD} = \frac{z}{c}.$$

[31] A partir de que

$$\frac{CR}{CD} = \frac{z}{c},$$

tenemos que

$$CD = \frac{CR \cdot c}{z} = \frac{c}{z}\left(y + \frac{bx}{z}\right) = \frac{c}{z}\,y + \frac{bc}{z^2}\,x.$$

[32] Es claro que si $EA = k$ y A es un punto de la línea EB, entonces $BE = k + x$.

[33] Puesto que los ángulos del triángulo ESB son dados, puede establecerse que

$$\frac{BE}{BS} = \frac{z}{d}$$

o también que

$$\frac{k+x}{BS} = \frac{z}{d}.$$

En consecuencia, operando obtenemos que

$$BS = \frac{(k+x)d}{z}$$

y puesto que $CS = BS + BC$, sustituyendo tenemos que

$$CS = \frac{(k+x)d}{z} + y.$$

De aquí tenemos que

$$CS = \frac{zy + dk + dx}{z}.$$

[34]
$$\frac{CS}{CF} = \frac{z}{e}.$$

[35] A partir de la proporción anterior tenemos que

$$CF = \frac{CSe}{z}.$$

Por ello, tenemos que

$$CF = \frac{e}{z}\left(\frac{zy + dk + dx}{z^2}\right).$$

A partir de aquí se obtiene que

$$CF = \frac{e}{z}\,y + \frac{de}{z^2} \times \frac{dek}{z^2}.$$

[36]
$$\frac{BG}{BT} = \frac{z}{f}; \quad BT = \frac{BG \cdot f}{z}.$$

Considerando el valor de BG, tenemos que

$$BT = \frac{f}{z}(l - x).$$

Como $CT = BC + BT$, tendremos que

$$CT = y + \frac{fl}{z} - \frac{fx}{z}.$$

Pero como los ángulos del triángulo TCH son dados, también es posible establecer que

$$\frac{CT}{CH} = \frac{z}{g}.$$

En consecuencia,

$$CH = \frac{CT \cdot g}{z};$$

sustituyendo el valor de CT, tenemos que

$$CH = \frac{g}{z}\left(y + \frac{fl}{z} - \frac{fx}{z}\right),$$

de donde tenemos que

$$CH = \frac{g}{z}y - \frac{fg}{z^2}x + \frac{fgl}{z^2}$$

[37] Segundo grado.

[38] No superior al cuarto grado.

[39] No superior al sexto grado.

[40] Por contraposición con éstas, Descartes habla de las líneas complejas o compuestas en la medida en que su trazado mecánico se aparta de la regla y el compás. Más adelante explicará el instrumento que facilita el trazado de algunas curvas.

[41] Tal es la condición para que sean introducidas en la geometría las curvas.

[42] Véase su tratamiento en J. Vuillemin, *ob. cit.*, pp. 146 y ss. Para Descartes, aunque los antiguos hubiesen estudiado estas líneas, no deberán incluirse por cuanto su descripción requiere dos movimientos que carecen de relación mensurable y, en consecuencia, no serían representables mediante funciones algebraicas. Bourbaki reconoce a Descartes haber establecido la distinción entre funciones algebraicas y funciones trascendentes, paralela a la establecida entre curvas «geométricas» y las «mecánicas» (*Elementos de historia de las matemáticas*. Madrid, 1976, p. 106).

[43] Las construcciones que permite este instrumento facilitan la obtención de las ecuaciones de diversas curvas. En el caso de la primera (AD) su ecuación sería $x^4 = a^2(x^2 + y^2)$. Para obtener tal ecuación basta con considerar las líneas que describe el punto $B(AB)$ y el punto $D(AD)$ al abrirse el ángulo XYZ.

Entonces tomamos que $YA = YB = a$; que $YC = x$, que $CD = y$ y que $YD = z$. Es claro que los triángulos YBC y el YCD son semejantes y que, por tanto, puede establecerse que

$$\frac{YF}{YC} = \frac{YC}{YB}$$

es decir, que

$$\frac{z}{x} = \frac{x}{a}.$$

Operando algebraicamente obtenemos que

$$z = \frac{x^2}{a}.$$

El teorema de Pitágoras permite afirmar que $YD^2 = YC^2 + CD^2$. Sustituyendo, de acuerdo con las igualdades establecidas inicialmente, tenemos que $z^2 = x^2 + y^2$. Dado el valor de z, tenemos que

$$\left(\frac{x^2}{a}\right)^2 = x^2 + y^2.$$

La ecuación, pues, será $x^4 = a^2(x^2 + y^2)$.

En segundo lugar puede determinarse analíticamente la curva AF. Para ello $YA = YB = a$, $YE = x$, $EF = y$, $YF = z$. De nuevo la semejanza de triángulos permite establecer que

$$\frac{YE}{YE} = \frac{YE}{YD}.$$

o según las igualdades iniciales que

$$\frac{z}{x} = \frac{x}{YD}.$$

El valor de YD tendremos que es igual a

$$\frac{x^2}{z}.$$

A su vez los triángulos YDE e YCD son semejantes y, por tanto, tenemos que

$$\frac{YE}{YD} = \frac{YD}{YC};$$

así pues,

$$YC = \frac{YD^2}{YE} = \frac{x^4}{z^2 x} = \frac{x^3}{z^2},$$

También podemos establecer que

$$\frac{YD}{YC} = \frac{YC}{A}$$

y que

$$YC^2 = \left(\frac{x^3}{z^2}\right)^2 = \frac{x^2}{z}\, a.$$

Obtenemos de este modo que

$$z^3 = \frac{x^4}{a}$$

y, en consecuencia, que

$$z = \sqrt[3]{\frac{x^4}{a}}$$

De nuevo en el triángulo YFE tenemos que es rectángulo y en base al teorema de Pitágoras, tenemos que $z^2 = y^2 + x^2$. La ecuación de la curva AF será, pues,

$$\frac{x^8}{a^2} = xx + y^2,$$

lo cual nos lleva a $x^8 = a^2(x^2 + y^2)^3$.

De igual modo podría calcularse la ecuación de AH. Se vería que su complejidad es creciente respecto de la segunda, como ésta lo era respecto de la primera. La mayor composición o complejidad no es razón para excluirlas de la Geometría como tampoco lo es el medio que se ha utilizado para su construcción.

[44] En consecuencia, no admite la oposición trazada por los griegos, reteniendo como válida y legítima la construcción mecánica, cuya utilidad tratará de mostrar.

[45] De este modo hace explícito el criterio con el que complementa la conclusión del párrafo anterior.

[46] Esto es, al producto.

[47] En la traducción alternamos la expresión del texto con las utilizadas por Descartes. En este caso podría decirse «cuando las dos magnitudes o una de ellas alcanza la tercera o cuarta dimensión».

[48] Al igual que en la nota anterior podríamos indicar siguiendo la terminología de Descartes, que «contiene la quinta o sexta dimensión».

[48bis] Según Descartes, que se funda para establecer relaciones en el teorema relacionado con la semejanza de triángulos, tendríamos que

$$\frac{NL}{KL} = \frac{CB}{KB};$$

de acuerdo con la signatura adoptada por Descartes, tendríamos que

$$\frac{e}{b} = \frac{y}{KB};$$

de aquí tendríamos que $c \cdot KB = b \cdot y$; de donde

$$KB = \frac{yb}{c}.$$

Estas simples operaciones han de suponerse en la mayor parte de los casos. Así

$$BL = BK - KL; \quad BL = \frac{by}{c} - b.$$

En el caso de

$$AL = x + \frac{by}{c} - b.$$

Descartes afirma, finalmente, que $GA \cdot BL = AL \cdot CB$; efectuando la oportuna sustitución se obtiene que

$$\frac{aby}{c} - ba = yx + \frac{by^2}{c} - by.$$

Tal igualdad puede establecerse del modo siguiente:

$$\frac{by^2}{c} + yx - by - \frac{aby}{c} + ba = 0;$$

en consecuencia, puede establecerse que

$$by^2 + ycx - bcy - aby + bac = 0;$$

expresión que permite establecer que

$$y^2 + \frac{cyx}{b} - cy - ay + ca = 0;$$

por ello, puede afirmarse que

$$y^2 = cy - \frac{cx}{b}\, y + ay - ac.$$

[49] Equivale a afirmar que aquéllas cuyas ecuaciones son de sexto grado en el mismo género que aquéllas cuya ecuación son de quinto. De modo semejante en otros casos. Como indica Descartes la razón reside en que existe una regla en virtud de la cual es posible efectuar tal reducción. Es, en virtud de la mayor o menor potencia de la ecuación como se clasifican las curvas.

[49bis] Un estudio de la Concorde, atribuida a Nicomedes, en Vuillemin, ob. cit., p. 148 ss.

[50] Esto es, en lenguaje de Descartes no alcanza sino al cuadrado.

[50bis] Ver en la edc. de Schooten, p. 206.

[51] Obtenido algebraicamente el valor de BC, su longitud será establecida geométricamente.

[52] Esto es, $m^2 + ox$.

[53] Esto es, $+m^2 - ox$.

[54] Esto es, $-m^2 + ox$.

[55] El procedimiento de construcción no permite determinar la naturaleza de una curva, pero no debe ser rechazado. El lector de *La Dióptrica* (D. X) ha podido comprobar la utilidad de tal tipo de construcciones.

[56] En el caso anterior se trata de las normales y en éste de las tangentes.

[57] $PM = PA \cdot MA$; $PM = v - y$.

[58] Por ser PMC un triángulo rectángulo, tenemos que $PC^2 = MC^2 + MP^2$. De acuerdo con las igualdades establecidas: $s^2 = x^2 + (v-y)^2$. Operando tenemos que $s^2 = x^2 + v^2 - 2vy + y^2$. De igual modo tendríamos que $x = \sqrt{s^2 - v^2 + 2vy - y^2}$.

[59] Pero como $(v-y)^2 = s^2 - x^2$, obtenemos que $y = v + \sqrt{s^2 - x^2}$. En consecuencia, sustituyendo estos valores de x e y en la ecuación podrá obtenerse la normal. Una vez que se ha determinado el punto C y que son conocidos x e y puede obtenerse v y s.

[60] Véase en A-T, VI, 393 ss.

[61] La ecuación de la normal, determinado C, permitirá que mediante x e y se pueda calcular el valor tanto de V como de s. Si P no se encuentra en el lugar requerido para la ecuación de la normal, entonces el círculo con centro en P y radio PC cortará necesariamente a la línea CE en otro punto.

[62] Sobre las diferencias con Fermat a este respecto, véase en Vuillemin, *ob. cit.*, pp. 59 ss.

[62bis] Aducimos esta expresión equivalente a la dada por Descartes:

$$GP = b\,\frac{-bcd^2 + bcde - bd^2z - ce^2z}{bde + cd^2 + d^2z - e^2z}.$$

[63] El libro posee un esquema claro y progresivo: en la primera parte se refiere a una serie de consideraciones generales sobre las ecuaciones; llama la atención el orden y claridad de su exposición lograda en base a la gradual respuesta que va facilitando de cuestiones diversas, v.gr. ofrecida la definición de ecuación procede a facilitar «algunas reglas» que integran y conforman el alcance operativo de su método; aplicando tales reglas podemos conocer el número de raíces de una ecuación, distinguir las raíces verdaderas de las falsas, discernir si una cantidad dada constituye el valor de una raíz, determinar cuántas raíces verdaderas pueden darse en una ecuación, etc. La concreción de estas reglas se contrapone a la vaga generalidad de los preceptos formulados en la segunda parte de *El Discurso del Método*. Por ello su comprensión es más problemática, como advierte al inicio de *La Geometría*, pero es la que permitirá decidir sobre su método a la vez que dirigirá adecuadamente nuestra razón al operar en este campo de problemas, evitando los errores y alcanzando la mayor simplicidad posible: en la segunda parte trata la correspondencia de las ecuaciones y problemas planos; en la tercera, ya analiza las cuestiones referentes a los sólidos.

[64] La geometría analítica tiene para las raíces negativas, situándolas a la izquierda de la intersección de las coordenadas, una ilustración de la que carecían los antiguos estudiosos del álgebra. Tal aportación del álgebra a la geometría trajo

entre otras esta aportación, adecuadamente valorada en su época. (Véase en Vuillemin, *ob. cit.*, p. 129, n. 6.)

[65] Vuillemin (*ob. cit.*) comenta en la nota VII (pp. 154-159) la resolución de la ecuación de tercer grado en base a los ejemplos, literal y numérico, que Descartes aduce y desarrollará en estas páginas (A-T, 455-456). El desarrollo de estas cuestiones justifica afirmaciones vertidas por Descartes en el libro II al tratar la clasificación de las curvas en géneros. Por otra parte, como indica Vuillemin, «el procedimiento de la división por un binomio no es sino un artificio secundario, cuya razón de ser está fundada en el método de las indeterminadas» (*ob. cit.*, p. 157).

[66] La ecuación la hemos presentado con los signos correspondientes, aunque hemos recogido la traducción del texto. El lector comprenderá que la ecuación era presentada del modo siguiente:
$x^4 \cdot pxx \cdot qx \cdot r = 0$.

[66bis] Véase en A-T, VI, 449.

[67] Descartes inicia la exposición de la construcción que cumple

$$x = \frac{1}{2}a + \sqrt{\frac{1}{2}a^2 + \frac{1}{4}c^2} - \sqrt{\frac{1}{4}c^2 - \frac{1}{2}a^2 + \frac{1}{2}a\sqrt{a^2+c^2}},$$

a la vez que son dados el cuadrado de lado BD y BN. Sería preciso sobre la prolongación del lado opuesto a BD, esto es AC, hallar un punto E tal que uniéndolo con B diera lugar a un segmento al ser cortado por CD que fuese igual a BN, esto es, que $FE = BN$. Utilizando solamente la regla y el compás, podríamos unir N y D; trazando con tal radio (ND) y con centro en D a NG, tendríamos que $ND = DG$. Si trazamos una circunferencia con el diámetro BG, veríamos que cortaría la prolongación del lado AC en el punto E, que sería el punto buscado.

Considérese que el ángulo BEG, es recto y que $DN = DG$; en consecuencia, puesto que DN es la hipotenusa del triángulo DBN, tendríamos que $DN^2 = DG^2 = NB^2 + BD^2$. Por otra parte, dado que ECF y EHG son semejantes podemos establecer que

$$\frac{CF}{HG} = \frac{CE}{EH}$$

y, en consecuencia, que $CF \cdot EH = HG \cdot CE$; por tanto,

$$HG = \frac{CF \cdot EH}{CE}.$$

De igual modo, basándose en la semejanza de los triángulos ECF y BDF tenemos que

$$\frac{CF}{FD} = \frac{CE}{BD}$$

y como $BD=CD=EH$, se establece en base al desarrollo de ambas igualdades que

$$CE=\frac{CF\cdot BD}{FD}=\frac{CF\cdot BD}{HG} \quad \text{y que} \quad HG=FD.$$

Por ello, pueden establecer las siguientes equivalencias:

$$DG^2=(DH+HG)^2=(CE+HG)^2=CE^2+HG^2+2CE\cdot HG=$$

$$CE^2+HG^2+2CE\,\frac{CF\cdot EH}{CE}=CE^2+FD^2+2CF\cdot CO=$$

$$CE^2+(CD-CF)^2+2CF\cdot CD=(CE^2+CF^2)+CD^2=FE^2+CD^2.$$

En base al desarrollo y comparación de estas igualdades se ha obtenido que $FE=BN$. Para establecer la correspondencia analítica de tal construcción si $DF=x$ y $BD=a$, tendríamos:

$$CF=a-x;$$

como

$$\frac{CF}{FE}=\frac{FD}{BF}, \quad \text{o} \quad \frac{a-x}{c}=\frac{x}{BF}.$$

Por tanto,

$$BF=\frac{cx}{a-x}.$$

A su vez como $BF^2=x^2+a^2$ y $x^2+a^2=\dfrac{c^2x^2}{x^2-2ax-a^2}$; si se multiplican ambos miembros por

$x^2-2ax+a^2$

se obtendría

$$x^4-2ax^2+(2a^2-c^2)x^2-2a^3xa^4=0.$$

Por tanto tendríamos que

$$DF=x=\frac{1}{2}a+\sqrt{\frac{1}{4}a^2+\frac{1}{4}c^2}-\sqrt{\frac{1}{4}c^2-\frac{1}{4}a^2+\frac{1}{2}a\sqrt{a^2+c^2}}$$

Descartes varía de incógnita. El análisis lo continúa y comenta Vuillemin, *ob. cit.*, p. 167 ss.

The Project Gutenberg EBook of La géométrie, by René Descartes

This eBook is for the use of anyone anywhere at no cost and with almost no restrictions whatsoever. You may copy it, give it away or re-use it under the terms of the Project Gutenberg License included with this eBook or online at www.gutenberg.org

Title: La géométrie

Author: René Descartes

Editor: A Hermann

Release Date: August 23, 2008 [EBook #26400]

Language: French

Character set encoding: ISO-8859-1

*** START OF THIS PROJECT GUTENBERG EBOOK LA GÉOMÉTRIE ***

Produced by K.F. Greiner, Joshua Hutchinson, Keith Edkins and the Online Distributed Proofreading Team at http://www.pgdp.net (This file was produced from images generously made available by Cornell University Digital Collections)

LA

G É O M É T R I E

DE

RENÉ DESCARTES

NOUVELLE ÉDITION

PARIS

A. HERMANN, LIBRAIRIE SCIENTIFIQUE

8–rue de la Sorbonne–8

MDCCCLXXXVI

AVERTISSEMENT

Peu de livres ont autant contribué que la Géométrie de Descartes au progrès des sciences Mathématiques. Aussi croyons-nous rendre service à la science en en publiant une nouvelle édition. Nous avons d'ailleurs été encouragé dans cette voie par plusieurs savants, et particulièrement par l'un de nos philosophes les plus distingués, M. de Blignières, gendre de l'illustre Liouville, qui a bien voulu contribuer pour une part importante aux frais d'impression.

A. H.

LA GÉOMÉTRIE[1]

LIVRE PREMIER

DES PROBLÈMES QU'ON PEUT CONSTRUIRE SANS Y EMPLOYER QUE DES CERCLES ET DES LIGNES DROITES.

Tous les problèmes de géométrie se peuvent facilement réduire à tels termes, qu'il n'est besoin par après que de connoître la longueur de quelques lignes droites pour les construire.

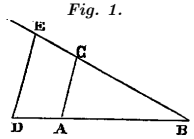
Comment le calcul d'arithmétique se rapporte aux opérations de géométrie.

Et comme toute l'arithmétique n'est composée que de quatre ou cinq opérations, qui sont, l'addition, la soustraction, la multiplication, la division, et l'extraction des racines, qu'on peut prendre pour une espèce de division, ainsi n'a-t-on autre chose à faire en géométrie touchant les lignes qu'on cherche pour les préparer à être connues, que leur en ajouter d'autres, ou en ôter ; ou bien en ayant une, que je nommerai l'unité pour la rapporter d'autant mieux aux nombres, et qui peut ordinairement être prise à discrétion, puis en ayant encore deux autres, en trouver une quatrième qui soit à l'une de ces deux comme l'autre est à l'unité, ce qui est le même que la multiplication ; ou bien en trouver une quatrième qui soit à l'une de ces deux comme l'unité est à l'autre, ce qui est le même que la division ; ou enfin trouver une ou deux, ou plusieurs moyennes proportionnelles entre l'unité et quelque autre ligne, ce qui est le même que tirer la racine carrée ou cubique, etc. Et je ne craindrai pas d'introduire ces termes d'arithmétique en la géométrie, afin de me rendre plus intelligible.

La multiplication.

Soit, par exemple, $A\,B$ *(fig. 1)* l'unité, et qu'il faille multiplier $B\,D$ par $B\,C$, je n'ai qu'à joindre les points A et C, puis tirer $D\,E$ parallèle à $C\,A$, et $B\,E$ est

Fig. 1.



le produit de cette multiplication.

(1)Pour en faciliter la lecture, nous avons substitué à quelques signes employés par Descartes d'autres signes universellement adoptés, toutes les fois que ces changements n'en apportoient pas dans le *principe* de la notation. Le lecteur en sera prévenu.

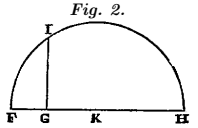
La division.

Ou bien, s'il faut diviser $B\,E$ par $B\,D$, ayant joint les points E et D, je tire $A\,C$ parallèle à $D\,E$, et $B\,C$ est le produit de cette division.

L'extraction de la racine carrée.

Ou s'il faut tirer la racine carrée de $G\,H$ *(fig. 2)*, je lui ajoute en ligne droite

Fig. 2.



$F\,G$, qui est l'unité, et divisant $F\,H$ en deux parties égales au point K, du centre K je tire le cercle $F\,I\,H$, puis élevant du point G une ligne droite jusques à I à angles droits sur $F\,H$, c'est $G\,I$ la racine cherchée. Je ne dis rien ici de la racine cubique, ni des autres, à cause que j'en parlerai plus commodément ci-après.

Comment on peut user de chiffres en géométrie.

Mais souvent on n'a pas besoin de tracer ainsi ces lignes sur le papier, et il suffit de les désigner par quelques lettres, chacune par une seule. Comme pour ajouter la ligne $B\,D$ à $G\,H$, je nomme l'une a et l'autre b, et écris $a+b$; et $a-b$ pour soustraire b de a ; et ab pour les multiplier l'une par l'autre ; et $\frac{a}{b}$ pour diviser a par b ; et aa ou a^2 pour multiplier a par soi-même(2) ; et a^3 pour le multiplier encore une fois par a, et ainsi à l'infini ; et $\sqrt{a^2+b^2}$, pour tirer la racine carrée de a^2+b^2 ; et $\sqrt{C\,.\,a^3-b^3+ab^2}$, pour tirer la racine cubique de $a^3-b^3+ab^2$, et ainsi des autres.

Où il est à remarquer que par a^2, ou b^3, ou semblables, je ne conçois ordinairement que des lignes toutes simples, encore que pour me servir des noms usités en l'algèbre je les nomme des carrés ou des cubes, etc.

Il est aussi à remarquer que toutes les parties d'une même ligne se doivent ordinairement exprimer par autant de dimensions l'une que l'autre, lorsque l'unité n'est point déterminée en la question, comme ici a^3 en contient autant que ab^2 or b^3 dont se compose la ligne que j'ai nommée

$$\sqrt{C\,.\,a^3-b^3+ab^2};$$

mais que ce n'est pas de même lorsque l'unité est déterminée, à cause qu'elle peut être sous-entendue partout où il y a trop ou trop peu de dimensions : comme s'il faut tirer la racine cubique de a^2b^2-b, il faut penser que la quantité a^2b^2 est divisée une fois par l'unité, et que l'autre quantité b est multipliée deux fois par la même.

Au reste, afin de ne pas manquer à se souvenir des noms de ces lignes, il en faut toujours faire un registre séparé à mesure qu'on les pose ou qu'on les change, écrivant par exemple(3) :

$A\,B=1$, c'est-à-dire $A\,B$ égal à 1.

(2)Cependant Descartes répète presque toujours les facteurs égaux lorsqu'ils ne sont qu'au nombre de deux. Nous avons ici constamment adopté la notation a^2.

(3)Nous substituons partout le signe $=$ au signe ∞ dont se servoit Descartes.

$G\,H = a$.

$B\,D = b$, etc.

Comment il faut venir aux équations qui servent à résoudre les problèmes.

Ainsi, voulant résoudre quelque problème, on doit d'abord le considérer comme déjà fait, et donner des noms à toutes les lignes qui semblent nécessaires pour le construire, aussi bien à celles qui sont inconnues qu'aux autres. Puis, sans considérer aucune différence entre ces lignes connues et inconnues, on doit parcourir la difficulté selon l'ordre qui montre le plus naturellement de tous en quelle sorte elles dépendent mutuellement les unes des autres, jusques à ce qu'on ait trouvé moyen d'exprimer une même quantité en deux façons, ce qui se nomme une équation ; car les termes de l'une de ces deux façons sont égaux à ceux de l'autre. Et on doit trouver autant de telles équations qu'on a supposé de lignes qui étoient inconnues. Ou bien, s'il ne s'en trouve pas tant, et que nonobstant on n'omette rien de ce qui est désiré en la question, cela témoigne qu'elle n'est pas entièrement déterminée. Et lors on peut prendre à discrétion des lignes connues pour toutes les inconnues auxquelles ne correspond aucune équation. Après cela, s'il en reste encore plusieurs, il se faut servir par ordre de chacune des équations qui restent aussi, soit en la considérant toute seule, soit en la comparant avec les autres, pour expliquer chacune de ces lignes inconnues, et faire ainsi, en les démêlant, qu'il n'en demeure qu'une seule égale à quelque autre qui soit connue, ou bien dont le carré, ou le cube, ou le carré de carré, ou le sursolide, ou le carré de cube, etc., soit égal à ce qui se produit par l'addition ou soustraction de deux ou plusieurs autres quantités, dont l'une soit connue, et les autres soient composées de quelques moyennes proportionnelles entre l'unité et ce carré, ou cube, ou carré de carré, etc., multipliées par d'autres connues. Ce que j'écris en cette sorte :

$$
\begin{aligned}
& z = b, \\
\text{ou } & z^2 = -az + b^2, \\
\text{ou } & z^3 = +az^2 + b^2 z - c^3, \\
\text{ou } & z^4 = az^3 - c^3 z + d^4, \text{ etc.};
\end{aligned}
$$

c'est-à-dire z, que je prends pour la quantité inconnue, est égale à b ; ou le carré de z est égal au carré de b moins a multiplié par z ; ou le cube de z est égal à a multiplié par le carré de z plus le carré de b multiplié par z moins le cube de c ; et ainsi des autres.

Et on peut toujours réduire ainsi toutes les quantités inconnues à une seule, lorsque le problème se peut construire par des cercles et des lignes droites, ou aussi par des sections coniques, ou même par quelque autre ligne qui ne soit que d'un ou deux degrés plus composée. Mais je ne m'arrête point à expliquer ceci plus en détail, à cause que je vous ôterois le plaisir de l'apprendre de vous-même, et l'utilité de cultiver votre esprit en vous y exerçant, qui est à mon avis la principale qu'on puisse tirer de cette science. Aussi que je n'y remarque rien de si difficile que ceux qui seront un peu versés en la géométrie commune et en l'algèbre, et qui prendront garde à tout ce qui est en ce traité, ne puissent trouver.

C'est pourquoi je me contenterai ici de vous avertir que, pourvu qu'en démêlant ces équations, on ne manque point à se servir de toutes les divisions qui seront possibles, on aura infailliblement les plus simples termes auxquels la question puisse être réduite.

Quels sont les problèmes plans.

Et que si elle peut être résolue par la géométrie ordinaire, c'est-à-dire en ne se servant que de lignes droites et circulaires tracées sur une superficie plate, lorsque la dernière équation aura été entièrement démêlée, il n'y restera tout au plus qu'un carré inconnu, égal à ce qui se produit de l'addition ou soustraction de sa racine multipliée par quelque quantité connue, et de quelque autre quantité aussi connue.

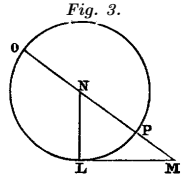
Comment ils se résolvent.

Et lors cette racine, ou ligne inconnue, se trouve aisément ; car si j'ai par exemple

$$z^2 = az + b^2,$$

je fais le triangle rectangle $N\,L\,M$ *(fig. 3)*, dont le côté $L\,M$ est égal à b, racine carrée de la quantité connue b^2, et l'autre $L\,N$ est $\frac{1}{2}a$, la moitié de l'autre

Fig. 3.



quantité connue qui étoit multipliée par z, que je suppose être la ligne inconnue ; puis prolongeant $M\,N$, la base de ce triangle, jusques à O, en sorte que $N\,O$ soit égale à $N\,L$, la toute $O\,M$ est z, la ligne cherchée ; et elle s'exprime en cette sorte :

$$z = \frac{1}{2}a + \sqrt{\frac{1}{4}a^2 + b^2}.$$

Que si j'ai $y^2 = -ay + b^2$, et que y soit la quantité qu'il faut trouver, je fais le même triangle rectangle $N\,L\,M$, et de sa base $M\,N$ j'ôte $N\,P$ égale à $N\,L$, et le reste $P\,M$ est y, la racine cherchée. De façon que j'ai

$$y = -\frac{1}{2}a + \sqrt{\frac{1}{4}a^2 + b^2}.$$

Et tout de même si j'avois

$$x^4 = -ax^2 + b^2,$$

$P\,M$ seroit x^2, et j'aurois

$$x = \sqrt{-\frac{1}{2}a + \sqrt{\frac{1}{4}a^2 + b^2}};$$

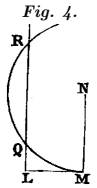
et ainsi des autres.

Enfin, si j'ai

$$z^2 = az - b^2,$$

je fais $N\,M$ *(fig. 4)* égale à $\frac{1}{2}a$, et $L\,M$ égale à b, comme devant ; puis, au lieu

Fig. 4.



de joindre les points $L\,N$, je tire $L\,Q\,R$ parallèle à $M\,N$, et du centre N, par L, ayant décrit un cercle qui la coupe aux points Q et R, la ligne cherchée z est $L\,Q$, ou bien $L\,R$; car en ce cas elle s'exprime en deux façons, à savoir

$$z = \frac{1}{2}a + \sqrt{\frac{1}{4}a^2 - b^2},$$

et

$$z = \frac{1}{2}a - \sqrt{\frac{1}{4}a^2 - b^2}.$$

Et si le cercle, qui ayant son centre au point N passe par le point M, ne coupe ni ne touche la ligne droite $L\,Q\,R$, il n'y a aucune racine en l'équation, de façon qu'on peut assurer que la construction du problème proposé est impossible.

Au reste, ces mêmes racines se peuvent trouver par une infinité d'autres moyens, et j'ai seulement voulu mettre ceux-ci, comme fort simples, afin de faire voir qu'on peut construire tous les problèmes de la géométrie ordinaire sans faire autre chose que le peu qui est compris dans les quatre figures que j'ai expliquées. Ce que je ne crois pas que les anciens aient remarqué ; car autrement ils n'eussent pas pris la peine d'en écrire tant de gros livres où le seul ordre de leurs propositions nous fait connoître qu'ils n'ont point eu la vraie méthode pour les trouver toutes, mais qu'ils ont seulement ramassé celles qu'ils ont rencontrées.

Exemple tiré de Pappus.

Et on peut le voir aussi fort clairement de ce que Pappus a mis au commencement de son septième livre, où après s'être arrêté quelque temps à dénombrer tout ce qui avoit été écrit en géométrie par ceux qui l'avoient précédé, il parle enfin d'une question qu'il dit que ni Euclide, ni Apollonius, ni aucun autre, n'avoient su entièrement résoudre ; et voici ses mots [4] :

Quem autem dicit (Apollonius) in tertio libro locum ad tres et quatuor lineas ab Euclide perfectum non esse, neque ipse perficere poterat, neque aliquis alius ;

[4] Je cite plutôt la version latine que le texte grec, afin que chacun l'entende plus aisément.

sed neque paululum quid addere iis, quæ Euclides scripsit, per ea tantum conica, quæ usque ad Euclidis tempora præmonstrata sunt, etc.

Et un peu après il explique ainsi quelle est cette question :

At locus ad tres et quatuor lineas, in quo (Apollonius) magnifice se jactat, et ostentat, nulla habita gratia ei, qui prius scripserat, est hujusmodi. Si positione datis tribus rectis lineis ab uno et eodem puncto, ad tres lineas in datis angulis rectæ lineæ ducantur, et data sit proportio rectanguli contenti duabus ductis ad quadratum reliquæ : punctum contingit positione datum solidum locum, hoc est unam ex tribus conicis sectionibus. Et si ad quatuor rectas lineas positione datas in datis angulis lineæ ducantur ; et rectanguli duabus ductis contenti ad contentum duabus reliquis proportio data sit : similiter punctum datam coni sectionem positione continget. Si quidem igitur ad duas tantum locus planus ostensus est. Quod si ad plures quam quatuor, punctum continget locos non adhuc cognitos, sed lineas tantum dictas ; quales autem sint, vel quam habeant proprietatem, non constat : earum unam, neque primam, et quæ manifestissima videtur, composuerunt ostendentes utilem esse. Propositiones autem ipsarum hæ sunt.

Si ab aliquo puncto ad positione datas rectas lineas quinque ducantur rectæ lineæ in datis angulis, et data sit proportio solidi parallelepipedi rectanguli, quod tribus ductis lineis continetur ad solidum parallelepipedum rectangulum, quod continetur reliquis duabus, et data quapiam linea, punctum positione datam lineam continget. Si autem ad sex, et data sit proportio solidi tribus lineis contenti ad solidum, quod tribus reliquis continetur ; rursus punctum continget positione datam lineam. Quod si ad plures quam sex, non adhuc habent dicere, an data sit proportio cujuspiam contenti quatuor lineis ad id quod reliquis continetur, quoniam non est aliquid contentum pluribus quam tribus dimensionibus.

Où je vous prie de remarquer en passant que le scrupule que faisoient les anciens d'user des termes de l'arithmétique en la géométrie, qui ne pouvoit procéder que de ce qu'ils ne voyoient pas assez clairement leur rapport, causoit beaucoup d'obscurité et d'embarras en la façon dont ils s'expliquoient ; car Pappus poursuit en cette sorte :

Acquiescunt autem his, qui paulo ante talia interpretati sunt ; neque unum aliquo pacto comprehensibile significantes quod his continetur. Licebit autem per conjunctas proportiones hæc, et dicere, et demonstrare universe in dictis proportionibus, atque his in hunc modum. Si ab aliquo puncto ad positione datas rectas lineas ducantur rectæ lineæ in datis angulis, et data sit proportio conjuncta ex ea, quam habet una ductarum ad unam, et altera ad alteram, et alia ad aliam, et reliqua ad datam lineam, si sint septem ; si vero octo, et reliqua ad reliquam : punctum continget positione datas lineas. Et similiter quotcumque sint impares vel pares multitudine, cum hæc, ut dixi, loco ad quatuor lineas respondeant, nullum igitur posuerunt ita ut linea nota sit, etc.

La question donc qui avoit été commencée à résoudre par Euclide et poursuivie par Apollonius, sans avoir été achevée par personne, étoit telle : Ayant trois ou quatre, ou plus grand nombre de lignes droites données par position ; premièrement on demande un point duquel on puisse tirer autant d'autres lignes droites, une sur chacune des données, qui fassent avec elles des angles donnés,

et que le rectangle contenu en deux de celles qui seront ainsi tirées d'un même point, ait la proportion donnée avec le carré de la troisième, s'il n'y en a que trois ; ou bien avec le rectangle des deux autres, s'il y en a quatre ; ou bien, s'il y en a cinq, que le parallélipipède composé de trois ait la proportion donnée avec le parallélipipède composé des deux qui restent, et d'une autre ligne donnée ; ou s'il y en a six, que le parallélipipède composé de trois ait la proportion donnée avec le parallélipipède des trois autres ; ou s'il y en a sept, que ce qui se produit lorsqu'on en multiplie quatre l'une par l'autre, ait la raison donnée avec ce qui se produit par la multiplication des trois autres, et encore d'une autre ligne donnée ; ou s'il y en a huit, que le produit de la multiplication de quatre ait la proportion donnée avec le produit des quatre autres ; et ainsi cette question se peut étendre à tout autre nombre de lignes. Puis à cause qu'il y a toujours une infinité de divers points qui peuvent satisfaire à ce qui est ici demandé, il est aussi requis de connoître et de tracer la ligne dans laquelle ils doivent tous se trouver. Et Pappus dit que lorsqu'il n'y a que trois ou quatre lignes droites données, c'est en une des trois sections coniques ; mais il n'entreprend point de la déterminer ni de la décrire, non plus que d'expliquer celles où tous ces points se doivent trouver, lorsque la question est proposée en un plus grand nombre de lignes. Seulement il ajoute que les anciens en avoient imaginé une qu'ils montroient y être utile, mais qui sembloit la plus manifeste, et qui n'étoit pas toutefois la première. Ce qui m'a donné occasion d'essayer si, par la méthode dont je me sers, on peut aller aussi loin qu'ils ont été.

Réponse à la question de Pappus.

Et premièrement j'ai connu que cette question n'étant proposée qu'en trois, ou quatre, ou cinq lignes, on peut toujours trouver les points cherchés par la géométrie simple, c'est-à-dire en ne se servant que de la règle et du compas, ni ne faisant autre chose que ce qui a déjà été dit ; excepté seulement lorsqu'il y a cinq lignes données, si elles sont toutes parallèles : auquel cas, comme aussi lorsque la question est proposée en 6, ou 7, ou 8, ou 9 lignes, on peut toujours trouver les points cherchés par la géométrie des solides, c'est-à-dire en y employant quelqu'une des trois sections coniques ; excepté seulement lorsqu'il y a neuf lignes données, si elles sont toutes parallèles : auquel cas, derechef, et encore en 10, 11, 12 ou 13 lignes, on peut trouver les points cherchés par le moyen d'une ligne courbe qui soit d'un degré plus composée que les sections coniques ; excepté en treize, si elles sont toutes parallèles : auquel cas, et en 14, 15, 16 et 17, il y faudra employer une ligne courbe encore d'un degré plus composée que la précédente, et ainsi à l'infini.

Puis j'ai trouvé aussi que lorsqu'il n'y a que trois ou quatre lignes données, les points cherchés se rencontrent tous, non seulement en l'une des trois sections coniques, mais quelquefois aussi en la circonférence d'un cercle ou en une ligne droite ; et que lorsqu'il y en a cinq, ou six, ou sept, ou huit, tous ces points se rencontrent en quelqu'une des lignes qui sont d'un degré plus composées que les sections coniques, et il est impossible d'en imaginer aucune qui ne soit utile à cette question ; mais ils peuvent aussi derechef se rencontrer en une section conique, ou en un cercle, ou en une ligne droite. Et s'il y en a 9, ou 10, ou 11, ou 12, ces points se rencontrent en une ligne qui ne peut être que d'un degré plus composée que les précédentes ; mais toutes celles qui sont d'un degré plus

composées y peuvent servir, et ainsi à l'infini.

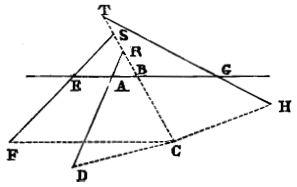
Au reste, la première et la plus simple de toutes, après les sections coniques, est celle qu'on peut décrire par l'intersection d'une parabole et d'une ligne droite, en la façon qui sera tantôt expliquée. En sorte que je pense avoir entièrement satisfait à ce que Pappus nous dit avoir été cherché en ceci par les anciens ; et je tâcherai d'en mettre la démonstration en peu de mots, car il m'ennuie déjà d'en tant écrire.

Soient *(fig. 5)* AB, AD, EF, GH, etc., plusieurs lignes données par position, et qu'il faille trouver un point, comme C, duquel ayant tiré d'autres lignes droites sur les données, comme CB, CD, CF et CH, en sorte que les angles CBA, CDA, CFE, CHG, etc., soient donnés, et que ce qui est produit par la multiplication d'une partie de ces lignes soit égal à ce qui est produit par la multiplication des autres, ou bien qu'ils aient quelque autre proportion donnée, car cela ne rend point la question plus difficile.

Comment on doit poser les termes pour venir à l'équation de cet exemple.

Premièrement, je suppose la chose comme déjà faite, et pour me démêler de

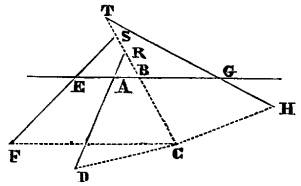
Fig. 5.



la confusion de toutes ces lignes je considère l'une des données, et l'une de celles qu'il faut trouver, par exemple AB et CB, comme les principales et auxquelles je tâche de rapporter ainsi toutes les autres. Que le segment de la ligne AB, qui est entre les points A et B, soit nommé x ; et que BC soit nommé y ; et que toutes les autres lignes données soient prolongées jusques à ce qu'elles coupent ces deux aussi prolongées, s'il est besoin, et si elles ne leur sont point parallèles ; comme vous voyez ici qu'elles coupent la ligne AB aux points A, E, G, et BC aux points R, S, T. Puis à cause que tous les angles du triangle ARB sont donnés, la proportion qui est entre les côtés AB et BR est aussi donnée, et je la pose comme de z à b, de façon que AB *(fig. 6)* étant x, BR sera $\frac{bx}{z}$, et la toute CR sera $y+\frac{bx}{z}$, à cause que le point B tombe entre C et R ; car si R tomboit entre C et B, CR seroit $y-\frac{bx}{z}$; et si C tomboit entre B et R, CR seroit $-y+\frac{bx}{z}$. Tout de même les trois angles du triangle DRC sont donnés, et par conséquent aussi la proportion qui est entre les côtés CR et CD, que je pose

comme de z à c, de façon que $C\,R$ étant $y + \frac{bx}{z}$, $C\,D$ sera $\frac{cy}{z} + \frac{bcx}{z^2}$. Après cela, pourceque les lignes $A\,B$, $A\,D$ et $E\,F$ sont données par position, la distance qui est entre les points A et E est aussi donnée, et si on la nomme k, on aura $E\,B$ égal à $k + x$; mais ce seroit $k - x$ si le point B tomboit entre E et A ; et $-k + x$ si E tomboit entre A et B. Et pourceque les angles du triangle $E\,S\,B$ sont tous donnés, la proportion de $B\,E$ à $B\,S$ est aussi donnée, et je la pose comme de z à

Fig. 6.



d, si bien que $B\,S$ est $\frac{dk + dx}{z}$, et la toute $C\,S$ est $\frac{zy + dk + dx}{z}$; mais ce seroit $\frac{zy - dk - dx}{z}$, si le point S tomboit entre B et C ; et ce seroit $\frac{-zy + dk + dx}{z}$, si C tomboit entre B et S. De plus les trois angles du triangle $F\,S\,C$ sont donnés, et ensuite la proportion de $C\,S$ à $C\,F$, qui soit comme de z à e, et la toute $C\,F$ sera $\frac{ezy + dek + dex}{z^2}$. En même façon $A\,G$ que je nomme l est donnée, et $B\,G$ est $l - x$, et à cause du triangle $B\,G\,T$, la proportion de $B\,G$ à $B\,T$ est aussi donnée, qui soit comme de z à f, et $B\,T$ sera $\frac{fl - fx}{z}$, et $C\,T = \frac{zy + fl - fx}{z}$. Puis derechef la proportion de $C\,T$ à $C\,H$ est donnée à cause du triangle $T\,C\,H$, et la posant comme de z à g, on aura $C\,H = \frac{gzy + fgl - fgx}{z^2}$.

Et ainsi vous voyez qu'en tel nombre de lignes données par position qu'on puisse avoir, toutes les lignes tirées dessus du point C à angles donnés, suivant la teneur de la question, se peuvent toujours exprimer chacune par trois termes, dont l'un est composé de la quantité inconnue y, multipliée ou divisée par quelque autre connue ; et l'autre de la quantité inconnue x, aussi multipliée ou divisée par quelque autre connue ; et le troisième d'une quantité toute connue ; excepté seulement si elles sont parallèles, ou bien à la ligne $A\,B$, auquel cas le terme composé de la quantité x sera nul ; ou bien à la ligne $C\,B$, auquel cas celui qui est composé de la quantité y sera nul, ainsi qu'il est trop manifeste pour que je m'arrête à l'expliquer. Et pour les signes + et − qui se joignent à ces termes, ils peuvent être changés en toutes les façons imaginables.

Puis vous voyez aussi que, multipliant plusieurs de ces lignes l'une par l'autre, les quantités x et y qui se trouvent dans le produit n'y peuvent avoir que chacune autant de dimensions qu'il y a eu de lignes à l'explication desquelles elles servent,

qui ont été ainsi multipliées ; en sorte qu'elles n'auront jamais plus de deux dimensions en ce qui ne sera produit que par la multiplication de deux lignes ; ni plus de trois, en ce qui ne sera produit que par la multiplication de trois, et ainsi à l'infini.

Comment on trouve que ce problème est plan, lorsqu'il n'est point proposé en plus de cinq lignes.

De plus, à cause que pour déterminer le point C, il n'y a qu'une seule condition qui soit requise, à savoir que ce qui est produit par la multiplication d'un certain nombre de ces lignes soit égal, ou, ce qui n'est de rien plus malaisé, ait la proportion donnée à ce qui est produit par la multiplication des autres ; on peut prendre à discrétion l'une des deux quantités inconnues x ou y, et chercher l'autre par cette équation, en laquelle il est évident que, lorsque la question n'est point posée en plus de cinq lignes, la quantité x, qui ne sert point à l'expression de la première, peut toujours n'y avoir que deux dimensions ; de façon que, prenant une quantité connue pour y, il ne restera que $x^2 = +$ ou $- ax +$ ou $- b^2$; et ainsi on pourra trouver la quantité x avec la règle et le compas, en la façon tantôt expliquée. Même, prenant successivement infinies diverses grandeurs pour la ligne y, on en trouvera aussi infinies pour la ligne x, et ainsi on aura une infinité de divers points, tels que celui qui est marqué C, par le moyen desquels on décrira la ligne courbe demandée.

Il se peut faire aussi, la question étant proposée en six ou plus grand nombre de lignes, s'il y en a entre les données qui soient parallèles à $A\,B$ ou $B\,C$, que l'une des deux quantités x ou y n'ait que deux dimensions en l'équation, et ainsi qu'on puisse trouver le point C avec la règle et le compas. Mais au contraire si elles sont toutes parallèles, encore que la question ne soit proposée qu'en cinq lignes, ce point C ne pourra ainsi être trouvé, à cause que la quantité x ne se trouvant point en toute l'équation, il ne sera plus permis de prendre une quantité connue pour celle qui est nommée y, mais ce sera celle qu'il faudra chercher. Et pourcequ'elle aura trois dimensions, on ne le pourra trouver qu'en tirant la racine d'une équation cubique, ce qui ne se peut généralement faire sans qu'on y emploie pour le moins une section conique. Et encore qu'il y ait jusques à neuf lignes données, pourvu qu'elles ne soient point toutes parallèles, on peut toujours faire que l'équation ne monte que jusques au carré de carré ; au moyen de quoi on la peut aussi toujours résoudre par les sections coniques, en la façon que j'expliquerai ci-après. Et encore qu'il y en ait jusques à treize, on peut toujours faire qu'elle ne monte que jusques au carré de cube ; ensuite de quoi on la peut résoudre par le moyen d'une ligne, qui n'est que d'un degré plus composée que les sections coniques, en la façon que j'expliquerai aussi ci-après. Et ceci est la première partie de ce que j'avois ici à démontrer ; mais avant que je passe à la seconde, il est besoin que je dise quelque chose en général de la nature des lignes courbes.

LIVRE SECOND

DE LA NATURE DES LIGNES COURBES.

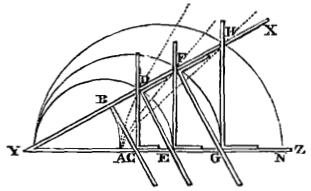
Quelles sont les lignes courbes qu'on peut recevoir en géométrie.

Les anciens ont fort bien remarqué qu'entre les problèmes de géométrie, les uns sont plans, les autres solides et les autres linéaires, c'est-à-dire que les uns peuvent être construits en ne traçant que des lignes droites et des cercles ; au lieu que les autres ne le peuvent être, qu'on n'y emploie pour le moins quelque section conique ; ni enfin les autres, qu'on n'y emploie quelque autre ligne plus composée. Mais je m'étonne de ce qu'ils n'ont point outre cela distingué divers degrés entre ces lignes plus composées, et je ne saurois comprendre pourquoi ils les ont nommées mécaniques plutôt que géométriques. Car de dire que c'ait été à cause qu'il est besoin de se servir de quelque machine pour les décrire, il faudroit rejeter par même raison les cercles et les lignes droites, vu qu'on ne les décrit sur le papier qu'avec un compas et une règle, qu'on peut aussi nommer des machines. Ce n'est pas non plus à cause que les instruments qui servent à les tracer, étant plus composés que la règle et le compas, ne peuvent être si justes ; car il faudroit pour cette raison les rejeter des mécaniques, où la justesse des ouvrages qui sortent de la main est désirée, plutôt que de la géométrie, où c'est seulement la justesse du raisonnement qu'on recherche, et qui peut sans doute être aussi parfaite touchant ces lignes que touchant les autres. Je ne dirai pas aussi que ce soit à cause qu'ils n'ont pas voulu augmenter le nombre de leurs demandes, et qu'ils se sont contentés qu'on leur accordât qu'ils pussent joindre deux points donnés par une ligne droite, et décrire un cercle d'un centre donné qui passât par un point donné ; car ils n'ont point fait de scrupule de supposer outre cela, pour traiter des sections coniques, qu'on pût couper tout cône donné par un plan donné. Et il n'est besoin de rien supposer pour tracer toutes les lignes courbes que je prétends ici d'introduire, sinon que deux ou plusieurs lignes puissent être mues l'une par l'autre, et que leurs intersections en marquent d'autres ; ce qui ne me paroît en rien plus difficile. Il est vrai qu'ils n'ont pas aussi entièrement reçu les sections coniques en leur géométrie, et je ne veux pas entreprendre de changer les noms qui ont été approuvés par l'usage ; mais il est, ce me semble, très clair que, prenant comme on fait pour géométrique ce qui est précis et exact, et pour mécanique ce qui ne l'est pas, et considérant la géométrie comme une science qui enseigne généralement à connoître les mesures de tous les corps, on n'en doit pas plutôt exclure les lignes les plus composées que les plus simples, pourvu qu'on les puisse imaginer être décrites par un mouvement continu, ou par plusieurs qui s'entre-suivent, et dont les derniers soient entièrement réglés par ceux qui les précèdent ; car par ce moyen on peut toujours avoir une connoissance exacte de leur mesure. Mais peut-être que ce qui a empêché les anciens géomètres de recevoir celles qui étoient plus composées que les sections coniques, c'est que les premières qu'ils ont considérées, ayant par hasard été la spirale, la quadratrice et semblables, qui n'appartiennent véritablement qu'aux mécaniques, et ne sont point du nombre

de celles que je pense devoir ici être reçues, à cause qu'on les imagine décrites par deux mouvements séparés, et qui n'ont entre eux aucun rapport qu'on puisse mesurer exactement ; bien qu'ils aient après examiné la conchoïde, la cissoïde, et quelque peu d'autres qui en sont, toutefois à cause qu'ils n'ont peut-être pas assez remarqué leurs propriétés, ils n'en ont pas fait plus d'état que des premières ; ou bien c'est que, voyant qu'ils ne connoissoient encore que peu de choses touchant les sections coniques, et qu'il leur en restait même beaucoup, touchant ce qui se peut faire avec la règle et le compas, qu'ils ignoroient, ils ont cru ne devoir point entamer de matière plus difficile. Mais pourceque j'espère que dorénavant ceux qui auront l'adresse de se servir du calcul géométrique ici proposé, ne trouveront pas assez de quoi s'arrêter touchant les problèmes plans ou solides, je crois qu'il est à propos que je les invite à d'autres recherches, où ils ne manqueront jamais d'exercice.

Voyez les lignes $A\,B$, $A\,D$, $A\,F$ et semblables *(fig. 7)*, que je suppose avoir été décrites par l'aide de l'instrument $Y\,Z$, qui est composé de plusieurs règles

Fig. 7.



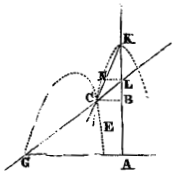
tellement jointes que celle qui est marquée $Y\,Z$ étant arrêtée sur la ligne $A\,N$, on peut ouvrir et fermer l'angle $X\,Y\,Z$, et que lorsqu'il est tout fermé, les points B, C, D, E, F, G, H sont tous assemblés au point A ; mais qu'à mesure qu'on l'ouvre, la règle $B\,C$, qui est jointe à angles droits avec $X\,Y$ au point B, pousse vers Z la règle $C\,D$, qui coule sur $Y\,Z$ en faisant toujours des angles droits avec elle ; et $C\,D$ pousse $D\,E$, qui coule tout de même sur $Y\,X$ en demeurant parallèle à $B\,C$; $D\,E$ pousse $E\,F$, $E\,F$ pousse $F\,G$, celle-ci pousse $G\,H$, et on en peut concevoir une infinité d'autres qui se poussent consécutivement en même façon, et dont les unes fassent toujours les mêmes angles avec $Y\,X$ et les autres avec $Y\,Z$. Or, pendant qu'on ouvre ainsi l'angle $X\,Y\,Z$, le point B décrit la ligne $A\,B$, qui est un cercle ; et les autres points D, F, H, où se font les intersections des autres règles, décrivent d'autres lignes courbes $A\,D$, $A\,F$, $A\,H$, dont les dernières sont par ordre plus composées que la première, et celle-ci plus que le cercle ; mais je ne vois pas ce qui peut empêcher qu'on ne conçoive aussi nettement et aussi distinctement la description de cette première que du cercle, ou du moins que des sections coniques ; ni ce qui peut empêcher qu'on ne conçoive la seconde, et la troisième, et toutes les autres qu'on peut décrire, aussi bien que la première ; ni par conséquent qu'on ne les reçoive toutes en même façon pour servir aux

spéculations de géométrie.

Je pourrois mettre ici plusieurs autres moyens pour tracer et concevoir des lignes courbes qui seroient de plus en plus composées par degrés à l'infini ; mais pour comprendre ensemble toutes celles qui sont en la nature, et les distinguer par ordre en certains genres, je ne sache rien de meilleur que de dire que tous les points de celles qu'on peut nommer géométriques, c'est-à-dire qui tombent sous quelque mesure précise et exacte, ont nécessairement quelque rapport à tous les points d'une ligne droite, qui peut être exprimée par quelque équation, en tous par une même ; et que, lorsque cette équation ne monte que jusqu'au rectangle de deux quantités indéterminées, ou bien au carré d'une même, la ligne courbe est du premier et plus simple genre, dans lequel il n'y a que le cercle, la parabole, l'hyperbole et l'ellipse qui soient comprises ; mais que lorsque l'équation monte jusqu'à la troisième ou quatrième dimension des deux, ou de l'une des deux quantités indéterminées (car il en faut deux pour expliquer ici le rapport d'un point à un autre), elle est du second ; et que lorsque l'équation monte jusqu'à la cinquième ou sixième dimension, elle est du troisième ; et ainsi des autres à l'infini. Comme si je veux savoir de quel genre est la ligne $E\,C$ *(fig. 8)*, que j'imagine être décrite par l'intersection de la règle $G\,L$ et du plan rectiligne

La façon de distinguer toutes les lignes courbe en certains genres, et de connoître le rapport qu'ont tous leurs points à ceux des lignes droites.

Fig. 8.



$C\,N\,K\,L$, dont le côté $K\,N$ est indéfiniment prolongé vers C, et qui, étant mu sur le plan de dessous en ligne droite, c'est-à-dire en telle sorte que son diamètre $K\,L$ se trouve toujours appliqué sur quelque endroit de la ligne $B\,A$ prolongée de part et d'autre, fait mouvoir circulairement cette règle $G\,L$ autour du point G, à cause qu'elle lui est tellement jointe qu'elle passe toujours par le point L. Je choisis une ligne droite comme $A\,B$, pour rapporter à ses divers points tous ceux de cette ligne courbe $E\,C$; et en cette ligne $A\,B$ je choisis un point comme A, pour commencer par lui ce calcul. Je dis que je choisis et l'un et l'autre, à cause qu'il est libre de les prendre tels qu'on veut ; car encore qu'il y ait beaucoup de choix pour rendre l'équation plus courte et plus aisée, toutefois en quelle façon qu'on les prenne, on peut toujours faire que la ligne paroisse de même genre, ainsi qu'il est aisé à démontrer. Après cela prenant un point à discrétion dans la courbe, comme C, sur lequel je suppose que l'instrument qui sert à la décrire est appliqué, je tire de ce point C la ligne $C\,B$ parallèle à $G\,A$, et pourceque $C\,B$ et $B\,A$ sont deux quantités indéterminées et inconnues, je les nomme l'une

y et l'autre x ; mais afin de trouver le rapport de l'une à l'autre, je considère aussi les quantités connues qui déterminent la description de cette ligne courbe, comme $G\,A$, que je nomme a, $K\,L$ que je nomme b, et $N\,L$, parallèle à $G\,A$, que je nomme c ; puis je dis, comme $N\,L$ est à $L\,K$, ou c à b, ainsi $C\,B$ ou y est à $B\,K$, qui est par conséquent $\frac{b}{c}y$: et $B\,L$ est $\frac{b}{c}y-b$, et $A\,L$ est $x+\frac{b}{c}y-b$. De plus, comme $C\,B$ est à $L\,B$, ou y à $\frac{b}{c}y-b$, ainsi $A\,G$ ou a est à $L\,A$ ou $x+\frac{b}{c}y-b$; de façon que, multipliant la seconde par la troisième, on produit $\frac{ab}{c}y-ab$ qui est égale à $xy+\frac{b}{c}y^2-by$, qui se produit en multipliant la première par la dernière : et ainsi l'équation qu'il falloit trouver est

$$y^2 = cy - \frac{cx}{b}y + ay - ac,$$

de laquelle on connoît que la ligne $E\,G$ est du premier genre, comme en effet elle n'est autre qu'une hyperbole.

Que si, en l'instrument qui sert à la décrire, on fait qu'au lieu de la ligne droite $C\,N\,K$, ce soit cette hyperbole, ou quelque autre ligne courbe du premier genre, qui termine le plan $C\,N\,K\,L$, l'intersection de cette ligne et de la règle $G\,L$ décrira, au lieu de l'hyperbole $E\,C$, une autre ligne courbe qui sera d'un second genre. Comme si $C\,N\,K$ est un cercle dont L soit le centre, on décrira la première conchoïde des anciens ; et si c'est une parabole dont le diamètre soit $K\,B$, on décrira la ligne courbe que j'ai tantôt dit être la première et la plus simple pour la question de Pappus, lorsqu'il n'y a que cinq lignes droites données par position ; mais si au lieu d'une de ces lignes courbes du premier genre, c'en est une du second qui termine le plan $C\,N\,K\,L$, on en décrira, par son moyen, une du troisième, ou si c'en est une du troisième, on en décrira une du quatrième, et ainsi à l'infini, comme il est fort aisé à connoître par le calcul. Et en quelque autre façon qu'on imagine la description d'une ligne courbe, pourvu qu'elle soit du nombre de celles que je nomme géométriques, on pourra toujours trouver une équation pour déterminer tous ses points en cette sorte.

Au reste, je mets les lignes courbes qui font monter cette équation jusqu'au carré, au même genre que celles qui ne la font monter que jusqu'au cube ; et celles dont l'équation monte au carré de cube, au même genre que celles dont elle ne monte qu'au sursolide, et ainsi des autres : dont la raison est qu'il y a règle générale pour réduire au cube toutes les difficultés qui vont au carré de carré, et au sursolide toutes celles qui vont au carré de cube ; de façon qu'on ne les doit point estimer plus composées.

Mais il est à remarquer qu'entre les lignes de chaque genre, encore que la plupart soient également composées, en sorte qu'elles peuvent servir à déterminer les mêmes points et construire les mêmes problèmes, il y en a toutefois aussi quelques unes qui sont plus simples, et qui n'ont pas tant d'étendue en leur puissance ; comme entre celles du premier genre, outre l'ellipse, l'hyperbole et la parabole, qui sont également composées, le cercle y est aussi compris, qui manifestement est plus simple ; et entre celles du second genre, il y a la conchoïde

vulgaire, qui a son origine du cercle ; et il y en a encore quelques autres qui, bien qu'elles n'aient pas tant d'étendue que la plupart de celles du même genre, ne peuvent toutefois être mises dans le premier.

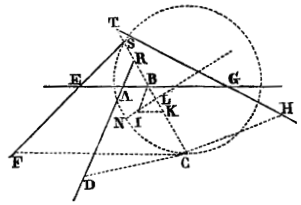
Suite de l'explication de la question de Pappus, mise au livre précédent.

Or, après avoir ainsi réduit toutes les lignes courbes à certains genres, il m'est aisé de poursuivre en la démonstration de la réponse que j'ai tantôt faite à la question de Pappus ; car premièrement, ayant fait voir ci-dessus que, lorsqu'il n'y a que trois ou quatre lignes droites données, l'équation qui sert à déterminer les points cherchés ne monte que jusqu'au carré, il est évident que la ligne courbe où se trouvent ces points est nécessairement quelqu'une de celles du premier genre, à cause que cette même équation explique le rapport qu'ont tous les points des lignes du premier genre à ceux d'une ligne droite ; et que lorsqu'il n'y a point plus de huit lignes droites données, cette équation ne monte que jusqu'au carré de carré tout au plus, et que par conséquent la ligne cherchée ne peut être que du second genre, ou au-dessous ; et que lorsqu'il n'y a point plus de douze lignes données, l'équation ne monte que jusqu'au carré de cube, et que par conséquent la ligne cherchée n'est que du troisième genre, ou au-dessous ; et ainsi des autres. Et même à cause que la position des lignes droites données peut varier en toutes sortes, et par conséquent faire changer tant les quantités connues que les signes + et − de l'équation, en toutes les façons imaginables, il est évident qu'il n'y a aucune ligne courbe du premier genre qui ne soit utile à cette question, quand elle est proposée en quatre lignes droites ; ni aucune du second qui n'y soit utile, quand elle est proposée en huit ; ni du troisième, quand elle est proposée en douze ; et ainsi des autres : en sorte qu'il n'y a pas une ligne courbe qui tombe sous le calcul et puisse être reçue en géométrie, qui n'y soit utile pour quelque nombre de lignes.

Solution de cette question quand elle n'est proposée qu'en trois ou quatre lignes.

Mais il faut ici plus particulièrement que je détermine et donne la façon de trouver la ligne cherchée qui sert en chaque cas, lorsqu'il n'y a que trois ou quatre lignes droites données ; et on verra, par même moyen, que le premier genre des lignes courbes n'en contient aucunes autres que les trois sections coniques et le cercle.

Fig. 9.



Reprenons les quatre lignes $A\,B$, $A\,D$, $E\,F$ et $G\,H$ *(fig. 9)* données ci-dessus, et qu'il faille trouver une autre ligne en laquelle il se rencontre une infinité de

points tels que C, duquel ayant tiré les quatre lignes $C\,B$, $C\,D$, $C\,F$ et $C\,H$, à angles donnés sur les données, $C\,B$ multipliée par $C\,F$ produit une somme égale à $C\,D$ multipliée par $C\,H$; c'est-à-dire, ayant fait

$$C\,B = y, \quad C\,D = \frac{czy + bcx}{z^2}, \quad C\,F = \frac{ezy + dek + dex}{z^2},$$
$$\text{et} \quad C\,H = \frac{gzy + fgl - fgx}{z^2},$$

l'équation est(5)

$$y^2 = \frac{(cfglz - dckz^2)y - (dez^2 + cfgz - bcgz)xy + bcfglx - bcfgx^2}{ez^3 - cgz^2}.$$

au moins en supposant ez plus grand que cg, car s'il étoit moindre il faudroit changer tous les signes + et −. Et si la quantité y se trouvoit nulle ou moindre que rien en cette équation, lorsqu'on a supposé le point C en l'angle $D\,A\,G$, il faudroit le supposer aussi en l'angle $D\,A\,E$, ou $E\,A\,R$, ou $R\,A\,G$, en changeant les signes + et − selon qu'il serait requis à cet effet. Et si en toutes ces quatre positions la valeur de y se trouvoit nulle, la question seroit impossible au cas proposé. Mais supposons-la ici être possible, et pour en abréger les termes, au lieu des quantités $\frac{cfglz - dekz^2}{ez^3 - cgz^2}$, écrivons $2m$; et au lieu de $\frac{dez^2 + cfgz - bcgz}{ez^3 - cgz^2}$, écrivons $\frac{2n}{z}$; et ainsi nous aurons

$$y^2 = 2my - \frac{2n}{z}xy + \frac{bcfglx - bcfgx^2}{ez^3 - cgz^2},$$

dont la racine est

$$y = m - \frac{nx}{z} + \sqrt{m^2 - \frac{2mnx}{z} + \frac{n^2x^2}{z^2} + \frac{bcfglx - bcfgx^2}{ez^3 - cgz^2}};$$

et de rechef pour abréger, au lieu de $-\frac{2mn}{z} + \frac{bcfgl}{ez^3 - cgz^2}$, écrivons o ; et au lieu de $\frac{n^2}{z^2} - \frac{bcfg}{ez^2 - cgz^2}$, écrivons $\frac{p}{m}$; car ces quantités étant toutes données, nous les pouvons nommer comme il nous plaît : et ainsi nous avons

$$y = m - \frac{n}{z}x + \sqrt{m^2 + ox + \frac{p}{m}x^2},$$

qui doit être la longueur de la ligne $B\,C$, en laissant $A\,B$ ou x indéterminée. Et il est évident que la question n'étant proposée qu'en trois ou quatre lignes, on

(5)Les termes contenus entre deux parenthèses sont placés l'un sous l'autre dans les anciennes éditions, comme, par exemple, $\left.\begin{matrix} -dekz^2 \\ +cfglz \end{matrix}\right\} y$.

peut toujours avoir de tels termes, excepté que quelques uns d'eux peuvent être nuls, et que les signes + et - peuvent diversement être changés.

Après cela je fais $K\,I$ égale et parallèle à $B\,A$, en sorte qu'elle coupe de $B\,C$ la partie $B\,K$ égale à m, à cause qu'il y a ici $+m$; et je l'aurois ajoutée en tirant cette ligne $I\,K$ de l'autre côté, s'il y avoit eu $-m$; et je ne l'aurois point du tout tirée, si la quantité m eût été nulle. Puis je tire aussi $I\,L$, en sorte que la ligne $I\,K$ est à $K\,L$ comme z est à n ; c'est-à-dire que $I\,K$ étant x, $K\,L$ est $\frac{n}{z}x$. Et par même moyen je connois aussi la proportion qui est entre $K\,L$ et $I\,L$, que je pose comme entre n et a : si bien que $K\,L$ étant $\frac{n}{z}x$, $I\,L$ est $\frac{a}{z}x$. Et je fais que le point K soit entre L et C, à cause qu'il y a ici $-\frac{n}{z}x$; au lieu que j'aurois mis L entre K et C, si j'eusse eu $+\frac{n}{z}x$; et je n'eusse point tiré cette ligne $I\,L$, si $\frac{n}{z}x$ eût été nulle.

Or, cela fait, il ne me reste plus pour la ligne $L\,C$ que ces termes

$$LC = \sqrt{m^2 + ox + \frac{p}{m}x^2},$$

d'où je vois que s'ils étoient nuls, ce point C se trouveroit en la ligne droite $I\,L$; et que s'ils étoient tels que la racine s'en pût tirer, c'est-à-dire que m^2 et $\frac{p}{m}x^2$ étant marqués d'un même signe + ou −, o^2 fût égal à $4pm$, ou bien que les termes m^2 et ox, ou ox et $\frac{p}{m}x^2$ fussent nuls, ce point C se trouveroit en une autre ligne droite qui ne seroit pas plus malaisée à trouver que $I\,L$. Mais lorsque cela n'est pas, ce point C est toujours en l'une des trois sections ou en un cercle dont l'un des diamètres est en la ligne $I\,L$, et la ligne $L\,C$ est l'une de celles qui s'appliquent par ordre à ce diamètre ; ou au contraire $L\,C$ est parallèle au diamètre auquel celle qui est en la ligne $I\,L$ est appliquée par ordre ; à savoir si le terme $\frac{p}{m}x^2$ est nul, cette section conique est une parabole ; et s'il est marqué du signe +, c'est une hyperbole ; et enfin s'il est marqué du signe −, c'est une ellipse, excepté seulement si la quantité a^2m est égale à pz^2, et que l'angle $I\,L\,C$ soit droit, auquel cas on a un cercle au lieu d'une ellipse. Que si cette section est une parabole, son côté droit est égal à $\frac{oz}{a}$, et son diamètre est toujours en la ligne $I\,L$; et pour trouver le point N, qui en est le sommet, il faut faire $I\,N$ égale à $\frac{am^2}{oz}$; et que le point I soit entre L et N, si les termes sont $+m^2+ox$; ou bien que le point L soit entre I et N, s'ils sont $+m^2-ox$; ou bien il faudroit que N fût entre I et L, s'il y avoit $-m^2+ox$. Mais il ne peut jamais y avoir $-m^2$, en la façon que les termes ont ici été posés. Et enfin le point N seroit le même que le point I si la quantité m^2 étoit nulle ; au moyen de quoi il est aisé de trouver cette parabole par le premier problème du premier livre d'Apollonius.

Que si la ligne demandée est un cercle, ou une ellipse, ou une hyperbole, il faut premièrement chercher le point M qui en est le centre, et qui est toujours en la ligne droite $I\,L$; ou on le trouve en prenant $\frac{aom}{2pz}$ pour $I\,M$, en sorte que si

la quantité o est nulle, ce centre est justement au point I. Et si la ligne cherchée est un cercle ou une ellipse, on doit prendre le point M du même côté que le point L, au respect du point I, lorsqu'on a $+ox$; et lorsqu'on a $-ox$, on le doit prendre de l'autre. Mais tout au contraire, en l'hyperbole, si on a $-ox$, ce centre M doit être vers L ; et si on a $+ox$, il doit être de l'autre côté. Après cela le côté droit de la figure doit être

$$\sqrt{\frac{o^2z^2}{a^2} + \frac{4mpz^2}{a^2}},$$

lorsqu'on a $+m^2$, et que la ligne cherchée est un cercle ou une ellipse ; ou bien lorsqu'on a $-m^2$, et que c'est une hyperbole ; et il doit être

$$\sqrt{\frac{o^2z^2}{a^2} - \frac{4mpz^2}{a^2}},$$

si la ligne cherchée, étant un cercle ou une ellipse, on a $-m^2$; ou bien si étant une hyperbole, et la quantité o^2 étant plus grande que $4mp$, on a $+m^2$. Que si la quantité m^2 est nulle, ce côté droit est $\frac{oz}{a}$; et si ox est nulle, il est

$$\sqrt{\frac{4mpz^2}{a^2}}.$$

Puis, pour le côté traversant, il faut trouver une ligne qui soit à ce côté droit comme a^2m est à pz^2 ; à savoir si ce côté droit est

$$\sqrt{\frac{o^2z^2}{a^2} + \frac{4mpz^2}{a^2}},$$

le traversant est

$$\sqrt{\frac{a^2o^2m^2}{p^2z^2} + \frac{4a^2m^3}{pz^2}}.$$

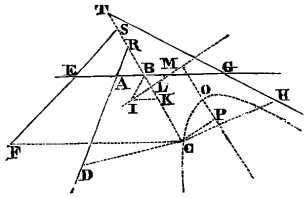
Et en tous ces cas le diamètre de la section est en la ligne $I\,M$, et $L\,C$ est l'une de celles qui lui est appliquée par ordre. Si bien que, faisant $M\,N$ égale à la moitié du côté traversant, et le prenant du même côté du point M qu'est le point L, on a le point N pour le sommet de ce diamètre ; ensuite de quoi il est aisé de trouver la section par les second et troisième problèmes du premier livre d'Apollonius.

Mais quand cette section étant une hyperbole, on a $+m^2$, et que la quantité o^2 est nulle ou plus petite que $4pm$, on doit tirer du centre M la ligne $M\,O\,P$ parallèle à $L\,C$, et $C\,P$ parallèle à $L\,M$, et faire $M\,O$ égale à

$$\sqrt{m^2 - \frac{o^2m}{4p}},$$

ou bien la faire égale à m si la quantité ox est nulle ; puis considérer le point O

Fig. 10.



comme le sommet de cette hyperbole, dont le diamètre est $O\,P$, et $C\,P$ la ligne qui lui est appliquée par ordre, et son côté droit est

$$\sqrt{\frac{4a^4m^4}{p^2z^4} - \frac{a^4o^2m^3}{p^3z^4}},$$

et son côté traversant est

$$\sqrt{4m^2 - \frac{o^2m}{p}};$$

excepté quand ox est nulle, car alors le côté droit est $\dfrac{2a^2m^2}{pz^2}$, et le traversant est $2m$; et ainsi il est aisé de la trouver par le troisième problème du premier livre d'Apollonius.

Démonstration de tout ce qui vient d'être expliqué.

Et les démonstrations de tout ceci sont évidentes ; car composant un espace des quantités que j'ai assignées pour le côté droit, et le traversant, et pour le segment du diamètre $N\,L$ ou $O\,P$, suivant la teneur du 11^{e}, du 12^{e}et du 13^{e}théorème du premier livre d'Apollonius, on trouvera tous les mêmes termes dont est composé le carré de la ligne $C\,P$, ou $C\,L$, qui est appliquée par ordre à ce diamètre. Comme en cet exemple, ôtant $I\,M$ qui est $\dfrac{aom}{2pz}$, de $N\,M$ qui est

$$\frac{am}{2pz}\sqrt{o^2 + 4mp},$$

j'ai $I\,N$, à laquelle ajoutant $I\,L$ qui est $\dfrac{a}{z}x$, j'ai $N\,L$ qui est

$$\frac{a}{z}x - \frac{aom}{2pz} + \frac{am}{2pz}\sqrt{o^2 + 4mp};$$

et ceci étant multiplié par $\dfrac{z}{a}\sqrt{o^2 + 4mp}$, qui est le côté droit de la figure, il vient

$$x\sqrt{o^2 + 4mp} - \frac{om}{2p}\sqrt{o^2 + 4mp} + \frac{mo^2}{2p} + 2m^2,$$

pour le rectangle, duquel il faut ôter un espace qui soit au carré de $N\,L$ comme le côté droit est au traversant, et ce carré de $N\,L$ est

$$\frac{a^2}{z^2}x^2 - \frac{a^2om}{pz^2}x + \frac{a^2m}{pz^2}x\sqrt{o^2 + 4mp} + \frac{a^2o^2m^2}{2p^2z^2} + \frac{a^2m^3}{pz^2} - \frac{a^2om^2}{2p^2z^2}\sqrt{o^2 + 4mp},$$

qu'il faut diviser par a^2m et multiplier par pz^2, à cause que ces termes expliquent la proportion qui est entre le côté traversant et le droit, et il vient

$$\frac{p}{m}x^2 - ox + x\sqrt{o^2 + 4mp} + \frac{o^2m}{2p} - \frac{om}{2p}\sqrt{o^2 + 4mp} + m^2,$$

ce qu'il faut ôter du rectangle précédent, et on trouve

$$m^2 + ox - \frac{p}{m}x^2$$

pour le carré de $C\,L$, qui par conséquent est une ligne appliquée par ordre dans une ellipse, ou dans un cercle, au segment du diamètre $N\,L$.

Et si on veut expliquer toutes les quantités données par nombres, en faisant par exemple $E\,A = 3$, $A\,G = 5$, $A\,B = B\,R$, $B\,S = \dfrac{1}{2}B\,E$, $G\,B = B\,T$, $C\,D = \dfrac{3}{2}C\,R$, $C\,F = 2C\,S$, $C\,H = \dfrac{2}{3}C\,T$, et que l'angle $A\,B\,R$ soit de 60 degrés, et enfin que le rectangle des deux $C\,B$ et $C\,F$ soit égal au rectangle des deux autres $C\,D$ et $C\,H$; car il faut avoir toutes ces choses afin que la question soit entièrement déterminée ; et avec cela, supposant $A\,B = x$, et $C\,B = y$, on trouve par la façon ci-dessus expliquée.

$$y^2 = 2y - xy + 5x - x^2,$$

$$y = 1 - \frac{1}{2}x + \sqrt{1 + 4x - \frac{3}{4}x^2},$$

si bien que $B\,K$ doit être 1, $K\,L$ doit être la moitié de $K\,I$; et pourceque l'angle $I\,K\,L$ ou $A\,B\,R$ est de 60 degrés, et $K\,I\,L$ qui est la moitié de $K\,I\,B$ ou $I\,K\,L$, de 30, $I\,L\,K$ est droit. Et pourceque $I\,K$ ou $A\,B$ est nommée x, $K\,L$ est $\dfrac{1}{2}x$, et $I\,L$ est $x\sqrt{\dfrac{3}{4}}$, et la quantité qui étoit tantôt nommée z est 1, celle qui étoit a est $\sqrt{\dfrac{3}{4}}$, celle qui étoit m est 1, celle qui étoit o est 4, et celle qui étoit p est $\dfrac{3}{4}$, de façon qu'on a $\sqrt{\dfrac{16}{3}}$ pour $I\,M$, et $\sqrt{\dfrac{19}{3}}$ pour $N\,M$; et pourceque a^2m, qui est $\dfrac{3}{4}$, est ici égal à pz^2, et que l'angle $I\,L\,C$ est droit, on trouve que la ligne courbe $N\,C$ est un cercle. Et on peut facilement examiner tous les autres cas en même sorte.

Au reste, à cause que les équations qui ne montent que jusqu'au carré sont toutes comprises en ce que je viens d'expliquer, non seulement le problème des anciens en trois et quatre lignes est ici entièrement achevé, mais aussi tout ce qui appartient à ce qu'ils nommoient la composition des lieux solides, et par conséquent aussi à celle des lieux plans, à cause qu'ils sont compris dans les solides : car ces lieux ne sont autre chose, sinon que, lorsqu'il est question de trouver quelque point auquel il manque une condition pour être entièrement déterminé, ainsi qu'il arrive en cet exemple, tous les points d'une même ligne peuvent être pris pour celui qui est demandé : et si cette ligne est droite ou circulaire, on la nomme un lieu plan ; mais si c'est une parabole, ou une hyperbole, ou une ellipse, on la nomme un lieu solide : et toutefois et quantes que cela est, on peut venir à une équation qui contient deux quantités inconnues, et est pareille à quelqu'une de celles que je viens de résoudre. Que si la ligne qui détermine ainsi le point cherché est d'un degré plus composée que les sections coniques, on la peut nommer, en même façon, un lieu sursolide, et ainsi des autres. Et s'il manque deux conditions à la détermination de ce point, le lieu où il se trouve est une superficie, laquelle peut être tout de même ou plate, ou sphérique, ou plus composée. Mais le plus haut but qu'aient eu les anciens en cette matière a été de parvenir à la composition des lieux solides ; et il semble que tout ce qu'Apollonius a écrit des sections coniques n'a été qu'à dessein de la chercher.

Quels sont les lieux plans et solides, et la façon de les trouver.

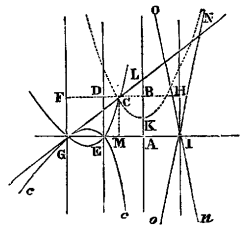
De plus, on voit ici que ce que j'ai pris pour le premier genre des lignes courbes n'en peut comprendre aucunes autres que le cercle, la parabole, l'hyperbole et l'ellipse, qui est tout ce que j'avois entrepris de prouver.

Que si la question des anciens est proposée en cinq lignes qui soient toutes parallèles, il est évident que le point cherché sera toujours en une ligne droite ; mais si elle est proposée en cinq lignes, dont il y en ait quatre qui soient parallèles, et que la cinquième les coupe à angles droits, et même que toutes les lignes tirées du point cherché les rencontrent aussi à angles droits, et enfin que le parallélipipède composé de trois des lignes ainsi tirées sur trois de celles qui sont parallèles soit égal au parallélipipède composé des deux lignes tirées, l'une sur la quatrième de celles qui sont parallèles, et l'autre sur celle qui les coupe à angles droits, et d'une troisième ligne donnée, ce qui est, ce semble, le plus simple cas qu'on puisse imaginer après le précédent, le point cherché sera en la ligne courbe qui est décrite par le mouvement d'une parabole, en la façon ci-dessus expliquée.

Quelle est la première et la plus simple de toutes les lignes courbes en la question des anciens quand elle est proposée en cinq lignes.

Soient par exemple les lignes données $A\,B$, $I\,H$, $E\,D$, $G\,F$, et $G\,A$ *(fig. 11)*, et qu'on demande le point C, en sorte que tirant $C\,B$, $C\,F$, $C\,D$, $C\,H$ et $C\,M$ à angles droits sur les données, le parallélipipède des trois $C\,F$, $C\,D$, $C\,H$ soit égal à celui des deux autres $C\,B$ et $C\,M$, et d'une troisième qui soit $A\,I$. Je pose $C\,B = y$, $C\,M = x$, $A\,I$ ou $A\,E$ ou $G\,E = a$; de façon que le point C étant entre les lignes $A\,B$ et $D\,E$, j'ai $C\,F = 2a - y$, $C\,D = a - y$, et $C\,H = y + a$; et multipliant ces trois l'une par l'autre, j'ai $y^3 - 2ay^2 - a^2y + 2a^3$ égal au produit des trois autres, qui est axy. Après cela je considère la ligne courbe $C\,E\,G$, que j'imagine être décrite par l'intersection de la parabole $C\,K\,N$, qu'on fait mouvoir en telle sorte que son diamètre $K\,L$ est toujours sur la ligne droite $A\,B$,

Fig. 11.



et de la règle $G\,L$ qui tourne cependant autour du point G en telle sorte qu'elle passe toujours dans le plan de cette parabole par le point L. Et je fais $K\,L = a$, et le côté droit principal, c'est-à-dire celui qui se rapporte à l'essieu de cette parabole, aussi égal à a, et $G\,A = 2a$, et $C\,B$ ou $M\,A = y$, et $C\,M$ ou $A\,B = x$. Puis à cause des triangles semblables $G\,M\,C$ et $C\,B\,L$, $G\,M$ qui est $2a - y$, est à $M\,C$ qui est x, comme $C\,B$ qui est y, est à $B\,L$ qui est par conséquent $\frac{xy}{2a-y}$. Et pourceque $K\,L$ est a, $B\,K$ est $a - \frac{xy}{2a-y}$, ou bien $\frac{2a^2 - ay - xy}{2a-y}$. Et enfin pourceque ce même $B\,K$, étant un segment du diamètre de la parabole, est à $B\,C$ qui lui est appliquée par ordre, comme celle-ci est au côté droit qui est a, le calcul montre que $y^3 - 2ay^2 - a^2y + 2a^3$ est égal à axy ; et par conséquent que le point C est celui qui étoit demandé. Et il peut être pris en tel endroit de la ligne $C\,E\,G$ qu'on veuille choisir, ou aussi en son adjointe $c\,E\,G\,c$, qui se décrit en même façon, excepté que le sommet de la parabole est tourné vers l'autre côté, ou enfin en leurs contreposées $N\,I\,o$, $n\,I\,O$, qui sont décrites par l'intersection que fait la ligne $G\,L$ en l'autre côté de la parabole $K\,N$.

Or encore que les parallèles données $A\,B$, $I\,H$, $E\,D$, et $G\,F$, ne fussent point également distantes, et que $G\,A$ ne les coupât point à angles droits, ni aussi les lignes tirées du point C vers elles, ce point C ne laisseroit pas de se trouver toujours en une ligne courbe qui seroit de même nature : et il s'y peut aussi trouver quelquefois, encore qu'aucune des lignes données ne soient parallèles. Mais si lorsqu'il y en a quatre ainsi parallèles, et une cinquième qui les traverse, et que le parallélipipède de trois des lignes tirées du point cherché, l'une sur cette cinquième, et les deux autres sur deux de celles qui sont parallèles, soit égal à celui des deux tirées sur les deux autres parallèles, et d'une autre ligne donnée : ce point cherché est en une ligne courbe d'une autre nature, à savoir en une qui est telle, que toutes les lignes droites appliquées par ordre à son diamètre étant égales à celles d'une section conique, les segments de ce diamètre qui sont entre le sommet et ces lignes ont même proportion à une certaine ligne donnée, que cette ligne donnée a aux segments du diamètre de la section conique, auxquels

les pareilles lignes sont appliquées par ordre. Et je ne saurois véritablement dire que cette ligne soit moins simple que la précédente, laquelle j'ai cru toutefois devoir prendre pour la première, à cause que la description et le calcul en sont en quelque façon plus faciles.

Pour les lignes qui servent aux autres cas, je ne m'arrêterai point à les distinguer par espèces, car je n'ai pas entrepris de dire tout ; et, ayant expliqué la façon de trouver une infinité de points par où elles passent, je pense avoir assez donné le moyen de les décrire.

Quelles sont les lignes courbes qu'on décrit en trouvant plusieurs de leurs points, qui peuvent être reçues en géométrie.

Même il est à propos de remarquer qu'il y a grande différence entre cette façon de trouver plusieurs points pour tracer une ligne courbe, et celle dont on se sert pour la spirale et ses semblables ; car par cette dernière on ne trouve pas indifféremment tous les points de la ligne qu'on cherche, mais seulement ceux qui peuvent être déterminés par quelque mesure plus simple que celle qui est requise pour la composer ; et ainsi, à proprement parler, on ne trouve pas un de ses points, c'est-à-dire pas un de ceux qui lui sont tellement propres qu'ils ne puissent être trouvés que par elle ; au lieu qu'il n'y a aucun point dans les lignes qui servent à la question proposée, qui ne se puisse rencontrer entre ceux qui se déterminent par la façon tantôt expliquée. Et pourceque cette façon de tracer une ligne courbe, en trouvant indifféremment plusieurs de ses points, ne s'étend qu'à celles qui peuvent aussi être décrites par un mouvement régulier et continu, on ne la doit pas entièrement rejeter de la géométrie.

Quelles sont aussi celles qu'on décrit avec une corde, qui peuvent y être reçues.

Et on n'en doit pas rejeter non plus celle où on se sert d'un fil ou d'une corde repliée pour déterminer l'égalité ou la différence de deux ou plusieurs lignes droites qui peuvent être tirées de chaque point de la courbe qu'on cherche, à certains autres points, ou sur certaines autres lignes à certains angles, ainsi que nous avons fait en la Dioptrique pour expliquer l'ellipse et l'hyperbole ; car encore qu'on n'y puisse recevoir aucunes lignes qui semblent à des cordes, c'est-à-dire qui deviennent tantôt droites et tantôt courbes, à cause que la proportion qui est entre les droites et les courbes n'étant pas connue, et même, je crois, ne le pouvant être par les hommes, on ne pourroit rien conclure de là qui fût exact et assuré. Toutefois à cause qu'on ne se sert de cordes en ces constructions que pour déterminer des lignes droites dont on connoît parfaitement la longueur, cela ne doit point faire qu'on les rejette.

Que pour trouver toutes les propriétés des lignes courbes il suffit de savoir le rapport qu'ont tous leurs points à ceux des lignes droites, et la façon de tirer d'autres lignes qui les coupent en tous ces points à angles droits.

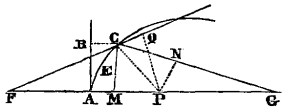
Or de cela seul qu'on sait le rapport qu'ont tous les points d'une ligne courbe à tous ceux d'une ligne droite, en la façon que j'ai expliquée, il est aisé de trouver aussi le rapport qu'ils ont à tous les autres points et lignes données ; et ensuite de connoître les diamètres, les essieux, les centres et autres lignes ou points à qui chaque ligne courbe aura quelque rapport plus particulier ou plus simple qu'aux autres ; et ainsi d'imaginer divers moyens pour les décrire, et d'en choisir les plus faciles ; et même on peut aussi, par cela seul, trouver quasi tout ce qui peut être déterminé touchant la grandeur de l'espace qu'elles comprennent, sans qu'il soit besoin que j'en donne plus d'ouverture. Et enfin pour ce qui est de toutes les autres propriétés qu'on peut attribuer aux lignes courbes, elles ne dépendent que de la grandeur des angles qu'elles font avec quelques autres lignes. Mais lorsqu'on peut tirer des lignes droites qui les coupent à angles droits, aux points où elles sont rencontrées par celles avec qui elles font les angles qu'on veut

mesurer, ou, ce que je prends ici pour le même, qui coupent leurs contingentes, la grandeur de ces angles n'est pas plus malaisée à trouver que s'ils étoient compris entre deux lignes droites. C'est pourquoi je croirai avoir mis ici tout ce qui est requis pour les éléments des lignes courbes, lorsque j'aurai généralement donné la façon de tirer des lignes droites qui tombent à angles droits sur tels de leurs points qu'on voudra choisir. Et j'ose dire que c'est ceci le problème le plus utile et le plus général, non seulement que je sache, mais même que j'aie jamais désiré de savoir en géométrie.

Façon générale pour trouver des lignes droites, qui coupent les courbes données ou leurs contingentes, à angles droits.

Soit $C\,E$ *(fig. 12)* la ligne courbe, et qu'il faille tirer une ligne droite par le point C, qui fasse avec elle des angles droits. Je suppose la chose déjà faite,

Fig. 12.



et que la ligne cherchée est $C\,P$, laquelle je prolonge jusqu'au point P, où elle rencontre la ligne droite $G\,A$, que je suppose être celle aux points de laquelle on rapporte tous ceux de la ligne $C\,E$; en sorte que faisant $M\,A$ on $C\,B = y$, et $C\,M$ ou $B\,A = x$, j'ai quelque équation qui explique le rapport qui est entre x et y ; puis je fais $P\,C = s$, et $P\,A = v$, ou $P\,M = v - y$; et à cause du triangle rectangle $P\,M\,C$, j'ai s^2, qui est le carré de la base, égal à $x^2 + v^2 - 2vy + y^2$, qui sont les carrés des deux côtés ; c'est-à-dire j'ai

$$x = \sqrt{s^2 - v^2 + 2vy - y^2},$$

ou bien

$$y = v + \sqrt{s^2 - x^2};$$

et par le moyen de cette équation, j'ôte de l'autre équation, qui m'explique le rapport qu'ont tous les points de la courbe $C\,E$ à ceux de la droite $G\,A$, l'une des deux quantités indéterminées x ou y ; ce qui est aisé à faire en mettant partout

$$\sqrt{s^2 - v^2 + 2vy - y^2}$$

au lieu de x, et le carré de cette somme au lieu de x^2, et son cube au lieu de x^3, et ainsi des autres, si c'est x que je veuille ôter ; ou bien si c'est y, en mettant en son lieu

$$v + \sqrt{s^2 - x^2},$$

et le carré ou le cube, etc., de cette somme au lieu de y^2 ou y^3, etc. De façon qu'il reste toujours après cela une équation en laquelle il n'y a plus qu'une seule quantité indéterminée x ou y. Comme si $C\,E$ est une ellipse, et que $M\,A$ soit le segment de son diamètre, auquel $C\,M$ soit appliquée par ordre, et qui ait r

pour son côté droit et q pour le traversant, on a, par le treizième théorème du premier livre d'Apollonius, $x^2 = ry - \frac{r}{q}y^2$, d'où ôtant x^2, il reste

$$s^2 - v^2 + 2vy - y^2 = ry - \frac{r}{q}y^2,$$

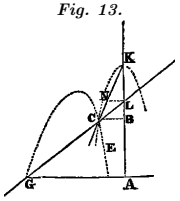
ou bien

$$y^2 + \frac{qry - 2qvy + qv^2 - qs^2}{q - r} = 0;$$

car il est mieux en cet endroit de considérer ainsi ensemble toute la somme que d'en faire une partie égale à l'autre.

Tout de même si $C\,E$ *(fig. 13)* est la ligne courbe décrite par le mouvement d'une parabole en la façon ci-dessus expliquée (page 13), et qu'on ait posé b

Fig. 13.



pour $G\,A$, c pour $K\,L$, et d pour le côté droit du diamètre $K\,L$ en la parabole, l'équation qui explique le rapport qui est entre x et y est

$$y^3 - by^2 - cdy + bcd + dxy = 0,$$

d'où ôtant x on a

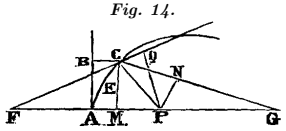
$$y^3 - by^2 - cdy + bcd + dy\sqrt{s^2 - v^2 + 2vy - y^2} = 0;$$

et remettant en ordre ces termes par le moyen de la multiplication, il vient

$$y^6 - 2by^5 + (b^2 - 2cd + d^2)y^4 + (4bcd - 2d^2v)y^3 + (c^2d^2 - d^2s^2 + d^2v^2 - 2b^2cd)y^2 - 2bc^2d^2y + b^2c^2d^2 = 0,$$

et ainsi des autres. Même, encore que les points de la ligne courbe ne se rapportassent pas en la façon que j'ai dit à ceux d'une ligne droite, mais en toute autre qu'on sauroit imaginer, on ne laisse pas de pouvoir toujours avoir une telle équation. Comme si $C\,E$ *(fig. 14)* est une ligne qui ait tel rapport aux trois points F, G et A, que les lignes droites tirées de chacun de ses points comme C jusques au point F, surpassent la ligne $F\,A$ d'une quantité qui ait certaine proportion donnée à une autre quantité dont $G\,A$ surpasse les lignes

Fig. 14.



tirées des mêmes points jusques à G. Faisons $G\,A = b$, $A\,F = c$, et prenant à discrétion le point C dans la courbe, que la quantité dont $C\,F$ surpasse $F\,A$, soit à celle dont $G\,A$ surpasse $G\,C$, comme d à e ; en sorte que si cette quantité qui est indéterminée se nomme z, $C\,F$ est $c + z$, et $G\,C$ est $b - \frac{e}{d}z$. Puis posant $M\,A = y$, $G\,M$ est $b - y$, et $F\,M$ est $c + y$, et à cause du triangle rectangle $C\,M\,G$, ôtant le carré de $G\,M$ du carré de $G\,C$, on a le carré de $C\,M$, qui est

$$\frac{e^2}{d^2}z^2 - \frac{2be}{d}z + 2by - y^2;$$

puis ôtant le carré de $F\,M$ du carré de $C\,F$, on a encore le carré de $C\,M$ en d'autres termes, à savoir $z^2 + 2cz - 2cy - y^2$; et ces termes étant égaux aux précédents, ils font connoître y ou $M\,A$, qui est

$$\frac{d^2z^2 + 2cd^2z - e^2z^2 + 2bdez}{2bd^2 + 2cd^2},$$

et substituant cette somme au lieu de y dans le carré de $C\,M$, on trouve qu'il s'exprime en ces termes :

$$\frac{bd^2z^2 + ce^2z^2 + 2bcd^2z - 2bcdez}{bd^2 + cd^2} - y^2.$$

Puis supposant que la ligne droite $P\,C$ rencontre la courbe à angles droits au point C, et faisant $P\,C = s$ et $P\,A = v$ comme devant, PM est $v - y$; et à cause du triangle rectangle $P\,C\,M$, on a $s^2 - v^2 + 2vy - y^2$ pour le carré de $C\,M$, ou derechef, ayant au lieu de y substitué la somme qui lui est égale, il vient

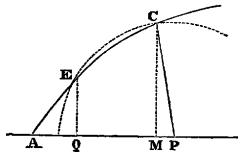
$$z^2 + \frac{2bcd^2z - 2bcdez - 2cd^2vz - 2bdevz - bd^2s^2 + bd^2v^2 - cd^2s^2 + cd^2v^2}{bd^2 + ce^2 + e^2v - d^2v} = 0$$

pour l'équation que nous cherchions.

Or après qu'on a trouvé une telle équation, au lieu de s'en servir pour connoître les quantités x, ou y, ou z, qui sont déjà données, puisque le point C est donné, on la doit employer à trouver v ou s, qui déterminent le point P qui est demandé. Et à cet effet il faut considérer que si ce point P est tel qu'on le désire, le cercle dont il sera le centre, et qui passera par le point C, y touchera la ligne courbe $C\,E$ sans la couper ; mais que si ce point P est tant soit peu plus proche ou plus éloigné du point A qu'il ne doit, ce cercle coupera la courbe, non

seulement au point C, mais aussi nécessairement en quelque autre. Puis il faut aussi considérer que lorsque ce cercle coupe la ligne courbe $C\,E$, l'équation par laquelle on cherche la quantité x ou y, ou quelque autre semblable, en supposant $P\,A$ et $P\,C$ être connues, contient nécessairement deux racines qui sont inégales. Car par exemple, si ce cercle coupe la courbe aux points C et E *(fig. 15)*, ayant tiré $E\,Q$ parallèle à $C\,M$, les noms des quantités indéterminées x et y conviendront aussi bien aux lignes $E\,Q$ et $Q\,A$ qu'à $C\,M$ et $M\,A$; puis $P\,E$ est égale à $P\,C$ à cause du cercle, si bien que cherchant les lignes $E\,Q$ et $Q\,A$, par $P\,E$ et $P\,A$ qu'on suppose comme données, on aura la même équation que si on cherchoit

Fig. 15.



$C\,M$ et $M\,A$ par $P\,C$, $P\,A$; d'où il suit évidemment que la valeur de x ou de y, ou de telle autre quantité qu'on aura supposée, sera double en cette équation, c'est-à-dire qu'il y aura deux racines inégales entre elles, et dont l'une sera $C\,M$, l'autre $E\,Q$, si c'est x qu'on cherche, ou bien l'une sera $M\,A$ et l'autre $Q\,A$, si c'est y ; et ainsi des autres. Il est vrai que si le point E ne se trouve pas du même côté de la courbe que le point C, il n'y aura que l'une de ces deux racines qui soit vraie, et l'autre sera renversée ou moindre que rien : mais plus ces deux points C et E sont proches l'un de l'autre, moins il y a de différence entre ces deux racines ; et enfin elles sont entièrement égales, s'ils sont tous deux joints en un, c'est-à-dire si le cercle qui passe par C y touche la courbe $C\,E$ sans la couper.

De plus il faut considérer que lorsqu'il y a deux racines égales en une équation, elle a nécessairement la même forme que si on multiplie par soi-même la quantité qu'on y suppose être inconnue, moins la quantité connue qui lui est égale, et qu'après cela, si cette dernière somme n'a pas tant de dimensions que la précédente, on la multiplie par une autre somme qui en ait autant qu'il lui en manque, afin qu'il puisse y avoir séparément équation entre chacun des termes de l'une et chacun des termes de l'autre.

Comme par exemple, je dis que la première équation trouvée ci-dessus, à savoir

$$y^2 + \frac{qry - 2qvy + qv^2 - qs^2}{q - r},$$

doit avoir la même forme que celle qui se produit en faisant e égal à y, et multipliant $y - e$ par soi-même, d'où il vient $y^2 - 2ey + e^2$, en sorte qu'on peut comparer séparément chacun de leurs termes, et dire que puisque le premier qui est y^2, est tout le même en l'une qu'en l'autre, le second qui est en l'une

$\frac{qry - 2qvy}{q - r}$, est égal au second de l'autre qui est $-2ey$; d'où cherchant la quantité v qui est la ligne $P\,A$, on a $v = e - \frac{r}{q}e + \frac{1}{2}r$, ou bien à cause que nous avons supposé e égal à y, on a $v = y - \frac{r}{q}y + \frac{1}{2}r$. Et ainsi on pourroit trouver s par le troisième terme $e^2 = \frac{qv^2 - qs^2}{q - r}$; mais pourceque la quantité v détermine assez le point P, qui est le seul que nous cherchions, on n'a pas besoin de passer outre.

Tout de même la seconde équation trouvée ci-dessus, à savoir

$$y^6 - 2by^5 + (b^2 - 2cd + d^2)y^4 + (4bcd - 2d^2v)y^3 + (c^2d^2 - 2b^2cd + d^2v^2 - d^2s^2)y^2 - 2bc^2d^2y + b^2c^2d^2,$$

doit avoir même forme que la somme qui se produit lorsqu'on multiplie

$$y^2 - 2ey + e^2 \quad \text{par} \quad y^4 + fy^3 + g^2y^2 + h^3y + k^4$$

qui est

$$y^6 + (f - 2e)y^5 + (g^2 - 2ef + e^2)y^4 + (h^3 - 2eg^2 + e^2f)y^3 + (k^4 - 2eh^3 + e^2g^2)y^2 + (e^2h^3 - 2ek^4)y + e^2k^4;$$

de façon que de ces deux équations j'en tire six autres qui servent à connoître les six quantités f, g, h, k, v et s. D'où il est fort aisé à entendre que, de quelque genre que puisse être la ligne courbe proposée, il vient toujours par cette façon de procéder autant d'équations qu'on est obligé de supposer de quantités qui sont inconnues. Mais pour démêler par ordre ces équations, et trouver enfin la quantité v, qui est la seule dont on a besoin, et à l'occasion de laquelle on cherche les autres, il faut premièrement par le second terme chercher f, la première des quantités inconnues de la dernière somme, et on trouve

$$f = 2e - 2b.$$

Puis par le dernier, il faut chercher k, la dernière des quantités inconnues de la même somme, et on trouve

$$k^4 = \frac{b^2c^2d^2}{e^2}.$$

Puis par le troisième terme, il faut chercher g, la seconde quantité, et on a

$$g^2 = 3e^2 - 4be - 2cd + b^2 + d^2.$$

Puis par la pénultième, il faut chercher h, la pénultième quantité, qui est

$$h^3 = \frac{2b^2c^2d^2}{e^3} - \frac{2bc^2d^2}{e^2}.$$

Et ainsi il faudroit continuer suivant ce même ordre jusques à la dernière, s'il y en avoit d'avantage en cette somme ; car c'est chose qu'on peut toujours faire en même façon.

Puis, par le terme qui suit en ce même ordre, qui est ici le quatrième, il faut chercher la quantité v, et on a

$$v = \frac{2e^3}{d^2} - \frac{3be^2}{d^2} + \frac{b^2e}{d^2} - \frac{2ce}{d} + e + \frac{2bc}{d} + \frac{bc^2}{e^2} - \frac{b^2c^2}{e^3};$$

ou mettant y au lieu de e qui lui est égal, on a

$$v = \frac{2y^3}{d^2} - \frac{3by^2}{d^2} + \frac{b^2y}{d^2} - \frac{2cy}{d} + y + \frac{2bc}{d} + \frac{bc^2}{y^2} - \frac{b^2c^2}{y^3}$$

pour la ligne $A\,P$.

Et ainsi la troisième équation, qui est

$$z^2 + \frac{2bcd^2z - 2bcdez - 2cd^2vz - 2bdevz - bd^2s^2 + bd^2v^2 - cd^2s^2 + cd^2v^2}{bd^2 + ce^2 + e^2v - d^2v}$$

a la même forme que

$$z^2 - 2fz + f^2,$$

en supposant f égal à z, si bien qu'il y a derechef équation entre $-2f$ ou $-2z$, et

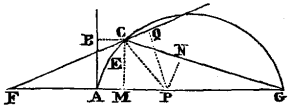
$$\frac{2bcd^2 - 2bcde - 2cd^2v - 2bdev}{bd^2 + ce^2 + e^2v - d^2v},$$

d'où on connoît que la quantité v est

$$\frac{bcd^2 - bcde + bd^2z + ce^2z}{cd^2 + bde - e^2z + d^2z}.$$

C'est pourquoi, composant la ligne $A\,P$ *(fig. 16)* de cette somme égale à v, dont toutes les quantités sont connues, et tirant du point P ainsi trouvé, une

Fig. 16.



ligne droite vers C, elle y coupe la courbe $C\,E$ à angles droits ; qui est ce qu'il falloit faire. Et je ne vois rien qui empêche qu'on n'étende ce problème en même façon à toutes les lignes courbes qui tombent sous quelque calcul géométrique.

Même il est à remarquer, touchant la dernière somme, qu'on prend à discrétion pour remplir le nombre des dimensions de l'autre somme lorsqu'il y en

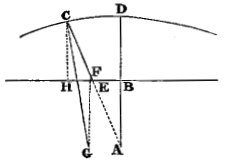
manque, comme nous avons pris tantôt $y^4 + fy^3 + g^2y^2 + h^3y + k^4$, que les signes $+$ et $-$ y peuvent être supposés tels qu'on veut, sans que la ligne v ou $A\,P$ se trouve diverse pour cela, comme vous pourrez aisément voir par expérience ; car s'il falloit que je m'arrêtasse à démontrer tous les théorèmes dont je fais quelque mention, je serois contraint d'écrire un volume beaucoup plus gros que je ne désire. Mais je veux bien en passant vous avertir que l'invention de supposer deux équations de même forme, pour comparer séparément tous les termes de l'une à ceux de l'autre, et ainsi en faire naître plusieurs d'une seule, dont vous avez vu ici un exemple, peut servir à une infinité d'autres problèmes, et n'est pas l'une des moindres de la méthode dont je me sers.

Je n'ajoute point les constructions par lesquelles on peut décrire les contingentes ou les perpendiculaires cherchées, ensuite du calcul que je viens d'expliquer, à cause qu'il est toujours aisé de les trouver, bien que souvent on ait besoin d'un peu d'adresse pour les rendre courtes et simples.

Exemple de la construction de ce problème en la conchoïde.

Comme par exemple, si $D\,C$ *(fig. 17)* est la première conchoïde des anciens,

Fig. 17.



dont A soit le pôle et $B\,H$ la règle, en sorte que toutes les lignes droites qui regardent vers A, et sont comprises entre la courbe $C\,D$ et la droite $B\,H$, comme $D\,B$ et $C\,E$, soient égales, et qu'on veuille trouver la ligne $C\,G$ qui la coupe au point C à angles droits, on pourroit, en cherchant dans la ligne $B\,H$ le point par où cette ligne $C\,G$ doit passer, selon la méthode ici expliquée, s'engager dans un calcul autant ou plus long qu'aucun des précédents : et toutefois la construction qui devroit après en être déduite est fort simple ; car il ne faut que prendre $C\,F$ en la ligne droite $C\,A$, et la faire égale à $C\,H$ qui est perpendiculaire sur $H\,B$; puis du point F tirer $F\,G$ parallèle à $B\,A$ et égale à $E\,A$; au moyen de quoi on a le point G, par lequel doit passer $C\,G$ la ligne cherchée.

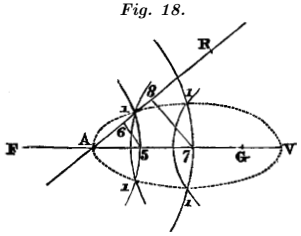
Explication de quatre nouveaux genres d'ovales qui servent à l'optique.

Au reste, afin que vous sachiez que la considération des lignes courbes ici proposée n'est pas sans usage, et qu'elles ont diverses propriétés qui ne cèdent en rien à celles des sections coniques, je veux encore ajouter ici l'explication de certaines ovales que vous verrez être très utiles pour la théorie de la catoptrique et de la dioptrique. Voici la façon dont je les décris :

Premièrement, ayant tiré les lignes droites $F\,A$ et $A\,R$ *(fig. 18)*, qui s'entrecoupent au point A, sans qu'il importe à quels angles, je prends en l'une le point F à discrétion, c'est-à-dire plus ou moins éloigné du point A, selon que je veux faire ces ovales plus ou moins grandes, et de ce point F, comme centre, je décris

un cercle qui passe quelque peu au-delà du point A, comme par le point 5 ; puis de ce point 5 je tire la ligne droite 5 6, qui coupe l'autre au point 6, en sorte que $A\,6$ soit moindre que $A\,5$ selon telle proportion donnée qu'on veut, à savoir

Fig. 18.

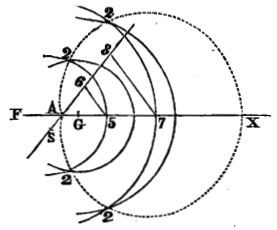


selon celle qui mesure les réfractions si on s'en veut servir pour la dioptrique. Après cela je prends aussi le point G en la ligne $F\,A$ du côté où est le point 5, à discrétion, c'est-à-dire en faisant que les lignes $A\,F$ et $G\,A$ ont entre elles telle proportion donnée qu'on veut. Puis je fais $R\,A$ égale à $G\,A$ en la ligne $A\,6$, et du centre G décrivant un cercle dont le rayon soit égal à $R\,6$, il coupe l'autre cercle de part et d'autre au point 1, qui est l'un de ceux par où doit passer la première des ovales cherchées. Puis derechef du centre F je décris un cercle qui passe un peu au-deçà ou au-delà du point 5, comme par le point 7, et ayant tiré la ligne droite 7 8 parallèle à 5 6, du centre G je décris un autre cercle dont le rayon est égal à la ligne $R\,8$, et ce cercle coupe celui qui passe par le point 7 au point 1, qui est encore l'un de ceux de la même ovale ; et ainsi on en peut trouver autant d'autres qu'on voudra, en tirant derechef d'autres lignes parallèles à 7 8, et d'autres cercles des centres F et G.

Pour la seconde ovale il n'y a point de différence, sinon qu'au lieu de $A\,R$ *(fig. 19)* il faut de l'autre côté du point A prendre $A\,S$ égal à $A\,G$, et que le rayon du cercle décrit du centre G, pour couper celui qui est décrit du centre F et qui passe par le point 5, soit égal à la ligne $S\,6$, ou qu'il soit égal à $S\,8$, si c'est pour couper celui qui passe par le point 7, et ainsi des autres ; au moyen de quoi ces cercles s'entre-coupent aux points marqués 2, 2, qui sont ceux de cette seconde ovale $A\,2\,X$.

Pour la troisième et la quatrième, au lieu de la ligne $A\,G$ il faut prendre $A\,H$ *(fig. 21 et 22)* de l'autre côté du point A, à savoir du même qu'est le point F ; et il y a ici de plus à observer que cette ligne $A\,H$ doit être plus grande que $A\,F$, laquelle peut même être nulle, en sorte que le point F se rencontre où est le point A en la description de toutes ces ovales. Après cela les lignes $A\,R$ et $A\,S$ étant égales à $A\,H$, pour décrire la troisième ovale $A\,3\,Y$, je fais un cercle du centre H, dont le rayon est égal à $S\,6$, qui coupe au point 3 celui du centre F, qui passe par le point 5 ; et un autre dont le rayon est égal à $S\,8$, qui coupe celui qui passe par le point 7 au point aussi marqué 3, et ainsi des autres. Enfin, pour

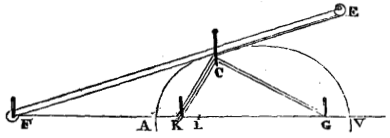
Fig. 19.



la dernière ovale, je fais des cercles du centre H, dont les rayons sont égaux aux lignes $R\,6$, $R\,8$, et semblables, qui coupent les autres cercles aux points marqués 4.

On pourroit encore trouver une infinité d'autres moyens pour décrire ces mêmes ovales ; comme par exemple, on peut tracer la première $A\,V$ *(fig. 20)*, lorsqu'on suppose les lignes $F\,A$ et $A\,G$ être égales, si on divise la toute $F\,G$ au

Fig. 20.



point L, en sorte que $F\,L$ soit à $L\,G$ comme $A\,5$ à $A\,6$, c'est-à-dire qu'elles aient la proportion qui mesure les réfractions. Puis ayant divisé $A\,L$ en deux parties égales au point K, qu'on fasse tourner une règle comme $E\,F$ autour du point F, en pressant du doigt C la corde $E\,C$, qui étant attachée au bout de cette règle vers E, se replie de C vers K, puis de K derechef vers C, et de C vers G, où son autre bout soit attaché, en sorte que la longueur de cette corde soit composée de celle des lignes $G\,A$, plus $A\,L$, plus $F\,E$, moins $A\,F$; et ce sera le mouvement du point C qui décrira cette ovale, à l'imitation de ce qui a été dit en la dioptrique de l'ellipse et de l'hyperbole ; mais je ne veux point m'arrêter plus long-temps sur ce sujet.

Or, encore que toutes ces ovales semblent être quasi de même nature, elles sont néanmoins de quatre divers genres, chacun desquels contient sous soi une infinité d'autres genres, qui derechef contiennent chacun autant de diverses espèces que fait le genre des ellipses ou celui des hyperboles ; car selon que la proportion

qui est entre les lignes $A5$, $A6$, ou semblables, est différente, le genre subalterne de ces ovales est différent ; puis selon que la proportion qui est entre les lignes AF et AG ou AH est changée, les ovales de chaque genre subalterne changent d'espèce ; et selon que AG ou AH est plus ou moins grande, elles sont diverses en grandeur ; et si les lignes $A5$ et $A6$ sont égales, au lieu des ovales du premier genre ou du troisième, on ne décrit que des lignes droites ; mais au lieu de celles du second on a toutes les hyperboles possibles, et au lieu de celles du dernier toutes les ellipses.

Les propriétés de ces ovales touchant les réflexions et les réfractions.

Outre cela, en chacune de ces ovales il faut considérer deux parties qui ont diverses propriétés : à savoir en la première, la partie qui est vers A *(fig. 18)*, fait que les rayons qui étant dans l'air viennent du point F, se retournent tous vers le point G, lorsqu'ils rencontrent la superficie convexe d'un verre dont la superficie est $1A1$, et dans lequel les réfractions se font telles que, suivant ce qui a été dit en la Dioptrique, elles peuvent toutes être mesurées par la proportion qui est entre les lignes $A5$ et $A6$ ou semblables, par l'aide desquelles on a décrit cette ovale.

Mais la partie qui est vers V fait que les rayons qui viennent du point G se réfléchiroient tous vers F, s'ils y rencontroient la superficie concave d'un miroir dont la figure fût $1V1$, et qui fût de telle matière qu'il diminuât la force de ces rayons selon la proportion qui est entre les lignes $A5$ et $A6$; car de ce qui a été démontré en la Dioptrique, il est évident que, cela posé, les angles de la réflexion seroient inégaux, aussi bien que sont ceux de la réfraction, et pourroient être mesurés en même sorte.

En la seconde ovale la partie $2A2$ *(fig. 19)* sert encore pour les réflexions dont on suppose les angles être inégaux ; car étant en la superficie d'un miroir composé de même matière que le précédent, elle feroit tellement réfléchir tous les rayons qui viendroient du point G, qu'ils sembleroient après être réfléchis venir du point F. Et il est à remarquer qu'ayant fait la ligne AG beaucoup plus grande que AF, ce miroir seroit convexe au milieu vers A, et concave aux extrémités ; car telle est la figure de cette ligne, qui en cela représente plutôt un cœur qu'une ovale.

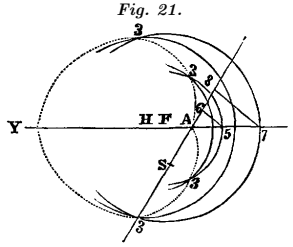
Mais son autre partie $X2$ sert pour les réfractions, et fait que les rayons qui étant dans l'air tendent vers F, se détournent vers G en traversant la superficie d'un verre qui en ait la figure.

La troisième ovale sert toute aux réfractions, et fait que les rayons qui étant dans l'air tendent vers F *(fig. 21)*, se vont rendre vers H dans le verre, après qu'ils ont traversé sa superficie dont la figure est $A3Y3$, qui est convexe partout, excepté vers A où elle est un peu concave, en sorte qu'elle a la figure d'un cœur aussi bien que la précédente ; et la différence qui est entre les deux parties de cette ovale consiste en ce que le point F est plus proche de l'une que n'est le point H, et qu'il est plus éloigné de l'autre que ce même point H.

En même façon la dernière ovale sert toute aux réflexions, et fait que si les rayons qui viennent du point H *(fig. 22)* rencontroient la superficie concave d'un miroir de même matière que les précédents, et dont la figure fût $A4Z4$, ils se réfléchiroient tous vers F.

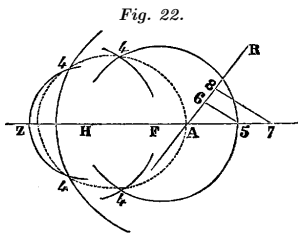
De façon qu'on peut nommer les points F et G ou H les points brûlants de

Fig. 21.



ces ovales, à l'exemple de ceux des ellipses et des hyperboles, qui ont été ainsi nommés en la Dioptrique.

Fig. 22.



Démonstration des propriétés de ces ovales touchant les réflexions et réfractions.

J'omets quantité d'autres réfractions et réflexions qui sont réglées par ces mêmes ovales, car n'étant que les converses ou les contraires de celles-ci, elles en peuvent facilement être déduites. Mais il ne faut pas que j'omette la démonstration de ce que j'ai dit ; et à cet effet prenons, par exemple, le point C *(fig. 16)* à discrétion en la première partie de la première de ces ovales ; puis tirons la ligne droite CP qui coupe la courbe au point C à angles droits, ce qui est facile par le problème précédent ; car prenant b pour AG, c pour AF, $c+z$ pour CF, et supposant que la proportion qui est entre d et e, que je prendrai ici toujours pour celle qui mesure les réfractions du verre proposé, désigne aussi celle qui est entre les lignes $A5$ et $A6$ ou semblables, qui ont servi pour décrire cette ovale, ce qui donne $b-\frac{e}{d}z$ pour CG, on trouve que la ligne AP est

$$\frac{bcd^2-bcde+bd^2z+ce^2z}{cd^2+bde-e^2z+d^2z},$$

ainsi qu'il a été montré ci-dessus (p. 29). De plus, du point P ayant tiré PQ à

angles droits sur la droite $C\,F$, et $P\,N$ aussi à angles droits sur $C\,G$, considérons que si $P\,Q$ est à $P\,N$ comme d est à e, c'est-à-dire comme les lignes qui mesurent les réfractions du verre convexe $A\,C$, le rayon qui vient du point F au point C, doit tellement s'y courber en entrant dans ce verre, qu'il s'aille rendre après vers G, ainsi qu'il est très évident de ce qui a été dit en la Dioptrique. Puis enfin voyons par le calcul s'il est vrai que $P\,Q$ soit à $P\,N$ comme d est à e. Les triangles rectangles $P\,Q\,F$ et $C\,M\,F$ sont semblables ; d'où il suit que $C\,F$ est à $C\,M$ comme $F\,P$ est à $P\,Q$, et par conséquent que $P\,F$ étant multipliée par $C\,M$ et divisée par $C\,F$ est égale à $P\,Q$. Tout de même les triangles rectangles $P\,N\,G$ et $C\,M\,G$ sont semblables ; d'où il suit que $G\,P$ multipliée par $C\,M$ et divisée par $C\,G$ est égale à $P\,N$. Puis à cause que les multiplications ou divisions qui se font de deux quantités par une même ne changent point la proportion qui est entre elles, si $P\,F$ multipliée par $C\,M$ et divisée par $C\,F$, est à $G\,P$ multipliée aussi par $C\,M$ et divisée par $C\,G$, comme d est à e, en divisant l'une et l'autre de ces deux sommes par $C\,M$, puis les multipliant toutes deux par $C\,F$ et derechef par $C\,G$, il reste $E\,P$ multipliée par $C\,G$ qui doit être à $G\,P$ multipliée par $C\,F$, comme d est à e. Or par la construction $F\,P$ est

$$c + \frac{bcd^2 - bcde + bd^2z + ce^2z}{cd^2 + bde - e^2z + d^2z},$$

ou bien

$$F\,P = \frac{bcd^2 + c^2d^2 + bd^2z + cd^2z}{cd^2 + bde - e^2z + d^2z},$$

et $C\,G$ est $b - \frac{e}{d}z$; si bien que, multipliant $F\,P$ par $C\,G$, il vient

$$\frac{b^2cd^2 + bc^2d^2 + b^2d^2z + bcd^2z - bcdez - c^2dez - bdez^2 - cdez^2}{cd^2 + bde - e^2z + d^2z},$$

puis $G\,P$ est

$$b - \frac{bcd^2 - bcde + bd^2z + ce^2z}{cd^2 + bde - e^2z + d^2z},$$

ou bien

$$G\,P = \frac{b^2de + bcde - be^2z - ce^2z}{cd^2 + bde - e^2z + d^2z},$$

et $C\,F$ est $c + z$; si bien qu'en multipliant $G\,P$ par $C\,F$ il vient

$$\frac{b^2cde + bc^2de + b^2dez + bcdez - bce^2z - c^2e^2z - be^2z^2 - ce^2z^2}{cd^2 + bde - e^2z + d^2z}.$$

Et pourceque la première de ces sommes divisée par d est la même que la seconde divisée par e, il est manifeste que $F\,P$ multipliée par $C\,G$, est à $G\,P$ multipliée par $C\,F$, c'est-à-dire que $P\,Q$ est à $P\,N$ comme d est à e, qui est tout ce qu'il falloit démontrer.

Et sachez que cette même démonstration s'étend à tout ce qui a été dit des autres réfractions ou réflexions qui se font dans les ovales proposées, sans qu'il

y faille changer aucune chose que les signes + et − du calcul ; c'est pourquoi chacun les peut aisément examiner de soi-même, sans qu'il soit besoin que je m'y arrête.

Mais il faut maintenant que je satisfasse à ce que j'ai omis en la Dioptrique, lorsqu'après avoir remarqué qu'il peut y avoir des verres de plusieurs diverses figures qui fassent aussi bien l'un que l'autre que les rayons venant d'un même point de l'objet s'assemblent tous en un autre point après les avoir traversés ; et qu'entre ces verres, ceux qui sont fort convexes d'un côté et concaves de l'autre ont plus de force pour brûler que ceux qui sont également convexes des deux côtés ; au lieu que tout au contraire ces derniers sont les meilleurs pour les lunettes. Je me suis contenté d'expliquer ceux que j'ai cru être les meilleurs pour la pratique, en supposant la difficulté que les artisans peuvent avoir à les tailler. C'est pourquoi, afin qu'il ne reste rien à souhaiter touchant la théorie de cette science, je dois expliquer encore ici la figure des verres qui, ayant l'une de leurs superficies autant convexe ou concave qu'on voudra, ne laissent pas de faire que tous les rayons qui viennent vers eux d'un même point, ou parallèles, s'assemblent après en un même point ; et celles des verres qui font le semblable, étant également convexes des deux côtés, ou bien la convexité de l'une de leurs superficies ayant la proportion donnée à celle de l'autre.

Comment on peut faire un verre autant convexe ou concave, en l'une de ses superficies, qu'on voudra qui rassemble à un point donné tous les rayons qui viennent d'un autre point donné.

Posons pour le premier cas, que les points G, Y, C et F *(fig. 23 et 24)* étant donnés, les rayons qui viennent du point G ou bien qui sont parallèles à $G\,A$ se doivent assembler au point F, après avoir traversé un verre si concave, que Y étant le milieu de sa superficie intérieure ; l'extrémité en soit au point C, en sorte que la corde $C\,M\,C$ et la flèche $Y\,M$ de l'arc $C\,Y\,C$ sont données. La question va là, que premièrement il faut considérer de laquelle des ovales expliquées la

Fig. 23.

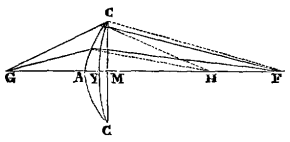
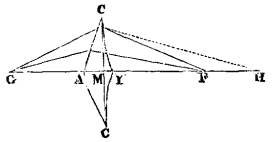


Fig. 24.



superficie du verre $Y\,C$ doit avoir la figure, pour faire que tous les rayons qui étant dedans tendent vers un même point, comme vers H, qui n'est pas encore connu, s'aillent rendre vers un autre, à savoir vers F, après en être sortis. Car il n'y a aucun effet touchant le rapport des rayons, changé par réflexion ou réfraction d'un point à un autre, qui ne puisse être causé par quelqu'une de ces ovales ; et on voit aisément que celui-ci le peut être par la partie de la troisième ovale qui a tantôt été marquée $3\,A\,3$ *(fig. 21)*, ou par celle de la même qui a été marquée $3\,Y\,3$, ou enfin par la partie de la seconde qui a été marquée $2\,X\,2$ *(fig. 19)*. Et pourceque ces trois tombent ici sous même calcul, on doit, tant pour l'une que pour l'autre, prendre Y *(fig. 23 et 24)* pour leur sommet, C pour l'un des points de leur circonférence, et F pour l'un de leurs points brûlants ;

après quoi il ne reste plus à chercher que le point H qui doit être l'autre point brûlant. Et on le trouve en considérant que la différence qui est entre les lignes $F\,Y$ et $F\,C$ doit être à celle qui est entre les lignes $H\,Y$ et $H\,C$ comme d est à e, c'est-à-dire comme la plus grande des lignes qui mesurent les réfractions du verre proposé est à la moindre, ainsi qu'on peut voir manifestement de la description de ces ovales. Et pourceque les lignes $F\,Y$ et $F\,C$ sont données, leur différence l'est aussi, et ensuite celle qui est entre $H\,Y$ et $H\,C$, pourceque la proportion qui est entre ces deux différences est donnée. Et de plus, à cause que $Y\,M$ est donnée, la différence qui est entre $M\,H$ et $H\,C$ l'est aussi ; et enfin pourceque $C\,M$ est donnée, il ne reste plus qu'à trouver $M\,H$ le côté du triangle rectangle $C\,M\,H$ dont on a l'autre côté $C\,M$, et on a aussi la différence qui est entre $C\,H$ la base et $M\,H$ le côté demandé ; d'où il est aisé de le trouver : car si on prend k pour l'excès de $C\,H$ sur $M\,H$, et n pour la longueur de la ligne $C\,M$, on aura $\frac{n^2}{2k} - \frac{1}{2}k$ pour $M\,H$. Et après avoir ainsi le point H, s'il se trouve plus loin du point Y *(fig. 24)* que n'en est le point F, la ligne $C\,Y$ doit être la première partie de l'ovale du troisième genre, qui a tantôt été nommée $3\,A\,3$ *(fig. 21)*. Mais si $H\,Y$ *(fig. 23)* est moindre que $F\,Y$: ou bien elle surpasse $H\,F$ de tant, que leur différence est plus grande à raison de la toute $F\,Y$ que n'est e la moindre des lignes qui mesurent les réfractions comparée avec d la plus grande, c'est-à-dire que faisant $H\,F = c$, et $H\,Y = c + h$, dh est plus grande que $2ce + eh$, et lors $C\,Y$ doit être la seconde partie de la même ovale du troisième genre, qui a tantôt été nommée $3\,Y\,3$ *(fig. 21)* : ou bien dh est égale ou moindre que $2ce + eh$, et lors $C\,Y$ *(fig. 23)* doit être la seconde partie de l'ovale du second genre, qui a ci-dessus été nommée $2\,X\,2$ *(fig. 19)* : et enfin si le point H *(fig. 23)* est le même que le point F, ce qui n'arrive que lorsque $F\,Y$ et $F\,C$ sont égales, cette ligne $Y\,C$ est un cercle.

Après cela il faut chercher $C\,A\,C$ l'autre superficie de ce verre, qui doit être une ellipse dont H soit le point brûlant, si on suppose que les rayons qui tombent dessus soient parallèles ; et lors il est aisé de la trouver. Mais si on suppose qu'ils viennent du point G, ce doit être la première partie d'une ovale du premier genre dont les deux points brûlants soient G et H, et qui passe par le point C ; d'où on trouve le point A pour le sommet de cette ovale, en considérant que $G\,C$ doit être plus grande que $G\,A$ d'une quantité qui soit à celle dont $H\,A$ surpasse $H\,C$, comme d à e ; car ayant pris k pour la différence qui est entre $C\,H$ et $H\,M$, si on suppose x pour $A\,M$, on aura $x - k$ pour la différence qui est entre $A\,H$ et $C\,H$; puis si on prend g pour celle qui est entre $G\,C$ et $G\,M$ qui sont données, on aura $g + x$ pour celle qui est entre $G\,C$ et $G\,A$; et pourceque cette dernière $g + x$ est à l'autre $x - k$ comme d est à e, on a

$$ge + ex = dx - dk,$$

ou bien $\frac{ge + dk}{d - e}$ pour la ligne x ou $A\,M$, par laquelle on détermine le point A qui étoit cherché.

Posons maintenant pour l'autre cas, qu'on ne donne que les points G, C et F *(fig. 24)*, avec la proportion qui est entre les lignes $A\,M$ et $Y\,M$, et qu'il faille

trouver la figure du verre $A\,C\,Y$ qui fasse que tous les rayons qui viennent du point G s'assemblent au point F.

Comment on peut faire un verre qui ait le même effet que le précédent, et que la convexité de l'une de ses superficies ait la proportion donnée avec celle de l'autre.

On peut derechef ici se servir de deux ovales dont l'une $A\,C$ ait G et H pour ses points brûlants, et l'autre $C\,Y$ ait F et H pour les siens. Et pour les trouver, premièrement, supposant le point H, qui est commun à toutes deux, être connu, je cherche $A\,M$ par les trois points G, C, H, en la façon tout maintenant expliquée, à savoir, prenant k pour la différence qui est entre $C\,H$ et $H\,M$, et g pour celle qui est entre $G\,C$ et $G\,M$, et $A\,C$ étant la première partie de l'ovale du premier genre, j'ai $\frac{ge + dk}{d - e}$ pour $A\,M$; puis je cherche aussi $M\,Y$ par les trois points F, C, H, en sorte que $C\,Y$ soit la première partie d'une ovale du troisième genre ; et prenant y pour $M\,Y$, et f pour la différence qui est entre $C\,F$ et $F\,M$, j'ai $f + y$ pour celle qui est entre $C\,F$ et $F\,Y$; puis ayant déjà k pour celle qui est entre $C\,H$ et $H\,M$, j'ai $k + y$ pour celle qui est entre $C\,H$ et $H\,Y$, que je sais devoir être à $f + y$ comme e est à d, à cause de l'ovale du troisième genre, d'où je trouve que y ou $M\,Y$ est $\frac{fe - dk}{d - e}$; puis joignant ensemble les deux quantités trouvées pour $A\,M$ et $M\,Y$, je trouve $\frac{ge + fe}{d - e}$ pour la toute $A\,Y$: d'où il suit que, de quelque côté que soit supposé le point H, cette ligne $A\,Y$ est toujours composée d'une quantité qui est à celle dont les deux ensemble GC et $C\,F$ surpassent la toute $G\,F$, comme e, la moindre des deux lignes qui servent à mesurer les réfractions du verre proposé, est à $d - e$ la différence qui est entre ces deux lignes, ce qui est un assez beau théorème. Or, ayant ainsi la toute $A\,Y$, il la faut couper selon la proportion que doivent avoir ses parties $A\,M$ et $M\,Y$; au moyen de quoi, pourcequ'on a déjà le point M, on trouve aussi les points A et Y, et ensuite le point H par le problème précédent. Mais auparavant il faut regarder si la ligne $A\,M$ ainsi trouvée est plus grande que $\frac{ge}{d - e}$, ou plus petite, ou égale. Car si elle est plus grande, on apprend de là que la courbe $A\,C$ doit être la première partie d'une ovale du premier genre, et $C\,Y$ la première d'une du troisième, ainsi qu'elles ont été ici supposées ; au lieu que si elle est plus petite, cela montre que c'est $C\,Y$ qui doit être la première partie d'une ovale du premier genre, et que $A\,C$ doit être la première d'une du troisième ; enfin si $A\,M$ est égale à $\frac{ge}{d - e}$, les deux courbes $A\,C$ et $C\,Y$ doivent être deux hyperboles.

On pourroit étendre ces deux problèmes à une infinité d'autres cas que je ne m'arrête pas à déduire, à cause qu'ils n'ont eu aucun usage en la dioptrique.

On pourroit aussi passer outre et dire (lorsque l'une des superficies du verre est donnée, pourvu qu'elle ne soit que toute plate, ou composée de sections coniques ou de cercles) comment on doit faire son autre superficie, afin qu'il transmette tous les rayons d'un point donné à un autre point aussi donné ; car ce n'est rien de plus difficile que ce que je viens d'expliquer, ou plutôt c'est chose beaucoup plus facile à cause que le chemin en est ouvert. Mais j'aime mieux que d'autres le cherchent, afin que s'ils ont encore un peu de peine à le trouver, cela leur fasse d'autant plus estimer l'invention des choses qui sont ici démontrées.

Au reste je n'ai parlé en tout ceci que des lignes courbes qu'on peut décrire sur une superficie plate ; mais il est aisé de rapporter ce que j'en ai dit à toutes celles qu'on sauroit imaginer être formées par le mouvement régulier des points de quelque corps dans un espace qui a trois dimensions : à savoir, en tirant deux perpendiculaires de chacun des points de la ligne courbe qu'on veut considérer, sur deux plans qui s'entre-coupent à angles droits, l'une sur l'un et l'autre sur l'autre ; car les extrémités de ces perpendiculaires décrivent deux autres lignes courbes, une sur chacun de ces plans, desquelles on peut en la façon ci-dessus expliquée déterminer tous les points et les rapporter à ceux de la ligne droite qui est commune à ces deux plans, au moyen de quoi ceux de la courbe qui a trois dimensions sont entièrement déterminés. Même si on veut tirer une ligne droite qui coupe cette courbe au point donné à angles droits, il faut seulement tirer deux autres lignes droites dans les deux plans, une en chacun, qui coupent à angles droits les deux lignes courbes qui y sont aux deux points où tombent les perpendiculaires qui viennent de ce point donné ; car ayant élevé deux autres plans, un sur chacune de ces lignes droites, qui coupe à angles droits le plan où elle est, on aura l'intersection de ces deux plans pour la ligne droite cherchée. Et ainsi je pense n'avoir rien omis des éléments qui sont nécessaires pour la connoissance des lignes courbes.

Comment on peut appliquer ce qui a été dit ici des lignes courbes, décrites sur une superficie plate, à celles qui se décrivent dans un espace qui a trois dimensions.

LIVRE TROISIÈME

DE LA CONSTRUCTION DES PROBLÈMES QUI SONT SOLIDES OU PLUS QUE SOLIDES.

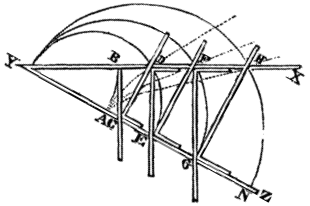
Encore que toutes les lignes courbes qui peuvent être décrites par quelque mouvement régulier doivent être reçues en la géométrie, ce n'est pas à dire qu'il soit permis de se servir indifféremment de la première qui se rencontre pour la construction de chaque problème, mais il faut avoir soin de choisir toujours la plus simple par laquelle il soit possible de le résoudre. Et même il est à remarquer que par les plus simples on ne doit pas seulement entendre celles qui peuvent le plus aisément être décrites, ni celles qui rendent la construction ou la démonstration du problème proposé plus facile, mais principalement celles qui sont du plus simple genre qui puisse servir à déterminer la quantité qui est cherchée.

De quelles lignes courbes on peut se servir en la construction de chaque problème.

Comme, par exemple, je ne crois pas qu'il y ait aucune façon plus facile pour

Exemple touchant l'invention de plusieurs moyennes proportionnelles.

25.



trouver autant de moyennes proportionnelles qu'on veut, ni dont la démonstration soit plus évidente, que d'y employer les lignes courbes qui se décrivent par l'instrument XYZ *(fig. 25)* ci-dessus expliqué. Car, voulant trouver deux moyennes proportionnelles entre YA et YE, il ne faut que décrire un cercle dont le diamètre soit YE, et pourceque ce cercle coupe la courbe AD au point D, YD est l'une des moyennes proportionnelles cherchées, dont la démonstration se voit à l'œil par la seule application de cet instrument sur la ligne YD ; car, comme YA ou YB, qui lui est égale, est à YC, ainsi YC est à YD, et YD à YE.

Tout de même pour trouver quatre moyennes proportionnelles entre YA et YG, ou pour en trouver six entre YA et YN, il ne faut que tracer le cercle YFG qui, coupant AF au point F, détermine la ligne droite YF qui est l'une de ces quatre proportionnelles ; ou YHN qui, coupant AH au point H, détermine YH l'une des six ; et ainsi des autres.

Mais pourceque la ligne courbe AD est du second genre, et qu'on peut trouver deux moyennes proportionnelles par les sections coniques qui sont du

premier ; et aussi pourcequ'on peut trouver quatre ou six moyennes proportionnelles par des lignes qui ne sont pas de genres si composés que sont $A\,F$ et $A\,H$, ce seroit une faute en géométrie que de les y employer. Et c'est une faute aussi, d'autre côté, de se travailler inutilement à vouloir construire quelque problème par un genre de lignes plus simple que sa nature ne permet.

De la nature des équations.

Or, afin que je puisse ici donner quelques règles pour éviter l'une et l'autre de ces deux fautes, il faut que je die quelque chose en général de la nature des équations, c'est-à-dire des sommes composées de plusieurs termes partie connus et partie inconnus dont les uns sont égaux aux autres, ou plutôt qui, considérés tous ensemble, sont égaux à rien : car ce sera souvent le meilleur de les considérer en cette sorte.

Combien il peut y avoir de racines en chaque équation.

Sachez donc qu'en chaque équation, autant que la quantité inconnue a de dimensions, autant peut-il y avoir de diverses racines, c'est-à-dire de valeurs de cette quantité ; car, par exemple, si on suppose x égale à 2, ou bien $x-2$ égal à rien ; et derechef $x=3$, ou bien $x-3=0$; en multipliant ces deux équations

$$x-2=0, \quad \text{et} \quad x-3=0,$$

l'une par l'autre, on aura

$$x^2-5x+6=0,$$

ou bien

$$x^2=5x-6,$$

qui est une équation en laquelle la quantité x vaut 2 et tout ensemble vaut 3. Que si derechef on fait

$$x-4=0,$$

et qu'on multiplie cette somme par

$$x^2-5x+6=0,$$

on aura

$$x^3-9x^2+26x-24=0,$$

qui est une autre équation en laquelle x, ayant trois dimensions, a aussi trois valeurs, qui sont 2, 3 et 4.

Quelles sont les fausses racines.

Mais souvent il arrive que quelques unes de ces racines sont fausses ou moindres que rien ; comme si on suppose que x désigne aussi le défaut d'une quantité qui soit 5, on a

$$x+5=0,$$

qui, étant multiplié par

$$x^3-9x^2+26x-24=0$$

fait

$$x^4-4x^3-19x^2+106x-120=0$$

pour une équation en laquelle il y a quatre racines, à savoir trois vraies qui sont 2, 3, 4, et une fausse qui est 5.

Comment on peut diminuer le nombre des dimensions d'une équation lorsqu'on connoît quelqu'une de ses racines.

Et on voit évidemment de ceci que la somme d'une équation qui contient plusieurs racines peut toujours être divisée par un binôme composé de la quantité inconnue moins la valeur de l'une des vraies racines, laquelle que ce soit, ou plus la valeur de l'une des fausses ; au moyen de quoi on diminue d'autant ses dimensions.

Comment on peut examiner si quelque quantité donnée est la valeur d'une racine.

Et réciproquement que si la somme d'une équation ne peut être divisée par un binôme composé de la quantité inconnue + ou − quelque autre quantité, cela témoigne que cette autre quantité n'est la valeur d'aucune de ses racines. Comme cette dernière

$$x^4-4x^3-19x^2+106x-120=0$$

peut bien être divisée par $x-2$, et par $x-3$, et par $x-4$, et par $x+5$, mais non point par $x+$ ou $-$ aucune autre quantité ; ce qui montre qu'elle ne peut avoir que les quatre racines 2, 3, 4 et 5.

Combien il peut y avoir de vraies racines en chaque équation.

On connoît aussi de ceci combien il peut y avoir de vraies racines et combien de fausses en chaque équation : à savoir il y en peut avoir autant de vraies que les signes + et − s'y trouvent de fois être changés, et autant de fausses qu'il s'y trouve de fois deux signes + ou deux signes − qui s'entre-suivent. Comme en la dernière, à cause qu'après $+x^4$ il y a $-4x^3$, qui est un changement du signe + en −, et après $-19x^2$ il y a $+106x$, et après $+106x$ il y a -120, qui sont encore deux autres changements, on connoît qu'il y a trois vraies racines ; et une fausse, à cause que les deux signes − de $4x^3$ et $19x^2$ s'entre-suivent.

Comment on fait que les fausses racines d'une équation deviennent vraies, et les vraies fausses.

De plus, il est aisé de faire en une même équation que toutes les racines qui étoient fausses deviennent vraies, et par même moyen que toutes celles qui étoient vraies deviennent fausses, à savoir en changeant tous les signes + ou − qui sont en la seconde, en la quatrième, en la sixième, ou autres places qui se désignent par les nombres pairs, sans changer ceux de la première, de la troisième, de la cinquième, et semblables qui se désignent par les nombres impairs. Comme si, au lieu de

$$+x^4-4x^3-19x^2+106x-120=0,$$

on écrit

$$+x^4+4x^3-19x^2-106x-120=0,$$

on a une équation en laquelle il n'y a qu'une vraie racine qui est 5, et trois fausses qui sont 2, 3 et 4.

Comment on peut augmenter ou diminuer les racines d'une équation sans les connoître.

Que si, sans connoître la valeur des racines d'une équation, on la veut augmenter ou diminuer de quelque quantité connue, il ne faut qu'au lieu du terme inconnu en supposer un autre qui soit plus ou moins grand de cette même quantité, et le substituer partout en la place du premier.

Comme si on veut augmenter de 3 la racine de cette équation

$$x^4+4x^3-19x^2-106x-120=0,$$

il faut prendre y au lieu de x, et penser que cette quantité y est plus grande que x de 3, en sorte que $y-3$ est égal à x ; et au lieu de x^2 il faut mettre le carré

de $y-3$, qui est y^2-6y+9 ; et au lieu de x^3 il faut mettre son cube qui est $y^3-9y^2+27y-27$; et enfin au lieu de x^4 il faut mettre son carré de carré qui est $y^4-12y^3+54y^2-108y+81$. Et ainsi, décrivant la somme précédente en substituant partout y au lieu de x, on a

$$\begin{array}{rrrrr}
y^4 & -12y^3 & +54y^2 & -108y & +\ 81 \\
 & +\ 4y^3 & -36y^2 & +108y & -108 \\
 & & -19y^2 & +114y & -171 \\
 & & & -106y & +318 \\
 & & & & -120 \\
\hline
y^4 & -\ 8y^3 & -\ y^2 & +\ 8y * (6) & = 0,
\end{array}$$

ou bien

$$y^3-8y^2-y+8=0,$$

où la vraie racine qui étoit 5 est maintenant 8, à cause du nombre 3 qui lui est ajouté.

Que si on veut au contraire diminuer de trois la racine de cette même équation, il faut faire $y+3=x$, et $y^2+6y+9=x^2$ et ainsi des autres, de façon qu'au lieu de

$$x^4+4x^3-19x^2-106x-120=0,$$

on met

$$\begin{array}{rrrrr}
y^4 & +12y^3 & +54y^2 & +108y & +\ 81 \\
 & +\ 4y^3 & +36y^2 & +108y & +108 \\
 & & -19y^2 & -114y & -171 \\
 & & & -106y & -318 \\
 & & & & -120 \\
\hline
y^4 & +16y^3 & +71y^2 & -\ 4y & -420 = 0.
\end{array}$$

Qu'en augmentant les vraies racines on diminue les fausses, et au contraire.

Et il est à remarquer qu'en augmentant les vraies racines d'une équation on diminue les fausses de la même quantité, ou au contraire en diminuant les vraies on augmente les fausses ; et que si on diminue, soit les unes, soit les autres, d'une quantité qui leur soit égale, elles deviennent nulles ; et que si c'est d'une quantité qui les surpasse, de vraies elles deviennent fausses, ou de fausses vraies. Comme ici, en augmentant de 3 la vraie racine qui étoit 5, on a diminué de 3 chacune des fausses, en sorte que celle qui étoit 4 n'est plus que 1, et celle qui étoit 3 est nulle, et celle qui étoit 2 est devenue vraie et est 1, à cause que $-2+3$ fait $+1$: c'est pourquoi en cette équation

$$y^3-8y^2-y+8=0$$

(6)Descartes indique l'absence d'un terme par le signe * mis à la place de ce terme ; nous l'ôterons comme inutile, de même que le facteur 1 qu'il laisse quelquefois.

il n'y a plus que trois racines, entre lesquelles il y en a deux qui sont vraies, 1 et 8, et une fausse qui est aussi 1 ; et en cette autre

$$y^4+16y^3+71y^2-4y-420=0,$$

il n'y en a qu'une vraie qui est 2, à cause que $+5-3$ fait $+2$, et trois fausses qui sont 5, 6 et 7.

Comment on peut ôter le second terme d'une équation.

Or par cette façon de changer la valeur des racines sans les connoître on peut faire deux choses qui auront ci-après quelque usage. La première est qu'on peut toujours ôter le second terme de l'équation qu'on examine, à savoir en diminuant les vraies racines de la quantité connue de ce second terme divisée par le nombre des dimensions du premier, si l'un de ces deux termes étant marqué du signe $+$, l'autre est marqué du signe $-$; ou bien en l'augmentant de la même quantité, s'ils out tous deux le signe $+$ ou tous deux le signe $-$. Comme pour ôter le second terme de la dernière équation qui est

$$y^4+16y^3+71y^2-4y-420=0,$$

ayant divisé 16 par 4, à cause des quatre dimensions du terme y^4, il vient derechef 4 ; c'est pourquoi je fais $z-4=y$, et j'écris

$$\begin{array}{rrrrr}
z^4 & -16z^3 & +\ 96z^2 & -256z & +\ 256 \\
 & +16z^3 & -192z^2 & +768z & -1024 \\
 & & +\ 71z^2 & -568z & +1136 \\
 & & & -\ 4z & +\ 16 \\
 & & & & -\ 420 \\
\hline
z^4 & & -\ 25z^2 & -\ 60z & -\ 36 = 0
\end{array}$$

où la vraie racine qui étoit 2 est 6, à cause qu'elle est augmentée de 4 ; et les fausses, qui étoient 5, 6 et 7, ne sont plus que 1, 2 et 3, à cause qu'elles sont diminuées chacune de 4.

Tout de même si on veut ôter le second terme de

$$x^4-2ax^3+(2a^2-c^2)x^2-2a^3x+a^4=0,$$

pourceque divisant $2a$ par 4 il vient $\frac{1}{2}a$, il faut faire $z + \frac{1}{2}a = x$, et écrire

$$\begin{array}{rrrrr}
z^4 & +2az^3 & +\frac{3}{2}a^2z^2 & +\frac{1}{2}a^3z & +\frac{1}{16}a^4 \\
 & -2az^3 & -3a^2z^2 & -\frac{3}{2}a^3z & -\frac{1}{4}a^4 \\
 & & +2a^2z^2 & +2a^3z & +\frac{1}{2}a^4 \\
 & & -c^2z^2 & -ac^2z & -\frac{1}{4}a^2c^2 \\
 & & & -2a^3z & -a^4 \\
 & & & & +a^4 \\
\hline
\multicolumn{5}{l}{z^4 + \left(\frac{1}{2}a^2 - c^2\right)z^2 - (a^3 + ac^2)z + \frac{5}{16}a^4 - \frac{1}{4}a^2c^2 = 0}
\end{array}$$

et si on trouve après la valeur de z, en lui ajoutant $\frac{1}{2}a$ on aura celle de x.

Comment on peut faire que toutes les fausses racines d'une équation deviennent vraies sans que les vraies deviennent fausses.

La seconde chose qui aura ci-après quelque usage est qu'on peut toujours, en augmentant la valeur des vraies racines d'une quantité qui soit plus grande que n'est celle d'aucune des fausses, faire qu'elles deviennent toutes vraies, en sorte qu'il n'y ait point deux signes + ou deux signes − qui s'entre-suivent, et outre cela que la quantité connue du troisième terme soit plus grande que le carré de la moitié de celle du second. Car encore que cela se fasse lorsque ces fausses racines sont inconnues, il est aisé néanmoins de juger à peu près de leur grandeur et de prendre une quantité qui les surpasse d'autant ou de plus qu'il n'est requis à cet effet. Comme si on a

$$x^6 + nx^5 - 6n^2x^4 + 36n^3x^3 - 216n^4x^2 + 1296n^5x - 7776n^6 = 0,$$

en faisant $y - 6n = x$ on trouvera

$$\begin{array}{rrrr

où les racines étant 2, 3 et 4, on connoît de là que celles de l'autre d'auparavant étoient $\frac{2}{3}$, 1 et $\frac{4}{3}$, et que celles de la première étoient

$$\frac{2}{9}\sqrt{3}, \quad \frac{1}{3}\sqrt{3}, \quad \text{et} \quad \frac{4}{9}\sqrt{3}.$$

Comment on rend la quantité connue de l'un des termes d'une équation égale à telle autre qu'on veut.

Cette opération peut aussi servir pour rendre la quantité connue de quelqu'un des termes de l'équation égale à quelque autre donnée, comme si ayant

$$x^3 - b^2x + c^3 = 0,$$

on veut avoir en sa place une autre équation en laquelle la quantité connue du terme qui occupe la troisième place, à savoir celle qui est ici b^2 soit $3a^2$, il faut supposer $y = x\sqrt{\frac{3a^2}{b^2}}$, puis écrire

$$y^3 - 3a^2y + \frac{3a^3c^3}{b^3}\sqrt{3} = 0.$$

Que les racines tant vraies que fausses peuvent être réelles ou imaginaires.

Au reste, tant les vraies racines que les fausses ne sont pas toujours réelles, mais quelquefois seulement imaginaires, c'est-à-dire qu'on peut bien toujours en imaginer autant que j'ai dit en chaque équation, mais qu'il n'y a quelquefois aucune quantité qui corresponde à celles qu'on imagine ; comme encore qu'on en puisse imaginer trois en celle-ci,

$$x^3 - 6x^2 + 13x - 10 = 0,$$

il n'y en a toutefois qu'une réelle qui est 2, et pour les deux autres, quoiqu'on les augmente ou diminue, ou multiplie en la façon que je viens d'expliquer, on ne saurait les rendre autres qu'imaginaires.

La réduction des équations cubiques, lorsque le problème est plan.

Or quand, pour trouver la construction de quelque problème, on vient à une équation en laquelle la quantité inconnue a trois dimensions, premièrement, si les quantités connues qui y sont contiennent quelques nombres rompus, il les faut réduire à d'autres entiers par la multiplication tantôt expliquée ; et s'ils en contiennent de sourds, il faut aussi les réduire à d'autres rationnaux autant qu'il sera possible, tant par cette même multiplication que par divers autres moyens qui sont assez faciles à trouver. Puis examinant par ordre toutes les quantités qui peuvent diviser sans fraction le dernier terme, il faut voir si quelqu'une d'elles, jointe avec la quantité inconnue par le signe + ou −, peut composer un binôme qui divise toute la somme ; et si cela est, le problème est plan, c'est-à-dire il peut être construit avec la règle et le compas ; car, ou bien la quantité connue de ce binôme est la racine cherchée, ou bien l'équation étant divisée par lui se réduit à deux dimensions, en sorte qu'on en peut trouver après la racine par ce qui a été dit au premier livre.

Par exemple, si on a

$$y^6 - 8y^4 - 124y^2 - 64 = 0,$$

le dernier terme qui est 64 peut être divisé sans fraction par 1, 2, 4, 8, 16, 32, 64 ; c'est pourquoi il faut examiner par ordre si cette équation ne peut point être divisée par quelqu'un des binômes $y^2 - 1$ ou $y^2 + 1$, $y^2 - 2$ ou $y^2 + 2$, $y^2 - 4$, etc. ; et on trouve qu'elle peut l'être par $y^2 - 16$ en cette sorte :

$$\begin{array}{r}
+y^6 - 8y^4 - 124y^2 - 64 = 0 \\
\underline{-y^6 - 8y^4 - 4y^2 - 16} \\
0 - 16y^4 - 128y^2 \\
\underline{-16 \quad -16} \\
+y^4 + 8y^2 + 4 = 0
\end{array}$$

La façon de diviser une équation par un binôme qui contient sa racine.

Je commence par le dernier terme, et divise -64 par -16, ce qui fait $+4$ que j'écris dans le quotient ; puis je multiplie $+4$ par $+y^2$, ce qui fait $+4y^2$; c'est pourquoi j'écris $-4y^2$ en la somme qu'il faut diviser, car il y faut toujours écrire le signe + ou − tout contraire à celui que produit la multiplication ; et joignant $-124y^2$ avec $-4y^2$, j'ai $-128y^2$ que je divise derechef par -16, et j'ai $+y^4$ pour mettre dans le quotient ; et en le multipliant par y^2, j'ai $-8y^4$ pour joindre avec le terme qu'il faut diviser, qui est aussi $-8y^4$; et ces deux ensemble font $-16y^4$ que je divise par -16, ce qui fait $+y^4$ pour le quotient et $-y^6$ pour joindre avec $+y^6$, ce qui fait 0 et montre que la division est achevée. Mais s'il étoit resté quelque quantité, ou bien qu'on n'eût pu diviser sans fraction quelqu'un des termes précédents, ou eût par là reconnu qu'elle ne pouvoit être faite.

Tout de même si on a

$$\left.\begin{array}{l} y^6 + a^2 \\ \quad -2c^2 \end{array}\right\} \left.\begin{array}{l} y^4 - a^4 \\ \quad +c^4 \end{array}\right\} \left.\begin{array}{l} y^2 - a^6 \\ \quad -2a^4c^2 \\ \quad -a^2c^4 \end{array}\right\} = 0,$$

le dernier terme se peut diviser sans fraction par a, a^2, $a^2 + c^2$, $a^3 + ac^2$, et semblables ; mais il n'y en a que deux qu'on ait besoin de considérer, à savoir a^2 et $a^2 + c^2$, car les autres, donnant plus ou moins de dimensions dans le quotient qu'il n'y en a en la quantité connue du pénultième terme, empêcheroient que la division ne s'y pût faire. Et notez que je ne compte ici les dimensions de y^6 que pour trois, à cause qu'il n'y a point de y^5, ni de y^3, ni de y en toute la somme. Or en examinant le binôme $y^2 - a^2 - c^2 = 0$, on trouve que la division se peut faire par lui en cette sorte :

$$\begin{array}{r}
\left.\begin{array}{l} +y^6 + a^2 \\ \underline{-y^6 - 2c^2} \\ 0 - 2a^2 \\ \quad +c^2 \end{array}\right\} y^4 \left.\begin{array}{l} -a^4 \\ +c^4 \\ -a^2c^2 \end{array}\right\} y^2 \left.\begin{array}{l} -a^6 \\ -2a^4c^2 \\ -a^2c^4 \\ \hline -a^2 - c^2 \end{array}\right\} = 0 \\
\hline
\qquad -a^2 - c^2 \qquad -a^2 - c^2 \\
\hline
+y^4 \qquad \left.\begin{array}{l} +2a^2 \\ -c^2 \end{array}\right\} y^2 \left.\begin{array}{l} +a^4 \\ +a^2c^2 \end{array}\right\} = 0,
\end{array}$$

Mais lorsqu'on ne trouve aucun binôme qui puisse ainsi diviser toute la somme de l'équation proposée, il est certain que le problème qui en dépend est solide ; et ce n'est pas une moindre faute après cela de tâcher à le construire sans y employer que des cercles et des lignes droites, que ce seroit d'employer des sections coniques à construire ceux auxquels on n'a besoin que de cercles : car enfin tout ce qui témoigne quelque ignorance s'appelle faute.

Quels problèmes sont solides lorsque l'équation est cubique.

Que si on a une équation dont la quantité inconnue ait quatre dimensions, il faut en même façon, après en avoir ôté les nombres sourds et rompus, s'il y en a, voir si on pourra trouver quelque binôme qui divise toute la somme en le composant de l'une des quantités qui divisent sans fraction le dernier terme. Et si on en trouve un, ou bien la quantité connue de ce binôme est la racine cherchée, ou du moins, après cette division, il ne reste en l'équation que trois dimensions, ensuite de quoi il faut derechef l'examiner en la même sorte. Mais lorsqu'il ne se trouve point de tel binôme, il faut, en augmentant ou diminuant la valeur de la racine, ôter le second terme de la somme en la façon tantôt expliquée, et après la réduire à une autre qui ne contienne que trois dimensions ; ce qui se fait en cette sorte : au lieu de

La réduction des équations qui ont quatre dimensions, lorsque le problème est plan. Et quels sont ceux qui sont solides.

$$+ x^4 \dots\ px^2 \dots\ qx \dots\ r = 0,$$

il faut écrire

$$+ y^6 \dots\ 2py^4 + (p^2 \dots\ 4r)y^2 - q^2 = 0.$$

Et pour les signes + ou − que j'ai omis, s'il y a eu $+p$ en la précédente équation, il faut mettre en celle-ci $+2p$, ou s'il y a eu $-p$, il faut mettre $-2p$; et au contraire s'il y a eu $+r$, il faut mettre $-4r$, ou s'il y a eu $-r$, il faut mettre $+4r$; et soit qu'il y ait eu $+q$ ou $-q$, il faut toujours mettre $-q^2$ et $+p^2$, au moins si on suppose que x^4 et y^6 sont marqués du signe +, car ce seroit tout le contraire si on y supposoit le signe −.

Par exemple, si on a

$$+ x^4 - 4x^2 - 8x + 35 = 0,$$

il faut écrire en son lieu

$$y^6 - 8y^4 - 124y^2 - 64 = 0,$$

car la quantité que j'ai nommée p étant -4, il faut mettre $-8y^4$ pour $2py^4$; et celle que j'ai nommée r étant 35, il faut mettre $(16-140)y^2$, c'est-à-dire $-124y^2$ au lieu de $(p^2-4r)y^2$; et enfin q étant 8, il faut mettre -64 pour $-q^2$.

Tout de même, au lieu de

$$+ x^4 - 17x^2 - 20x - 6 = 0,$$

il faut écrire

$$+ y^6 - 34y^4 + 313y^2 - 400 = 0,$$

car 34 est double de 17, et 313 en est le carré joint au quadruple de 6, et 400 est le carré de 20.

Tout de même aussi au lieu de

$$+ z^4 + \left(\frac{1}{2}a^2 - c^2\right) z^2 - (a^3 + ac^2)z - \frac{5}{16}a^4 - \frac{1}{4}a^2c^2 = 0,$$

il faut écrire

$$y^6 + (a^2 - 2c^2)y^4 + (c^4 - a^4)y^2 - a^6 - 2a^4c^2 - a^2c^4 = 0;$$

car p est $\frac{1}{2}a^2 - c^2$, et p^2 est $\frac{1}{4}a^4 - a^2c^2 + c^4$, et $4r$ est $-\frac{5}{4}a^4 + a^2c^2$, et enfin $-q^2$ est $-a^6 - 2a^4c^2 - a^2c^4$.

Après que l'équation est ainsi réduite à trois dimensions, il faut chercher la valeur de y^2 par la méthode déjà expliquée ; et si elle ne peut être trouvée, on n'a point besoin de passer outre, car il suit de là infailliblement que le problème est solide. Mais si on la trouve, on peut diviser par son moyen la précédente équation en deux autres, en chacune desquelles la quantité inconnue n'aura que deux dimensions et dont les racines seront les mêmes que les siennes ; à savoir, au lieu de

$$+ x^4 \dots\ px^2 \dots\ qx \dots\ r = 0,$$

il faut écrire ces deux autres

$$+ x^2 - yx + \frac{1}{2}y^2 \dots\ \frac{1}{2}p \dots\ \frac{q}{2y} = 0,$$

et

$$+ x^2 + yx + \frac{1}{2}y^2 \dots\ \frac{1}{2}p \dots\ \frac{q}{2y} = 0.$$

Et pour les signes + et − que j'ai omis, s'il y a $+p$ en l'équation précédente, il faut mettre $+\frac{1}{2}p$ en chacune de celles-ci, et $-\frac{1}{2}p$ s'il y a en l'autre $-p$; mais il faut mettre $+\frac{q}{2y}$ en celle où il y a $-yx$, et $-\frac{q}{2y}$ en celle où il y a $+yx$, lorsqu'il y a $+q$ en la première ; et au contraire, s'il y a $-q$, il faut mettre $-\frac{q}{2y}$ en celle où il y a $-yx$, et $+\frac{q}{2y}$ en celle où il y a $+yx$. Ensuite de quoi il est aisé de connoître toutes les racines de l'équation proposée, et par conséquent de construire le problème dont elle contient la solution, sans y employer que des cercles et des lignes droites.

Par exemple, à cause que faisant

$$y^6 - 34y^4 + 313y^2 - 400 = 0$$

pour

$$x^4 - 17x^2 - 20x - 6 = 0,$$

on trouve que y^2 est 16, on doit, au lieu de cette équation

$$+ x^4 - 17x^2 - 20x - 6 = 0,$$

écrire ces deux autres

$$+x^2-4x-3=0,$$

et

$$+x^2+4x+2=0,$$

car y est 4, $\frac{1}{2}y^2$ est 8, p est 17, et q est 20, de façon que

$$+\frac{1}{2}y^2-\frac{1}{2}p-\frac{q}{2y}\text{ fait }-3,\text{ et }+\frac{1}{2}y^2-\frac{1}{2}p+\frac{q}{2y}\text{ fait }+2.$$

Et tirant les racines de ces deux équations, on trouve toutes les mêmes que si on les tiroit de celle où est x^4, à savoir, on en trouve une vraie qui est $\sqrt{7}+2$, et trois fausses qui sont

$$\sqrt{7}-2,\quad 2+\sqrt{2},\quad\text{et}\quad 2-\sqrt{2},$$

Ainsi ayant

$$x^4-4x^2-8x+35=0,$$

pourceque la racine de

$$y^6-8y^4-124y^2-64=0$$

est derechef 16, il faut écrire

$$x^2-4x+5=0$$

et

$$x^2+4x+7=0.$$

Car ici

$$+\frac{1}{2}y^2-\frac{1}{2}p-\frac{q}{2y}\text{ fait 5, et }+\frac{1}{2}y^2-\frac{1}{2}p+\frac{q}{2y}\text{ fait 7.}$$

Et pourcequ'on ne trouve aucune racine, ni vraie ni fausse, en ces deux dernières équations, on connoît de là que les quatre de l'équation dont elles procèdent sont imaginaires, et que le problème pour lequel on l'a trouvée est plan de sa nature, mais qu'il ne sauroit en aucune façon être construit, à cause que les quantités données ne peuvent se joindre.

Tout de même ayant

$$z^4+\left(\frac{1}{2}a^2-c^2\right)z^2-(a^3+ac^2)z+\frac{5}{16}a^4-\frac{1}{4}a^2c^2=0,$$

pourcequ'on trouve a^2+c^2 pour y^2, il faut écrire

$$z^2-\sqrt{a^2+c^2}z+\frac{3}{4}a^2-\frac{1}{2}a\sqrt{a^2+c^2}=0,$$

et

$$z^2+\sqrt{a^2+c^2}z+\frac{3}{4}a^2+\frac{1}{2}a\sqrt{a^2+c^2}=0,$$

car y est $\sqrt{a^2+c^2}$, et $+\frac{1}{2}y^2+\frac{1}{2}p$ est $\frac{3}{4}a^2$, et $\frac{q}{2y}$ est $\frac{1}{2}a\sqrt{a^2+c^2}$, d'où on connoît que la valeur de z est

$$\frac{1}{2}\sqrt{a^2+c^2}+\sqrt{-\frac{1}{2}a^2+\frac{1}{4}c^2+\frac{1}{2}a\sqrt{a^2+c^2}},$$

ou bien

$$\frac{1}{2}\sqrt{a^2+c^2}-\sqrt{-\frac{1}{2}a^2+\frac{1}{4}c^2+\frac{1}{2}a\sqrt{a^2+c^2}}.$$

Et pourceque nous avions fait ci-dessus $z+\frac{1}{2}a=x$, nous apprenons que la quantité x, pour la connoissance de laquelle nous avons fait toutes ces opérations, est

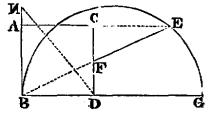
$$+\frac{1}{2}+\sqrt{\frac{1}{4}a^2+\frac{1}{4}c^2}-\sqrt{\frac{1}{4}c^2-\frac{1}{2}a^2+\frac{1}{2}a\sqrt{a^2+c^2}}.$$

Exemple de l'usage de ces réductions.

Mais afin qu'on puisse mieux connoître l'utilité de cette règle il faut que je l'applique à quelque problème.

Si le carré $A\,D$ *(fig. 26)* et la ligne $B\,N$ étant donnés, il faut prolonger le côté $A\,C$ jusques à E, en sorte que $E\,F$, tirée de E vers B, soit égale à $N\,B$: on apprend de Pappus, qu'ayant premièrement prolongé $B\,D$ jusques à G, en

Fig. 26.



sorte que $D\,G$ soit égale à $D\,N$, et ayant décrit un cercle dont le diamètre soit $B\,G$, si on prolonge la ligne droite $A\,C$, elle rencontrera la circonférence de ce cercle au point E qu'on demandoit. Mais pour ceux qui ne sauroient point cette construction, elle seroit assez difficile à rencontrer ; et, en la cherchant par la méthode ici proposée, ils ne s'aviseroient jamais de prendre DG pour la quantité inconnue, mais plutôt $C\,F$ ou $F\,D$, à cause que ce sont elles qui conduisent le plus aisément à l'équation ; et lors ils en trouveroient une qui ne seroit pas facile à démêler sans la règle que je viens d'expliquer. Car posant a pour $B\,D$ ou $C\,D$, et c pour $E\,F$, et x pour $D\,F$, on a $C\,F=a-x$, et comme $C\,F$ ou $a-x$ est à $F\,E$ ou c, ainsi $F\,D$ ou x est à $B\,F$, qui par conséquent est $\frac{cx}{a-x}$. Puis à cause du triangle rectangle $B\,D\,F$ dont les côtés sont l'un x et l'autre a, leurs carrés, qui sont x^2+a^2, sont égaux à celui de la base, qui est $\frac{c^2x^2}{x^2-2ax+a^2}$; de façon que, multipliant le tout par $x-2ax+a^2$, on trouve que l'équation est

$$x^4-2ax^3+2a^2x^2-2a^3x+a^4=c^2x^2,$$

ou bien

$$x^4 - 2ax^3 + (2a^2 - c^2)x^2 - 2a^3x + a^4 = 0;$$

et on connoît par les règles précédentes que sa racine, qui est la longueur de la ligne $D\,F$, est

$$\frac{1}{2}a + \sqrt{\frac{1}{4}a^2 + \frac{1}{4}c^2} - \sqrt{\frac{1}{4}c^2 - \frac{1}{2}a^2 + \frac{1}{2}a\sqrt{a^2 + c^2}}.$$

Que si on posoit $B\,F$, ou $C\,E$, ou $B\,E$, pour la quantité inconnue, on viendroit derechef à une équation en laquelle il y auroit quatre dimensions, mais qui seroit plus aisée à démêler, et on y viendroit assez aisément ; au lieu que si c'étoit $D\,G$ qu'on supposât, on viendroit beaucoup plus difficilement à l'équation, mais aussi elle seroit très simple. Ce que je mets ici pour vous avertir que, lorsque le problème proposé n'est point solide, si en le cherchant par un chemin on vient à une équation fort composée, on peut ordinairement venir à une plus simple en le cherchant par un autre.

Je pourrois encore ajouter diverses règles pour démêler les équations qui vont au cube ou au carré de carré, mais elles seroient superflues ; car lorsque les problèmes sont plans on en peut toujours trouver la construction par celles-ci.

Règle générale pour réduire les équations qui passent le carré de carré.

Je pourrois aussi en ajouter d'autres pour les équations qui montent jusques au sursolide, ou au carré de cube, ou au-delà, mais j'aime mieux les comprendre toutes en une, et dire en général que, lorsqu'on a tâché de les réduire à même forme que celles d'autant de dimensions qui viennent de la multiplication de deux autres qui en ont moins, et qu'ayant dénombré tous les moyens par lesquels cette multiplication est possible, la chose n'a pu succéder par aucun, on doit s'assurer qu'elles ne sauroient être réduites à de plus simples ; en sorte que si la quantité inconnue a trois ou quatre dimensions, le problème pour lequel on la cherche est solide, et si elle en a cinq ou six, il est d'un degré plus composé, et ainsi des autres.

Au reste, j'ai omis ici les démonstrations de la plupart de ce que j'ai dit, à cause qu'elles m'ont semblé si faciles que, pourvu que vous preniez la peine d'examiner méthodiquement si j'ai failli, elles se présenteront à vous d'elles-mêmes ; et il sera plus utile de les apprendre en cette façon qu'en les lisant.

Façon générale pour construire tous les problèmes solides réduits à une équation de trois ou quatre dimensions.

Or, quand on est assuré que le problème proposé est solide, soit que l'équation par laquelle on le cherche monte au carré de carré, soit qu'elle ne monte que jusques au cube, on peut toujours en trouver la racine par l'une des trois sections coniques, laquelle que ce soit, ou même par quelque partie de l'une d'elles, tant petite qu'elle puisse être, en ne se servant au reste que de lignes droites et de cercles. Mais je me contenterai ici de donner une règle générale pour les trouver toutes par le moyen d'une parabole, à cause qu'elle est en quelque façon la plus simple.

Premièrement, il faut ôter le second terme de l'équation proposée, s'il n'est déjà nul, et ainsi la réduire à telle forme

$$z^3 = \dots\ apz\ \dots\ a^2q,$$

si la quantité inconnue n'a que trois dimensions ; ou bien à telle

$$z^4 = \dots\ apz^2\ \dots\ a^2qz\ \dots\ a^3r,$$

si elle en a quatre ; ou bien, en prenant a pour l'unité, à telle

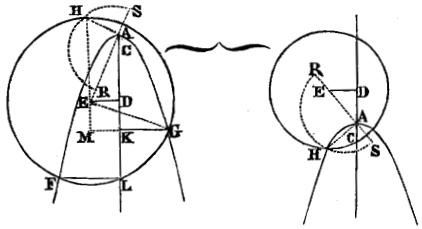
$$z^3 = \dots\ pz\ \dots\ q,$$

et à telle

$$z^4 = \dots\ pz^2\ \dots\ qz\ \dots\ r.$$

Après cela, supposant que la parabole $F\,A\,G$ *(fig. 27)* est déjà décrite, et que son essieu est $A\,C\,D\,K\,L$, et que son côté droit est a ou 1 dont $A\,C$ est la moitié, et enfin que le point C est au dedans de cette parabole, et que A en est le sommet ; il faut faire $C\,D = \frac{1}{2}p$, et la prendre du même côté qu'est le

Fig. 27.



point A au regard du point C, s'il y a $+p$ en l'équation ; mais s'il y a $-p$, il faut la prendre de l'autre côté. Et du point D, ou bien, si la quantité p étoit nulle, du point C *(fig. 28)* il faut élever une ligne à angles droits jusques à E, en sorte qu'elle soit égale à $\frac{1}{2}q$. Et enfin du centre E il faut décrire le cercle $F\,G$ dont le demi-diamètre soit $A\,E$ si l'équation n'est que cubique, en sorte que la quantité r soit nulle. Mais quand il y a $+r$ il faut dans cette ligne $A\,E$ *(fig. 27)* prolongée prendre d'un côté $A\,R$ égale à r, et de l'autre $A\,S$ égale au côté droit de la parabole qui est 1 ; et ayant décrit un cercle dont le diamètre soit $R\,S$, il faut faire $A\,H$ perpendiculaire sur $A\,E$, laquelle $A\,H$ rencontre ce cercle $R\,H\,S$ au point H qui est celui par où l'autre cercle $F\,H\,G$ doit passer. Et quand il y a $-r$, il faut, après avoir ainsi trouvé la ligne $A\,H$ *(fig. 29)*, inscrire $A\,I$ qui lui soit égale, dans un autre cercle dont $A\,E$ soit le diamètre, et lors c'est par le point I que doit passer $F\,I\,G$ le premier cercle cherché. Or ce cercle $F\,G$ peut couper ou toucher la parabole en un, ou deux, ou trois, ou quatre points, desquels tirant des perpendiculaires sur l'essieu, on a toutes les racines de l'équation tant vraies que fausses. A savoir si la quantité q est marquée du

Fig. 28.

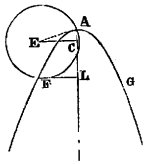
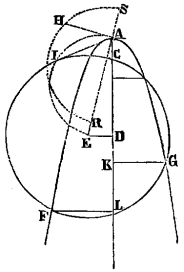


Fig. 29.



signe +, les vraies racines seront celles de ces perpendiculaires qui se trouveront du même côté de la parabole que E le centre du cercle, comme $F\,L$; et les autres, comme $G\,K$, seront fausses. Mais au contraire, si cette quantité q est marquée du signe −, les vraies seront celles de l'autre côté, et les fausses ou moindres que rien seront du côté où est E le centre du cercle. Et enfin si ce cercle ne coupe ni ne touche la parabole en aucun point, cela témoigne qu'il n'y a aucune racine ni vraie ni fausse en l'équation, et qu'elles sont toutes imaginaires. En sorte que cette règle est la plus générale et la plus accomplie qu'il soit possible de souhaiter.

Et la démonstration en est fort aisée ; car si la ligne $G\,K$ (*fig. 27*), trouvée par cette construction, se nomme z, $A\,K$ sera z^2, à cause de la parabole en laquelle $G\,K$ doit être moyenne proportionnelle entre $A\,K$ et le côté droit qui est 1 ; puis, si de $A\,K$ j'ôte $A\,C$ qui est $\frac{1}{2}$, et $C\,D$ qui est $\frac{1}{2}p$, il reste $D\,K$ ou $E\,M$ qui est $z^2 - \frac{1}{2}p - \frac{1}{2}$, dont le carré est

$$z^4 - pz^2 - z^2 + \frac{1}{4}p^2 + \frac{1}{2}p + \frac{1}{4};$$

et à cause que $D\,E$ ou $K\,\mathrm{M}$ est $\frac{1}{2}q$, la toute $G\,M$ est $z + \frac{1}{2}q$, dont le carré est

$$z^2 + qz + \frac{1}{4}q^2;$$

et assemblant ces deux carrés on a

$$z^4 - pz^2 + qz + \frac{1}{4}q^2 + \frac{1}{4}p^2 + \frac{1}{2}p + \frac{1}{4},$$

pour le carré de la ligne $G\,E$, à cause qu'elle est la base du triangle rectangle $E\,M\,G$.

Mais à cause que cette même ligne $G\,E$ est le demi-diamètre du cercle $F\,G$, elle se peut encore expliquer en d'autres termes, à savoir $E\,D$ étant $\frac{1}{2}q$, et $A\,D$ étant $\frac{1}{2}p + \frac{1}{2}$, $A\,E$ est

$$\sqrt{\frac{1}{4}q^2 + \frac{1}{4}p^2 + \frac{1}{2}p + \frac{1}{4}},$$

à cause de l'angle droit $A\,D\,E$; puis $H\,A$ étant moyenne proportionnelle entre $A\,S$ qui est 1 et $A\,R$ qui est r, elle est $\sqrt{r}$; et à cause de l'angle droit $E\,A\,H$, le carré de $H\,E$ ou $E\,G$ est

$$\frac{1}{4}q^2 + \frac{1}{4}p^2 + \frac{1}{2}p + \frac{1}{4} + r;$$

si bien qu'il y a équation entre cette somme et la précédente, ce qui est le même que

$$z^4 = pz^2 - qz + r,$$

et par conséquent la ligne trouvée $G\,K$ qui a été nommée z est la racine de cette équation, ainsi qu'il falloit démontrer. Et si vous appliquez ce même calcul à tous les autres cas de cette règle en changeant les signes + et − selon l'occasion, vous y trouverez votre compte en même sorte, sans qu'il soit besoin que je m'y arrête.

L'invention de deux moyennes proportionnelles.

Si on veut donc, suivant cette règle, trouver deux moyennes proportionnelles entre les lignes a et q *(fig. 28)*, chacun sait que posant z pour l'une, comme a est à z, ainsi z à $\frac{z^2}{a}$, et $\frac{z^2}{a}$ à $\frac{z^3}{a^2}$; de façon qu'il y a équation entre q et $\frac{z^3}{a^2}$, c'est-à-dire

$$z^3 = a^2 q.$$

Et la parabole $F\,A\,G$ étant décrite, avec la partie de son essieu $A\,C$ qui est $\frac{1}{2}a$ la moitié du côté droit, il faut du point C élever la perpendiculaire $C\,E$ égale à $\frac{1}{2}q$, et du centre E par A, décrivant le cercle $A\,F$, on trouve $F\,L$ et $L\,A$ pour les deux moyennes cherchées.

La façon de diviser un angle en trois.

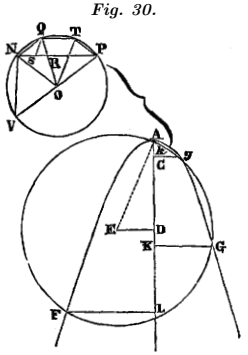
Tout de même si on veut diviser l'angle $N\,O\,P$ *(fig. 30)*, ou bien l'arc ou portion de cercle $N\,Q\,P\,T$ en trois parties égales, faisant $N\,O = 1$ pour le rayon du cercle, et $N\,P = q$ pour la subtendue de l'arc donné, et $N\,Q = z$ pour la subtendue du tiers de cet arc, l'équation vient

$$z^3 = 3z - q.$$

Car ayant tiré les lignes $N\,Q$, $O\,Q$, $O\,T$, et faisant $Q\,S$ parallèle à $T\,O$, on voit que comme $N\,O$ est à $N\,Q$, ainsi $N\,Q$ à $Q\,R$, et $Q\,R$ à $R\,S$; en sorte que $N\,O$ étant 1, et $N\,Q$ étant z, $Q\,R$ est z^2, et $R\,S$ est z^3 ; et à cause qu'il s'en faut seulement $R\,S$ ou z^3 que la ligne $N\,P$ qui est q ne soit triple de $N\,Q$ qui est z, on a

$$q = 3z - z^3,$$

Fig. 30.



ou bien

$$z^3 = 3z - q.$$

Puis la parabole $F\,A\,G$ étant décrite, et $C\,A$ la moitié de son côté droit principal étant $\frac{1}{2}$, si on prend $C\,D = \frac{3}{2}$, et la perpendiculaire $D\,E = \frac{1}{2}q$, et que du centre E par A on décrive le cercle $F\,A\,g\,G$, il coupe cette parabole aux trois points F, g et G, sans compter le point A qui en est le sommet ; ce qui montre qu'il y a trois racines en cette équation, à savoir les deux $G\,K$ et $g\,k$ qui sont vraies, et la troisième qui est fausse, à savoir $F\,L$. Et de ces deux vraies c'est $g\,k$ la plus petite qu'il faut prendre pour la ligne $N\,Q$ qui étoit cherchée ; car l'autre $G\,K$ est égale à $N\,V$ la subtendue de la troisième partie de l'arc $N\,V\,P$, qui avec l'autre arc $N\,Q\,P$ achève le cercle. Et la fausse $F\,L$ est égale à ces deux ensembles $Q\,N$ et $N\,V$, ainsi qu'il est aisé à voir par le calcul.

Que tous les problèmes solides se peuvent réduire à ces deux constructions.

Il seroit superflu que je m'arrêtasse à donner ici d'autres exemples, car tous les problèmes qui ne sont que solides se peuvent réduire à tel point qu'on n'a aucun besoin de cette règle pour les construire, sinon en tant qu'elle sert à trouver deux moyennes proportionnelles, ou bien à diviser un angle en trois parties égales, ainsi que vous connoîtrez en considérant que leurs difficultés peuvent toujours être comprises en des équations qui ne montent que jusques au carré de carré ou au cube, et que toutes celles qui montent au carré de carré se réduisent au carré par le moyen de quelques autres qui ne montent que jusques au cube, et enfin qu'on peut ôter le second terme de celles-ci ; en sorte qu'il n'y

en a point qui ne se puisse réduire à quelqu'une de ces trois formes :

$$z^3 = -pz + q,$$
$$z^3 = +pz + q,$$
$$z^3 = +pz - q.$$

Or si on a $z^3 = -pz + q$, la règle dont Cardan attribue l'invention à un nommé Scipio Ferreus nous apprend que la racine est

$$\sqrt{C. + \frac{1}{2}q + \sqrt{\frac{1}{4}q^2 + \frac{1}{27}p^3}} - \sqrt{C. - \frac{1}{2}q + \sqrt{\frac{1}{4}q^2 + \frac{1}{27}p^3}}.$$

Comme aussi lorsqu'on a $z^3 = +pz+q$, et que le carré de la moitié du dernier terme est plus grand que le cube du tiers de la quantité connue du pénultième, une pareille règle nous apprend que la racine est

$$\sqrt{C. + \frac{1}{2}q + \sqrt{\frac{1}{4}q^2 - \frac{1}{27}p^3}} + \sqrt{C. + \frac{1}{2}q - \sqrt{\frac{1}{4}q^2 - \frac{1}{27}p^3}}.$$

D'où il paroît qu'on peut construire tous les problèmes dont les difficultés se réduisent à l'une de ces deux formes, sans avoir besoin des sections coniques pour autre chose que pour tirer les racines cubiques de quelques quantités données, c'est-à-dire pour trouver deux moyennes proportionnelles entre ces quantités et l'unité.

Puis, si on a $z^3 = +pz + q$, et que le carré de la moitié du dernier terme ne soit point plus grand que le cube du tiers de la quantité connue du pénultième, en supposant le cercle $N\,Q\,P\,V$ dont le demi-diamètre $N\,O$ soit $\sqrt{\frac{1}{3}p}$, c'est-à-dire la moyenne proportionnelle entre le tiers de la quantité donnée p et l'unité, et supposant aussi la ligne $N\,P$ inscrite dans ce cercle qui soit $\frac{3q}{p}$, c'est-à-dire qui soit à l'autre quantité donnée q comme l'unité est au tiers de p, il ne faut que diviser chacun des deux arcs $N\,Q\,P$ et $N\,V\,P$ en trois parties égales, et on aura $N\,Q$ la subtendue du tiers de l'un, et $N\,V$ la subtendue du tiers de l'autre, qui jointes ensemble composeront la racine cherchée.

Enfin si on a $z^3 = pz - q$, en supposant derechef le cercle $N\,Q\,P\,V$ dont le rayon $N\,O$ soit $\sqrt{\frac{1}{3}p}$, et l'inscrite $N\,P$ soit $\frac{3q}{p}$, $N\,Q$ la subtendue du tiers de l'arc $N\,Q\,P$ sera l'une des racines cherchées, et $N\,V$ la subtendue du tiers de l'autre arc sera l'autre. Au moins, si le carré de la moitié du dernier terme n'est point plus grand que le cube du tiers de la quantité connue du pénultième ; car s'il étoit plus grand, la ligne $N\,P$ ne pourroit être inscrite dans le cercle, à cause qu'elle seroit plus longue que son diamètre, ce qui seroit cause que les deux vraies racines de cette équation ne seroient qu'imaginaires, et qu'il n'y en

auroit de réelle que la fausse, qui, suivant la règle de Cardan, seroit

$$\sqrt{C.\frac{1}{2}q+\sqrt{\frac{1}{4}q^2-\frac{1}{27}p^2}}+\sqrt{C.\frac{1}{2}q-\sqrt{\frac{1}{4}q^2-\frac{1}{27}p^2}}.$$

La façon d'exprimer la valeur de toutes les racines des cubiques, et ensuite de toutes celles qui ne montent que jusques au carré de carré.

Au reste, il est à remarquer que cette façon d'exprimer la valeur des racines par le rapport qu'elles ont aux côtés de certains cubes dont il n'y a que le contenu qu'on connoisse, n'est en rien plus intelligible ni plus simple que de les exprimer par le rapport qu'elles ont aux subtendues de certains arcs ou portions de cercles dont le triple est donné ; en sorte que toutes celles des équations cubiques qui ne peuvent être exprimées par les règles de Cardan, le peuvent être autant ou plus clairement par la façon ici proposée.

Car si, par exemple, on pense connoître la racine de cette équation

$$z^3 = -qz+p,$$

à cause qu'on sait qu'elle est composée de deux lignes dont l'une est le côté d'un cube duquel le contenu est $\frac{1}{2}q$, ajouté au côté d'un carré duquel derechef le contenu est $\frac{1}{4}q^2-\frac{1}{27}p^3$, et l'autre est le côté d'un autre cube dont le contenu est la différence qui est entre $\frac{1}{2}q$ et le côté de ce carré dont le contenu est $\frac{1}{4}q^2-\frac{1}{27}p^3$, qui est tout ce qu'on en apprend par la règle de Cardan. Il n'y a point de doute qu'on ne connoisse autant ou plus distinctement la racine de celle-ci

$$z^3 = +qz-p,$$

en la considérant inscrite dans un cercle dont le demi-diamètre est $\sqrt{\frac{1}{3}p}$, et sachant qu'elle y est la subtendue d'un arc dont le triple a pour sa subtendue $\frac{3q}{p}$. Même ces termes sont beaucoup moins embarrassés que les autres, et ils se trouveront beaucoup plus courts si on veut user de quelque chiffre particulier pour exprimer ces subtendues, ainsi qu'on fait du chiffre $\sqrt{C.}$ pour exprimer le côté des cubes.

Et on peut aussi ensuite de ceci exprimer les racines de toutes les équations qui montent jusques au carré de carré par les règles ci-dessus expliquées ; en sorte que je ne sache rien de plus à désirer en cette matière : car enfin la nature de ces racines ne permet pas qu'on les exprime en termes plus simples, ni qu'on les détermine par aucune construction qui soit ensemble plus générale et plus facile.

Il est vrai que je n'ai pas encore dit sur quelles raisons je me fonde pour oser ainsi assurer si une chose est possible ou ne l'est pas. Mais si on prend garde comment, par la méthode dont je me sers, tout ce qui tombe sous la considération des géomètres se réduit à un même genre de problèmes, qui est de chercher la valeur des racines de quelque équation, on jugera bien qu'il n'est pas

Pourquoi les problèmes solides ne peuvent être construits sans les sections coniques, ni ceux qui sont plus composés sans quelques autres lignes plus composées.

malaisé de faire un dénombrement de toutes les voies par lesquelles on les peut trouver, qui soit suffisant pour démontrer qu'on a choisi la plus générale et la plus simple. Et particulièrement pour ce qui est des problèmes solides, que j'ai dit ne pouvoir être construits sans qu'on y emploie quelque ligne plus composée que la circulaire, c'est chose qu'on peut assez trouver de ce qu'ils se réduisent tous à deux constructions, en l'une desquelles il faut avoir tout ensemble les deux points qui déterminent deux moyennes proportionnelles entre deux lignes données, et en l'autre les deux points qui divisent en trois parties égales un arc donné ; car d'autant que la courbure du cercle ne dépend que d'un simple rapport de toutes ses parties au point qui en est le centre, on ne peut aussi s'en servir qu'à déterminer un seul point entre deux extrêmes, comme à trouver une moyenne proportionnelle entre deux lignes droites données, ou diviser en deux un arc donné ; au lieu que la courbure des sections coniques, dépendant toujours de deux diverses choses, peut aussi servir à déterminer deux points différents.

Mais pour cette même raison il est impossible qu'aucun des problèmes qui sont d'un degré plus composés que les solides, et qui présupposent l'invention de quatre moyennes proportionnelles, ou la division d'un angle en cinq parties égales, puissent être construits par aucune des sections coniques. C'est pourquoi je croirai faire en ceci tout le mieux qui se puisse, si je donne une règle générale pour les construire, en y employant la ligne courbe qui se décrit par l'intersection d'une parabole et d'une ligne droite en la façon ci-dessus expliquée ; car j'ose assurer qu'il n'y en a point de plus simple en la nature qui puisse servir à ce même effet, et vous avez vu comme elle suit immédiatement les sections coniques en cette question tant cherchée par les anciens, dont la solution enseigne par ordre toutes les lignes courbes qui doivent être reçues en géométrie.

Façon générale pour construire tous les problèmes réduits à une équation qui n'a point plus de six dimensions.

Vous savez déjà comment, lorsqu'on cherche les quantités qui sont requises pour la construction de ces problèmes, on les peut toujours réduire à quelque équation qui ne monte que jusques au carré de cube ou au sursolide. Puis vous savez aussi comment, en augmentant la valeur des racines de cette équation, on peut toujours faire qu'elles deviennent toutes vraies, et avec cela que la quantité connue du troisième terme soit plus grande que le carré de la moitié de celle du second ; et enfin comment, si elle ne monte que jusques au sursolide, on la peut hausser jusques au carré de cube, et faire que la place d'aucun de ces termes ne manque d'être remplie. Or, afin que toutes les difficultés dont il est ici question puissent être résolues par une même règle, je désire qu'on fasse toutes ces choses, et par ce moyen qu'on les réduise toujours à une équation de telle forme,

$$y^6-py^5+qy^4-ry^3+sy^2-ty+u=0,$$

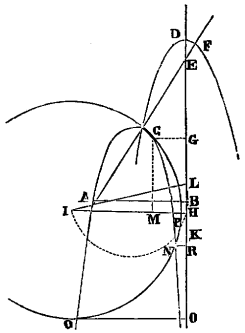
et en laquelle la quantité nommée q soit plus grande que le carré de la moitié de celle qui est nommée p. Puis ayant fait la ligne $B\,K$ *(fig. 31)* indéfiniment longue des deux côtés, et du point B ayant tiré la perpendiculaire $A\,B$ dont la longueur soit $\frac{1}{2}p$, il faut dans un plan séparé décrire une parabole, comme

$C\,D\,F$, dont le côté droit principal soit

$$\sqrt{\frac{t}{\sqrt{u}}+q-\frac{1}{4}p^2},$$

que je nommerai n pour abréger. Après cela, il faut poser le plan dans lequel est

Fig. 31.

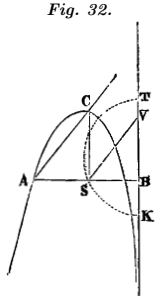


cette parabole sur celui où sont les lignes $A\,B$ et $B\,K$, en sorte que son essieu $D\,E$ se rencontre justement au-dessus de la ligne droite $B\,K$; et ayant pris la partie de cet essieu qui est entre les points E et D égale à $\frac{2\sqrt{u}}{pn}$, il faut appliquer sur ce point E une longue règle en telle façon qu'étant aussi appliquée sur le point A du plan de dessous, elle demeure toujours jointe à ces deux points pendant qu'on haussera ou baissera la parabole tout le long de la ligne $B\,K$ sur laquelle son essieu est appliqué ; au moyen de quoi l'intersection de cette parabole et de cette règle, qui se fera au point C, décrira la ligne courbe $A\,C\,N$, qui est celle dont nous avons besoin de nous servir pour la construction du problème proposé. Car après qu'elle est ainsi décrite, si on prend le point L en la ligne $B\,K$, du côté vers lequel est tourné le sommet de la parabole, et qu'on fasse $B\,L$ égale à $D\,E$, c'est-à-dire à $\frac{2\sqrt{u}}{pn}$; puis du point L vers B qu'on prenne en la même ligne $B\,K$ la ligne $L\,H$ égale à $\frac{t}{2n\sqrt{u}}$, et que du point H ainsi trouvé on tire à angles droits du côté qu'est la courbe $A\,C\,N$ la ligne $H\,I$ dont la longueur soit $\frac{r}{2n^2}+\frac{\sqrt{u}}{n^2}+\frac{pt}{4n^2\sqrt{u}}$, qui pour abréger sera nommé $\frac{m}{n^2}$; et après, ayant joint les points L et I, qu'on décrive le cercle $L\,P\,I$ dont $I\,L$ soit le diamètre, et

qu'on inscrive en ce cercle la ligne $L\,P$ dont la longueur soit $\sqrt{\frac{s+p\sqrt{u}}{n^2}}$; puis enfin du centre I, par le point P ainsi trouvé, qu'on décrive le cercle $P\,C\,N$. Ce cercle coupera ou touchera la ligne courbe $A\,C\,N$ en autant de points qu'il y aura de racines en l'équation, en sorte que les perpendiculaires tirées de ces points sur la ligne $B\,K$, comme $C\,G$, $N\,R$, $Q\,O$, et semblables, seront les racines cherchées, sans qu'il y ait aucune exception ni aucun défaut en cette règle. Car si la quantité s étoit si grande à proportion des autres p, q, r, t, et u, que la ligne $L\,P$ se trouvât plus grande que le diamètre du cercle $L\,I$, en sorte qu'elle n'y pût être inscrite, il n'y auroit aucune racine en l'équation proposée qui ne fût imaginaire ; non plus que si le cercle $I\,P$ étoit si petit qu'il ne coupât la courbe $A\,C\,N$ en aucun point. Et il la peut couper en six différents, ainsi qu'il peut y avoir six diverses racines en l'équation. Mais lorsqu'il la coupe en moins, cela témoigne qu'il y a quelques-unes de ces racines qui sont égales entre elles, ou bien qui ne sont qu'imaginaires.

Que si la façon de tracer la ligne $A\,C\,N$ par le mouvement d'une parabole vous semble incommode, il est aisé de trouver plusieurs autres moyens pour la décrire : comme si, ayant les mêmes quantités que devant pour $A\,B$ et $B\,L$ *(fig. 32)*, et la même pour $B\,K$ qu'on avoit posée pour le côté droit principal de la parabole, on décrit le demi-cercle $K\,S\,T$ dont le centre soit pris à discrétion dans la ligne $B\,K$, en sorte qu'il coupe quelque part la ligne $A\,B$ comme au point S ; et que du point T où il finit on prenne vers K la ligne $T\,V$ égale à $B\,L$; puis ayant tiré la ligne $S\,V$, qu'on en tire une autre qui lui soit parallèle par le

Fig. 32.



point A, comme $A\,C$, et qu'on en tire aussi une autre par S qui soit parallèle à $B\,K$, comme $S\,C$, le point C où ces deux parallèles se rencontrent sera l'un de ceux de la ligne courbe cherchée. Et on en peut trouver en même sorte autant d'autres qu'on en désire.

Or la démonstration de tout ceci est assez facile ; car, appliquant la règle $A\,E$ *(fig. 31)* avec la parabole $F\,D$ sur le point C, comme il est certain qu'elles

peuvent y être appliquées ensemble, puisque ce point C est en la courbe $A\,C\,N$ qui est décrite par leur intersection, si $C\,G$ se nomme y, $G\,D$ sera $\frac{y^2}{n}$, à cause que le côté droit qui est n est à $C\,G$ comme $C\,G$ à $G\,D$; et ôtant $D\,E$ qui est $\frac{2\sqrt{u}}{pn}$, de $G\,D$, on a $\frac{y^2}{n} - \frac{2\sqrt{u}}{pn}$ pour $G\,E$. Puis, à cause que $A\,B$ est à $B\,E$ comme $C\,E$ est à $G\,E$, $A\,B$ étant $\frac{1}{2}p$, $B\,E$ est $\frac{py}{2n} - \frac{\sqrt{u}}{ny}$.

Et tout de même en supposant que le point C *(fig. 32)* de la courbe a été trouvé par l'intersection des lignes droites $S\,C$ parallèle à $B\,K$, et $A\,C$ parallèle à $S\,V$, $S\,B$ qui est égale à $C\,G$ est y; et $B\,K$ étant égale au côté droit de la parabole que j'ai nommé n, $B\,T$ est $\frac{y^2}{n}$, car comme $K\,B$ est à $B\,S$, ainsi $B\,S$ est à $B\,T$. Et $T\,V$ étant la même que $B\,L$, c'est-à-dire $\frac{2\sqrt{u}}{pn}$, $B\,V$ est $\frac{y^2}{n} - \frac{2\sqrt{u}}{pn}$; et comme $S\,B$ est à $B\,V$, ainsi $A\,B$ est à $B\,E$, qui est par conséquent $\frac{py}{2n} - \frac{\sqrt{u}}{ny}$ comme devant, d'où on voit que c'est une même ligne courbe qui se décrit en ces deux façons.

Après cela, pourceque $B\,L$ et $D\,E$ *(fig. 31)* sont égales, $D\,L$ et $B\,E$ le sont aussi; de façon qu'ajoutant $L\,H$ qui est $\frac{t}{2n\sqrt{u}}$, à $D\,L$ qui est $\frac{py}{2n} - \frac{\sqrt{u}}{ny}$, on a la toute $D\,H$ qui est

$$\frac{py}{2n} - \frac{\sqrt{u}}{ny} + \frac{t}{2n\sqrt{u}};$$

et en ôtant $G\,D$ qui est $\frac{y^2}{n}$, on a $G\,H$ qui est

$$\frac{py}{2n} - \frac{\sqrt{u}}{ny} + \frac{t}{2n\sqrt{u}} - \frac{y^2}{n},$$

ce que j'écris par ordre en cette sorte,

$$G\,H = \frac{-y^3 + \frac{1}{2}py^2 + \frac{ty}{2\sqrt{u}} - \sqrt{u}}{ny},$$

et le carré de $G\,H$ est

$$\frac{y^6 - py^5 + \left(\frac{1}{4}p^2 - \frac{t}{\sqrt{u}}\right)y^4 + \left(2\sqrt{u} + \frac{pt}{2\sqrt{u}}\right)y^3 + \left(\frac{t^2}{4u} - p\sqrt{u}\right)y^2 - ty + u}{n^2y^2}.$$

Et en quelque autre endroit de cette ligne courbe qu'on veuille imaginer le point C, comme vers N ou vers Q, on trouvera toujours que le carré de la ligne droite qui est entre le point H et celui où tombe la perpendiculaire du point C sur $B\,H$, peut être exprimé en ces mêmes termes et avec les mêmes signes $+$ et $-$.

De plus, $H\,I$ étant $\frac{m}{n^2}$, et $L\,H$ étant $\frac{t}{2n\sqrt{u}}$, $I\,L$ est

$$\sqrt{\frac{m^2}{n^4} + \frac{t^2}{4n^2u}},$$

à cause de l'angle droit $I\,H\,L$; et $L\,P$ étant

$$\sqrt{\frac{s}{n^2} + \frac{p\sqrt{u}}{n^2}},$$

$I\,P$ ou $I\,C$ est

$$\sqrt{\frac{m^2}{n^4} + \frac{t^2}{4n^2u} - \frac{s}{n^2} - \frac{p\sqrt{u}}{n^2}},$$

à cause aussi de l'angle droit $I\,P\,L$. Puis ayant fait $C\,M$ perpendiculaire sur $I\,H$, $I\,M$ est la différence qui est entre $H\,I$ et $H\,M$ ou $C\,G$, c'est-à-dire entre $\frac{m}{n^2}$ et y, en sorte que son carré est toujours

$$\frac{m^2}{n^4} - \frac{2my}{n^2} + y^2,$$

qui étant ôté du carré de $I\,C$, il reste

$$\frac{t^2}{4n^2u} - \frac{s}{n^2} - \frac{p\sqrt{u}}{n^2} + \frac{2my}{n^2} - y^2$$

pour le carré de $C\,M$, qui est égal au carré de $G\,H$ déjà trouvé. Ou bien en faisant que cette somme soit divisée comme l'autre par n^2y^2, on a

$$\frac{-n^2y^4 + 2my^3 - p\sqrt{u}y^2 - sy^2 + \frac{t^2}{4u}y^2}{n^2y^2};$$

puis remettant $\frac{t}{\sqrt{u}}y^4 + qy^4 - \frac{1}{4}p^2y^4$ pour n^2y^4, et $ry^3 + 2\sqrt{u}y^3 + \frac{pt}{2\sqrt{u}}y^3$ pour $2my^3$; et multipliant l'une et l'autre somme par n^2y^2 on a

$$y^6 - py^5 + \left(\frac{1}{4}p^2 - \frac{t}{\sqrt{u}}\right)y^4 + \left(2\sqrt{u} + \frac{pt}{2\sqrt{u}}\right)y^3 + \left(\frac{t^2}{4u} - p\sqrt{u}\right)y^2 - ty + u$$

égal à

$$\left(\frac{1}{4}p^2 - q - \frac{t}{\sqrt{u}}\right)y^4 + \left(r + 2\sqrt{u} + \frac{pt}{2\sqrt{u}}\right)y^3 + \left(\frac{t^2}{4u} - s - p\sqrt{u}\right)y^2,$$

c'est-à-dire qu'on a

$$y^6 - py^5 + qy^4 - ry^3 + sy^2 - ty + u = 0.$$

D'où il paroît que les lignes $C\,G$, $N\,R$, $Q\,O$, et semblables, sont les racines de cette équation, qui est ce qu'il falloit démontrer.

L'invention de quatre moyennes proportionnelles.

Ainsi donc si on veut trouver quatre moyennes proportionnelles entre les lignes a et b, ayant posé x pour la première, l'équation est

$$x^5 - a^4 b = 0, \quad \text{ou bien} \quad x^6 - a^4 b x = 0.$$

Et faisant $y - a = x$, il vient

$$y^6 - 6ay^5 + 15a^2y^4 - 20a^3y^3 + 15a^4y^2 - \left(6a^5 + a^4b\right)y + a^6 + a^5b = 0;$$

c'est pourquoi il faut prendre $3a$ pour la ligne $A\,B$, et

$$\sqrt{\frac{6a^3 + a^2b}{\sqrt{a^2 + ab}} + 6a^2}$$

pour $B\,K$ ou le côté droit de la parabole, que j'ai nommé n, et $\frac{a}{3n}\sqrt{a^2 + ab}$ pour $D\,E$ ou $B\,L$. Et après avoir décrit la ligne courbe $A\,C\,N$ sur la mesure de ces trois, il faut faire

$$L\,H = \frac{6a^3 + a^2b}{2n\sqrt{a^2 + ab}}$$

et

$$H\,I = \frac{10a^3}{n^2} + \frac{a^2}{n^2}\sqrt{a^2 + ab} + \frac{18a^4 + 3a^3b}{2n^2\sqrt{a^2 + ab}},$$

et

$$L\,P = \frac{a}{n}\sqrt{15a^2 + 6a\sqrt{a^2 + ab}};$$

car le cercle, qui ayant son centre au point I passera par le point P ainsi trouvé, coupera la courbe aux deux points C et N, desquels ayant tiré les perpendiculaires $N\,R$ et $C\,G$, si la moindre $N\,R$ est ôtée de la plus grande $C\,G$, le reste sera x, la première des quatre moyennes proportionnelles cherchées.

Il est aisé en même façon de diviser un angle en cinq parties égales, et d'inscrire une figure de onze ou treize côtés égaux dans un cercle, et de trouver une infinité d'autres exemples de cette règle.

Toutefois il est à remarquer qu'en plusieurs de ces exemples il peut arriver que le cercle coupe si obliquement la parabole du second genre, que le point de leur intersection soit difficile à reconnoître, et ainsi que cette construction ne soit pas commode pour la pratique ; à quoi il seroit aisé de remédier en composant d'autres règles à l'imitation de celle-ci, comme on en peut composer de mille sortes.

Mais mon dessein n'est pas de faire un gros livre, et je tâche plutôt de comprendre beaucoup en peu de mots, comme on jugera peut-être que j'ai fait, si on considère qu'ayant réduit à une même construction tous les problèmes d'un même genre, j'ai tout ensemble donné la façon de les réduire à une infinité d'autres diverses, et ainsi de résoudre chacun d'eux en une infinité de façons ;

puis outre cela, qu'ayant construit tous ceux qui sont plans en coupant d'un cercle une ligne droite, et tous ceux qui sont solides en coupant aussi d'un cercle une parabole, et enfin tous ceux qui sont d'un degré plus composés en coupant tout de même d'un cercle une ligne qui n'est que d'un degré plus composée que la parabole, il ne faut que suivre la même voie pour construire tous ceux qui sont plus composés à l'infini : car, en matière de progressions mathématiques, lorsqu'on a les deux ou trois premiers termes, il n'est pas malaisé de trouver les autres. Et j'espère que nos neveux me sauront gré, non seulement des choses que j'ai ici expliquées, mais aussi de celles que j'ai omises volontairement, afin de leur laisser le plaisir de les inventer.

FIN.

TABLE DES MATIÈRES

LIVRE PREMIER

DES PROBLÈMES QU'ON PEUT CONSTRUIRE SANS Y EMPLOYER QUE DES CERCLES ET DES LIGNES DROITES.

Comment le calcul d'arithmétique se rapporte aux opérations de géométrie 1
Comment se font géométriquement la multiplication, la division et l'extraction de la racine carrée 1
Comment on peut user de chiffres en géométrie 2
Comment il faut venir aux équations qui servent à résoudre les problèmes 3
Quels sont les problèmes plans, et comment ils se résolvent 4
Exemple tiré de Pappus 5
Réponse à la question de Pappus 7
Comment on doit poser les termes pour venir à l'équation en cet exemple 8
Comment on trouve que ce problème est plan lorsqu'il n'est point proposé en plus de cinq lignes 10

LIVRE SECOND

DE LA NATURE DES LIGNES COURBES.

Quelles sont les lignes courbes qu'on peut recevoir en géométrie 11
La façon de distinguer toutes ces lignes courbes en certains genres, et de connoître le rapport qu'ont tous leurs points à ceux des lignes droites 13
Suite de l'explication de la question de Pappus mise au livre précédent 15
Solution de cette question quand elle n'est proposée qu'en trois ou quatre lignes 15
Démonstration de cette solution 19
Quels sont les lieux plans et solides, et la façon de les trouver tous 21
Quelle est la première et la plus simple de toutes les lignes courbes qui servent à la question des anciens quand elle est proposée en cinq lignes 21
Quelles sont les lignes courbes qu'on décrit en trouvant plusieurs de leurs points qui peuvent être reçus en géométrie 23
Quelles sont aussi celles qu'on décrit avec une corde qui peuvent y être reçues 23
Que, pour trouver toutes les propriétés des lignes courbes, il suffit de savoir le rapport qu'ont tous leurs points à ceux des lignes droites ; et la façon de tirer d'autres lignes qui les coupent en tous ces points à angles droits 23
Façon générale pour trouver des lignes droites qui coupent les courbes données ou leurs contingentes à angles droits 24
Exemple de cette opération en une ellipse et en une parabole du second genre 24
Autre exemple en un ovale du second genre 25
Exemple de la construction de ce problème en la conchoïde 30
Explication de quatre nouveaux genres d'ovales qui servent à l'optique 30
Les propriétés de ces ovales touchant les réflexions et les réfractions ... 33
Démonstration de ces propriétés 34
Comment on peut faire un verre autant convexe ou concave en l'une de ses superficies qu'on voudra, qui rassemble à un point donné tous les rayons qui viennent d'un autre point donné 36
Comment on en peut faire un qui fasse le même, et que la convexité de l'une de ses superficies ait la proportion donnée avec la convexité ou concavité de l'autre 38
Comment on peut rapporter tout ce qui a été dit des lignes courbes décrites sur une superficie plate, à celles qui se décrivent dans un espace qui a trois dimensions, ou bien sur une superficie courbe 39

LIVRE TROISIÈME

DE LA CONSTRUCTION DES PROBLÈMES SOLIDES OU PLUS QUE SOLIDES.

De quelles lignes courbes on peut se servir en la construction de chaque problème 40
Exemple touchant l'invention de plusieurs moyennes proportionnelles . 40
De la nature des équations 41
Combien il peut y avoir de racines en chaque équation 41
Quelles sont les fausses racines 41
Comment on peut diminuer le nombre des dimensions d'une équation, lorsqu'on connoît quelqu'une de ses racines 42
Comment on peut examiner si quelque quantité donnée est la valeur d'une racine 42
Combien il peut y avoir de vraies racines dans chaque équation 42
Comment on fait que les fausses racines deviennent vraies, et les vraies fausses 42
Comment on peut augmenter ou diminuer les racines d'une équation .. 42
Qu'en augmentant ainsi les vraies racines on diminue les fausses, ou au contraire 43
Comment on peut ôter le second terme d'une équation 44
Comment on fait que les fausses racines deviennent vraies sans que les vraies deviennent fausses 45
Comment on fait que toutes les places d'une équation soient remplies . 46
Comment on peut multiplier ou diviser les racines d'une équation 46
Comment on ôte les nombres rompus d'une équation 46
Comment on rend la quantité connue de l'un des termes d'une équation égale à telle autre qu'on veut 47
Que les racines, tant vraies que fausses, peuvent être réelles ou imaginaires 47
La réduction des équations cubiques lorsque le problème est plan 47
La façon de diviser une équation par un binôme qui contient sa racine 48
Quels problèmes sont solides lorsque l'équation est cubique 49
La réduction des équations qui ont quatre dimensions lorsque le problème est plan ; et quels sont ceux qui sont solides 49
Exemple de l'usage de ces réductions 52
Règle générale pour réduire toutes les équations qui passent le carré de carré 53
Façon générale pour construire tous les problèmes solides réduits à une équation de trois ou quatre dimensions 53
L'invention de deux moyennes proportionnelles 56

La division de l'angle en trois 56
Que tous les problèmes solides se peuvent réduire à ces deux constructions 57
La façon d'exprimer la valeur de toutes les racines des équations cubiques, et ensuite de toutes celles qui ne montent que jusques au carré de carré 59
Pourquoi les problèmes solides ne peuvent être construits sans les sections coniques, ni ceux qui sont plus composés sans quelques autres lignes plus composées 60
Façon générale pour construire tous les problèmes réduits à une équation qui n'a point plus de six dimensions 60
L'invention de quatre moyennes proportionnelles 65

FIN DE LA TABLE.

Typographical Errors corrected in Project Gutenberg edition

p. 42 the last factor "et par $x-5$" in original, amended to "et par $x+5$" to match the preceding and following discussion.

p. 61 "entre les points E et D égle à $\frac{2\sqrt{u}}{pn}$" in original, amended to "égale".

End of the Project Gutenberg EBook of La géométrie, by René Descartes

*** END OF THIS PROJECT GUTENBERG EBOOK LA GÉOMÉTRIE ***

***** This file should be named 26400-pdf.pdf or 26400-pdf.zip *****
This and all associated files of various formats will be found in:
http://www.gutenberg.org/2/6/4/0/26400/

Produced by K.F. Greiner, Joshua Hutchinson, Keith Edkins
and the Online Distributed Proofreading Team at
http://www.pgdp.net (This file was produced from images
generously made available by Cornell University Digital
Collections)

Updated editions will replace the previous one--the old editions
will be renamed.

Creating the works from public domain print editions means that no
one owns a United States copyright in these works, so the Foundation
(and you!) can copy and distribute it in the United States without
permission and without paying copyright royalties. Special rules,
set forth in the General Terms of Use part of this license, apply to
copying and distributing Project Gutenberg-tm electronic works to
protect the PROJECT GUTENBERG-tm concept and trademark. Project
Gutenberg is a registered trademark, and may not be used if you
charge for the eBooks, unless you receive specific permission. If you
do not charge anything for copies of this eBook, complying with the
rules is very easy. You may use this eBook for nearly any purpose
such as creation of derivative works, reports, performances and
research. They may be modified and printed and given away--you may do
practically ANYTHING with public domain eBooks. Redistribution is
subject to the trademark license, especially commercial
redistribution.

*** START: FULL LICENSE ***

THE FULL PROJECT GUTENBERG LICENSE
PLEASE READ THIS BEFORE YOU DISTRIBUTE OR USE THIS WORK

To protect the Project Gutenberg-tm mission of promoting the free
distribution of electronic works, by using or distributing this work
(or any other work associated in any way with the phrase "Project
Gutenberg"), you agree to comply with all the terms of the Full Project
Gutenberg-tm License (available with this file or online at
http://gutenberg.org/license).

Section 1. General Terms of Use and Redistributing Project Gutenberg-tm electronic works

1.A. By reading or using any part of this Project Gutenberg-tm electronic work, you indicate that you have read, understand, agree to and accept all the terms of this license and intellectual property (trademark/copyright) agreement. If you do not agree to abide by all the terms of this agreement, you must cease using and return or destroy all copies of Project Gutenberg-tm electronic works in your possession. If you paid a fee for obtaining a copy of or access to a Project Gutenberg-tm electronic work and you do not agree to be bound by the terms of this agreement, you may obtain a refund from the person or entity to whom you paid the fee as set forth in paragraph 1.E.8.

1.B. "Project Gutenberg" is a registered trademark. It may only be used on or associated in any way with an electronic work by people who agree to be bound by the terms of this agreement. There are a few things that you can do with most Project Gutenberg-tm electronic works even without complying with the full terms of this agreement. See paragraph 1.C below. There are a lot of things you can do with Project Gutenberg-tm electronic works if you follow the terms of this agreement and help preserve free future access to Project Gutenberg-tm electronic works. See paragraph 1.E below.

1.C. The Project Gutenberg Literary Archive Foundation ("the Foundation" or PGLAF), owns a compilation copyright in the collection of Project Gutenberg-tm electronic works. Nearly all the individual works in the collection are in the public domain in the United States. If an individual work is in the public domain in the United States and you are located in the United States, we do not claim a right to prevent you from copying, distributing, performing, displaying or creating derivative works based on the work as long as all references to Project Gutenberg are removed. Of course, we hope that you will support the Project Gutenberg-tm mission of promoting free access to electronic works by freely sharing Project Gutenberg-tm works in compliance with the terms of this agreement for keeping the Project Gutenberg-tm name associated with the work. You can easily comply with the terms of this agreement by keeping this work in the same format with its attached full Project Gutenberg-tm License when you share it without charge with others.

1.D. The copyright laws of the place where you are located also govern what you can do with this work. Copyright laws in most countries are in a constant state of change. If you are outside the United States, check the laws of your country in addition to the terms of this agreement before downloading, copying, displaying, performing, distributing or creating derivative works based on this work or any other Project Gutenberg-tm work. The Foundation makes no representations concerning the copyright status of any work in any country outside the United States.

1.E. Unless you have removed all references to Project Gutenberg:

1.E.1. The following sentence, with active links to, or other immediate access to, the full Project Gutenberg-tm License must appear prominently whenever any copy of a Project Gutenberg-tm work (any work on which the phrase "Project Gutenberg" appears, or with which the phrase "Project Gutenberg" is associated) is accessed, displayed, performed, viewed, copied or distributed:

This eBook is for the use of anyone anywhere at no cost and with almost no restrictions whatsoever. You may copy it, give it away or re-use it under the terms of the Project Gutenberg License included with this eBook or online at www.gutenberg.org

1.E.2. If an individual Project Gutenberg-tm electronic work is derived from the public domain (does not contain a notice indicating that it is posted with permission of the copyright holder), the work can be copied and distributed to anyone in the United States without paying any fees or charges. If you are redistributing or providing access to a work with the phrase "Project Gutenberg" associated with or appearing on the work, you must comply either with the requirements of paragraphs 1.E.1 through 1.E.7 or obtain permission for the use of the work and the Project Gutenberg-tm trademark as set forth in paragraphs 1.E.8 or 1.E.9.

1.E.3. If an individual Project Gutenberg-tm electronic work is posted with the permission of the copyright holder, your use and distribution must comply with both paragraphs 1.E.1 through 1.E.7 and any additional terms imposed by the copyright holder. Additional terms will be linked to the Project Gutenberg-tm License for all works posted with the permission of the copyright holder found at the beginning of this work.

1.E.4. Do not unlink or detach or remove the full Project Gutenberg-tm License terms from this work, or any files containing a part of this work or any other work associated with Project Gutenberg-tm.

1.E.5. Do not copy, display, perform, distribute or redistribute this electronic work, or any part of this electronic work, without prominently displaying the sentence set forth in paragraph 1.E.1 with active links or immediate access to the full terms of the Project Gutenberg-tm License.

1.E.6. You may convert to and distribute this work in any binary, compressed, marked up, nonproprietary or proprietary form, including any word processing or hypertext form. However, if you provide access to or distribute copies of a Project Gutenberg-tm work in a format other than "Plain Vanilla ASCII" or other format used in the official version posted on the official Project Gutenberg-tm web site (www.gutenberg.org), you must, at no additional cost, fee or expense to the user, provide a copy, a means of exporting a copy, or a means of obtaining a copy upon

request, of the work in its original "Plain Vanilla ASCII" or other form. Any alternate format must include the full Project Gutenberg-tm License as specified in paragraph 1.E.1.

1.E.7. Do not charge a fee for access to, viewing, displaying, performing, copying or distributing any Project Gutenberg-tm works unless you comply with paragraph 1.E.8 or 1.E.9.

1.E.8. You may charge a reasonable fee for copies of or providing access to or distributing Project Gutenberg-tm electronic works provided that

- You pay a royalty fee of 20% of the gross profits you derive from the use of Project Gutenberg-tm works calculated using the method you already use to calculate your applicable taxes. The fee is owed to the owner of the Project Gutenberg-tm trademark, but he has agreed to donate royalties under this paragraph to the Project Gutenberg Literary Archive Foundation. Royalty payments must be paid within 60 days following each date on which you prepare (or are legally required to prepare) your periodic tax returns. Royalty payments should be clearly marked as such and sent to the Project Gutenberg Literary Archive Foundation at the address specified in Section 4, "Information about donations to the Project Gutenberg Literary Archive Foundation."

- You provide a full refund of any money paid by a user who notifies you in writing (or by e-mail) within 30 days of receipt that s/he does not agree to the terms of the full Project Gutenberg-tm License. You must require such a user to return or destroy all copies of the works possessed in a physical medium and discontinue all use of and all access to other copies of Project Gutenberg-tm works.

- You provide, in accordance with paragraph 1.F.3, a full refund of any money paid for a work or a replacement copy, if a defect in the electronic work is discovered and reported to you within 90 days of receipt of the work.

- You comply with all other terms of this agreement for free distribution of Project Gutenberg-tm works.

1.E.9. If you wish to charge a fee or distribute a Project Gutenberg-tm electronic work or group of works on different terms than are set forth in this agreement, you must obtain permission in writing from both the Project Gutenberg Literary Archive Foundation and Michael Hart, the owner of the Project Gutenberg-tm trademark. Contact the Foundation as set forth in Section 3 below.

1.F.

1.F.1. Project Gutenberg volunteers and employees expend considerable effort to identify, do copyright research on, transcribe and proofread public domain works in creating the Project Gutenberg-tm collection. Despite these efforts, Project Gutenberg-tm electronic works, and the medium on which they may be stored, may contain "Defects," such as, but not limited to, incomplete, inaccurate or corrupt data, transcription errors, a copyright or other intellectual property infringement, a defective or damaged disk or other medium, a computer virus, or computer codes that damage or cannot be read by your equipment.

1.F.2. LIMITED WARRANTY, DISCLAIMER OF DAMAGES - Except for the "Right of Replacement or Refund" described in paragraph 1.F.3, the Project Gutenberg Literary Archive Foundation, the owner of the Project Gutenberg-tm trademark, and any other party distributing a Project Gutenberg-tm electronic work under this agreement, disclaim all liability to you for damages, costs and expenses, including legal fees. YOU AGREE THAT YOU HAVE NO REMEDIES FOR NEGLIGENCE, STRICT LIABILITY, BREACH OF WARRANTY OR BREACH OF CONTRACT EXCEPT THOSE PROVIDED IN PARAGRAPH F3. YOU AGREE THAT THE FOUNDATION, THE TRADEMARK OWNER, AND ANY DISTRIBUTOR UNDER THIS AGREEMENT WILL NOT BE LIABLE TO YOU FOR ACTUAL, DIRECT, INDIRECT, CONSEQUENTIAL, PUNITIVE OR INCIDENTAL DAMAGES EVEN IF YOU GIVE NOTICE OF THE POSSIBILITY OF SUCH DAMAGE.

1.F.3. LIMITED RIGHT OF REPLACEMENT OR REFUND - If you discover a defect in this electronic work within 90 days of receiving it, you can receive a refund of the money (if any) you paid for it by sending a written explanation to the person you received the work from. If you received the work on a physical medium, you must return the medium with your written explanation. The person or entity that provided you with the defective work may elect to provide a replacement copy in lieu of a refund. If you received the work electronically, the person or entity providing it to you may choose to give you a second opportunity to receive the work electronically in lieu of a refund. If the second copy is also defective, you may demand a refund in writing without further opportunities to fix the problem.

1.F.4. Except for the limited right of replacement or refund set forth in paragraph 1.F.3, this work is provided to you 'AS-IS' WITH NO OTHER WARRANTIES OF ANY KIND, EXPRESS OR IMPLIED, INCLUDING BUT NOT LIMITED TO WARRANTIES OF MERCHANTIBILITY OR FITNESS FOR ANY PURPOSE.

1.F.5. Some states do not allow disclaimers of certain implied warranties or the exclusion or limitation of certain types of damages. If any disclaimer or limitation set forth in this agreement violates the law of the state applicable to this agreement, the agreement shall be interpreted to make the maximum disclaimer or limitation permitted by the applicable state law. The invalidity or unenforceability of any provision of this agreement shall not void the remaining provisions.

1.F.6. INDEMNITY - You agree to indemnify and hold the Foundation, the trademark owner, any agent or employee of the Foundation, anyone providing copies of Project Gutenberg-tm electronic works in accordance with this agreement, and any volunteers associated with the production, promotion and distribution of Project Gutenberg-tm electronic works, harmless from all liability, costs and expenses, including legal fees, that arise directly or indirectly from any of the following which you do or cause to occur: (a) distribution of this or any Project Gutenberg-tm work, (b) alteration, modification, or additions or deletions to any Project Gutenberg-tm work, and (c) any Defect you cause.

Section 2. Information about the Mission of Project Gutenberg-tm

Project Gutenberg-tm is synonymous with the free distribution of electronic works in formats readable by the widest variety of computers including obsolete, old, middle-aged and new computers. It exists because of the efforts of hundreds of volunteers and donations from people in all walks of life.

Volunteers and financial support to provide volunteers with the assistance they need, is critical to reaching Project Gutenberg-tm's goals and ensuring that the Project Gutenberg-tm collection will remain freely available for generations to come. In 2001, the Project Gutenberg Literary Archive Foundation was created to provide a secure and permanent future for Project Gutenberg-tm and future generations. To learn more about the Project Gutenberg Literary Archive Foundation and how your efforts and donations can help, see Sections 3 and 4 and the Foundation web page at http://www.pglaf.org.

Section 3. Information about the Project Gutenberg Literary Archive Foundation

The Project Gutenberg Literary Archive Foundation is a non profit 501(c)(3) educational corporation organized under the laws of the state of Mississippi and granted tax exempt status by the Internal Revenue Service. The Foundation's EIN or federal tax identification number is 64-6221541. Its 501(c)(3) letter is posted at http://pglaf.org/fundraising. Contributions to the Project Gutenberg Literary Archive Foundation are tax deductible to the full extent permitted by U.S. federal laws and your state's laws.

The Foundation's principal office is located at 4557 Melan Dr. S. Fairbanks, AK, 99712., but its volunteers and employees are scattered throughout numerous locations. Its business office is located at 809 North 1500 West, Salt Lake City, UT 84116, (801) 596-1887, email business@pglaf.org. Email contact links and up to date contact information can be found at the Foundation's web site and official

page at http://pglaf.org

For additional contact information:
Dr. Gregory B. Newby
Chief Executive and Director
gbnewby@pglaf.org

Section 4. Information about Donations to the Project Gutenberg Literary Archive Foundation

Project Gutenberg-tm depends upon and cannot survive without wide spread public support and donations to carry out its mission of increasing the number of public domain and licensed works that can be freely distributed in machine readable form accessible by the widest array of equipment including outdated equipment. Many small donations ($1 to $5,000) are particularly important to maintaining tax exempt status with the IRS.

The Foundation is committed to complying with the laws regulating charities and charitable donations in all 50 states of the United States. Compliance requirements are not uniform and it takes a considerable effort, much paperwork and many fees to meet and keep up with these requirements. We do not solicit donations in locations where we have not received written confirmation of compliance. To SEND DONATIONS or determine the status of compliance for any particular state visit http://pglaf.org

While we cannot and do not solicit contributions from states where we have not met the solicitation requirements, we know of no prohibition against accepting unsolicited donations from donors in such states who approach us with offers to donate.

International donations are gratefully accepted, but we cannot make any statements concerning tax treatment of donations received from outside the United States. U.S. laws alone swamp our small staff.

Please check the Project Gutenberg Web pages for current donation methods and addresses. Donations are accepted in a number of other ways including checks, online payments and credit card donations. To donate, please visit: http://pglaf.org/donate

Section 5. General Information About Project Gutenberg-tm electronic works.

Professor Michael S. Hart is the originator of the Project Gutenberg-tm concept of a library of electronic works that could be freely shared with anyone. For thirty years, he produced and distributed Project Gutenberg-tm eBooks with only a loose network of volunteer support.

Project Gutenberg-tm eBooks are often created from several printed
editions, all of which are confirmed as Public Domain in the U.S.
unless a copyright notice is included. Thus, we do not necessarily
keep eBooks in compliance with any particular paper edition.

Most people start at our Web site which has the main PG search facility:

     http://www.gutenberg.org

This Web site includes information about Project Gutenberg-tm,
including how to make donations to the Project Gutenberg Literary
Archive Foundation, how to help produce our new eBooks, and how to
subscribe to our email newsletter to hear about new eBooks.

LA GEOMETRÍA DE DESCARTES

Todos los problemas de Geometría pueden reducirse fácilmente a términos tales, que no es necesario conocer de antemano más que la longitud de algunas líneas rectas para construirlos.
Descartes. *La Geometría* (G.AT.VI.369).

Descartes mediante un nuevo método hizo pasar de las tinieblas a la luz cuanto en las Matemáticas había permanecido inaccesible a los antiguos y todo cuanto los contemporáneos habían sido incapaces de descubrir; luego puso los cimientos inquebrantables de la Filosofía sobre los cuales es posible asentar la mayor parte de las verdades en el orden y con la certidumbre de las Matemáticas.
Spinoza. *Los Principios de la Filosofía cartesiana.*

Lo que ha inmortalizado el nombre de Descartes es la aplicación que ha sabido hacer del Álgebra a la Geometría, una idea de las más vastas y felices que haya tenido el espíritu humano, y que será siempre la llave de los más profundos descubrimientos no solamente en la Geometría, sino en todas las ciencias físico-matemáticas.
D'Alembert. *Discours Préliminaire de l'Encyclopédie* (Orbis, 1984, pp.84–85).

La Geometría analítica, mucho más que cualquiera de sus especulaciones metafísicas, inmortaliza el nombre de Descartes y constituye el máximo paso hecho en el progreso de las ciencias exactas.
J. Stuart Mill. (citado por E.Bell en *Les grands mathématiciens*. Payot, París, 1950. Cap.3. p.46).

Introducción.

- **Del Álgebra Geométrica griega a la Geometría Analítica Cartesiana.**
- **El Álgebra Geométrica del Libro II de *Los Elementos* de Euclides.**
- **Coordenadas en *Las Cónicas* de Apolonio.**
- **El Análisis Geométrico griego y la Geometría Analítica.**

***La Geometría* de Descartes.**

- **La formación de Descartes en La Flèche.**
- **Citas memorables de Descartes.**
- **Citas memorables sobre Descartes.**
- **Los sueños de Descartes y los orígenes de *La Geometría.***
- ***Las Regulae, El Discurso del Método y La Geometría.***
- **La percepción de Descartes sobre el eco científico de *La Geometría.***
- **El contenido de *La Geometría.***
  - **La construcción geométrico-algebraica de las operaciones aritméticas.**
  - **La notación matemática cartesiana.**
  - ***Análisis y Síntesis*: planteamiento y resolución de las ecuaciones.**
  - **Sistemas de referencia. *El Problema de Pappus*.**
  - **Las rectas normales a una curva. *El Método del círculo*.**
- ***La Geometría* de Descartes y la Geometría Analítica.**
- **La proyección histórica de la Geometría Analítica cartesiana.**
- **Bibliografía.**

Introducción:
Del Álgebra Geométrica griega a la Geometría Analítica Cartesiana

La Geometría Analítica es un poderoso instrumento de ataque de los problemas geométricos que utiliza como herramienta básica el Álgebra. La esencia de su aplicación en el plano es el establecimiento de una correspondencia entre los puntos del plano y pares ordenados de números reales, es decir, un sistema de coordenadas, lo que posibilita una asociación entre curvas del plano y ecuaciones en dos variables, de modo que cada curva del plano tiene asociada una ecuación f(x,y)=0 y, recíprocamente, para cada ecuación en dos variables está definida una curva que determina un conjunto de puntos en el plano, siempre respecto de un sistema de coordenadas. En particular queda establecida una asociación entre rectas del plano y ecuaciones de primer grado de la forma Ax+By+C=0. La Geometría Analítica es, pues, una especie de diccionario entre el Álgebra y la Geometría que asocia pares de números a puntos y ecuaciones a curvas. Pero esta asociación va más allá de lo gramatical ya que vincula también las sintaxis del Álgebra y de la Geometría, es decir, las relaciones, vínculos y operaciones entre los elementos de ambas. Así pues, para hallar geométricamente la intersección de dos curvas f(x,y)=0, g(x,y)=0 –problema geométrico– habría que resolver algebraicamente el sistema formado por ambas ecuaciones –problema algebraico–. Además, para cada curva f(x,y)=0, la Geometría Analítica establece también una correspondencia entre las propiedades algebraicas y analíticas de la ecuación f(x,y)=0 y las propiedades geométricas de la curva asociada. De hecho, estas propiedades geométricas son el trasunto geométrico de la estructura algebraica de la expresión f(x,y)=0 y se establecen mediante el cálculo literal que permite el Álgebra. En particular la tarea de probar un teorema o resolver un problema en Geometría se traslada de forma muy eficiente a probarlo o resolverlo en Álgebra utilizando el cálculo analítico.

Es indiscutible que Fermat y Descartes son los verdaderos artífices de la Geometría Analítica. Descartes publica en 1637 *La Geometría*, junto con *La Dióptrica y Los Meteoros* como apéndices de su *Discurso del Método* o éste como prólogo de aquellos opúsculos. El mismo año, Fermat envía al Padre Mersenne sus investigaciones de alrededor de 1629 contenidas en la memoria *Introducción a los Lugares Planos y Sólidos* (*Ad Locos Planos et Solidos Isagoge*). Las obras citadas de Descartes y Fermat contienen los fundamentos de la llamada más tarde Geometría Analítica.

Hay una gran unanimidad en considerar a *La Geometría* de Descartes como una de las obras más importantes en la historia del pensamiento matemático. Al utilizar el Álgebra simbólica como herramienta algorítmica básica, Descartes realiza una nueva lectura de la Geometría griega, que supera sus limitaciones y rebasa sus conquistas geométricas. A base de elaborar una excelente herramienta para enfrentar y resolver problemas geométricos antiguos y modernos, Descartes libera a la Geometría de la dependencia a la estructura geométrica de las figuras e introduce una forma de solución de los problemas basada en la aplicación del Análisis mediante la intervención del Álgebra.

El Análisis Geométrico griego utilizaba un equivalente de las coordenadas pero sólo empleaba Álgebra Geométrica. El *Arte Analítica* de Vieta desarrolla el Álgebra simbólica pero no usa coordenadas. Al aunar ambos instrumentos, coordenadas y Álgebra literal, Descartes alumbra la Geometría Analítica que establece un puente para transitar entre la Geometría y el Álgebra, al permitir asociar curvas y ecuaciones, a base de aplicar el Análisis algebraico de Vieta a los problemas de lugares geométricos de Apolonio y Pappus, definidos, en un sistema de coordenadas, por una ecuación indeterminada en dos incógnitas. La Geometría Analítica resultante, dotada del simbolismo literal, con toda la potencia algorítmica de la mecánica operatoria del cálculo, manipulación y simplificación que permite el Álgebra, sustituye las ingeniosas construcciones geométricas de la rígida y retórica *Álgebra Geométrica* de los griegos por sistemáticas operaciones algebraicas y se convierte en una poderosa herramienta de investigación, mediante la cual Descartes resuelve de forma brillante y asombrosa, numerosos problemas geométricos, clásicos y modernos, algunos realmente difíciles, como el trazado de normales a las curvas, el *Problema de Apolonio* y otros que se habían resistido a lo largo de la Historia como el famoso *Problema de Pappus*.

Aquí vamos a realizar un estudio crítico de la obra de Descartes, *La Geometría*. Para valorar la trascendencia de esta obra en la Historia de la Matemática, haremos una breve descripción de los métodos de la Geometría griega, no sólo porque por comparación podremos ponderar la eficiencia de los métodos cartesianos sino porque la motivación y el origen de la obra cartesiana arranca de su lectura por parte de Descartes y la crítica de sus limitaciones. También conviene remontarse a la Geometría griega para rastrear ciertos vestigios analíticos, entre ellos el uso rudimentario de coordenadas en Apolonio y la naturaleza del Análisis Geométrico griego como instrumento de investigación, ya que ambos elementos son primigenios antecedentes de la Geometría Analítica cartesiana.

La Geometría Analítica de Descartes es un salto revolucionario sin precedentes en la Historia de la Matemática. Para valorar en su justo valor el nuevo instrumento científico, así como para comprender cómo tuvo lugar su gestación es imprescindible conocer la naturaleza de la Geometría griega, condicionada por el veto al infinito que trajo la aparición de los inconmensurables, con la consiguiente estructuración rígida de la Matemática griega elemental en la enciclopédica obra de *Los Elementos* de Euclides, que establece de forma paradigmática un estilo sintético de exposición que oculta la vía heurística del descubrimiento, impulsa la Geometría al margen de la Aritmética, impide el desarrollo de un Álgebra en sentido algorítmico y simbólico y limita la introducción de nuevas curvas a su construcción mediante intersección de superficies o lugares geométricos definidos a través de relaciones de áreas o longitudes, en forma de proporción, y no por medio de ecuaciones.

La estructura que adopta esta Matemática se llama el *Álgebra Geométrica* de los griegos. Se trata de una especie de Geometría algebraica, resultado de la geometrización de los métodos algebraicos mesopotámicos, en la que los números son sustituidos por segmentos de recta y las operaciones entre ellos se llevan a cabo, mediante construcciones geométricas que obligan a mantener escrupulosamente la homogeneidad de los términos. Esta teoría constituye una potente técnica de resolución de ecuaciones –muy rigurosa aunque un tanto onerosa para nosotros–, que se llama el método de *Aplicación de las Áreas*. Mediante una sofisticada aplicación de este método, Apolonio construye con un inefable virtuosismo su famosa obra *Las Cónicas*, donde aparece una aplicación muy incipiente de las coordenadas. De todas estas cuestiones –que interesan al origen de la Geometría Analítica– se habla en los capítulos introductorios. Y también del método de *Análisis* de los griegos, del que la Geometría Analítica recibirá no sólo su nombre sino sobre todo sus procedimientos. Se trata en particular el concepto que sobre él tenían Platón y Pappus y la visión de Descartes sobre el mismo en las *Reglas para la dirección del espíritu*.

El objetivo fundamental de este trabajo es desentrañar las raíces de la Geometría Analítica en el pensamiento filosófico y matemático cartesianos. Por eso se ha dedicado una generosa extensión a la importante cuestión del anclaje de *La Geometría* de Descartes en su obra filosófica y en particular en *El Discurso del Método* y en las *Reglas para la dirección del espíritu*, obras donde se sitúa la metodología cartesiana, en particular los preceptos del *Análisis* y la *Síntesis* que Descartes aplicará de forma constante en *La Geometría*. Así pues, se analizan de forma exhaustiva los textos de esas obras de Descartes, que son esenciales para entender como se va fraguando la motivación y la estructuración de la metodología cartesiana de *La Geometría*.

Se estudia en un capítulo, de forma sucinta, el contenido general de La Geometría, pero en los siguientes capítulos se concretan, con gran extensión, los aspectos de la obra cartesiana que interesan a los orígenes de la Geometría Analítica: La construcción geométrico-algebraica de las operaciones aritméticas, la notación matemática cartesiana, la aplicación del Análisis y la Síntesis en el planteamiento y resolución de las ecuaciones, los sistemas de referencia en el estudio del *Problema de Pappus* y la construcción de las rectas normales a una curva mediante el método del círculo.

La Geometría Analítica es mucho más que una mera combinación de Álgebra y Geometría. Para poder circular del Álgebra a la Geometría y de la Geometría al Álgebra se necesita como ingredientes ineludibles no sólo el carácter algorítmico operatorio del Álgebra simbólica sino también la aplicación de las coordenadas. Una aproximación al uso de éstas ya tuvo lugar en la Geometría griega con Apolonio y Pappus, pero el Álgebra simbólica no

se desarrolla de forma satisfactoria hasta los trabajos de Vieta. Al vincular ambos elementos en los desarrollos de Descartes, emerge la Geometría Analítica de forma inexorable. Pero como en cualquier creación humana, la de Descartes es tributaria de importantes desarrollos matemáticos anteriores. Por eso, a lo largo de este estudio acerca origen de la Geometría Analítica en La Geometría de Descartes intentamos clarificar la influencia de los diversos hitos históricos –geométricos y algebraicos– sobre el hallazgo cartesiano. Esta es la razón por la cual hemos incluido aspectos de la Historia de la Geometría griega y de la evolución del Álgebra sin los cuales no se entendería la obra de Descartes y su enorme incidencia en la Historia de la Matemática.

A fin de concretar el significado de los términos, digamos que entendemos por Geometría Analítica lo que hemos descrito más arriba como su esencia y que ahora sintetizamos:

> «*La aplicación del Álgebra simbólica al estudio de problemas geométricos mediante la asociación de curvas y ecuaciones indeterminadas en un sistema de coordenadas*.»

Esta definición es la que nos ha guiado para entresacar de toda la obra cartesiana los elementos que se refieren a lo que dos siglos después de Descartes se llamó Geometría Analítica. Así pues, se empieza por la descripción de los preliminares geométrico-algebraicos que Descartes estudia con la denominación «*Cómo el cálculo de la aritmética se relaciona con las operaciones de geometría*», de una gran importancia, porque en este apartado Descartes soslaya el problema de la inconmensurabilidad, al asignar longitudes a los segmentos, previa la adopción de un segmento unidad a discreción, tras lo cual construye de forma efectiva las operaciones aritméticas dándoles significado geométrico. De esta forma Descartes elimina la limitación pitagórica de la inconmensurabilidad.

El siguiente punto esencial es la simplificación de la notación algebraica, una cuestión intrínsecamente vinculada a los métodos de la Geometría Analítica. Tanto es así, que en todo estudio histórico sobre la Geometría Analítica una parte importante la ocupa la evolución histórica del simbolismo, que alcanza su clímax en los aportes del propio Descartes a la notación algebraica, ingrediente esencial del descubrimiento cartesiano, que aparece primero en las *Regulae* y después en *La Geometría*. En el apartado de ésta titulado «*Cómo pueden emplearse letras en geometría*», Descartes considera un segmento de recta tanto como magnitud geométrica continua como una medida numérica, pero establece que la potencia de un segmento sigue siendo un segmento, de modo que cuadrado y cubo ya no son magnitudes planas o espaciales, sino la segunda o tercera potencia de un número. De este modo, las operaciones aritméticas quedan incluidas en un terreno estrictamente algebraico. Con ello Descartes elimina otra limitación, la euclídea de la homogeneidad.

Continúa el trabajo cartesiano con la aplicación de la metodología cartesiana del Análisis y la Síntesis en el planteamiento y resolución de ecuaciones que corresponden a los *problemas planos*, donde se desarrolla todo un protocolo de actuación –suponer el problema resuelto; dar nombre a todos los segmentos que parecen necesarios para representar los datos del problema, tanto los conocidos como los desconocidos; determinar la ecuación entre las longitudes conocidas y las desconocidas; resolver la ecuación resultante; construir geométricamente la solución–. Se trata de un verdadero método de resolución de problemas geométricos donde se transita de forma reversible de la Geometría al Álgebra y del Álgebra a al Geometría. En particular, Descartes exhibe de forma ostentosa eficientes métodos de resolución de ecuaciones y de construcción geométrica de las soluciones, que contrastan con la farragosidad del Álgebra Geométrica de *Los Elementos* de Euclides. Realmente aquí vemos la magnificencia y simplicidad de los métodos de *La Geometría* de Descartes en contraposición a la prolijidad y precariedad de la Geometría griega.

Sigue a continuación un tratamiento exhaustivo del histórico *Problema de Pappus* donde Descartes introduce el primer sistema de coordenadas de *La Geometría*. Este problema fue un indicador fehaciente, ante el público científico coetáneo, de la novedad y de la inusitada potencialidad del método analítico cartesiano en Geometría en un asunto geométrico que desbordó a lo largo de los siglos las posibilidades del Análisis geométrico griego.

Finalmente se considera el problema más querido y anhelado para Descartes, según sus propias palabras, la determinación de las rectas normales a una curva, donde se resuelve

de forma prodigiosa el problema de normales y tangentes, pero sobre todo se apunta a la asociación de curvas y ecuaciones que instaura los dos principios fundamentales de la llamada Geometría Analítica:

- La relación entre las coordenadas de los puntos de una curva –la ecuación de la curva– establece una correspondencia entre las propiedades algebraicas de la expresión de la ecuación y las propiedades geométricas de la curva asociada.
- La intersección de curvas –que es un problema geométrico– se reconduce a la resolución de sistemas de ecuaciones – que es un problema algebraico–.

La Geometría de Descartes traslada de la Geometría al Álgebra la resolución de los problemas geométricos y además convierte al Álgebra en un magnífico instrumento de exploración e investigación geométrica. Por ejemplo, la realización de ciertos cálculos y en particular la resolución de algunas ecuaciones vinculadas a la expresión de la curva, permiten la obtención de los elementos geométricos notables de la misma, es decir, diámetros, ejes, asíntotas, centros, etc. Incluso Descartes habla de su aplicación «*a la medida del espacio que abarcan*», queriendo indicar, tal vez, las cuadraturas.

Y más todavía, la propia expresión analítica de la ecuación de una curva es una incipiente aproximación al concepto de función.

Como una manifestación de la trascendencia de *La Geometría* de Descartes en la Historia de la Matemática, nos ha parecido conveniente describir la decisiva influencia de la obra cartesiana en el descubrimiento y desarrollo del Cálculo Infinitesimal.

Hacia el final se relaciona de forma muy sucinta la rápida evolución de la Geometría Analítica poscartesiana, hasta situarnos en el umbral de la Geometría Analítica moderna, la que se imparte hoy académicamente, salvo en lo que se refiere al instrumento vectorial.

Acabamos con unas reflexiones sobre la proyección histórica de la Geometría Analítica cartesiana, una potente herramienta que domina el pensamiento matemático desde la época de Descartes, que al aportar simplificación, generalización, mecanización, unificación, flexibilidad, versatilidad, claridad, economía, brevedad y difusión, se convierte en el lenguaje universal de las ciencias. La Geometría Analítica cambió el rostro de las Matemáticas y el semblante de la Educación matemática.

Además de la consulta a diversos textos de Historia de la Matemática, Historia de la Filosofía, Filosofía de la Ciencia y de la Matemática, artículos de revistas científicas, etc., el manantial bibliográfico fundamental utilizado ha sido, por una parte, Obras originales de los principales matemáticos griegos (Euclides, Apolonio, Diofanto y Pappus), y por otra de Vieta y Descartes. En concreto, dignas son de mención las ediciones de Blanchard, en francés, de Paul Ver Eecke de *Las Cónicas* de Apolonio, *La Aritmética* de Diofanto y *La Colección Matemática* de Pappus; y las mismas obras en ediciones en español de F.Vera incluidas en *Científicos griegos* (Aguilar, Madrid, 1970).

Los argumentos de los diversos capítulos se sustentan de forma esencial en los textos originales de Descartes. Aparte de diversas ediciones, en español, de *El Discurso del Método* y de las *Reglas para la dirección del espíritu*, la referencia esencial ha sido *Oeuvres de Descartes*, Publicadas por C.Adam et P.Tannery (Librairie philosophique J.Vrin, París, 1964-74), sobre todo el volumen VI que contiene *El Discours de la Méthode* y *La Géométrie* y el volumen X que contiene las *Regulae ad directionem ingenii*. Han sido también de una gran utilidad las ediciones de *La Geometría*, en español (Espasa-Calpe y Alfaguara), inglés (Dover) y en especial la magnífica edición en catalán del Institut d'Estudis Catalans (Barcelona, 1999) con introducción y notas de J. Pla y P. Viader.

La referencia concreta a un texto de Descartes se hará respecto a *Oeuvres de Descartes*, indicando la página a continuación de la partícula DM.AT,VI., G.AT.VI., o R.AT.X. respectivamente, según se trate de *El Discurso del Método*, *La Geometría* o las *Regulae*. Por ejemplo (G.AT,VI,372) indicará que el texto al que se hace alusión se encuentra en la página 372 del sexto tomo de las *Oeuvres* de Descartes, que contiene *La Geometría*.

El Álgebra Geométrica del Libro II de *Los Elementos* de Euclides

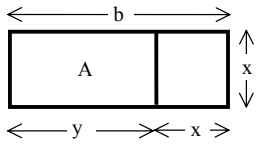
Como consecuencia de la aparición de las magnitudes inconmensurables, los griegos no podían reconocer la existencia de números irracionales, lo que les dificultaba el tratamiento numérico de longitudes, áreas, volúmenes y ángulos. Esta limitación operacional junto a un deficiente sistema de numeración que utilizaba las letras del alfabeto para representar los números enteros, con la consiguiente dificultad para realizar las operaciones, impedía asignar a las figuras geométricas números que midieran sus longitudes, áreas y volúmenes y por tanto los griegos tenían que tratar directamente con las figuras a modo de magnitudes. El abismo infranqueable que se había abierto entre número y magnitud continua impedía someter las magnitudes geométricas a manipulaciones algebraicas, como se hace con los números, lo que determinó la transformación del Álgebra oriental que los pitagóricos habían heredado de los babilonios en el *Álgebra Geométrica* del Libro II de *Los Elementos* de Euclides que juega un papel fundamental en la Geometría griega. Con gran habilidad en la práctica geométrica, los griegos hicieron de su *Álgebra Geométrica* un poderoso instrumento para la resolución de ecuaciones, mediante el método de la *Aplicación de las Áreas*, teoría que según Proclo sería de ascendencia pitagórica.

El *Álgebra Geométrica*, denominación acuñada por el historiador de la Matemática H.G. Zeuthen hacia 1886, viene a ser una geometrización de los métodos algebraicos practicados por los babilónicos, una especie de Geometría algebraica, en la que los números son sustituidos por segmentos de recta y las operaciones entre ellos se llevan a cabo mediante construcciones geométricas –respetando escrupulosamente la homogeneidad de los términos– de la siguiente forma:

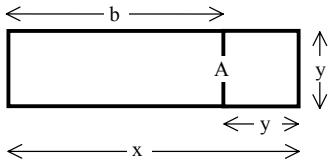
- La suma de dos números se obtiene prolongando sobre el primero un segmento igual al segundo.
- La diferencia de dos números se obtiene recortando del primero un segmento igual al segundo.
- El producto de dos números es el área del rectángulo cuyos lados tienen como longitudes esos números.
- El cociente de dos números es la razón de los segmentos que los representan (según los principios del libro V de *Los Elementos* de Euclides).
- La suma y la diferencia de productos se reemplaza por la adición y sustracción de rectángulos.
- La extracción de una raíz cuadrada se establece mediante la construcción de un cuadrado de área equivalente a la de un rectángulo dado (*Euclides* II.14).

Por ejemplo, el viejo problema mesopotámico en el que dada la suma o diferencia y el producto de los lados de un rectángulo, $x\cdot y=A$, $x\pm y=b$, se pedía hallar dichos lados, se interpretaba geométricamente de la siguiente forma:

$$\left.\begin{array}{l} x\cdot y = A \\ x+y=b \end{array}\right\} \; y=b-x \;,\; x\cdot(b-x)=A \;,\; bx-x^2=A$$

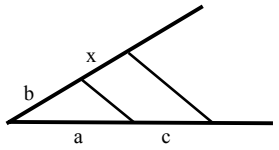


$$\left.\begin{array}{l} x\cdot y = A \\ x-y=b \end{array}\right\} \; x=b+y \;,\; y\cdot(b+y)=A \;,\; by+y^2=A$$



La solución geométrica lleva a la construcción sobre un segmento b de un rectángulo cuya altura desconocida x debe ser tal que el área del rectángulo en cuestión exceda del área dada A (en el caso de signo positivo) en el cuadrado de lado x; o difiera del área dada (en el caso de signo negativo) en el cuadrado de lado y.

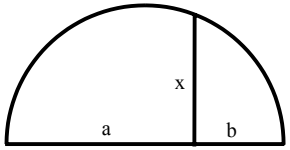
En su Álgebra Geométrica los griegos utilizaron principalmente dos métodos para resolver cierto tipo de ecuaciones, el método de las proporciones y el método de *Aplicación de las Áreas*. El método de las proporciones permite construir exactamente, como se hace hoy, un segmento de línea x dado por:



$$\frac{a}{b}=\frac{c}{x},\ a\cdot x = b\cdot c$$

Se aplica la cuarta proporcional (*Euclides* VI.12)

o bien:



$$\frac{a}{x}=\frac{x}{b},\ x^2 = a\cdot b$$

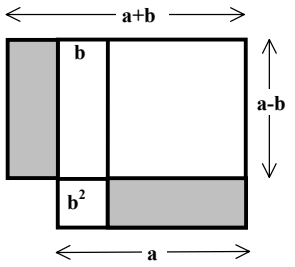
Se aplica la media proporcional (*Euclides* VI.13)

No obstante la inseguridad provocada en la Matemática griega por las magnitudes inconmensurables, conducía a evitar a toda costa el uso de razones en la Geometría elemental. Por eso el tratamiento de ecuaciones tan sencillas como $a\cdot x = bc$ y $x^2 = ab$, en forma de proporción, tiene lugar en el Libro VI de *Los Elementos* de Euclides, es decir, se retrasa hasta después de desarrollar la *Teoría de la Proporción* de Eudoxo en el libro V.

La parte más importante del *Álgebra Geométrica* de los griegos se encuentra en el Libro II de *Los Elementos* de Euclides. En la actualidad su contenido no juega ningún papel fundamental en los libros de texto modernos. Sin embargo en la Geometría griega ejerce una función primordial. La discrepancia radical entre los puntos de vista griego y moderno estriba en que hoy nosotros podemos disponer de un Álgebra simbólica y una Trigonometría, que han sustituido completamente a sus equivalentes geométricos clásicos, precisamente gracias a *La Geometría* de Descartes, que al aplicar la naturaleza algorítmica del Álgebra a los problemas geométricos alumbró su Geometría Analítica.

Mientras nosotros representamos las magnitudes con letras que se sobreentiende son números conocidos o desconocidos, con las cuales operamos mediante las reglas algorítmicas del Álgebra, los griegos representaban las magnitudes rectilíneas mediante segmentos de línea recta que debían obedecer a los axiomas y teoremas de la Geometría. Con estos elementos los griegos disponían de un Álgebra –Geométrica– que cumplía a todos los efectos las mismas funciones que nuestra moderna Álgebra simbólica. Cierto es que el Álgebra moderna con su cálculo literal facilita de forma considerable la manipulación de las operaciones y las relaciones entre magnitudes geométricas, pero no es menos cierto que con su Álgebra Geométrica los griegos eran mucho más hábiles que nosotros en la práctica geométrica. Y es que el *Álgebra Geométrica* griega sorprende al estudioso moderno por ser bastante difícil y artificiosa, pero los griegos la utilizaron con soltura y para ellos debió ser una herramienta de utilización necesaria, básica y cómoda.

Así por ejemplo la Proposición II.5 de *Los Elementos* de Euclides:



***Euclides* II.5**
$(a+b)(a-b)+b^2=a^2$

«*Si se divide una recta en partes iguales y desiguales, el rectángulo comprendido por las partes desiguales de la recta entera, más el cuadrado de la diferencia entre las dos partes, es equivalente al cuadrado de la mitad de la recta dada*»

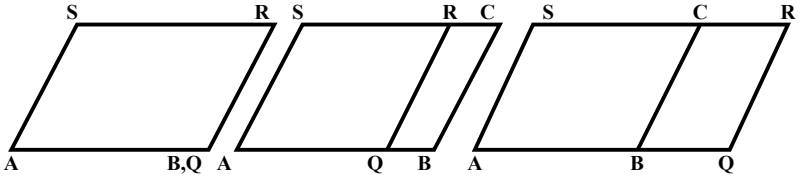
equivale –a pesar del circunloquio retórico– a la identidad algebraica:

$(a + b)\cdot(a–b) + b^2 = a^2$ ó $(a + b)\cdot(a – b) = a^2 – b^2$,

y no es más que la formulación geométrica de una de las leyes fundamentales de la Aritmética –suma por diferencia igual a diferencia de cuadrados–.

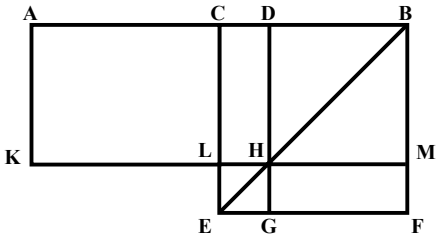
La evidencia visual del teorema aludido para un estudioso griego es muy superior a su contrapartida algebraica actual. Claro está que la demostración rigurosa de Euclides de esta proposición puede ocupar más de una página.

Para explicar de forma más efectiva el método de la aplicación de las áreas, consideremos un segmento de línea AB y un paralelogramo AQRS cuyo lado AQ está a lo largo de AB:



1. Cuando Q coincide con B, se dice que «*el paralelogramo AQRS se ha aplicado sobre el segmento AB*».
2. Cuando Q está entre A y B, se dice que «*el paralelogramo AQRS se ha aplicado sobre el segmento AB de forma elíptica o con defecto el paralelogramo QBCR*».
3. Cuando Q está en la prolongación de AB, se dice que «*el paralelogramo AQRS se ha aplicado sobre el segmento AB de forma hiperbólica o con exceso el paralelogramo QBCR*».

Volviendo a la Proposición II.5 de *Los Elementos* de Euclides, para su demostración consideremos la figura siguiente:



Sea AB el segmento de línea recta dado, dividido de forma igual por C y de forma desigual por D, la proposición establece que: $AD\cdot DB + CD^2 = CB^2$.

Tomando AB=2a, AC=a, CD=b, resulta la identidad algebraica:

$(a + b)\cdot(a–b) + b^2 = a^2$.

Simplificando el lenguaje retórico de Euclides tenemos:

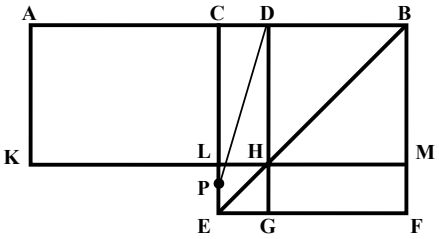
$AD\cdot DB + CD^2 = AKHD + LEGH = AKLC + CLHD + LEGH = CLMB + CLHD + LEGH = CLMB + HGFM + LEGH = CB^2$.

Digamos que más importante que la demostración exhibida, es el diagrama que utiliza Euclides en esta demostración porque es un esquema gráfico que jugarían un papel fundamental en la resolución geométrica de ecuaciones cuadráticas.

En efecto: sea resolver en la Geometría griega la ecuación $ax-x^2=b^2$, es decir, encontrar un segmento de línea x que cumpla la condición expresada por la ecuación $ax-x^2=b^2$, donde a,b son segmentos tales que a>2b.

Sea ahora AB=a, y sea C el punto medio de AB, levantemos por C una perpendicular CP de longitud igual a b. Con centro en P y radio a/2 tracemos una circunferencia que corte a AB en el punto D.

Construyamos sobre AB un rectángulo ABMK de anchura BM=BD y completemos el cuadrado BDHM. Este cuadrado es el área x^2 que cumple la condición expresada por la ecuación cuadrática. En lenguaje griego de la *Aplicación de las Áreas* se ha aplicado de *forma elíptica* al segmento AB=a un rectángulo AH de área $(a-x)\cdot x$, es decir $ax-x^2$, que es igual a un cuadrado dado b^2, y que es deficiente del rectángulo AM en un cuadrado DM. La demostración de este hecho viene dada por la Proposición Euclides II.5, según la cual el rectángulo ADHK es igual al polígono cóncavo CBFGHL, es decir, difiere de $(a/2)^2$ en el cuadrado LHGE cuyo lado es por construcción CD= $\left[\sqrt{(a/2)^2-b^2}\right]$.



Sintetizando los cálculos geométricos:

AB=a, AC=CB

CP=b, PD=a/2, LH=CD= $\left[\sqrt{(a/2)^2-b^2}\right]$.

Rectángulo ABMK (BM=BD)

Cuadrado BDHM = x^2 .

Euclides II.5: $AD\cdot DB+CD^2=CB^2$.

ADHK + LHGE = CBFE ,

ADHK = ABMK – BDHM ,

ADHK = CBFE – LHGE ,

ABMK–BDHM=CBFE–LHGE

$ax - x^2 = (a/2)^2 - \left[\sqrt{(a/2)^2-b^2}\right]^2 = b^2$.

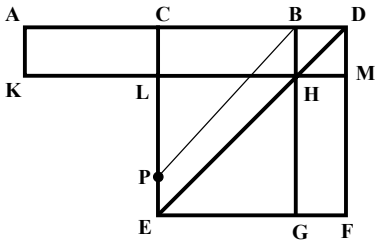
De manera similar se resuelve la ecuación cuadrática $ax+x^2=b^2$ mediante la Proposición II.6 de *Los Elementos* de Euclides:

> «*Si se divide una recta en dos partes iguales y se prolonga, el rectángulo comprendido por la recta entera, más la prolongación, y por la prolongación, junto con el cuadrado de la recta mitad, es equivalente al cuadrado de la recta formada por la recta mitad y la prolongación.*»

En este caso se trata de aplicar *de forma hiperbólica* a una línea recta dada AB=a, un rectángulo $AM=ax+x^2$, que sea igual a un cuadrado dado b^2 y que exceda al rectángulo AH en un cuadrado x^2 .

Sea C el punto medio de AB, levantemos por C una perpendicular CP de longitud igual a b. Con centro en C y radio PB tracemos una circunferencia que corte a AB en el punto D.

Esta vez la distancia es CD=PB= $\sqrt{(a/2)^2+b^2}$, y como por la proposición se sabe que el rectángulo $AM=ax+x^2$ más el cuadrado $LG=(a/2)^2$ es igual al cuadrado $CF=(a/2)+b^2$, se verifica la condición de la ecuación $ax+x^2=b^2$.



Sintetizando los cálculos geométricos:

AB=a, AC=CB, CP=b, CD=PB= $\left[\sqrt{(a/2)^2+b^2}\right]$.

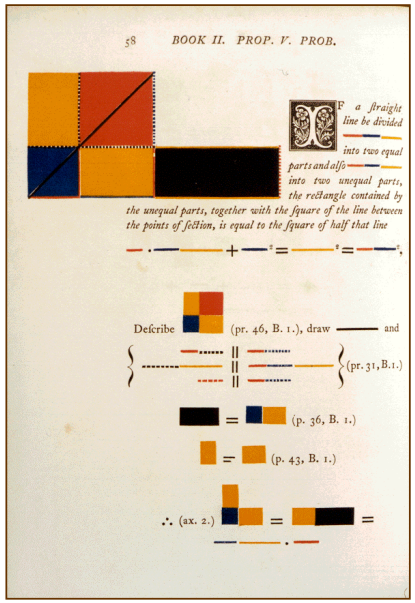
Rectángulo ADMK (BM=BD), Cuadrado BDHM = x^2 .

Euclides II.6: ADMK + LHGE = CDFE .

ADMK = CDFE – LHGE, ADMK= ABHK + BDMH , ABHK + BDMH = CDFE – LHGE

$ax + x^2 = \left[\sqrt{(a/2)^2+b^2}\right]^2 - (a/2)^2 = b^2$.

EL *ÁLGEBRA GEOMÉTRICA* Y LA *APLICACIÓN DE LAS ÁREAS* EN EL LIBRO II DE *LOS ELEMENTOS* DE EUCLIDES


58 BOOK II. PROP. V. PROB.

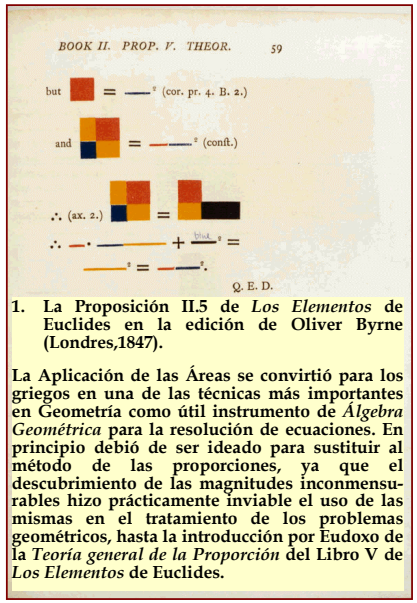
IF a straight line be divided into two equal parts and also into two unequal parts, the rectangle contained by the unequal parts, together with the square of the line between the points of section, is equal to the square of half that line

Describe (pr. 46, B. 1.), draw and (pr. 31, B. 1.)

(p. 36, B. 1.)

(p. 43, B. 1.)

∴ (ax. 2.)


BOOK II. PROP. V. THEOR. 59

but (cor. pr. 4. B. 2.)

and (conft.)

∴ (ax. 2.)

Q. E. D.

1. La Proposición II.5 de *Los Elementos* de Euclides en la edición de Oliver Byrne (Londres,1847).

La Aplicación de las Áreas se convirtió para los griegos en una de las técnicas más importantes en Geometría como útil instrumento de *Álgebra Geométrica* para la resolución de ecuaciones. En principio debió de ser ideado para sustituir al método de las proporciones, ya que el descubrimiento de las magnitudes inconmensurables hizo prácticamente inviable el uso de las mismas en el tratamiento de los problemas geométricos, hasta la introducción por Eudoxo de la *Teoría general de la Proporción* del Libro V de *Los Elementos* de Euclides.

Las bases firmes de *Teoría de la Proporción* permiten a Euclides en las Proposiciones 27, 28 y 29 del Libro VI dar una generalización del método de *Aplicación de las Áreas*, donde el libre uso del concepto de semejanza facilita la sustitución de los rectángulos del Libro II por paralelogramos, permitiendo aplicar a un segmento dado un paralelogramo igual a una figura rectilínea dada y que exceda o sea deficiente en un paralelogramo semejante a otro dado. Las construcciones correspondientes como las de las Proposiciones II.5, II.6 son en la práctica soluciones geométricas de las ecuaciones cuadráticas $ax\pm x^2=bx$, sometidas a la restricción geométrica equivalente a que el discriminante sea no negativo, es decir, las aludidas Proposiciones VI.27, VI.28 y VI.29 son una especie de contrapartida geométrica de la forma algebraica más generalizada de ecuaciones cuadráticas con raíz real y positiva.

Además, desde el punto de vista histórico la *Aplicación de las Áreas* está en el punto de partida de la teoría de Apolonio (hacia 200 a.C.) de las secciones cónicas. De hecho los tres nombres acuñados por Apolonio para las cónicas no degeneradas provienen de la denominación de los tres tipos de aplicación de las áreas: *elíptico* (dado un segmento construir sobre una parte de él o sobre él mismo extendido, un paralelogramo igual en área a una figura rectilínea dada y resultando deficiente en un paralelogramo semejante a uno dado), *hiperbólico* (idem. resultando excedente) y *parabólico* (idem. resultando igual).

Coordenadas en *Las Cónicas* de Apolonio

Las Cónicas de Apolonio es una de las obras más importantes de toda la Geometría griega. En ella quedan acuñadas, con significado, para la posteridad, los nombres de *elipse*, *parábola* e *hipérbola*, procedentes del lenguaje pitagórico de la *Aplicación de las Areas*. En el cambio de denominación de las cónicas por Apolonio subyace un cambio conceptual, toda vez que una vez construidas a través del cono, Apolonio maneja las cónicas mediante relaciones de áreas y longitudes, que expresan en cada caso la propiedad característica de definición de la curva de la que se obtienen sus propiedades intrínsecas. Apolonio fue capaz de vincular los aspectos estereométricos y planos de las cónicas, al mostrar que las secciones de los conos tenían importantes propiedades como lugares planos, traducibles en básicas expresiones geométricas –equivalentes a nuestras *ecuaciones*–, que permitían deducir, a su vez, otras innumerables propiedades de las cónicas. Es bajo esta visión sobre el trabajo de Apolonio que algunos historiadores modernos (Zeuthen, Coolidge, Loria y Heath) reclaman para los griegos, y empezando por Apolonio, la paternidad de la Geometría Analítica, al establecer como la esencia de esta rama de la Matemática el estudio de los lugares por medio de *ecuaciones*.

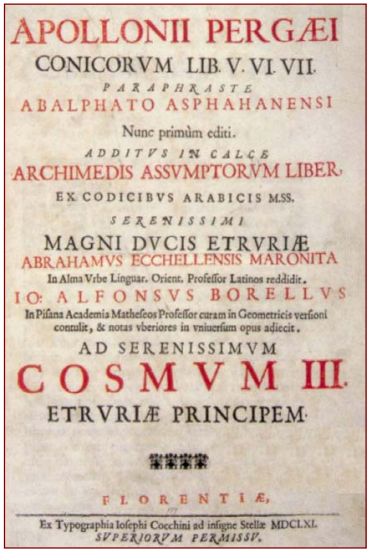
En el estudio de las cónicas, Apolonio considera ciertas *líneas de referencia* –diámetros conjugados o diámetro-tangente–, que juegan un papel de *coordenadas*. En el segundo caso al tomar un diámetro y una tangente en uno de sus extremos como rectas de referencia, las distancias medidas a lo largo del diámetro a partir del punto de tangencia son las *abscisas* y los segmentos paralelos a la tangente, interceptada por el diámetro y la curva, son las *ordenadas*. Para cada cónica, la conocida relación de áreas y longitudes en forma de proporción en el lenguaje del Álgebra Geométrica –propiedad geométrica de la curva equivalente a su definición como lugar geométrico– se traduce en una relación entre las *abscisas* y las correspondientes *ordenadas*, que Apolonio llamaba el *symptoma* de la curva y que no es sino la expresión retórica de la ecuación analítica de la curva, que en su evolución histórica daría lugar a la llamada *ecuación característica*. El lenguaje de Apolonio es sintético, utilizando con una pericia increíble la técnica pitagórica de la *Aplicación de las Áreas*, pero sus «*métodos de coordenadas*» guardan una gran similitud con los de la Geometría Analítica.

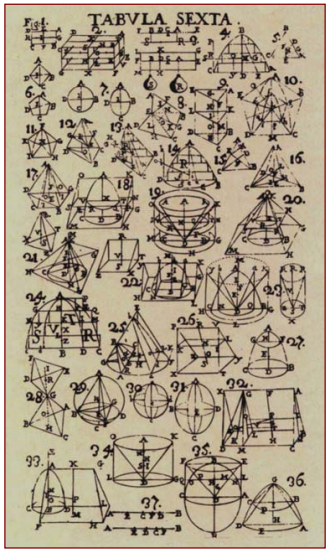
Debemos aquilatar, no obstante, ciertas afirmaciones sobre elementos precursores de la Geometría Analítica, porque al señalar tales atribuciones, más o menos fundadas o infundadas, siempre nos encontraremos con las serias limitaciones impuestas por el carácter geométrico-sintético de la Geometría griega y por la ausencia de un Álgebra simbólica en sentido algorítmico, que es un componente ineludible de una verdadera Geometría Analítica general, y que a fin de cuentas es lo que permite la real y mutua correspondencia entre curvas y ecuaciones.

Esto fue realmente lo que se plantearon y resolvieron Fermat y Descartes con el concurso del *Arte Analítica* de Vieta, al establecer que una ecuación arbitraria en dos cantidades indeterminadas determina, con respecto a un sistema dado de coordenadas, una curva.

Al analizar la posición histórica de Apolonio en el camino hacia la Geometría Analítica digamos que, a pesar de los conceptos y elementos geométricos introducidos, que parecen emular la presencia de *sistemas de referencia con coordenadas –abscisas* y *ordenadas*– que permiten expresar las *ecuaciones* de las cónicas, estos *sistemas de coordenadas* aparecían siempre superpuestos *a posteriori* a las curvas para estudiar sus propiedades. En la Geometría griega, las *coordenadas*, *variables* y *ecuaciones* no eran elementos de partida, sino conceptos subsidiarios derivados de situaciones geométricas concretas de curvas que determinan las *ecuaciones* sin que se dé la situación inversa, es decir, que las ecuaciones determinen las curvas, ya que éstas siempre se producían mediante una construcción estereométrica como secciones de un sólido –tal es el caso de las propias cónicas de Apolonio– o de forma cinemática como composición de movimientos –tal es el caso de la Espiral de Arquímedes o la cuadratriz de Dinostrato–, de forma que el conjunto de curvas manejadas por los griegos fue necesariamente muy limitado.

LAS CÓNICAS DE APOLONIO Y LA GEOMETRÍA ANALÍTICA


APOLLONII PERGÆI
CONICORVM LIB. V. VI. VII.
PARAPHRASTE
ABALPHATO ASPHAHANENSI
Nunc primùm editi.
ADDITVS IN CALCE
ARCHIMEDIS ASSVMPTORVM LIBER.
EX CODICIBVS ARABICIS MSS.
SERENISSIMI
MAGNI DVCIS ETRVRIÆ
ABRAHAMVS ECCHELLENSIS MARONITA
In Alma Vrbe Linguar. Orient. Professor Latinos reddidit.
IO: ALFONSVS BORELLVS
In Pisana Academia Matheseos Professor curam in Geometricis versioni contulit, & notas vberiores in vniuersum opus adiecit.
AD SERENISSIMVM
COSMVM III.
ETRVRIÆ PRINCIPEM.
FLORENTIÆ,
Ex Typographia Iosephi Cocchini ad insigne Stellæ MDCLXI.
SUPERIORUM PERMISSU.



Portada y página con ilustraciones de figuras geométricas de *Apollonii Pergaei Conicorum* Lib. V, VI, VII. Edición de Borelli. Florencia 1661. Biblioteca de la Universidad de Pavía.

***Las Cónicas* de Apolonio contienen muchos aspectos que anticipan elementos de la Geometría Analíticas. Como Descartes, Apolonio considera, ciertas *líneas de referencia* –diámetros conjugados o diámetro-tangente– que al jugar un papel de *coordenadas*, son asociados a la cónica dada, de modo que mediante Álgebra retórica son expresadas en función de esas líneas las propiedades geométricas de la curva equivalentes a su definición como lugares geométricos.**

C.Boyer escribe sobre Apolonio y la Geometría Analítica (en *Historia de las Matemáticas*, Alianza Universidad, Madrid, 1986, cap.17, p.208):

> **«*El hecho de que Apolonio, uno de los más grandes geómetras de la antigüedad, no consiguiese desarrollar de una manera efectiva la Geometría Analítica, se debe probablemente más a una pobreza en el número de curvas que de pensamiento; los métodos generales no son ni muy necesarios ni muy útiles cuando los problemas se refieren siempre a un número limitado de casos particulares. Por otra parte, es bien cierto que los primeros inventores de la Geometría Analítica tenían a su disposición todo el álgebra renacentista* [el Álgebra de los cosistas italianos y el Álgebra simbólica de Vieta], *mientras que Apolonio tuvo que trabajar con las herramientas del Álgebra Geométrica, mucho más rigurosa pero a la vez mucho más incómoda de manejar*».**

Hay que ponderar la magnífica obra de Apolonio, primer estadio en la Historia de la Matemática sobre la aplicación de coordenadas al estudio de las propiedades de las curvas; y aunque el discurso retórico sustituye al simbolismo y la construcción geométrica a las técnicas algebraicas, las relaciones de áreas y longitudes mediante las que Apolonio expresa las propiedades intrínsecas de la curva se traducen con gran facilidad al ulterior lenguaje del Álgebra simbólica de ecuaciones que permitirá la asociación de curvas y ecuaciones, esencia de la Geometría Analítica. Así pues, el trabajo de Apolonio inicia la singladura histórica hacia el desarrollo de la Geometría Analítica de Descartes.

Además, dos problemas históricos importantes de gran incidencia sobre la Geometría Analítica de Descartes tienen su origen en los trabajos de Apolonio:

1. ***El Problema de Apolonio* («*Dados tres elementos, punto, recta o circunferencia, trácese una circunferencia que sea tangente a cada uno de los tres*»)**
2. ***El Problema de Pappus* o «*lugar geométrico determinado por tres o cuatro rectas*»: «*Dadas tres (resp. cuatro) rectas en un plano, encuéntrese el lugar geométrico de un punto que se mueve de forma que el cuadrado de la distancia a una de las tres rectas es proporcional al producto de las distancias a las otras dos (resp. el producto de las distancias a dos de ellas es proporcional al producto de las distancias a las otras dos), si las distancias se miden en direcciones tales que formen ángulos dados con las líneas correspondientes*».**

El Análisis Geométrico griego y la Geometría Analítica

Los Elementos de Euclides establecieron en la Geometría griega un severo modelo de exposición y demostración que oculta el camino de la investigación hacia el descubrimiento. Surge de forma natural la pregunta acerca de cómo los geómetras griegos encontraban sus impresionantes resultados que después plasmaban en sus obras con un rigor impecable. Pues bien, es aquí donde interviene el *Análisis* como un procedimiento metodológico capital para el progreso de la Matemática, del que la Geometría Analítica heredará no sólo su nombre sino sobre todo sus procedimientos.

Proclo (411-485 d.C.) en sus *Comentarios al Libro I de los Elementos de Euclides* atribuye a Hipócrates de Quíos (hacia 450 a.C.) la invención del *Método Analítico* cuando lo define:

> «*La apagogé es una reducción de un problema o de un teorema a otro, que si es conocido o determinado, conduce a la solución de la cuestión propuesta*».

Pero siempre se ha imputado su paternidad a Platón –según ciertos pasajes del *Menón* (86e–87a), la *República* (510c) y la *Ética a Nicómaco* de Aristóteles (1095a)–, que lo formularía como un método pedagógicamente conveniente, viniendo a decir que cuando una cadena de razonamientos desde unas premisas a una conclusión no es obvia, se puede invertir el proceso; uno puede empezar por la proposición que ha de probarse y deducir de ella una conclusión que es conocida. Si entonces podemos invertir los pasos en esta cadena de razonamientos, el resultado (*Síntesis*) es una prueba legítima de la proposición. Es decir, mediante el *Análisis* se asume como cierto aquello que hay que probar y se razona con base en esta asunción hasta llegar a algo que forma parte de los principios o alcanzar un resultado cierto por haber sido previamente establecido. Si entonces podemos invertir la secuencia de los pasos anteriores se obtiene una demostración del teorema que había que probar. Así pues, el *Análisis* viene a ser un procedimiento sistemático de descubrir «*condiciones necesarias*» para que un teorema sea cierto, de modo que si por medio de la *Síntesis* se muestra que estas condiciones son también *suficientes*, se obtiene una demostración correcta de la proposición.

Conviene explicar un poco en qué medida la Geometría Analítica recibe su nombre precisamente del método de *Análisis* de los griegos. Como se ha dicho, el *Análisis* empieza «*asumiendo como cierto aquello que hay que probar*». Esto es precisamente un principio que aplica Descartes desde el comienzo de *La Geometría*. Por ejemplo en el segundo epígrafe del Libro I, titulado*: «Cómo se llega a las ecuaciones que sirven para resolver los problemas*», Descartes escribe (G.AT,VI,372):

> «*Así, si se quiere resolver algún problema, debe de antemano considerarse como ya resuelto,*[...]»

Descartes no sólo realizará una aplicación directa de los procedimientos del *Análisis* y la *Síntesi*s de los griegos sino que reformulados serán las dos reglas intermedias de las cuatro reglas del *El Discurso del Método* (DM.AT,VI,17-18). Una y otra vez en la multiplicidad de problemas que resuelve en *La Geometría*, Descartes empezará por suponer el problema resuelto. En concreto en dos de los problemas más importantes que trata, Descartes escribe literalmente:

> «*Primeramente yo supongo la cosa como ya hecha, ...*» (*Problema de Pappus* [G.AT,VI, 382]).
>
> «*Supongamos que la cosa está hecha*, ...» (rectas normales a una curva [G.AT,VI, 413]).

Naturalmente hay una diferencia notable entre la aplicación que del método de *Análisis* y *Síntesis* hacen los griegos y lo que realiza Descartes en lo que se ha llamado su Geometría Analítica. Éste es el asunto que queremos estudiar: a partir de algunos de los principios

metodológicos de la Geometría griega tiene lugar el nacimiento de algo completamente nuevo y revolucionario –*La Geometría* de Descartes– que consigue clausurar, en gran parte, el punto de partida – la propia Geometría griega–.

¿Qué poderoso instrumento utilizará y Descartes para alcanzar tal hazaña matemática? El Álgebra, una herramienta que no pudo disfrutar la Geometría griega porque la aparición súbita de los inconmensurables desvió la influencia de la Matemática babilónica, bien versada en Aritmética y en incipientes técnicas algebraicas, hacia la Geometría Sintética y el Álgebra Geométrica. Cuando Descartes, bajo la inspiración de Vieta (1540-1603), aplique todo el potencial algorítmico del Álgebra árabe, renacentista y del propio Vieta, el *Análisis* alcanzará su máximo poder heurístico para la resolución de los problemas geométricos –incluso los que se habían resistido de forma reiterada a los métodos clásicos, como el *Problema de Pappus* y el *Problema de Apolonio*–, a base de complementar el estudio analítico con la síntesis algebraica, lo que le permitirá mediante las ecuaciones pasar de la Geometría al Álgebra y del Álgebra a la Geometría.

La forma más esmerada del *Análisis* y la *Síntesi*s la aplica Pappus en el *Tesoro del Análisis*, describiendo como para comprobar la validez y encontrar la prueba de un teorema o resolver un problema –en general de construcción– se procede analíticamente, asumiendo por el momento que el teorema en cuestión es válido o que el problema está resuelto. Siguiendo entonces las implicaciones lógicas del teorema o la solución del problema, se llega a alcanzar una solución conocida que es verdadera o falsa. Si se trata de un teorema, de una falsa conclusión resulta la invalidez del teorema, y entonces del mismo *Análisis* resulta la refutación del teorema por reducción al absurdo; pero, si la conclusión obtenida a través del Análisis es verdadera, nada se puede decir de la validez del teorema. Es decir, el método de Análisis produce una cadena de inferencias que lleva de una premisa de valor verdadero desconocido a una conclusión de valor verdadero conocido; la falsedad de la conclusión implica la de la premisa, pero la verdad de la conclusión no dice nada acerca de la de la premisa, a menos que, como señalaba Platón, uno pueda dar la vuelta a la inferencia. La eficiencia del *Análisis* es doble, por una parte abundan los teoremas geométricos que tienen un recíproco válido, y por otra, cuando el recíproco de un teorema no es válido puede llegar a serlo añadiendo ciertas condiciones suplementarias, que eran llamadas por los griegos «*diorismos*». Gran parte de la investigación geométrica consistía en la búsqueda del *diorismo* adecuado para poder invertir una inferencia. Una vez que se ha hallado el *diorismo*, la inferencia invertida constituye una *Síntesis*, es decir la rigurosa demostración del teorema. Las considerables dificultades inherentes a la inversión de inferencias propiciaron que los grandes matemáticos griegos se expresaran en sus obras mediante formales demostraciones sintéticas de los resultados que habían obtenido aplicando el método de *Análisis*. Es decir, el Análisis geométrico griego era una fecunda heurística geométrica, el instrumento fundamental de investigación y creación matemática; pero, alcanzada tras el *Análisis*, la *Síntesis*, en presencia de la demostración sintética cualquier análisis era superfluo y como tal se suprimía de los grandes tratados. De esta forma, los griegos ocultaban la forma y el camino utilizados en la obtención de sus magníficos resultados matemáticos.

Cuando a partir del Renacimiento tiene lugar la recuperación, reconstrucción y divulgación del legado clásico griego, los matemáticos lo acogen con entusiasmo, pero preocupados porque el estilo sintético y apodíctico de exposición de la Geometría griega, y en particular de las obras de Euclides, Arquímedes y Apolonio, privaba a los investigadores de la forma en que habían sido descubiertos los resultados, manifiestan junto a su admiración, una cierta perplejidad y extrañeza. Incluso algunos (Torricelli, Barrow, Wallis,...) sospechaban sin fundamento que los griegos disponían de algún instrumento (¿el Álgebra?), un determinado tipo de Análisis Geométrico, pero que lo habían ocultado de forma tan perfecta que a los modernos matemáticos les había resultado más fácil inventar un nuevo Análisis –la Geometría Analítica– que recuperar el antiguo. Quizá es Descartes quien con mayor claridad muestra – en la Regla IV de las *Regulae*– la insatisfacción de una curiosidad frustrada por la ocultación de los métodos de descubrimiento de la Geometría griega.

ANÁLISIS Y ÁLGEBRA EN LA REGLA IV (AT.X.373-377) DE LAS *REGLAS PARA LA DIRECCIÓN DEL ESPÍRITU* DE DESCARTES

Retrato caricaturesco de Descartes escribiendo un libro y con el pie apoyado en una obra de Aristóteles. Grabado de C.Hellemans. Biblioteca Nacional. París.

Descartes subraya en la regla IV (*Regulae ad directionem ingenii***) que los antiguos geómetras utilizaban cierto *Análisis* para la resolución de todos los problemas geométricos –como se advierte en Pappus y Diofanto-, pero privaron de él a la posteridad con la expresión sintética que oculta los métodos de descubrimiento, y merecen por ello –al impedir la divulgación de los métodos de trabajo- la más acerba de las críticas.**

Descartes elogia, en cambio, a «*hombres de gran talento*» (¿Vieta?), que han recuperado el Análisis Geométrico de los antiguos y lo han desarrollado con los nuevos instrumentos del Álgebra –un arte que clarificado y liberado de su actual farragosidad podría cumplir una función similar a la del *Análisis* de los antiguos–. Con base en estos analistas Descartes destilará un auténtico Análisis Algebraico, que históricamente se desarrollará en la línea de una verdadera Geometría Analítica.

«**[...]** ***En las más fáciles de las ciencias, la Aritmética y la Geometría, vemos con toda claridad que los antiguos geómetras se han servido de cierto Análisis, que extendían a la resolución de todos los problemas, si bien privaron de él a la posteridad. Y ahora florece cierta clase de Aritmética que llaman Álgebra, para realizar sobre los números lo que los antiguos hacían sobre las figuras*** **[...]** ***Cuando por primera vez me dediqué a las disciplinas Matemáticas, de inmediato leí por completo la mayor parte de lo que suelen enseñar sus autores, y cultivé preferentemente la Aritmética y la Geometría, porque se las tenía por las más simples y como un camino para las demás. Pero no caían en mis manos autores que me satisficieran plenamente: leía cosas acerca de los números que yo comprobaba, habiendo hecho cálculos, ser verdaderas; y lo mismo respecto de las figuras;*** **[...]** ***Pero por qué esto era así, y cómo eran halladas, no parecían mostrarlo suficientemente a la mente,*** **[...]** ***Pero como después pensase por qué sucedía que antiguamente los primeros creadores de la Filosofía no quisieran admitir para el estudio de la sabiduría a nadie que no supiese Mathesis,*** **[...]**, ***tuve la sospecha de que ellos conocían cierta Mathesis muy diferente de la Matemática vulgar de nuestro tiempo*** **[...]** ***Y ciertamente me parece que vestigios de esta verdadera Mathesis aparecen en Pappus y Diofanto,*** **[...]** ***Y fácilmente creería que después fue ocultada por cierta audacia perniciosa por los mismos escritores; pues así como es cierto que lo han hecho muchos artistas con sus inventos, así ellos temieron quizá que, siendo tan fácil y sencilla, se envileciese después de divulgada; y para que les admirásemos prefirieron presentarnos en su lugar, como productos de su método, algunas verdades estériles deducidas con sutileza, en vez de enseñarnos el método mismo que hubiera hecho desaparecer por completo la admiración. Ha habido, finalmente, algunos hombres de gran talento que se han esforzado en este siglo por resucitarla; pues aquel arte no parece ser otra cosa, que lo que con nombre extranjero llaman Álgebra, con tal que pueda zafarse de las múltiples cifras e inexplicables figuras de que está recargado a fin de que no falte ya aquella claridad y facilidad suma que suponemos debe haber en la verdadera Mathesis*** **[...]**».

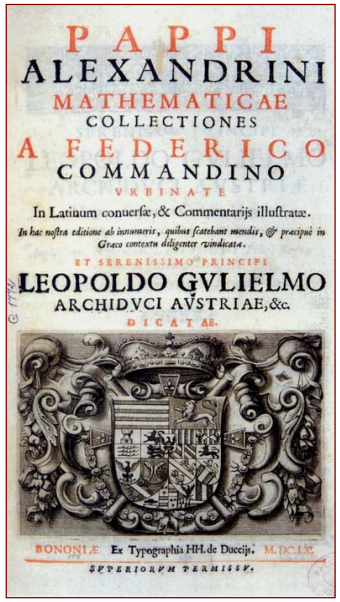
El texto de la IV Regla de Descartes es fundamental para poder entender la actitud mental de Descartes sobre su magno proyecto de reforma de la Filosofía, la Ciencia, y sobre todo de la Matemática de donde surgen las fuentes de su Geometría Analítica.

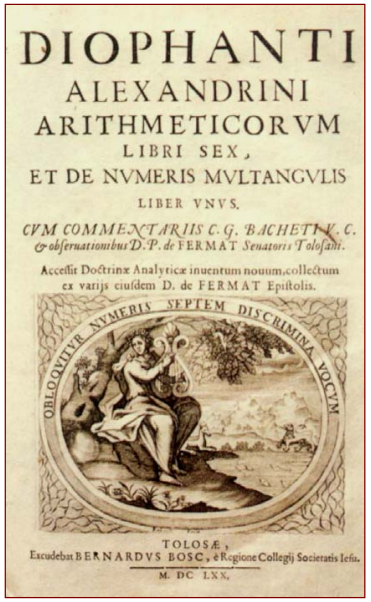
Descartes habla de la *Mathesis* como si se tratara de un saber aún más universal que la propia Matemática, y aplicable a todas las ciencias. En puridad, la *Mathesis* no se identifica, por tanto, con la Matemática, pero surge del espíritu, de la naturaleza, de los rasgos, del estilo, del modo, del proceder, de los métodos, etc., de las ciencias matemáticas –de la Geometría (Pappus), de la Aritmética (Diofanto) y del Álgebra (Vieta, uno «*de los hombres de gran talento*» que han resucitado la *Mathesis*)–. Precisamente uno de los instrumentos más potentes que se ha desarrollado en toda la Historia del Pensamiento matemático –la Geometría Analítica Cartesiana–, sin duda íntimamente vinculada a la *Mathesis,* surge de la aplicación del Álgebra simbólica de Vieta al estudio de los problemas del Análisis Geométrico de Pappus mediante ecuaciones indeterminadas, cuyo origen remoto, así como las raíces de la esencial simplificación de la notación cartesiana están en *La Aritmética* de Diofanto. Por eso Descartes rinde claro homenaje a los matemáticos griegos, Pappus y Diofanto, y de forma implícita también a Vieta, al atribuirles vestigios de la *Mathesis*.

Como señala Descartes, en la pléyade de geómetras griegos, Pappus fue una excepción, porque desarrolló una singular metodología en la forma de exposición, codificando todo un cuerpo de tratados analíticos de solución de problemas en el llamado *Tesoro del Análisis* del Libro VII de *La Colección Matemática*. En estos tratados queda patente el camino que sigue la investigación matemática ya que se procede a la reducción de un problema dado a un problema equivalente cuya solución era ya conocida. Encontramos en la obra de Pappus, además de infinidad de teoremas y problemas sobre Geometría superior –no incluida en *Los Elementos* de Euclides–, un gran número de cuestiones que debemos situar en las raíces históricas de la Geometría Analítica como son la más elaborada exposición sobre los métodos de *Análisis* y *Síntesis*, numerosas soluciones a los problemas clásicos –sobre todo la duplicación del cubo y la trisección del ángulo–, nuevos estudios y extensiones de propiedades de las secciones cónicas como lugares geométricos y la clasificación definitiva de los problemas geométricos en planos, sólidos y lineales –según sean resolubles, respectivamente, con rectas y circunferencias, cónicas u otras curvas superiores–, que perseguía la idea de ajustar la envergadura de los instrumentos geométricos a utilizar a la enjundia de los problemas geométricos a resolver, en la línea de aplicar siempre los medios más simples posibles, lo que será no sólo un rasgo distintivo de la Geometría Analítica de Descartes, sino un componente general de la mejor Matemática, que siempre exige elegancia y economía en el razonamiento.

Pero quizá el asunto más importante sea el tratamiento general del llamado *Problema de Pappus* o *lugar geométrico de n rectas*, que en su formulación más sencilla, para tres o cuatro rectas ya era conocido por Apolonio, siendo la solución una cónica, y que ha tenido un valor emblemático para la Historia de la Geometría Analítica. Pappus realiza un estudio exhaustivo del problema, propone la generalización a más de cuatro rectas y reconoce que independientemente del número de rectas involucradas en el problema, queda determinada una curva concreta. He aquí la observación más general sobre lugares geométricos de toda la Geometría griega, lo que implica, además, la consideración de infinitos tipos nuevos de curvas planas, algo esencial en un mundo geométrico tan limitado en cuanto a curvas planas. Pappus vacila a la hora de considerar el problema para más de seis líneas porque: «*no hay nada contenido en más de tres dimensiones*». De haber seguido en esa dirección, se habría dado un paso muy importante de anticipación de la Geometría Analítica, toda vez que ello hubiera propiciado un necesario tratamiento algebraico y no geométrico de los productos de líneas involucradas en el problema. Naturalmente los métodos sintéticos le desbordan a Pappus en el abordaje del problema. Cuando el nuevo Álgebra Simbólica de Vieta actúe sobre el Análisis Geométrico de los griegos aparecerá la Geometría Analítica cartesiana como poderoso instrumento algorítmico de ataque de los problemas geométricos difíciles como el propio *Problema de Pappus*, que fue la prueba de fuego que tuvo que pasar *La Geometría* de Descartes par demostrar la potencia de los nuevos métodos de la Geometría Analítica.

PAPPUS Y DIOFANTO
Y LA GEOMETRÍA ANALÍTICA





1. ***La Colección Matemática*** **de Pappus. Edición de F.Commandino. (Bolonia, 1670).**

2. ***La Aritmética de*** **Diofanto. Edición de 1670 de S. de Fermat con las observaciones de su padre P. de Fermat.**

La Colección Matemática **de Pappus tiene un gran valor histórico y didáctico. Pappus realiza una encomiable labor de compilación, comentario, restauración, organización, clasificación y generalización del conocimiento matemático superior de la antigüedad. La obra describe una multitud de trabajos matemáticos perdidos que constituyen lo que se llama** ***Tesoro del Análisis*****. Además, Pappus nos relata las vías que seguía la investigación geométrica, oculta en los grandes tratados clásicos debido a su estilo sintético, es decir, lo que los antiguos geómetras entendían por** ***Análisis*** **y** ***Síntesis*****.**

La obra de Pappus contiene soluciones nuevas a numerosos problemas clásicos, la clasificación definitiva de los problemas geométricos en planos, sólidos y lineales, estudios definitivos de las cónicas como lugares geométricos y una visión más general del famoso ***Problema de Pappus*****, todas ellas cuestiones de trascendental influencia sobre la evolución del Álgebra Geométrica y el Análisis Geométrico griegos hacia la Geometría Analítica de Descartes.**

Diofanto es el responsable, con su obra ***La Aritmética*****, de los primeros escarceos del Álgebra simbólica –el Álgebra sincopada–. A base de adoptar ciertas letras o expresiones como abreviaturas para las cantidades indeterminadas y sus potencias y para las operaciones más habituales, fragua un incipiente simbolismo antecedente de la notación algebraica que en su evolución a lo largo de los siglos, culminará con la simplificación notacional poderosamente definitiva que acuñará Descartes en La Geometría y que se convertirá en el alfabeto de la Matemática. Al ser el Álgebra simbólica un instrumento algorítmico ineludible de la Geometría Analítica, y Diofanto el primer iniciador de esta utilidad, debemos situar su obra, en una dirección conveniente hacia la generación de la Geometría Analítica.**

La Geometría de Descartes
La formación de Descartes en La Flèche

Descartes instaura una nueva época en la Matemática, la Ciencia y la Filosofía sin parangón en la Historia de la Cultura, donde el conocimiento cierto y seguro de la Matemática ejerce un poder y adquiere una universalidad, que se convierte en la base racional del pensamiento cartesiano y revoluciona todas las ciencias. Con sus fundamentales aportaciones en los campos de la Filosofía, la Geometría, la Óptica, la Mecánica y otros, Descartes da un aliento unitario y orgánico al pensamiento científico, construye una visión global del conocimiento y marca un nuevo rumbo en la Filosofía.

René Descartes nace en La Haye, en Touraine, el 31 de marzo de 1596. Su formación tiene lugar con gran autosatisfacción en el Colegio jesuita de La Flèche entre 1606 y 1616. Descartes siempre tuvo la conciencia de haber sido instruido en una de las mejores escuelas de Europa, como manifiesta en *El Discurso del Método* (DM.AT,VI,5).

Descartes había alcanzado en La Flèche un soberbio conocimiento de la Cultura clásica, que incluía un gran dominio del latín –incluso como lengua viva, en la que podía hablar y escribir–, griego e italiano, y había desarrollado una irrefrenable afición a la lectura como demuestran ciertos pasajes de *El Discurso del Método* (DM.AT,VI,5).

En cuanto a la Filosofía aprendida, siempre se mostró un tanto displicente (DM.AT,VI,6):

> «[...] *Mientras las Matemáticas me han hecho disfrutar he visto la Filosofía como un medio para hablar de manera superficialmente convincente de cualquier cosa y ganar la admiración de los menos cultos*.»

Es más, a juzgar por *El Discurso del Método*, parece que la Filosofía inicialmente no le interesara (DM.AT,VI,17):

> «*La Lógica, sus silogismos y la mayor parte de sus otras reglas sirven más bien para explicar a otro lo que uno sabe más que para aprenderlo*.»

En cuanto a su formación matemática el joven Descartes confiesa que quedó cautivado por la parte del curso de Filosofía referente a las Matemáticas que impartía el padre Françoise, que atendía no sólo a los aspectos teóricos de las Matemáticas, sino también a las artes mecánicas, los autómatas, la Óptica, la magia y la Astrología.

La propia *Ratio Studiorum* de los Jesuitas de 1586 establece:

> «*La Enseñanza de las Matemáticas conviene a los fines de la Orden, no sólo por el prestigio que dan a toda Academia, sino también en razón de su utilidad en todas las profesiones*.»

También en el folio 183 de los Archivos romanos de la Compañía de Jesús consta:

> «*La Enseñanza de las Matemáticas es de las más útiles no sólo porque contribuye a la precisión del razonamiento sino también porque procura conocimientos infinitamente ventajosos para el bien de la sociedad*.»

La enseñanza matemática de los Jesuitas tenía una orientación eminentemente práctica. Además del *Cuadrivium* pitagórico añadía nociones de Mecánica, Óptica, Acústica, Topografía, Perspectiva, Hidráulica y Balística, según el cuadro general de la Matemática práctica renacentista, y con una orientación hacia la ingeniería civil y militar, de interés para los jóvenes nobles que ocuparían cargos en la administración y en el ejército. Por eso cuando un Descartes ya maduro mira retrospectivamente, en el autobiográfico *Discurso del Método*, hacia sus años de formación comenta (DM.AT,VI,16):

> «*Las Matemáticas tienen invenciones muy sutiles y pueden utilizarse tanto para contentar a los curiosos como para facilitar todas las Artes y disminuir el trabajo de los hombres*.»

LA INFLUENCIA DE C.CLAVIUS SOBRE DESCARTES

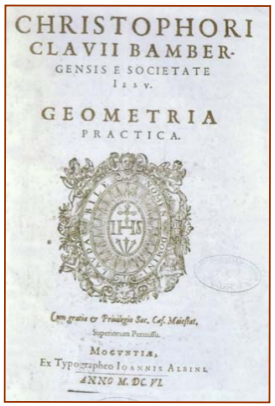
El jesuita alemán C.Clavius, profesor de Matemáticas en el Colegio Romano de Roma, fue el gran inspirador de la Enseñanza de la Matemática en la época de Descartes que incluía *La Ratio Studiorum* de los Jesuitas.

Clavius organizó un verdadero seminario de jóvenes matemáticos, destinados a proveer de profesores de Matemáticas a los colegios jesuitas.

Aparte de excelente profesor, Clavius se reveló como un magnífico escritor de libros de texto. Publicó en 1574 una célebre edición de *Los Elementos de Euclides*, que tuvo reediciones en 1589, 1591, 1603, 1607, 1612 y 1674, lo que da idea de su valor. También escribió magníficos manuales de *Aritmética Práctica* (1583), *Geometría Práctica* (1604), *Álgebra* (1608), una edición comentada de la *Sphera* de Sacrobosco (1591), un compendio de Trigonometría y Astronomía ampliamente utilizado y muy encomiado por Kepler y tuvo una importante intervención en la reforma gregoriana del calendario.

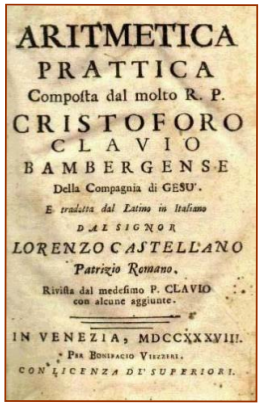
EVCLIDIS
ELEMENTORVM
LIBRI XV.

COLONIÆ.


CHRISTOPHORI
CLAVII BAMBER-
GENSIS E SOCIETATE
IESV.

GEOMETRIA
PRACTICA.

MOGVNTIÆ,
Ex Typographeo IOANNIS ALBINI.
ANNO M. DC. VI.


ARITMETICA
PRATTICA
Composta dal molto R. P.
CRISTOFORO
CLAVIO
BAMBERGENSE
Della Compagnia di GESU'.
E tradotta dal Latino in Italiano
DAL SIGNOR
LORENZO CASTELLANO
Patrizio Romano.
Riveduta dal medesimo P. CLAVIO
con alcune aggiunte.
IN VENEZIA, MDCCXXVIII.
CON LICENZA DE' SUPERIORI.

Ediciones de Clavius de *Los Elementos* de Euclides (Colonia 1591), *Geometría Práctica* (Maguncia, 1606) y *Aritmética Práctica* (Venecia, 1738)
Ya que la enseñanza en el Colegio de *La Flèche* estaba inspirada en la *Ratio Studiorum* de los Jesuitas, es de suponer que la doctrina Matemática recibida por Descartes en sus años de Formación se basaría en las primeras ediciones de estos manuales de Clavius. Pero su proyecto de reforma de la Geometría tuvo que partir necesariamente de un profundo conocimiento de las grandes obras de la Matemática Griega de Euclides, Apolonio, Diofanto y Pappus.

Con toda seguridad los Jesuitas usaban los manuales del más famoso de sus matemáticos C.Clavius, llamado «*El nuevo Euclides*» por sus coetáneos, quien con su enseñanza de Matemáticas en el Colegio Romano, la más prestigiosa institución docente de los Jesuitas, y sus publicaciones, contribuyó más que nadie, a dignificar y extender el papel de la Matemática en el Currículum general de la Enseñanza. Sus trabajos fueron recopilados en su famosa *Opera Mathematica*, publicada en cinco volúmenes en 1611. En esta obra, Clavius realiza una apasionada apología de la Matemática, contrastando la firmeza y la unanimidad de las opiniones de los matemáticos con la multiplicidad de visiones diferentes, y por tanto de incertidumbre, que habita en la mente de los filósofos, sensación que manifestará claramente Descartes en *El Discurso del Método* (DM.AT,VI, 7-8):

> «*Me complacía especialmente* [en mi juventud] *en las Matemáticas por la certeza y la evidencia de sus razonamientos,* [...] *De la Filosofía sólo diré que, habiendo sido cultivada por los espíritus más excelentes, y que sin embargo aún no hay nada de lo que no se discuta,* [...] *, y considerando cuantas opiniones diversas pueden haber acerca de un mismo tema,* [...].»

EL JOVEN DESCARTES

***Descartes joven*. Supuesto retrato del filósofo. Escuela Francesa del siglo XVII. Museo de los Agustinos de Toulouse.**

A través de una irresistible pasión por la lectura, la actividad intelectual adolescente de Descartes iba fraguando su pensamiento filosófico y matemático.

La sólida formación en las humanidades del mundo clásico y la consiguiente afición de Descartes a la Cultura griega, se extendía a los grandes tratados de la Matemática griega: *Los Elementos de Euclides*, las Obras de Arquímedes, *La Aritmética* de Diofanto y sobre todo *Las Cónicas* de Apolonio y *La Colección Matemática* de Pappus, obras que conocía en profundidad. Asimismo, Descartes debía estar al corriente de los desarrollos del Álgebra de los matemáticos italianos, Tartaglia, Cardano y Ferrari, y aunque confiesa que desconocía la obra de Vieta antes de escribir *El Discurso del Método*, es inconcebible que así fuera, ya que hay una manifiesta continuidad en la línea de pensamiento geométrico entre la obra de Vieta y *La Geometría* de Descartes.

Descartes sale de La Flèche con el mejor bagaje cultural para emprender su aventura intelectual. El dominio del latín y del griego le abren las puertas al saber de los clásicos y a toda la erudición renacentista; las Humanidades y la Retórica animaron una conversación interesante y el don de gentes; situándose en las mejores condiciones de integrarse en la agitada vida social y pública de su época y dedicarse al conocimiento del mundo, como manifiesta en *El Discurso del Método* (D.M.AT,VI, 9):

> **« [...], *gracias a Dios, no me encontraba en la situación de verme obligado a hacer de la Ciencia un oficio para alivio de mi fortuna,* [...] *Empleé el resto de mi juventud en viajar, en ver cortes y ejércitos,* [...], *en recoger experiencias diversas, en probarme a mí mismo, en reflexionar sobre lo que me ocurriera,* [...].»**

Hegel en su *Lecturas sobre la Historia de la Filosofía* describe al joven Descartes cono un ser vivaz e inquieto, con insaciable afán de conocimiento en todos los sistemas y formas de pensamiento, de modo que tres experiencias juveniles sucesivas jalonarían la forja de su espíritu: sus amplios estudios de juventud en La Flèche; su buceo en el gran libro del mundo, con otros hombres y otros pueblos; y el encanto o hechizo de las Matemáticas, cuya esencia impregnará todo su pensamiento. Tres experiencias que señalan tres caminos o vías en la búsqueda incesante de la verdad.

CITAS MEMORABLES DE DESCARTES

REGLAS PARA LA DIRECCIÓN DEL ESPÍRITU. (R.AT.X. 359–468)

1. Sólo la Aritmética y la Geometría están libres de todo defecto de falsedad e incertidumbre. [RII. 364].
2. Los que buscan el camino recto de la verdad no deben ocuparse de ningún objeto sobre el que no puedan tener una certidumbre semejante a las demostraciones de la Aritmética y de la Geometría. [RII. 366].
3. Es mucho más acertado no pensar jamás en buscar la verdad de las cosas que hacerlo sin método. [RIV. 371].
4. Ninguna ciencia puede obtenerse, sino mediante la intuición de la mente o la deducción. [RIV. 372].
5. Cultivé [en mi juventud] preferentemente la Aritmética y la Geometría, porque se las tenía por las ciencias más simples y como un camino para las demás. [RIV. 374].
6. El silogismo es completamente inútil para los que desean investigar la verdad de las cosas y sólo puede aprovechar , a veces, para exponer con mayor facilidad a los otros las razones ya conocidas. [RX. 406].

EL DISCURSO DEL MÉTODO (DM.AT.VI. 1–78)

1. Las Matemáticas tienen invenciones muy sutiles y pueden utilizarse tanto para contentar a los curiosos como para facilitar todas las Artes y disminuir el trabajo de los hombres. [6]
2. Mientras las Matemáticas me han hecho disfrutar he visto la Filosofía como un medio para hablar de manera superficialmente convincente de cualquier cosa y ganar la admiración de los menos cultos. [6]
3. Gustaba, sobre todo de las Matemáticas por la certeza y evidencia de sus razones. [7].
4. La Lógica, sus silogismos y la mayor parte de sus otras reglas sirven más bien para explicar a otro lo que uno sabe más que para aprenderlo. [17].
5. Esas largas cadenas trabadas de razones muy simples y fáciles, que los geómetras acostumbran a emplear para llegar a sus más difíciles demostraciones, me habían dado ocasión para imaginar que todas las cosas que entran en la esfera del conocimiento humano se encadenan de la misma manera. [19].
6. Entre todos los que han buscado la verdad en las ciencias, sólo los matemáticos han podido hallar algunas demostraciones, esto es, algunas razones ciertas y evidentes. [19].
7. No esperaba sacar de las demostraciones matemáticas más utilidad que acostumbrar mi espíritu a saciarse de verdades y a no contentarse con falsas razones. [19].

LA GEOMETRÍA (G.AT.VI. 369–485)

1. Todos los problemas de Geometría pueden reducirse fácilmente a términos tales, que no es necesario conocer de antemano más que la longitud de algunas líneas rectas para construirlos. [369].
2. Pero no me detengo a explicar esto con más detalle para no privar a cada uno del placer de aprenderlo por sí mismo, ni impedir el cultivo útil del propio espíritu ejercitándolo, que es, a mi parecer, la principal utilidad que puede obtenerse de esta ciencia [374].
3. Se pueden construir todos los problemas de la geometría ordinaria sin hacer más que lo poco que está comprendido en las cuatro figuras que he explicado. [376]
4. Para encontrar todas las propiedades de las líneas curvas basta con saber la relación que tienen todos sus puntos con los de las líneas rectas, [...] y conocer la manera de trazar otras líneas que las corten en todos esos puntos en ángulo recto. [...] Y me atrevo a decir que éste es el problema más útil y más general no sólo que yo conozca, sino aun que yo haya anhelado jamás conocer en Geometría. [412–413].
5. Y yo espero que nuestros descendientes me estarán agradecidos no sólo por las cosas que aquí he explicado, sino también por aquellas que he omitido voluntariamente a fin de dejarles el placer de descubrirlas. [485].

CITAS MEMORABLES SOBRE DESCARTES

1. A Descartes le fue revelada en sueños la clave mágica que le abría el acceso al tesoro de la naturaleza y que le colocaba en situación de poseer los verdaderos fundamentos de todas la ciencias. **A.Baillet. *La Vie de Monsieur Des-Cartes*.**
2. Acudiendo a la cita con su ejército, en la calma del invierno, combinaba en su mente los misterios de la naturaleza con las leyes de las Matemáticas, aspirando a desvelar los secretos de ambas. **Epitafio de Descartes por H. Pierre Chanot, 1650.**
3. Su alma siempre con sabiduría fecunda, hacia ver a los espíritus lo que se esconde a los ojos. Después de haber explicado el modelo del mundo reveló el misterio de los cielos. **Epitafio de Descartes por C. Huygens, 1650.**
4. Descartes mediante un nuevo método hizo pasar de las tinieblas a la luz cuanto en las Matemáticas había permanecido inaccesible a los antiguos y todo cuanto los contemporáneos habían sido incapaces de descubrir; luego puso los cimientos inquebrantables de la Filosofía sobre los cuales es posible asentar la mayor parte de las verdades en el orden y con la certidumbre de las Matemáticas. **Spinoza. *Los Principios de la Filosofía cartesiana.***
5. Lo que ha inmortalizado el nombre de este gran hombre, es la aplicación que ha sabido hacer del Álgebra a la Geometría, una idea de las más vastas y felices que haya tenido el espíritu humano, y que será siempre la llave de los más profundos descubrimientos no solamente en la Geometría, sino en todas las ciencias físico-matemáticas. **D'Alembert. *Discours Préliminaire de l'Encyclopédie* (Orbis, 1984, pp.84,85).**
6. La *Dióptrica* de Descartes es la más grande y la más bella aplicación que se haya hecho hasta ahora de la Geometría a la Física. **D'Alembert. *Discours Préliminaire de l'Encyclopédie* (Orbis, 1984, p.85):**
7. Descartes se caracterizaba por su espíritu vivaz e inquieto, que buscaba con insaciable afán todas las ramas del conocimiento humano, buceando en todos los sistemas y formas de pensamiento. **Hegel. *Lecturas sobre la Historia de la Filosofía.***
8. Sólo quien haya pensado real y detenidamente este escrito **[Las *Reglas para la dirección del espíritu*]**, radicalmente parco, hasta en sus rincones más recónditos y fríos, está en condiciones de tener una idea de lo que pasa en la ciencia moderna. **M.Heidegger. *Die Frage nach dem Ding*.**
9. El cartesianismo no debe nada esencial a ninguna doctrina de la antigüedad. **H.Bergson. *La Filosofía.***
10. No hay una Matemática, hay muchas Matemáticas. [...]. El espíritu antiguo creo su Matemática casi de la nada. El espíritu occidental, histórico, había aprendido la Matemática antigua, y la poseía, aunque sólo exteriormente y sin incorporarla a su intimidad; hubo, pues, de crear la suya modificando y mejorando, al parecer, pero en realidad aniquilando la Matemática euclidiana, que no le era adecuada. Pitágoras llevó acabo lo primero; Descartes lo segundo. Pero los dos actos son, en lo profundo, idénticos. **O.Spengler. *El sentido de los números*. (*La decadencia de Occidente*, p.144).**
11. La Geometría analítica, mucho más que cualquiera de sus especulaciones metafísicas, inmortaliza el nombre de Descartes y constituye el máximo paso hecho en el progreso de las ciencias exactas. **J. Stuart Mill. (citado por E.Bell en *Les grands mathématiciens*. Payot, París, 1950. Cap.3. p.46).**
12. La Geometría Analítica de Descartes ha afectado probablemente a la vida humana más profundamente, aunque menos violentamente, que la máquina de vapor o el aeroplano. **L.Hull. *Historia y Filosofía de la Ciencia*, 1981, p.268.**
13. La Geometría Analítica de Descartes cambió la faz de las Matemáticas. **M.Kline. *El pensamiento matemático de la Antigüedad a nuestros días*, 1992. vol.1, p.425.**

Los sueños de Descartes y los orígenes de *La Geometría*

El 10 de noviembre de 1618 ocurrió un evento de trascendental importancia en la vida de Descartes, el encuentro con Beeckman, un intelectual amante de la Física y la Matemática. Vagando por Breda, Descartes se tropezó con una gente que miraba un anuncio en el que un matemático retaba a que se resolviese un problema, cosa muy propia de la época. Como Descartes todavía no dominaba el holandés, suplicó a quien estaba al lado que se lo tradujese al latín o al francés. Resultó ser Beeckman, quien hablándole en latín, le explicó en qué consistía el problema: «¿*Cuán lejos caerá una piedra en una hora si se sabe cuán lejos cae en dos?*», y le dio su tarjeta de visita. Beeckman se quedó atónito cuando al día siguiente el joven francés se presentó en su casa, sin anunciarse, con la solución del problema, lo que inició una fructífera amistad, mantenida sobre todo de forma epistolar, plena de gratitud recíproca, como muestran sendas cartas de Descartes a Beeckman, donde se explaya en palabras de agradecimiento hacia él por ser el catalizador de la empresa intelectual de ordenamiento de sus reflexiones y concepciones científicas.

> «*Podéis estar seguro que antes olvidaría a las musas que a vos,* [...]*, ellas me han atado a vos con lazos de afecto*» (24/1/1619, AT,X, 162-163).

> «[...] V*os habéis sido el instigador, el motor primero de mis investigaciones,* [...]*, vos me sacudisteis la desidia, apartándome de la erudición inútil, conduciendo mi espíritu, que vagaba en ocupaciones ociosas, a otras mejores,* [...]» (23/4/1619).

Influido por Beeckman, Descartes emprende una serie de estudios matemáticos en relación con la trisección del ángulo y las ecuaciones cúbicas, y es consciente, tras los contactos con la literatura rosacruciana alemana, de que los aspectos algebraicos, en especial el simbolismo que apuntaba a convertirse en el lenguaje universal que permitiría el conocimiento y el dominio global de la realidad, entroncan con la tradición hermética y cabalística del arte luliano en íntima relación con la idea del saber universal.

EN UNA MANO LA PLUMA Y EN L A OTRA LA ESPADA

Descartes en las calles de París, por Chartan.

S. de Sacy, en su biografía de Descartes de 1956, refleja la imagen del filósofo como lo hace esta ilustración, describiendo «*al hombre izado entre una generación de aventureros,* [...]*, un mosquetero del alma,* [...] *en vagabundeo metódico*, [...]».

En el gran teatro del mundo, apareciendo como desherado y marginal de las clases sociales dominantes, desligado de la tradición y del marco familiar, Descartes «*sostiene en una mano la pluma y en la otra la espada*», en un continuo vaivén entre el afán de retiro y el estudio y su curiosidad por la vida mundana, alistándose en ejércitos, dedicado a la vida militar, como aventurero y rebelde, junto a los mercenarios de las guerras de Religión que asolaban Europa.

Provisto del bagaje intelectual del Renacimiento, dotado de una prodigiosa erudición alcanzada en *La Flèche* y de una brillante retórica, con una incontenible afición a la Matemática, trasmitida por el padre Françoise y por Beeckman, Descartes viaja y conoce mundo «*pues es casi lo mismo conversar con gentes de otros siglos que viajar*» (DM.AT,VI, 6) y se convierte en el *filósofo enmascarado* que persigue la sabiduría universal que anunciaban sus curiosas lecturas de adolescencia, de raíces lulianas, dentro de la tradición hermético-cabalística.

El interés por los rosacruces, que presumían de tener la clave de la sabiduría universal, impele a Descartes a abandonar Holanda hacia mayo de 1619 camino de Alemania, donde asiste a la coronación de Fernando II y se enrola en las tropas del Duque de Baviera.

A la llegada del invierno, Descartes se retira a alguno de los refugios militares, quizá en Ulm, donde vivía el matemático Faulhaber. En un ambiente propicio para la meditación, Descartes se plantea algunos problemas geométricos y la solución lograda le induce a buscar un método general para resolver cualquier problema de Geometría que se le presentase. Pero enseguida extrapola sus ideas y amplía tan ambicioso plan para concebir la posibilidad de encontrar un método para el descubrimiento de la verdad en cualquier rama de la ciencia. Así que en la mente de Descartes fue tomando cuerpo el ideal de un conocimiento integral, unificado sobre la totalidad global de las ciencias, que al tratar de lo divino y de lo humano, reuniría todas las ciencias con un simbolismo adecuado, intuyendo que el Álgebra y la Geometría, adecuadamente interpretadas e insertas en un simbolismo superior, podían convertirse o al menos apuntar hacia el tan proclamado saber universal.

Con estas ideas fijas en la mente, en la noche del 10 de noviembre de 1619, primer aniversario del encuentro con Beeckman, Descartes tuvo una concatenación de sueños que le dejaron una profunda impresión marcando un hito en su ulterior evolución espiritual.

La profunda experiencia visionaria fue plasmada por Descartes en un manuscrito de 1620, en latín, con el nombre de *Olympica*, ahora perdido, que parece ser fue ojeado por Leibniz en su estancia en París en 1675 y que fue traducido por Baillet, el biógrafo de Descartes (AT.X, 181-188). En una minuciosa descripción de los sueños y su interpretación, Descartes relata un itinerario simbólico en sus tres sueños: siente angustia en el primero, luces prometedoras en el segundo, hasta alcanzar la revelación de la verdad en el tercero, en el cual el espíritu de la Verdad quería «*abrirle los tesoros de todas las ciencias*».

Los *Olympica* comienza con estas palabras:

> «X *Novembris* 1619, *cum plenus forem Enthousiasmo et mirabilis scientiae fundamenta reperirem* ...»

> «X *de noviembre de* 1619, *cuando, lleno de entusiasmo, descubrí los fundamentos de una ciencia admirable*.»

En la interpretación mística que hace Descartes de sus sueños, a él se le ha revelado la unidad de la ciencia, ha sido ungido de un sagrado entusiasmo místico que le ha liberado de una crisis espiritual y le ha cargado de una gran responsabilidad en el alumbramiento de la verdad al tomar conciencia de una misión:

> ¿*Será ésta emprender la magna empresa de reforma de la Filosofía y consecuentemente de la Matemática*?

El espíritu de la verdad ha conducido a Descartes a una exaltación intelectual para alcanzar «*la visión de una ciencia nueva y admirable*», que tal vez debía de ser el conocimiento de todas las cosas de las que el espíritu humano es capaz, y que sus fundamentos consistirían en un método general –extraído de los procedimientos del pensamiento matemático– donde se experimenta la certeza y evidencia inherentes al verdadero saber.

Al año siguiente, el mismo día 10 de Noviembre, aniversario del encuentro con Beeckman y de los sueños, Descartes vuelve a tener una visión que le ilumina, escribiendo al margen del manuscrito de Los *Olympica* (AT.X.179):

> «*X Novembris 1620. Coepi intelligere fundamentum inventi mirabilis*.»

> «10 Noviembre 1620. *He empezado a entender el fundamento de un admirable descubrimiento*.»

LOS SUEÑOS DE DESCARTES Y *LA GEOMETRÍA*

Descartes representado como Fausto. *Opuscula posthuma, physica et mathematica*. J. Blaeu. Amsterdam, 1701.

La noche del domingo del 10 al 11 de noviembre de 1619, en un descanso en los cuarteles de invierno de los ejércitos de Maximiliano de Baviera, Descartes enfrebrecido sufre alucinaciones. Sintiendo una iluminación interior asiste «*lleno de entusiasmo*» a la revelación de «*los fundamentos de una ciencia admirable*». Descartes tiene tres sueños que le hacen tomar conciencia de su vocación filosófica. En palabras de su primer biógrafo A.Baillet (*La Vie de Monsieur Des-Cartes*):

> **«[...] *Le fue revelada la clave mágica que le abría el acceso al tesoro de la naturaleza y que le colocaba en situación de poseer los verdaderos fundamentos de todas la ciencias*».**

La descripción de los sueños de Descartes en los *Olympica* pudo ser un pretexto literario o un artificio poético-filosófico para explicar que se sentía predestinado a la búsqueda de la sabiduría universal. En todo caso la experiencia onírica de Descartes fue una intensa vivencia personal, un auténtico Pentecostés, que marcó su porvenir. A raíz de los sueños, Descartes, imbuido de entusiasmo y satisfacción, decidió ir en peregrinación al santuario de Santa María de Loreto y aunque cambió de residencia muchas veces entre 1619 y 1650, jamás se separó del manuscrito de sus sueños.

Los sueños de Descartes, de gran significado freudiano, marcaron una impronta inmarcesible en la orientación de su pensamiento. El rapto místico habría de servir a Descartes de cimiento de un sólido edificio racionalista, presidido por la «*unidad*» como emblema para entender el mundo: unidad de la Matemática a través de la fusión del Álgebra y la Geometría; unidad entre Física y Matemática; unidad de todas la ciencias; presidida por la unidad de método y criterio «*para bien conducir la razón y buscar la verdad en las ciencias*», que así subtitulará precisamente a su principal obra filosófica y científica: *El Discurso del Método*; en suma, unidad de todo el saber radicada en el espíritu proclamada con carácter de primariedad desde el comienzo de su último escrito de juventud: *Las Reglas para la dirección del espíritu* (*Regulae ad directionem ingenii*)

La importancia que *La Geometría* de Descartes tiene en la Historia de la Matemática, ha propiciado, a veces, la sublimación de la intuición de sus raíces en la mente de Descartes, de modo que algunos historiadores le atribuyen un origen casi legendario, según el cual el 10 de Noviembre de 1619, en su delirio onírico Descartes habría adivinado la unión del Álgebra y la Geometría en un solo cuerpo de doctrina: *La Geometría Analítica* –aunque más bien habría que hablar de Geometría Algebraica– y ante «*la visión de una ciencia nueva y admirable*» se habría sentido predestinado para construir un nuevo sistema filosófico, donde la Matemática ocuparía una situación privilegiada como llave del conocimiento y la sabiduría universal e instrumento de explicación racional de los fenómenos naturales, de modo que la iluminación de Descartes le procuraría una explicación global de la naturaleza física, es decir una Filosofía natural –una Física en sentido actual– basada en la Matemática. Así se explicarían los tres ensayos que al acompañar a *El Discurso del Método* justificarían de forma verdadera y global el método cartesiano.

Las Regulae, El Discurso del Método y La Geometría

La lectura de las *Regulae* y *El Discurso del Método* es un preliminar necesario, o al menos aconsejable, para entender la motivación y los presupuestos intelectuales de Descartes acerca de la Ciencia, de la universalización del razonamiento matemático como base del conocimiento racional y en particular de los orígenes y objetivos de *La Geometría*. Como señala Víctor Gómez Pin (Congreso de Ontología, 24/3–31/3 de 1996, San Sebastián-Barcelona):

> «*El Discurso del Método es tan sólo el prólogo añadido por Descartes a sus escritos científicos* [los tres ensayos *la Dióptrica, los Meteoros y la Geometría*]*, a fin de mostrar lo estéril que sería abordar éstos sin el hilo conductor de la problemática común, es decir, sin referencia a la unidad de la razón que tales escritos despliegan*.»

Descartes había estudiado las Matemáticas con gran fruición en su adolescencia y desde el primer momento apreció su indudable condición de certeza, pero sólo más tarde llegó a reparar en lo que él llama su verdadero uso hacia la gestación y desarrollo del *Método*. Es en los sueños de 1619 y cuando escribe, en 1620, en los *Olympica*, sobre «*los fundamentos de una ciencia admirable*», cuando empieza un primer estadio en la intuición del *Método;* el segundo estadio de puesta a punto del *Método* tiene lugar con las *Reglas para la dirección del espíritu de* 1628 (Las *Regulae*) y el tercero, de codificación, con *El Discurso del Método* de 1637.

Descartes busca un fundamento absoluto e inconmovible de la verdad en que basar el conocimiento científico sobre el que cimentar la vida y la acción. Pero ello no es posible alcanzarlo sin método. La Regla IV de las *Regulae* se titula precisamente: «*El método es necesario para la investigación de la verdad de las cosas*» (RIV.AT.X.371), y en ella Descartes alude de forma reiterada sobre el asunto:

> «[...] E*s mucho más acertado no pensar jamás en buscar la verdad de las cosas que hacerlo sin método*» (RIV.AT.X.371).
>
> «[...] *Entiendo por método reglas ciertas y fáciles, mediante las cuales el que las observe exactamente no tomará nunca nada falso por verdadero, y, no empleando inútilmente ningún esfuerzo de la mente, sino aumentando siempre su ciencia, llegará al conocimiento verdadero de todo aquello de que es capaz*» (RIV.AT.X.371–372).
>
> «*El método explica rectamente de qué modo ha de usarse la intuición de la mente para no caer en el error contrario a la verdad y cómo han de ser hechas las deducciones para llegar al conocimiento de todas las cosas,* [...] *ninguna ciencia puede obtenerse, sino mediante la intuición de la mente o la deducción*» (RIV.AT.X.372).

Así pues, las reglas del método se remiten al saber de la razón, pero de los textos cartesianos hay que colegir, como veremos, que en principio se trata de la razón matemática y que en origen las reglas del método lo son primariamente del saber matemático.

En el cultivo de las Matemáticas desde su juventud, Descartes atribuye a las verdades matemáticas una naturaleza esencialmente diferente a la de las verdades basadas en la experiencia. Según Descartes las proposiciones matemáticas no deben su verdad a la experiencia y no pueden ser desmentidas por ésta, es decir son «*verdades de razón*» con una validez universal y absoluta. Es el ámbito de la razón, sobre el que descansa la Matemática, al que acudirá Descartes para fundamentar su método que impone la certeza como condición epistemológica ineludible que excluye los conocimientos tan sólo probables. Así tiene lugar en la Regla II de las *Regulae* titulada (RII.AT.X.361):

> «*Conviene ocuparse tan sólo de aquellos objetos, sobre los que nuestros espíritus parezcan ser suficientes para obtener un conocimiento cierto e indudable*.»

La exigencia cartesiana de encontrar un conocimiento cierto y evidente que alcance y rija con seguridad la unidad del saber, hace recalar a Descartes en el modo del pensar matemático, elaborando el *Método* con base en su larga experiencia en las ciencias geométricas y aritméticas. Así se advierte de forma palmaria a lo largo de las *Regulae*:

> «[...] *Cuando por primera vez me dediqué a las disciplinas Matemáticas, de inmediato leí por completo la mayor parte de lo que suelen enseñar sus autores, y cultivé preferentemente la Aritmética y la Geometría, porque se las tenía por las más simples y como un camino para las demás*» (RIV.AT.X.374–375).

La Aritmética y la Geometría deben, pues, ejercer para Descartes una función propedéutica e indicativa porque en ellas se experimenta la certeza y evidencia requeridas para el verdadero saber; a ellas hay que reducirse, pues sólo ellas están libres de incertidumbre y falsedad, de modo que (RII.AT.X.363,364):

> «[...] *Si calculamos bien, de las ciencias ya descubiertas sólo quedan la Aritmética y la Geometría, a las que la observación de esta regla* [Regla II] *nos reduce*.»

> «[...] *Sólo la Aritmética y la Geometría están libres de todo defecto de falsedad e incertidumbre.*»

En el Colegio de La Flèche Descartes habría recibido una sólida formación matemática, pero más allá de esta ciencia, habría captado el espíritu mismo del saber matemático, que al aunarlo con su notable y peculiar penetración filosófica, alcanzaría la visión de las Matemáticas, por la certeza y evidencia de sus razones, como instrumento clave del descubrimiento de una técnica puramente especulativa –el *Método*– que sitúa al espíritu en posesión de la verdad y en posesión de sí mismo, experimentando lo que deviene el conocimiento humano cuando se le ahorma según el patrón de la evidencia matemática. Descartes se propone con *El Discurso del Método* y *La Geometría* una magna empresa de reforma de la Filosofía y de la Matemática, tomando esta ciencia como principio básico del fundamento de la sabiduría universal. Descartes adopta la demostración matemática frente al recurso a la autoridad y pondera la firmeza y certeza de la Matemática *versus* la incertidumbre de la Filosofía. Pero no todo es panegírico respecto de las Matemáticas, ya que Descartes se queja tanto del uso restringido que se hacía de la Matemática como de la forma misma de enseñarla y pone en un plano secundario el valor técnico de las Matemáticas como mera herramienta para las artes y los oficios mecánicos. En efecto, al aludir a su etapa de formación, Descartes escribe, en el *Discurso del Método*, respecto de las Matemáticas (DM.AT,VI,7):

> *Gustaba* [en mi juventud]*, sobre todo, de las Matemáticas por la certeza y evidencia de sus razonamientos, pero no había entendido todavía su verdadero uso y, pensando que sólo servían para las artes mecánicas, me sorprendía de que, siendo tan firmes sus fundamentos, no se hubiera construido sobre ellas nada más relevante*.»

Más adelante, en la segunda parte de la obra, Descartes continúa diciendo (DM.AT,VI,17):

> «*Había estudiado entre las partes de la Filosofía, la Lógica, y de las Matemáticas, el Análisis de los geómetras y el Álgebra, tres Artes o Ciencias, que debían, al parecer, contribuir algo a mi propósito.* [...] *Respecto al Análisis de los antiguos y el Álgebra de los modernos, aparte de que no se refieren sino a muy abstractas materias que no parecen ser de ningún uso, el primero está siempre tan constreñido a considerar las figuras, que no puede ejercitar el entendimiento sin fatigar mucho la imaginación, y en la última hay que sujetarse tanto a ciertas reglas y cifras, que se ha hecho un arte confuso y oscuro, bueno para enredar el espíritu, en lugar de una ciencia que lo cultive. Esto fue causa de que pensase que era necesario buscar algún otro método que, reuniendo las ventajas de estos tres, estuviese libre de sus defectos.*»

A juzgar por este texto, el valor propedéutico y pedagógico de la Aritmética y la Geometría en la concepción del *Método*, es asumido por Descartes una vez se hayan corregido las deficiencias y limitaciones de estas ciencias, es decir, una vez que Descartes haya transformado los antiguos instrumentos de la Geometría griega –el Álgebra Geométrica y el Análisis Geométrico– en lo que hoy llamamos la Geometría Analítica cartesiana, mediante la intervención del Álgebra literal y simbólica de Vieta sobre la Geometría, tras la drástica reforma y simplificación de la notación algebraica que el propio Descartes realizará, primero de forma provisional en la Regla XVI de las *Regulae* (RXVI.AT.X.455) y ya de forma definitiva en *La Geometría* (G.AT,VI,371). Efectivamente, uno de los atributos más importantes de la Geometría Analítica es que libera al investigador de la dependencia a ultranza de las figuras geométricas al reemplazar las ingeniosas construcciones geométricas de la Geometría griega por sistemáticas operaciones algebraicas, es decir, permite «*ejercitar el entendimiento sin fatigar mucho la imaginación*» (DM.AT,VI,17). Así concibe Descartes una ciencia matemática que se convierte en un saber más fácil y simple, y generalizable y valido para todo el ámbito de la cantidad. Pero no sólo esto, porque el modo de proceder y el espíritu de esta verdadera Matemática, experimentado y cultivado en el quehacer y en la investigación matemáticos, es lo que inspira las reglas del *Método* y el *Método* mismo.

Según Descartes, sólo la Aritmética y la Geometría no ofrecen dudas ni conocimientos probables. Pero esto no significa que sólo haya que aprender y ocuparse de estas ciencias sino que la certeza y los rasgos que encontramos en ellas es lo que debemos requerir en la búsqueda del camino que nos conduce a la verdad en general (RII.AT.X.366):

> «[...] *Mas de todo esto se ha de concluir no ciertamente que se han de aprender sólo la Aritmética y la Geometría, sino únicamente que aquellos que buscan el recto camino de la verdad no deben ocuparse de ningún objeto del que no puedan tener una certeza semejante a la de las demostraciones de la Aritmética y de la Geometría.*»

Aún más esclarecedor es el texto que sigue a las famosas cuatro reglas en el que Descartes reconoce el proceder de los geómetras en la inspiración de su método (DM.AT,VI,19):

> «*Esas largas cadenas trabadas de razones muy simples y fáciles, que los geómetras acostumbran a emplear para llegar a sus más difíciles demostraciones, me habían dado ocasión para imaginar que todas las cosas que entran en la esfera del conocimiento humano se encadenan de la misma manera,* [...]*, y considerando que entre todos los que antes han buscado la verdad en las ciencias, sólo los matemáticos han podido hallar algunas demostraciones, esto es, algunas razones ciertas y evidentes, no dudé de que debía comenzar por las mismas que ellos han examinado.*»

He aquí un texto muy significativo de la importancia del método matemático en el fundamento del pensamiento cartesiano, sobre todo el método seguido por los geómetras, que parten de las cosas más sencillas y fáciles de conocer para elevarse mediante «*largas cadenas de trabadas razones*» hasta alcanzar las cuestiones más difíciles y complejas. Descartes concebía que las entidades del conocimiento se encadenan como las proposiciones geométricas, que son, junto con las aritméticas, las únicas que gozan de certeza y evidencia, por tanto por ellas había que empezar como guía hacia el *Método*.

La naturaleza de la Matemática, tiene para Descartes un acusado carácter instrumental y pedagógico en la búsqueda y fundación de un saber científico unificado, universal, cierto y evidente, que Descartes denomina *Mathesis Universalis*, una cierta ciencia general, una determinada y precisa forma de saber extraída del modo, el estilo y el método de los saberes matemáticos. En este sentido Descartes se acerca al pensamiento platónico de la *República* que concebía la Matemática no sólo como imprescindible propedéutica en el ascenso hacia la Filosofía y fundamento de todo el saber humano, sino también como el camino inexcusable en la realización de la *Paidea*, entendida como cultivo y formación del espíritu humano en todas sus facetas. En efecto, así lo manifiesta Descartes en la Regla IV (RIV.AT.X.375-376):

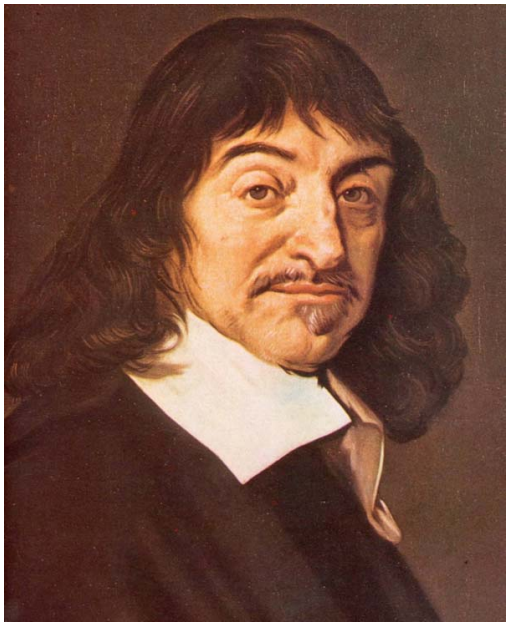
«[Pensé] *por qué sucedía que antiguamente los primeros creadores de la Filosofía no quisieran admitir para el estudio de la sabiduría a nadie que no supiese Mathesis, como si esta disciplina pareciese la más fácil y sobremanera necesaria de todas para educar los espíritus y prepararlos para comprender otras ciencias más altas,* [...].»

El último texto citado de *El Discurso del Método* continua con estas palabras (DM.AT,VI,20):

«[...] A*l advertir que, aunque* [las ciencias matemáticas] *tienen objetos diferentes, concuerdan todas en no considerar sino las relaciones o proporciones que se encuentran en tales objetos, pensé que más valía limitarse a examinar esas proporciones en general,* [...] *, pensé que, para considerarlas mejor particularmente, debía suponerlas en línea* [recta], *pues nada hallaba más simple ni que más distintamente pudiera representarse a mi imaginación y a mis sentidos. Y que para retenerlas o comprenderlas era necesario explicarlas mediante algunas cifras lo más cortas que fuera posible; de esta manera tomaría lo mejor del Análisis geométrico y del Álgebra y corregiría los defectos del uno por medio de la otra.*»

He aquí, en términos del propio Descartes, el origen y los fundamentos de *La Geometría*. Descartes toma la línea recta como representación de toda magnitud y, además, propone una reforma de la notación algebraica. De esta forma conservará del Análisis Geométrico el auxilio que recibe de la imaginación y del Álgebra –una vez reformada la notación– la mecanización operacional que permite su simbolismo. La proyección del Álgebra sobre el Análisis geométrico –Descartes dice «*corrigiendo sus defectos*»– producirá lo que llamamos su «*Geometría Analítica*».

EL MÁS FAMOSO RETRATO DE UN FILÓSOFO



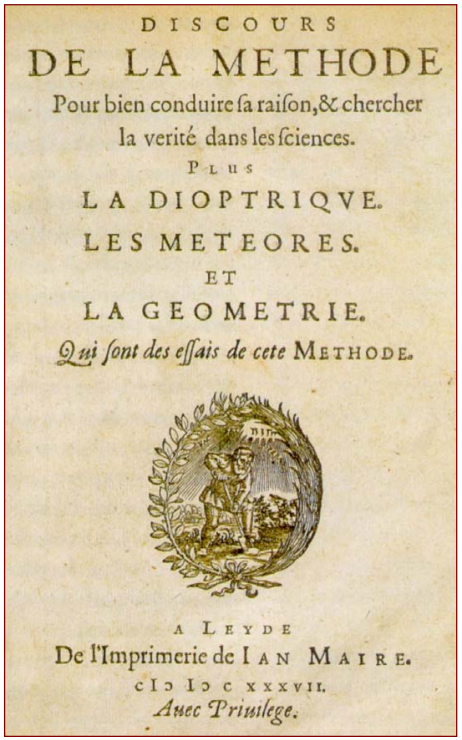
Retrato de Descartes atribuido a F.Hals. Museo de Louvre.

Tal vez es el retrato más célebre de un filósofo. aunque no se puede decir con certeza que sea Descartes ni que sea de F.Hals. Impresiona por su penetrante e inteligente mirada.

Descartes es considerado como el fundador de la Filosofía moderna. Prescindiendo de las bases ideológicas de sus antecesores, Descartes intenta construir, tanto en Filosofía como Matemática, un edificio completamente nuevo y sistemáticamente completo en sí mismo. La evolución ulterior del pensamiento nos indica que fue más afortunado en ésta que en aquélla. La originalidad de Descartes siempre ha sido muy ponderada en la historiografía, de modo que es casi una regla en la Historia de la Filosofía y de las Ciencias, la afirmación de que Descartes no debe casi nada a sus predecesores. Así lo asegura Bergson (*La Filosofía*, Larouse, París, 1916). con la frase:

«*El cartesianismo no debe nada esencial a ninguna doctrina de la antigüedad.*»

EL DISCURSO DEL MÉTODO Y LAS MATEMÁTICAS


DISCOURS
DE LA METHODE
Pour bien conduire sa raison,& chercher
la verité dans les sciences.
PLUS
LA DIOPTRIQVE.
LES METEORES.
ET
LA GEOMETRIE.
Qui sont des essais de cete METHODE.
A LEYDE
De l'Imprimerie de IAN MAIRE.
cIↃ IↃ C XXXVII.
Auec Priuilege.

La primera edición de *El Discurso del Método* con los tres ensayos *la Dióptrica, los Meteoros y la Geometría (Leyden, 1637).*

***El Discurso del Método* es la autobiografía intelectual de Descartes.**

Descartes encontró en la Matemática, un modelo paradigmático en la búsqueda de las primeras verdades absolutamente ciertas que pudieran servirle de base, apoyo y fundamento en la reconstrucción de todo el edificio científico y filosófico, por eso la Matemática devino en la base racional de su pensamiento

Cuando se habla del cartesianismo como método de la razón se debe entender «*método de la razón matemática*» en el sentido de que las reglas del método son extraídas por Descartes del saber y del conocimiento matemáticos, por una parte, y de la práctica, estilo y procedimientos matemáticos, por otra.

Concretamente Descartes habla de tres Artes o Ciencias que habían de contribuir a su propósito (DM.AT,VI,17):

- **La *Lógica* como parte de la Filosofía.**
- **El antiguo *Análisis de los geómetras.***
- **El *Álgebra* de los modernos.**

Así pues, utilizando la Matemática como paradigma en la indagación de la verdad, es decir, el *Análisis* de los geómetras y la *Síntesis* de los algebristas, Descartes establece el «*Método para conducir bien la razón y buscar la verdad en las ciencias*».

Descartes persigue ante todo en *El Discurso del Método* la búsqueda de un «*ars inveniendi*», es decir, un método que sirviera para descubrir verdades y no para probar lo que ya se ha hallado, defender tesis o exponer teorías, y aplicable en todas las ciencias y en particular en la Geometría. Como ya había sentado R. Bacon en el *Novum* Organum, la Lógica aristotélica era inútil para la invención científica porque el silogismo no es aplicable a los principios de las ciencias, ya que sólo sirve para imponer el asentimiento y no para aprehender la realidad. Esta misma actitud asume Descartes con respecto a la lógica tradicional tanto en *El Discurso del Método* como en las *Regulae*:

> **«*La Lógica, sus silogismos y la mayor parte de sus otras reglas sirven más bien para explicar a otro lo que uno sabe más que para aprenderlo*» (DM.AT,VI,17).**
>
> **«*El silogismo es completamente inútil para los que desean investigar la verdad de las cosas y sólo puede aprovechar, a veces, para exponer con mayor facilidad a los otros las razones ya conocidas*» (RX.AT.X.406).**

Descartes ya se había fijado, para su propósito, en la IV Regla de las *Regulae*, especialmente en la bondad del Análisis de los antiguos y del Álgebra de los modernos, cuando escribe:

> **«[...] *Los antiguos geómetras se han servido de cierto Análisis, que extendían a la resolución de todos los problemas, si bien privaron de él a la posteridad. Y ahora florece cierta clase de Aritmética que llaman Álgebra, para realizar sobre los números lo que los antiguos hacían sobre las figuras*» (RIV.AT.X.373).**
>
> **«*Ha habido, finalmente, algunos hombres de gran talento que se han esforzado en este siglo por resucitarla; pues aquel arte no parece ser otra cosa, que lo que con nombre extranjero llaman Álgebra, con tal que pueda zafarse de las múltiples cifras e inexplicables figuras de que está recargado a fin de que no falte ya aquella claridad y facilidad suma que suponemos debe haber en la verdadera Mathesis*» (RIV.AT.X.377).**

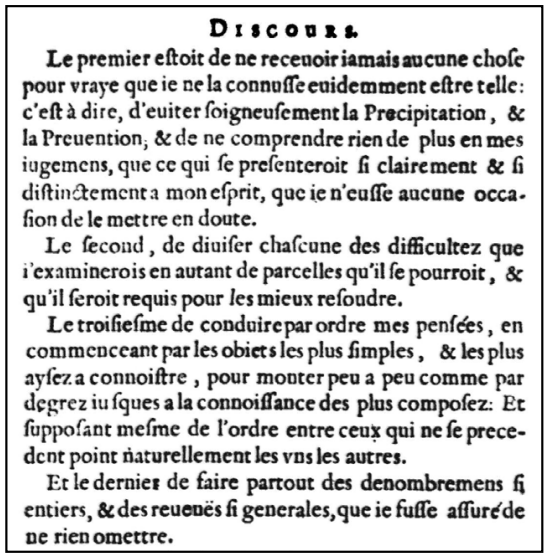
Todas estas cuestiones e inquietudes que Descartes refleja en *El Discurso del Método* y en las *Regulae*, con un lenguaje bello y claro, son esenciales de tener en cuenta para entender como se había ido fraguando en su mente adolescente no sólo el origen de *La Geometría* sino la idea de la sabiduría universal.

Descartes ilustra la introducción de los célebres cuatro preceptos o reglas –*evidencia, análisis, síntesis* y *verificación*– de *El Discurso del Método*, que aplicará cuidadosamente en toda especulación filosófica y matemática, con estas palabras(DM.AT,VI,18-19):

> «*Y como la multitud de leyes sirve a menudo de disculpa a los vicios, siendo un Estado mucho mejor regido cuando hay pocas pero muy estrictamente observadas, así también, en lugar del gran número de preceptos que encierra la lógica, creí que me bastarían los cuatro siguientes, siempre que tomara la firme y constante resolución de no dejar de observarlos ni una sola vez*.»

El primer precepto es el criterio de evidencia, que determina los primeros conocimientos y verdades, las ideas «*claras y distintas*» de los *Principia philosophiae* de 1644, siendo el «*cogito, ergo sum*» la primera verdad indudable y el punto de arranque de toda su Filosofía. Los otros tres preceptos corresponden, a la primera y segunda parte de la Regla V y a la Regla VII de *Las Regulae*. De acuerdo con el segundo precepto –*Regla del Análisis*– hay que dividir las dificultades hasta descubrir los elementos más simples, que se aprehenden por intuición. Según el tercero –*Regla de la Síntesis*–, debemos partir de tales objetos simples y ascender por deducción, poco a poco, hasta los más complejos. Ambos preceptos encierran el núcleo del método cartesiano. Por fin, hay que aplicar el cuarto precepto para hacer una revisión final, examinar cuidadosamente la cadena deductiva, ordenar y enumerar esos elementos simples de modo que estemos seguros de no omitir nada.

LAS CUATRO REGLAS DE *EL DISCURSO DEL MÉTODO*



Discours.

Le premier eſtoit de ne receuoir iamais aucune choſe pour vraye que ie ne la connuſſe euidemment eſtre telle: c'eſt à dire, d'euiter ſoigneuſement la Precipitation, & la Preuention; & de ne comprendre rien de plus en mes iugemens, que ce qui ſe preſenteroit ſi clairement & ſi diſtinctement a mon eſprit, que ie n'euſſe aucune occaſion de le mettre en doute.

Le ſecond, de diuiſer chaſcune des difficultez que i'examinerois en autant de parcelles qu'il ſe pourroit, & qu'il ſeroit requis pour les mieux reſoudre.

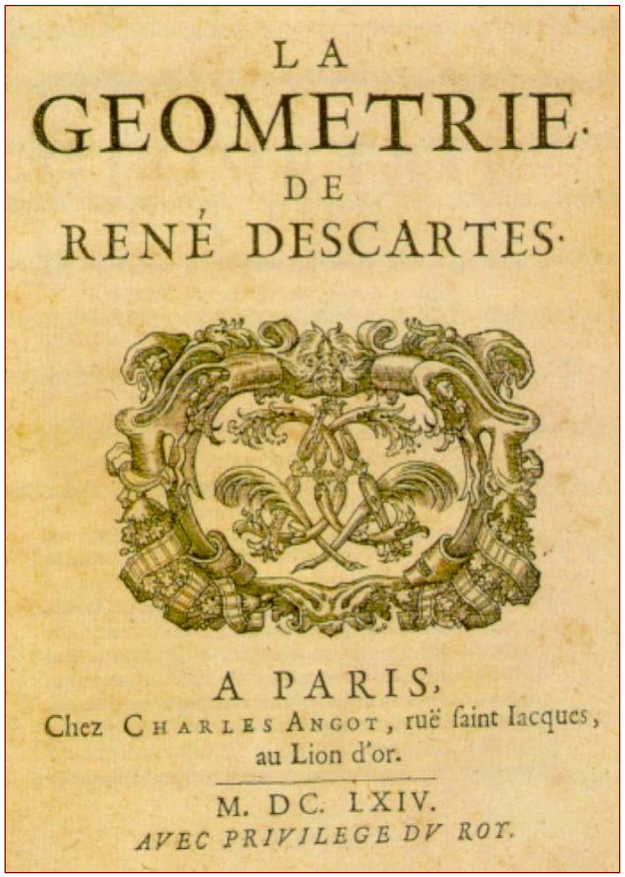
Le troiſieſme de conduire par ordre mes penſées, en commenceant par les obiets les plus ſimples, & les plus ayſez a connoiſtre, pour monter peu a peu comme par degrez iuſques a la connoiſſance des plus compoſez: Et ſuppoſant meſme de l'ordre entre ceux qui ne ſe precedent point naturellement les vns les autres.

Et le dernier de faire partout des denombremens ſi entiers, & des reuenës ſi generales, que ie fuſſe aſſuré de ne rien omettre.

Las cuatro reglas de *El Discurso del Método* en la primera edición de 1637:

1. **Evidencia: no aceptar como verdadero lo que no ofrezca plena evidencia, evitando la precipitación y la prevención.**
2. **Análisis: dividir las dificultades en tantos elementos como sea necesario para resolverlas completamente.**
3. **Síntesis: Llevar los pensamientos en orden, procediendo de lo simple a lo complejo.**
4. **Verificación: hacer suficientes enumeraciones y revisiones para tener la seguridad de no omitir nada.**

LA GEOMETRÍA DE DESCARTES



LA
GEOMETRIE.
DE
RENÉ DESCARTES.

A PARIS,
Chez CHARLES ANGOT, ruë ſaint Iacques,
au Lion d'or.

M. DC. LXIV.
AVEC PRIVILEGE DV ROY.

***La Geometría* de Descartes, edición separada de *El Discurso del Método* (París, 1664).**

***La Geometría* de Descartes es considerada con gran unanimidad como una de las obras fundamentales del pensamiento geométrico a lo largo de toda la Historia de la Matemática. Mediante el uso del Álgebra como herramienta algorítmica esencial, Descartes da una nueva lectura a la Geometría de los griegos, que supera sus limitaciones y trasciende sus conquistas geométricas a base de elaborar un magnífico instrumento de ataque de los problemas geométricos antiguos y modernos que libera a la Geometría de la dependencia y sometimiento a la estructura geométrica de la figura y su representación espacial y propone una forma de solución de los problemas basada en la aplicación del Análisis mediante la actuación del Álgebra, que supone el problema resuelto y establece una ordenada dependencia entre lo conocido y lo desconocido, hasta hallar el resultado buscado, de modo que las reglas del método cartesiano adquieren el sentido matemático de normas para la solución de los problemas geométricos mediante ecuaciones.**

De acuerdo con la idea de Descartes acerca de la Matemática como fundamento de la sabiduría universal, y en particular como base racional de todas las ciencias, *La Geometría* de Descartes perfecciona de forma muy notable el Álgebra esbozada en el Libro II de las *Regulae* al establecer el Análisis Algebraico no sólo como un instrumento que aplicado a la Geometría creará la Geometría Analítica sino como algo mucho más universal todavía, el lenguaje de expresión y por tanto la clave de todas las ciencias.

La percepción de Descartes sobre el eco científico de *La Geometría*

La Geometría de Descartes no puede entenderse de forma aislada ya que forma parte indisoluble del proyecto metodológico general de alcanzar la unidad de la Ciencia que Descartes intenta fijar en las *Regulae* y en *El Discurso del Método*. Descartes se propone con *El Discurso del Método* y los tres ensayos que lo acompañan, demostrar que ha alcanzado un nuevo método de especulación sobre la verdad científica, mejor que todo método anterior y que precisamente *La Geometría* demuestra este aserto. En este sentido escribe en una carta al Padre Mersenne de diciembre de 1637 (AT,I,478):

> «[...] *Con La Dióptrica y Los Meteoros he querido únicamente convencer de que mi método es mejor que el ordinario y creo que lo he demostrado con mi Geometría, al resolver en las primeras páginas una cuestión que, según Pappus, no había podido resolver ningún geómetra de la antigüedad* [...]»

Ya cerca de la fecha de publicación de *El Discurso del Método* y *La Geometría*, en marzo de 1637, Descartes escribe otra carta al Padre Mersenne para comunicarle el título y el contenido de la magna obra:

> «*Proyecto de una Ciencia universal que pueda elevar nuestra naturaleza a su mas alto grado de perfección. Además, La Dióptrica, los Meteoros y la Geometría, donde las más curiosas materias que el autor haya podido elegir, para dar prueba de la Ciencia universal que él autor propone, son explicadas de tal manera, que aun aquellos que no han estudiado puedan entenderlos.*»

Descartes elaboró *La Geometría* como un ejemplo de su método y fue uno de los gérmenes del mismo y aunque la labor matemática de Descartes, en cierto modo, no fue mas que un episodio en su tarea como filósofo, la Matemática era para él la base racional de su pensamiento y estaba convencido de que en el campo de las Matemáticas *La Geometría* contenía todo el elenco de conocimientos en esa área del saber. Así se lo había propuesto desde los primeros planteamientos de su magna empresa intelectual. En una carta a su amigo Beeckman del 26/3/1619, Descartes se expresa (AT.X,157):

> « [...]. *Y para no ocultaros nada de lo que es motivo de mi trabajo quisiera publicar no un Ars Brevis, como Lulio, sino una ciencia toda nueva, que permitiera resolver en general todos los problemas que pudieran presentarse.* [...] *Porque ciertos problemas pueden ser resueltos con líneas rectas o círculos, pero otros requieren otras líneas que puedan originarse por el movimiento continuo,* [...] *Finalmente otros problemas sólo pueden resolverse mediante líneas curvas engendradas por movimientos diferentes y no continuos. Espero poder demostrar qué problemas se resuelven de una manera y cuáles de otra, por lo cual no quedará casi nada por resolver en Geometría. ¡Que proyecto tan ambicioso! ¡Apenas concebible! Pero en el oscuro caos de esta ciencia, he podido vislumbrar yo no sé que luz, gracias a la cual las mas espesas tinieblas podrán disiparse.*»

La impresión del volumen, tras la concesión del privilegio real, se termina el 8 de junio de 1637, en Leyden, en la imprenta de Jan Maire, tirándose 3000 ejemplares en dos ediciones. Con tal de conocer mejor las opiniones y las críticas, Descartes decidió que en la portada no figurara el nombre del autor. Aun cuando el título fue simplificado quedó en la forma:

"*DISCOURS DE LA METHODE pour bien conduire sa raison et chercher la verite dans les sciences. Plus LA DIOPTRIQUE, LES MÉTÉORES ET LA GÉOMÉTRIE, qui sont essais de cette METHODE*"

"*DISCURSO DEL MÉTODO para conducir bien la razón y buscar la verdad en las ciencias. Además LA DIÓPTRICA, LOS METEOROS y LA GEOMETRÍA que son ensayos de este MÉTODO*".

A lo largo de la correspondencia con el Padre Mersenne Descartes asegura, una y otra vez, que *La Geometría* trata de dar un procedimiento general para resolver todos los problemas que no lo habían sido jamás, por ejemplo el *Problema de Apolonio y* el famoso *Problema de Pappus* que en su formulación general se había resistido a lo largo de toda la historia y que llegó a ser la prueba de fuego de la eficiencia del método cartesiano en Geometría.

Con estos antecedentes en la gestación de su magnífica obra filosófica y matemática, es fácil comprender la importancia que Descartes concede a *La Geometría*, lo que se puede calibrar por algunos de sus escritos de lo que son buena muestra los párrafos siguientes entresacados de su correspondencia y de su obra:

En la carta al Padre Mersenne de diciembre de 1637 Descartes escribe (AT,I, 478):

> «*Además, lo que he escrito en el segundo libro sobre la naturaleza y las propiedades de las líneas curvas y sobre el método para estudiarlas, está tan lejos de la geometría ordinaria, como lo está la retórica de Cicerón del abc de los chiquillos ...*
>
> *Ante la sugerencia de que lo que he escrito puede haber sido sacado fácilmente de Vieta, la realidad es que mi tratado es precisamente difícil de comprender porque he intentado no introducir en él nada de lo ya conocido por Vieta o por cualquier otro* [...] *Comienzo las reglas de mi álgebra con lo que Vieta escribe al final de su libro De emendatione aequationum* [...], e*s decir, que comienzo donde él abandonó.*»

En otra carta al Padre Mersenne de 31/3/1638:

> «*No describo todos los casos posibles, sino que, como los arquitectos que sólo indican lo que se debe hacer, dejando el trabajo manual a los albañiles y carpinteros,* [...] .*He prescindido en mi Geometría de muchas cosas que puedan servir para facilitar la práctica, y lo he hecho deliberadamente,* [...] *Había previsto que ciertas gentes, que se vanaglorian de saberlo todo, no hubieran dejado de decir que yo no había escrito nada que ellos no supieran, si lo hubiera hecho de forma mas inteligible.*»

En una carta a Alphonse Pollot:

> «*Es un gran honor para mi que os hayáis tomado la molestia de examinar mi Geometría, os guardo uno de los seis ejemplares que destino para los seis primeros que me parezcan que la entienden.*»

En una carta al Padre Mersenne de mayo de 1637:

> «*Tengo que decirle que honestamente creo que hay muy poca gente que pueda entenderla.*»

En una carta a Plempius de 3 de octubre de 1637 (AT,I,409):

> «*La Geometría tendrá un pequeño número de lectores, pues deben ser personas que no solamente estén al corriente de todo lo que se sabe en Geometría y Álgebra, sino que deben ser además laboriosos, ingeniosos y atentos.*»

Empieza el *Libro Primero* de *La Geometría* con un auténtico desafío geométrico en el que anuncia las coordenadas (G.AT,VI, 369):

> «*Todos los problemas de Geometría pueden reducirse fácilmente a términos tales, que no es necesario conocer de antemano más que la longitud de algunas líneas rectas para construirlos.*»

En el mismo Libro I de *La Geometría*, tras la construcción geométrica de la solución de las ecuaciones, inmediatamente antes del ataque al problema de Pappus, Descartes escribe:

«[...] *Se pueden construir todos los problemas de la geometría ordinaria sin hacer más que lo poco que está comprendido en las cuatro figuras que he explicado.*»

En la aplicación del *Análisis y* la *Síntesis* al planteamiento y resolución de las ecuaciones, Descartes indica (G.AT,VI, 374):

«[...] *Pero no me detengo a explicar esto con más detalle para no privar a cada uno del placer de aprenderlo por sí mismo, ni impedir el cultivo útil del propio espíritu ejercitándolo, que es, a mi parecer, la principal utilidad que puede obtenerse de esta ciencia*, [...].»

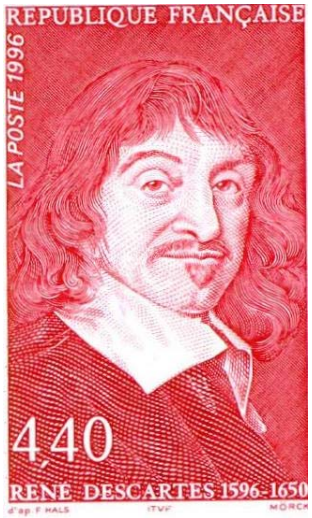
En el problema de la determinación de las rectas normales a una curva, Descartes enfatiza (G.AT,VI, 412-413):

«*Para encontrar todas las propiedades de las líneas curvas basta con saber la relación que tienen todos sus puntos con los de las líneas rectas,* [...] *y conocer la manera de trazar otras líneas que las corten en todos esos puntos en ángulo recto.* [...] *Y me atrevo a decir que éste es el problema más útil y más general no sólo que yo conozca, sino aun que yo haya anhelado jamás conocer en Geometría.*»

La *Geometría* termina con una frase antológica de las que hacen época (G.AT,VI, 485):

«[...] *Y yo espero que nuestros descendientes me estarán agradecidos no sólo por las cosas que aquí he explicado, sino también por aquellas que he omitido voluntariamente a fin de dejarles el placer de descubrirlas.*»

IMÁGENES DE DESCARTES EN LOS SELLOS DE CORREOS



1. **Emitido en Francia el 9 de junio de 1937, en conmemoración del tercer centenario de la publicación de *El Discours sur la Méthode.***
2. **Emitido en Mónaco en el 400 aniversario del nacimiento de Descartes.**
3. **Emitido en Francia en el 400 aniversario del nacimiento de Descartes.**

El contenido de *La Geometría*

La Geometría se compone de tres libros bien diferenciados y a la vez muy entrelazados; tiene en la edición original 120 páginas, con 48 figuras de las que son diferentes 30.

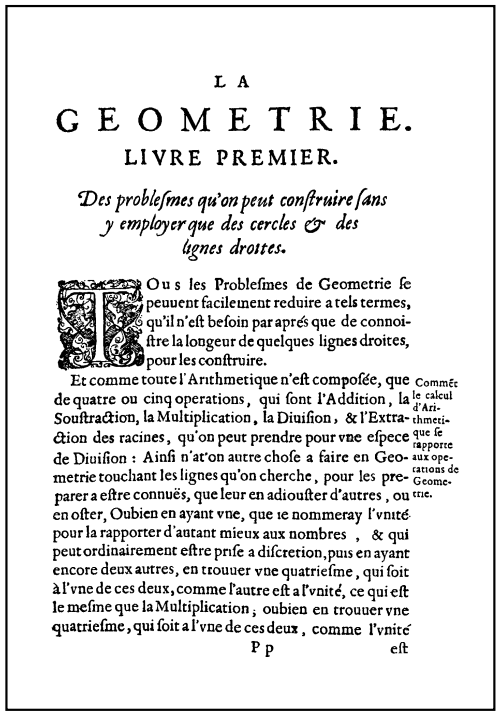
El *Libro Primero* de *La Geometría* trata «*De los Problemas que pueden construirse sin emplear más que círculos y líneas rectas*».

En este libro Descartes fija (basándose siempre en *El Discurso del Método*) la metodología cartesiana que aplicará a la traducción algebraica de los problemas geométricos clásicos, de modo que el libro contiene el núcleo de toda la formulación cartesiana de *La Geometría*, siempre íntimamente ligada al método.

Empieza el libro con una auténtica declaración de principios (G.AT,VI, 369):

«*Todos los problemas de Geometría pueden reducirse fácilmente a términos tales, que no es necesario conocer de antemano más que la longitud de algunas líneas rectas para construirlos.*»

Así pues, como las líneas rectas son lo que se nos presenta de la forma *más clara y distinta* en el campo de la Geometría, para resolver problemas geométricos, partiremos de ciertas líneas rectas –en realidad de algunos segmentos rectilíneos–; pero como un problema geométrico sólo está completamente resuelto –es decir, geométricamente resuelto– cuando se ha construido la solución, es preciso dar ésta en términos de segmentos que se deben construir.



LA
GEOMETRIE.
LIVRE PREMIER.

Des problesmes qu'on peut construire sans y employer que des cercles & des lignes droites.

TOus les Problesmes de Geometrie se peuuent facilement reduire a tels termes, qu'il n'eſt beſoin par aprés que de connoiſtre la longeur de quelques lignes droites, pour les conſtruire.

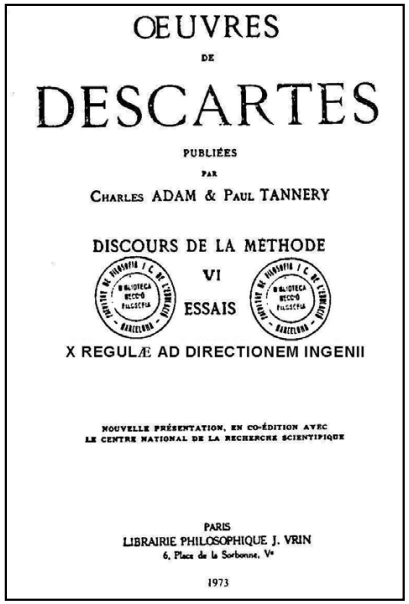
Comme le calcul d'Arithmetique ſe rapporte aux operations de Geometrie.

Et comme toute l'Arithmetique n'eſt compoſée, que de quatre ou cinq operations, qui ſont l'Addition, la Souſtraction, la Multiplication, la Diuiſion, & l'Extraction des racines, qu'on peut prendre pour vne eſpece de Diuiſion : Ainſi n'at'on autre choſe a faire en Geometrie touchant les lignes qu'on cherche, pour les preparer a eſtre connuës, que leur en adiouſter d'autres, ou en oſter, Oubien en ayant vne, que ie nommeray l'vnité pour la rapporter d'autant mieux aux nombres, & qui peut ordinairement eſtre priſe a diſcretion, puis en ayant encore deux autres, en trouuer vne quatrieſme, qui ſoit à l'vne de ces deux, comme l'autre eſt a l'vnité, ce qui eſt le meſme que la Multiplication, oubien en trouuer vne quatrieſme, qui ſoit a l'vne de ces deux, comme l'vnité

Pp eſt

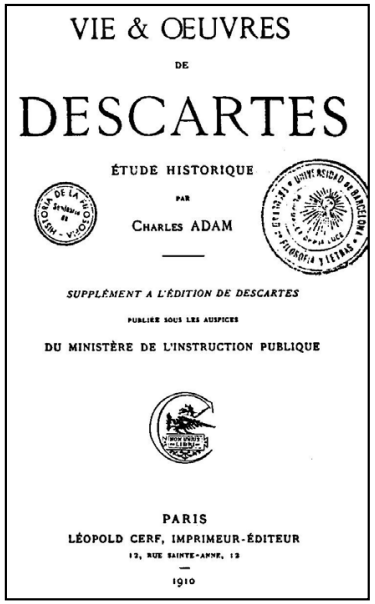
Primera página de la edición de 1637 de *La Geometría* de Descartes.

LAS PRINCIPALES REFERENCIAS SOBRE LA OBRA DE DESCARTES



Oeuvres de Descartes. **Publiées par C.Adam et P.Tannery. 12 volúmenes. Librairie philosophique J.Vrin. París. 1964-74.**

El Volumen VI contiene *El Discours de la Méthode, La Géométrie* **y los otros Ensayos. El Volumen X contiene las** *Regulae ad directionem ingenii* **entre otros tratados.**

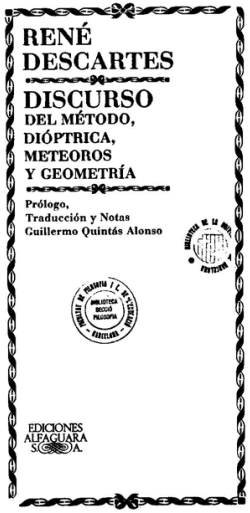


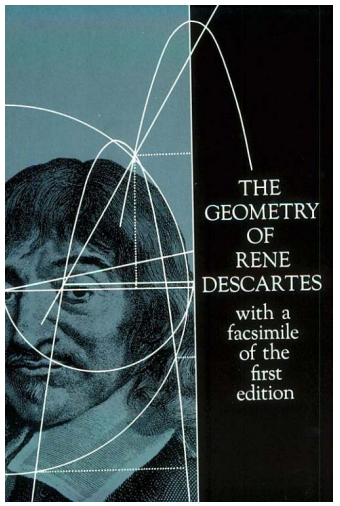
Vie et Oeuvres de Descartes. **Étude historique par C.Adam. París, 1910.**

Es una obra muy completa sobre la vida, viajes, polémicas y escritos de Descartes: *Metaphysique, Le Monde, Dioptrique, Géométrie, Discours de la Méthode, Polémiques, Méditations, Principes de la Philosophie, Passions de l'Âme, ...*

El primer punto consistirá en advertir que las operaciones aritméticas elementales entre segmentos producen siempre un nuevo segmento, por eso en los primeros capítulos Descartes expone los procedimientos ya conocidos para construir geométricamente las operaciones de la Aritmética: omite las construcciones de la suma y diferencia de segmentos y construye la multiplicación, la división y la extracción de raíces cuadradas, a base de introducir el concepto de segmento unidad. Así pues, Descartes pone de manifiesto que el producto de dos o de tres segmentos es otro segmento, así como que el cociente de dos segmentos también es otro segmento. Esta interpretación geométrico-algebraica de las operaciones aritméticas marca un hito en la Historia de la Matemática porque, por una parte soslaya la limitación pitagórica que la inconmensurabilidad había impuesto a la Geometría griega –la imposibilidad de asignar números a las figuras geométricas ante el fantasma de lo inconmensurable–, y por otra, permite romper con el problema de la homogeneidad dimensional, que había sido, sin duda hasta entonces, otra de las grandes limitaciones de la aplicación del Álgebra a la Geometría. Desde luego así había sido en la Geometría griega, pero incluso en la época de Descartes el producto de dos segmentos era un rectángulo, y el producto de tres segmentos un paralelepípedo, por tanto el producto de más de tres segmentos no tenía sentido y en consecuencia no se llevaba a efecto.

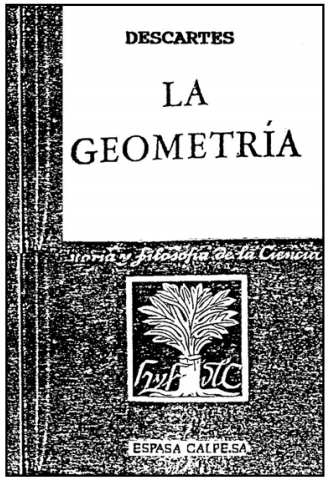
ALGUNAS EDICIONES DE *LA GEOMETRÍA* DE DESCARTES

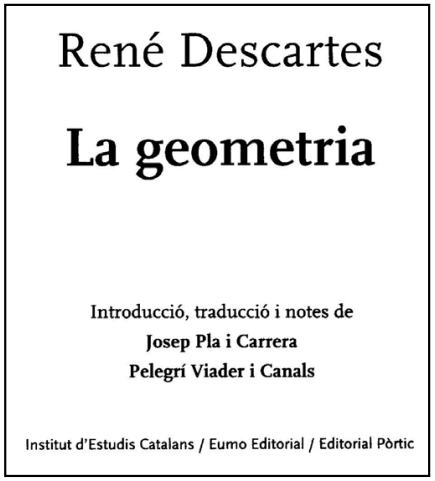




Ediciones de *La Geometría de Descartes utilizadas en este trabajo*:

1. **DESCARTES.** *La Geometría*. **Espasa-Calpe, Buenos Aires. 1947.**
2. **DESCARTES.** *The Geometry*. **Dover, New York, 1954.**
3. **DESCARTES.** *La Geometría*. **Alfaguara, Madrid, 1986.**
4. **DESCARTES.** *La Geometria*. **Institut d'Estudis Catalans, Barcelona, 1999.**





Tras la construcción geométrica de las operaciones, Descartes pasa a mostrar «*Cómo se llega a las ecuaciones que sirven para resolver los problemas*» y «*Cómo se resuelven*», lo que aplicará a dar solución al *Problema de Pappus* que trata en la forma más general, creando el método analítico de solución y discusión de los problemas matemáticos. Este problema había sido resuelto por Descartes en 1632 cuando Golius se lo propuso para que aplicara sobre él sus nuevos métodos, convirtiéndose en una auténtica piedra de toque que pone a prueba el nuevo método cartesiano, llegando en su resolución mucho más lejos que los geómetras griegos. El enunciado del problema está dado en latín («*cito la versión latina y no la griega con el fin de que pueda ser entendido con más facilidad*», escribe Descartes) y es la reproducción de la traducción de Commandino de *La Colección Matemática* de Pappus de 1589.

El *Libro Segundo* de *La Geometría* titulado «*De la naturaleza de las líneas curvas*» consta de cuatro partes bien diferenciadas:

a) *La naturaleza geométrica de las líneas curvas*, vinculada sobre todo a dos cuestiones íntimamente ligadas: los compases cartesianos y la teoría de la proporción continua. Mientras en el Libro I, sin olvidar los lugares geométricos, Descartes centra más la atención sobre puntos individualizados, en el Libro II se proyecta sobre el objeto geométrico *Curva*. Descartes mantiene la división clásica griega de los problemas geométricos en *planos, sólidos y lineales* (que se resuelven con ecuaciones de segundo, tercer y cuarto o mayor grado, respectivamente) y demuestra que los problemas planos se construyen con rectas y circunferencias, los sólidos con secciones cónicas y el resto con líneas más complejas, llamadas por los antiguos *curvas mecánicas*, aunque más correcto sería llamarlas *curvas geométricas*. Descartes tiene el propósito de poner un poco de orden en el estudio de las curvas de la Geometría de los griegos, que según él era un caos completo, secuela de la limitación platónica de la regla y el compás, al no ser capaces de distinguir las diversas clases de curvas por no poder dilucidar la naturaleza de las mismas. Esto es precisamente lo que se propone Descartes, a base de establecer qué curvas son las que se pueden admitir en Geometría.

b) *El Problema general de Pappus*, ahora tratado con las herramientas precisas para poder clasificar las diversas soluciones de los diversos planteamientos del mismo. Con el método cartesiano el clásico *Problema de Pappus* queda completamente resuelto.

c) *La construcción y propiedades de tangentes y normales* a una curva geométrica. Una vez concebida y definida, de forma clara y distinta, *la naturaleza geométrica de las líneas curvas*, Descartes introduce uno de los principios básicos de su método: «p*ara encontrar todas las propiedades de las líneas curvas basta con saber la relación que tienen todos sus puntos con los de las líneas rectas*, [...]», y establece cómo se puede utilizar la expresión algebraica (la ecuación de las curvas) para determinar los elementos geométricos más notables de las curvas (diámetros, ejes, centros, etc.) y, en particular, las normales, líneas cuya consideración y utilidad deriva de los problemas de la reflexión de la luz sobre las superficies curvas, y que literalmente es considerado por Descartes como el más importante problema geométrico que pueda ser concebido.

d) Finalmente Descartes estudia *los Óvalos* como curvas especiales que responden a consideraciones fijadas de las tangentes o normales. Descartes introduce cuatro amplias familias de curvas nuevas, de las que las cónicas son casos particulares.

El *Libro Tercero* de *La Geometría* trata «*De la construcción de los problemas que son sólidos o más que sólidos*» mediante el estudio de la resolución de ecuaciones, discusión de sus raíces, y relaciones entre los coeficientes. Descartes pretende ofrecer un método de resolución de cualquier ecuación algebraica. En realidad sólo llega a la resolución geométrica de un determinado tipo de ecuaciones de quinto y sexto grado, pero su método quiere ser general. Muestra que una ecuación puede tener tantas raíces como dimensiones tiene el grado (existe «*la posibilidad de imaginar tantas raíces como el grado del polinomio*»), da luego su famosa regla de los signos y adelanta el *Teorema de Ruffini* del factor. Descartes introduce como transformaciones de la variable las traslaciones y las

homotecias (no es el primero, Vieta y Harriot, respectivamente se habían adelantado a ellas), mediante las cuales consigue (como había hecho Vieta con anterioridad), reducir el segundo término y cuando es posible, racionalizar.

Después de fundamentar las operaciones y propiedades algebraicas necesarias, Descartes introduce el simple criterio de divisibilidad sobre el término independiente de la ecuación polinómica (como condición necesaria aunque no suficiente) para obtención de raíces enteras y a partir de aquí ir reduciendo el grado de la ecuación mediante el algoritmo de la división. La existencia de una raíz entera permite caracterizar el problema geométrico inicial que conduce a la ecuación polinómica en cuestión como un *problema plano*, siempre y cuando la raíz sea adecuada para el problema geométrico. Por ejemplo en el caso de ecuaciones cúbicas con raíz entera, el problema es *plano*, ya que tras efectuar la división, obtenemos una ecuación cuadrática, cuyas soluciones, si existen, se obtienen con regla y compás, de acuerdo con lo establecido en el libro I. En caso de ausencia de raíces enteras, se puede afirmar, sin duda alguna, que el problema geométrico inicial es sólido y Descartes establece entonces que sólo hay dos formas de resolver la ecuación mediante la intervención de los clásicos problemas: la *duplicación del cubo* y la *trisección del ángulo*. De hecho, todo problema cúbico es equivalente a uno de estos dos problemas geométricos (G.AT,VI, 471-475). He aquí un nuevo y magnífico éxito del método cartesiano aplicado a la Geometría: las ecuaciones del Álgebra son el reflejo lingüístico de los problemas de la Geometría.

Finalmente, Descartes plantea la siguiente cuestión: si en el caso de ecuaciones cuadráticas, las soluciones venían dadas por segmentos construibles con regla y compás (según el Libro I), es decir, el problema geométrico de donde procedían era *plano, ¿es posible construir las soluciones de las ecuaciones cúbicas y cuárticas con solución real*? Descartes resuelve de forma contundente el problema mediante la intersección de una circunferencia y una parábola convenientemente determinadas por la ecuación que se quiere resolver (G. AT,VI, 464-471).

DESCARTES EN *L'ENCYCLOPÉDIE* DE D'ALEMBERT Y DIDEROT

D'Alembert escribe sobre Descartes en el *Discours Préliminaire de l'Encyclopédie* (Orbis, Barcelona, 1984, pp.84-85):

«Descartes tenía todo lo que se necesitaba para cambiar la faz de la Filosofía: una fuerte imaginación, un espíritu muy consecuente, unos conocimientos procedentes de su interioridad más que de los libros, un gran coraje para combatir los prejuicios y una total independencia intelectual.

Se puede considerar a Descartes como geómetra o como filósofo. Las Matemáticas, a las que parece haberle prestado poca atención, hoy son sin embargo la parte más sólida y menos discutida de su gloria. El Álgebra creada, de alguna forma, por los italianos, prodigiosamente desarrollada por nuestro ilustre Vieta, ha recibido entre las manos de Descartes nuevos enriquecimientos. Uno de los más considerables es su Método de las indeterminadas, artificio muy ingenioso y muy sutil, que luego se ha podido aplicar a un gran número de investigaciones.

Pero lo que ha inmortalizado el nombre de este gran hombre es la aplicación que hizo del Álgebra a la Geometría, una idea de las más vastas y felices que el intelecto humano haya concebido jamás, y que será siempre la llave de los más profundos descubrimientos no solamente en la geometría sublime, sino en todas las ciencias físico-matemáticas.»

La construcción geométrico-algebraica de las operaciones aritméticas

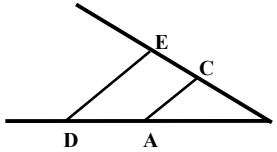
Veamos cómo dados un segmento unidad y dos segmentos a y b, Descartes construye, mediante circunferencias y rectas, el producto a·b, el cociente a/b y la raíz cuadrada $\sqrt{a}$, que resultan ser segmentos, porque se obtienen como una cuarta proporcional o como una media proporcional, de modo que establece que las operaciones aritméticas elementales de segmentos dan segmentos que se construyen mediante la regla y el compás.

Cómo el cálculo de la aritmética se relaciona con las operaciones de geometría.

> «*Y así como la aritmética no comprende más que cuatro o cinco operaciones, que son la adición, la sustracción, la multiplicación, la división y la extracción de raíces, que pueden tomarse como una especie de división, así también no hay otra cosa que hacer en geometría, respecto a las líneas que se buscan, para prepararlas a ser conocidas, que agregarles o quitarles otras, o bien, teniendo una, que llamaré la unidad para relacionarla lo más posible con los números, y que ordinariamente puede ser tomada a discreción, y teniendo luego otras dos, encontrar una cuarta que sea a una de esas dos, como la otra es a la unidad, que es lo mismo que la multiplicación; o bien encontrar una cuarta que sea a una de esas dos como la unidad es a la otra, lo que es lo mismo que la división; o, en fin, encontrar una, dos, o varias medias proporcionales entre la unidad y alguna otra línea, lo que es lo mismo que extraer la raíz cuadrada, o cúbica, etc. Y yo no temeré introducir estos términos de aritmética en la geometría, a fin de hacerme más inteligible*». (G.AT,VI, 369-371).

La limitación operacional que trajo la inconmensurabilidad impidió en la Geometría griega asignar a las figuras geométricas números que midieran sus longitudes, áreas y volúmenes. En este ámbito, la raíz cuadrada, por ejemplo, equivalía al problema geométrico de cuadrar un rectángulo, es decir, hallar el lado de un cuadrado equivalente a un rectángulo dado. Descartes rompe aquí también con el pasado y abre una nueva brecha al asignar longitudes a los segmentos (empezando por adoptar un segmento unidad del que había hablado en la Regla XVI, RXVI.AT.X.449), de modo que mientras en la Aritmética las únicas raíces cuadradas exactas que pueden obtenerse son las de los cuadrados perfectos, en Geometría, a partir de Descartes, puede hallarse un segmento que represente exactamente la raíz cuadrada de otro segmento dado, incluso cuando este segmento no sea conmensurable con la unidad.

Veamos la construcción efectiva de Descartes del producto, el cociente y la raíz cuadrada:



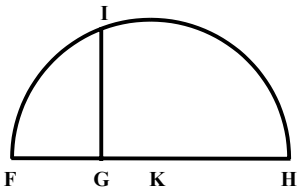
La multiplicación.

Sea, por ejemplo, AB la unidad, y que deba multiplicarse BD por BC; no tengo más que unir los puntos A y C, luego trazar DE paralela a CA, y BE es el producto de esta multiplicación.

Como en muchos problemas de *La Geometría*, Descartes aplica el *Teorema de Tales* (*Euclides*, VI.4) a la semejanza de triángulos. En este caso, como en el siguiente de la división, la semejanza de los triángulos ∆ BAC y ∆ BDE, que determinan BE/BD = BC/BA.

La división

O bien, si deben dividirse BE por BD, habiendo unido los puntos E y D, se traza AC paralela a DE y BC es el resultado de esa división.

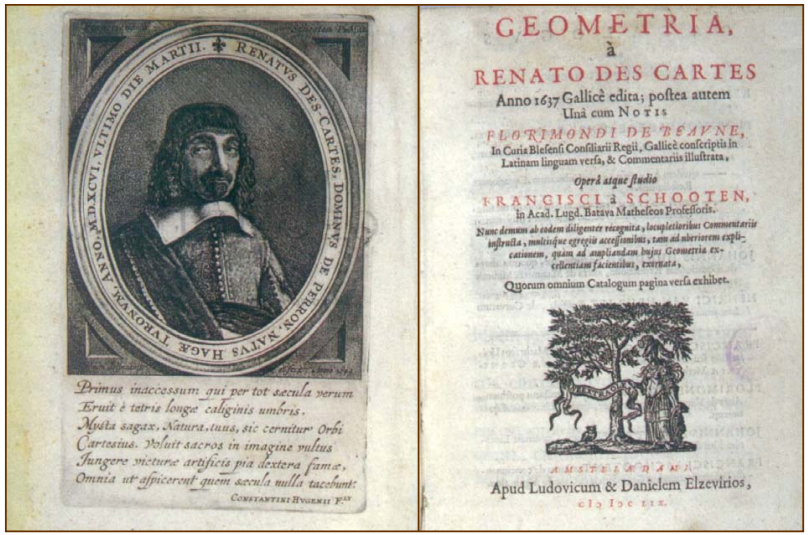


La extracción de la raíz cuadrada

O, si hay que extraer la raíz cuadrada de GH, se le agrega en línea recta FG, que es la unidad y dividiendo FH en dos partes iguales por el punto K, con ese punto como centro se traza el círculo FIH; luego elevando desde el punto G una línea recta, con ángulos rectos sobre FH, hasta I, es GI la raíz buscada. No digo nada aquí de la raíz cúbica, ni de las otras, pues de ellas trataré con más detalle más adelante.

En esta ocasión, Descartes usa el otro *Teorema de Tales* (*Euclides*, III.31) y el *Teorema de la Altura* (*Euclides*, VI.8).

No hay ninguna novedad en la traducción geométrica de las operaciones algebraicas elementales que hace Descartes pensando en la ulterior resolución de ecuaciones. De hecho sabemos que el *Álgebra Geométrica* de los griegos era una forma geométrico-sintética de resolver ecuaciones, y después de los griegos, los matemáticos árabes disponían de algoritmos de resolución de ecuaciones mediante ciertas construcciones geométricas, que traducían las operaciones algebraicas casi en los mismos términos que Descartes. La gran innovación cartesiana estriba en que Descartes las utiliza para resolver problemas geométricos, es decir, para hacer Geometría mediante el Álgebra y no al revés.



La edición de F. van Schooten, de 1659, de *La Geometría* de Descartes, parte de su largo título es:

***Geometria, à Renato Des Cartes : anno 1637 Gallicè edita; postea autem una cum notis / Florimondi de Beavne ... Gallicè conscriptis in Latinam Linguam versa, & commentariis illustrata, operâ atque studio Francisci à Schooten ...* -- Amstelaedami : apud Ludovicum & Danielem Elzevirios, 1659.**

En 1644, el matemático Frans van Schooten, editor de *La Geometría* de Descartes y autor de las figuras, graba la imagen *ad vitum* de un caballero con bigote y barba, circundando el retrato con la siguiente inscripción: Renatus Des Cartes ... Natus Hagae Turorum anno MDXCVI. Señalan los historiadores que esta inscripción es el único documento que atestigua la fecha de nacimiento de Descartes. El filósofo encontró este retrato «*muy bien hecho, aunque la barba y el vestido no se le parecen en nada*».

El retrato acompaña a la edición de van Schooten, de 1659, de *La Geometría* de Descartes.

La notación matemática cartesiana

Al comienzo de su trabajo matemático Descartes hace uso, bajo el influjo de Clavius, de la notación cósica. Pero ya en *Las Reglas para la dirección del espíritu* hay, quizá como secuela de la lectura de Vieta, una primera evolución hacia el simbolismo. En el título de la Regla XVI Descartes pondera el significado y la importancia que tiene una notación sencilla:

> «*En cuanto a las cosas que no requieren la atención presente de la mente, incluso si son necesarias para la conclusión, es mejor designarlas por medio de signos muy breves que por figuras completas: pues así la memoria no podrá fallar, mientras que además el pensamiento no se distraerá en retenerlas, cuando se dedique a deducir otras*» (RXVI.AT.X.454).

Descartes hace alusión al uso de la escritura (RXVI.AT.X.454–455):

> «[...] *Muy acertadamente el arte inventó la escritura, fiados en cuya ayuda nada en absoluto encomendaremos ya a la memoria, sino que, dejando a la fantasía en su totalidad libre para las ideas presentes, escribiremos en el papel cuanto haya de ser retenido*; *y ello por medio de signos muy breves* [...]. *A cuanto haya de ser contemplado para la solución de una dificultad, lo designaremos por medio de un signo único que puede ser formado al capricho de cada cual*.»

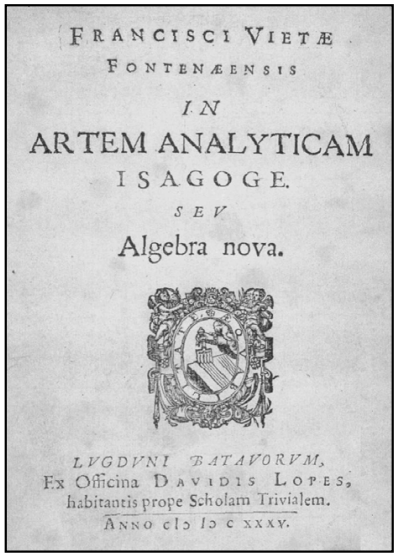
Más adelante, Descartes concreta el simbolismo a adoptar: (RXVI.AT.X.455):

> «*Mas, para mayor facilidad, utilizaremos las letras a,b,c, etc., para expresar las magnitudes ya conocidas, y las letras A,B,C, etc., para las incógnitas; las haremos preceder frecuentemente de los signos numéricos 1,2,3,4, etc., para expresar su multiplicidad, y les agregaremos también el número de sus relaciones que en ellas habrán de entenderse, así si escribo* $2a^3$, *será lo mismo que si dijera el duplo de la magnitud denotada por la letra a, que contiene tres relaciones. Con este artificio, no sólo resumiremos muchas palabras, sino que, mostraremos los términos de la dificultad tan puros y desnudos, que sin omitir nada útil, no se encuentre en ellos nada superfluo, que ocupe inútilmente la capacidad del espíritu, mientras la mente se vea obligada a abarcar a un tiempo muchas cosas.*»

He aquí la introducción de parámetros e incógnitas como ya había hecho Vieta con anterioridad, aunque Descartes no lo menciona sino que habla de su nueva forma de trabajar frente a los tradicionales calculistas (RXVI.AT.X.455–456, 458):

> «A *fin de que todo esto se entienda con mayor claridad, ha de observarse, en primer lugar, que los Calculistas acostumbran a designar cada una de las magnitudes por medio de varias unidades o por medio de algún número, y que nosotros en cambio en este lugar hacemos abstracción de los números mismos no menos que poco antes de las figuras geométricas. Hacemos esto tanto para evitar el tedio de un cálculo largo y superfluo, como principalmente para que las partes del objeto que atañe a la naturaleza de la dificultad permanezcan siempre distintas y no sean envueltas por números inútiles: así, si se busca la base de un triángulo rectángulo cuyos lados son 9 y 12, el Calculista dirá que aquella es* $\sqrt{225}$ *o 15; nosotros, sin embargo en lugar de 9 y 12, pondremos a y b, y encontraremos que la base es* $\sqrt{a^2+b^2}$ *, y aquellas dos partes* a^2 *y* b^2*, que en el número están confusas, permanecerán distintas*.»
>
> «*Todo esto* [diferenciación entre magnitudes conocidas –parámetros– y magnitudes desconocidas –incógnitas–] *lo distinguimos nosotros que buscamos un conocimiento evidente y distinto de las cosas, pero no los Calculistas, que se quedan satisfechos con tal que se les presente el resultado* [numérico] *buscado, aun cuando no se den cuenta de qué modo éste depende de los datos, en lo cual solo, sin embargo, consiste propiamente la ciencia*.»

EL *ARTE ANALÍTICA* DE VIETA


FRANCISCI VIETÆ
FONTENÆENSIS
IN
ARTEM ANALYTICAM
ISAGOGE.
SEV
Algebra nova.
LVGDVNI BATAVORVM,
Ex Officina DAVIDIS LOPES,
habitantis prope Scholam Trivialem.
ANNO CIↃ IↃ C XXXV.

Portada de la edición de 1635*In Artem Analyticam Isagoge* de Vieta. La primera edición es de 1591.

La obra de Vieta *In Artem Analyticam Isagoge* está inspirada profundamente en la obra de Diofanto y Pappus. En ella Vieta fundamenta los principios y las reglas del cálculo algebraico literal.

El *Arte Analítica* de Vieta perfecciona considerablemente el Álgebra sincopada de Diofanto y de los matemáticos árabes y renacentistas, e inicia el cálculo literal del Álgebra simbólica mediante la introducción de los parámetros, lo que le permite obtener la solución general de las ecuaciones mediante fórmulas que expresan las incógnitas en función de los parámetros. Ya que los parámetros no permiten obtener un resultado numérico concreto tras las operaciones combinatorias que conducen a la resolución de una ecuación, sino una solución simbólica, Vieta trasciende la *Logistica numerosa* ordinaria, aplicada al cálculo con números, y alcanza la *Logistica speciosa* que tiene que ver con las especies, entendiendo por éstas cualquier tipo de magnitud, en particular elementos geométricos como ángulos o longitudes. Esto quiere decir que las cantidades simbólicas del *Arte Analítica* al ser interpretadas como magnitudes geométricas y las operaciones simbólicas como procedimientos de construcción geométrica, permiten obtener la solución simbólica de las ecuaciones generales con significado geométrico, de modo que *el Arte Analítica* podía ser aplicado no sólo a los problemas numéricos sino también a problemas geométricos. De esta forma, el *Arte Analítica* de Vieta que tuvo su origen en el *Tesoro del Análisis* de Pappus, revierte sobre éste, de manera que su contenido es traducido al lenguaje simbólico del Arte, es decir, mediante el concurso del Algebra simbólica, Vieta puede reconstruir, en términos algebraicos, el Análisis Geométrico clásico, lo que prepara el terreno para el advenimiento de la Geometría Analítica de Descartes.

Como explica Vieta:

> «*La debilidad del antiguo Análisis residía en que se aplicaba sólo a los números, es decir, era una Logistica numerosa. Pero el Álgebra permite razonar sobre cualquier tipo de magnitud –número, segmento, ángulo, figura,...– de modo que lo que hay que hacer es considerar una Logistica speciosa, aplicable a cualquier especie de cantidad, que se podrá expresar de una manera genérica mediante letras, tanto si es una magnitud desconocida* **[incógnita]** *como conocida* **[parámetro]**, *ya que no hago diferencia entre ellas. Es más, consideraré las magnitudes desconocidas como si se conocieran y operando según las reglas del Arte Analítica, las desconocidas con las conocidas, obtendré aquellas en función de éstas. He aquí el fundamento de la obtención de soluciones generales de los problemas donde los antiguos sólo obtenían soluciones particulares.* »

Por la naturaleza del *Arte Analítica*, el Análisis algebraico-geométrico de Vieta es un estadio intermedio esencial en el camino que arranca del Álgebra Geométrica de los griegos y confluye en la Geometría Analíticas de Descartes.

Aunque nunca reconocerá la paternidad de Vieta en algunas de sus ideas fundamentales, Descartes debió inspirarse en él en la introducción del uso de letras para designar no sólo las cantidades desconocidas –incógnitas o variables– sino incluso las conocidas–parámetros–. Así se aplica un magnífico instrumento que permite obtener la solución general de los problemas mediante fórmulas que expresan las incógnitas en función de los parámetros. Como escribe Vieta: «*He aquí el fundamento de la obtención de soluciones generales donde los antiguos sólo daban soluciones particulares*». En este sentido los calculadores, a los que alude Descartes, siguen siendo antiguos, ya que: «*no se dan cuenta de qué modo el resultado depende de los datos*» (RXVI.AT.X.458). En efecto, cuando no se emplean más que números para designar cantidades conocidas, estos números se confunden y se disipan en el curso de las operaciones, de modo que en la obtención de la solución no queda ninguna traza de la línea seguida en las operaciones. Con la introducción de los parámetros de Vieta y Descartes, al contrario, la cantidad desconocida siendo despejada e igualada a las cantidades conocidas, se dispone de un cuadro general con todas las operaciones que es preciso hacer sobre los datos para obtener la solución. Así es posible construir una Teoría general de ecuaciones, de modo que, por ejemplo, se estudia, no ecuaciones cúbicas, sino *la ecuación cúbica*, es decir, la ecuación general de tercer grado expresaba en una incógnita y cuatro parámetros.

Vieta había dado un gran paso hacia el Álgebra simbólica, pero para las potencias permaneció en la tradición indicando *quad* para el cuadrado y *cub* para el cubo, *aeq* para la igualdad y *in* para el producto. Así la ecuación «$x^3+5bx^2-2cx=d$», Vieta la expresaría como: «A cub + B 5 in A quad – C plano 2 in A aeq D solido», donde los parámetros B,C y D deben ser tales que cada término de la ecuación sea tridimensional, ya que para Vieta, como herencia griega, las operaciones aritméticas están incluidas todavía en un terreno estrictamente geométrico, que mantiene la homogeneidad, siendo cuadrado una magnitud plana y cubo una magnitud espacial. Como se ve, el simbolismo en Vieta no es total, falta aplicar signos para la igualdad, el producto, las potencias, y otros, que de hecho ya existían en su época. De haberlo hecho, Vieta, por ejemplo, podría haber escrito *todas las ecuaciones cuadráticas* de la forma $BA^2+CA+D=0$, donde A es la incógnita y B,C,D, son los parámetros. Así pues, al manejar todavía simplemente abreviaturas el Álgebra de Vieta sigue siendo sincopada. Será Descartes quien introduce en la Regla XVI, como se ha visto, la convención actual para la codificación de los símbolos de incógnitas y potencias, que por primera vez en la Historia de la Matemática serán símbolos artificiales, arbitrarios [«*formado al capricho de cada cual*» (RXVI.AT.X.455)] y no abreviadores. El signo cartesiano, como notación «*no cósica*», no es considerado como imagen del concepto, sino que es una mera apoyatura operacional para captar y manipular dicho concepto.

El convenio establecido es perfeccionado por Descartes en *La Geometría*, donde en lugar de designar por A,B,C, las incógnitas, utiliza las últimas letras minúsculas x,y,z; y en cuanto a las potencias y raíces Descartes establece (G.AT,VI, 371):

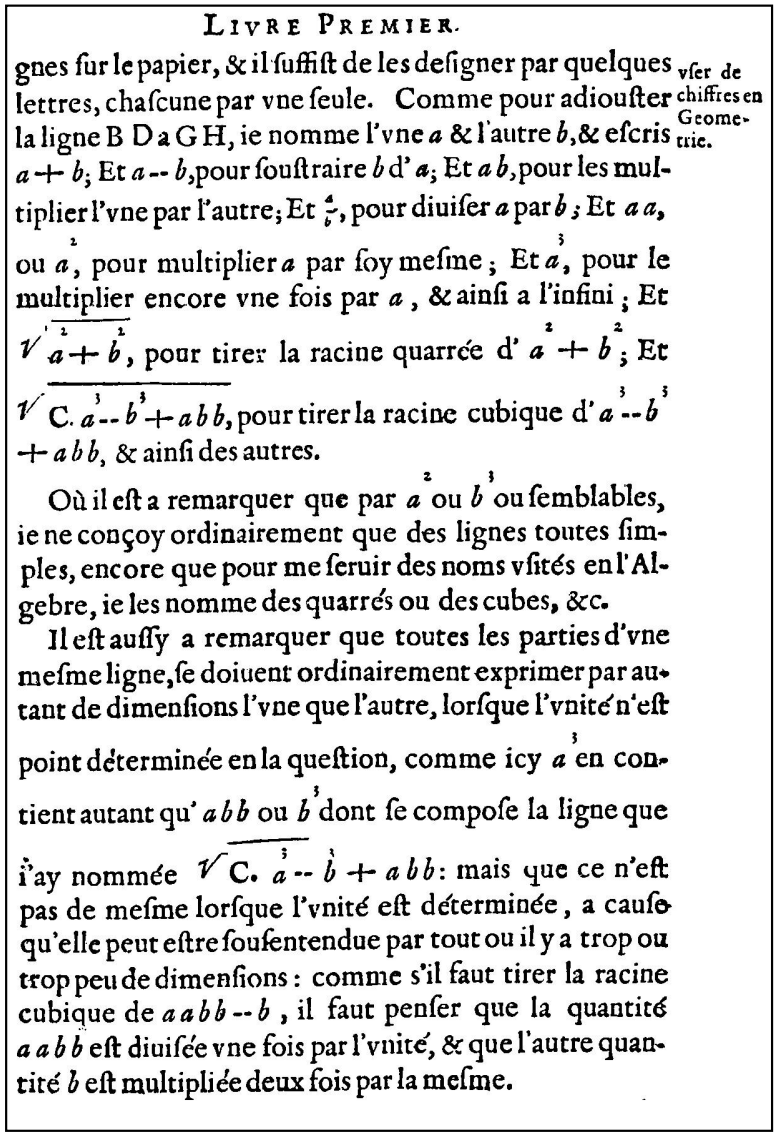
Cómo pueden emplearse letras en geometría.

«*Pero a menudo no hay necesidad de trazar esas líneas sobre el papel y basta con designarlas por ciertas letras, una sola para cada línea. Así, para sumar la línea BD a la GH, designo a la una a y a la otra b y escribo a + b ; y a - b para restar b de a; y ab para multiplicar la una por la otra; y* $\frac{a}{b}$ *para dividir a por b; y aa o* a^2 *para multiplicar a por sí misma; y* a^3 *para multiplicar otra vez por a, y así al infinito; y* $\sqrt{a^2+b^2}$ *para extraer la raíz cuadrada de* a^2+b^2*; y* $\sqrt{C.a^3-b^3+abb}$ *para extraer la raíz cúbica* a^3-b^3+abb *y así otras.*

Es de señalar que para a^2 *o* b^3 *u otras expresiones semejantes, yo no concibo ordinariamente mas que líneas simples, aunque para servirme de los nombres usados en álgebra, los designe por cuadrados, cubos, etc.*

Por último, a fin de no dejar de recordar los nombres de estas líneas, conviene siempre hacer una anotación separada, a medida que se las coloca o se las cambia, escribiendo, por ejemplo, AB ∞ 1, es decir AB igual a 1».

LA NOTACIÓN MATEMÁTICA CARTESIANA



LIVRE PREMIER.

gnes ſur le papier, & il ſuffiſt de les deſigner par quelques lettres, chaſcune par vne ſeule. Comme pour adiouſter la ligne B D a G H, ie nomme l'vne a & l'autre b, & eſcris $a+b$; Et $a-b$, pour ſouſtraire b d'a; Et ab, pour les multiplier l'vne par l'autre; Et $\frac{a}{b}$, pour diuiſer a par b; Et aa, ou a^2, pour multiplier a par ſoy meſme; Et a^3, pour le multiplier encore vne fois par a, & ainſi a l'infini; Et $\sqrt{a^2+b^2}$, pour tirer la racine quarrée d'a^2+b^2; Et $\sqrt{C.a^3-b^3+abb}$, pour tirer la racine cubique d'a^3-b^3+abb, & ainſi des autres. vſer de chiffres en Geometrie.

Où il eſt a remarquer que par a^2 ou b^3 ou ſemblables, ie ne conçoy ordinairement que des lignes toutes ſimples, encore que pour me ſeruir des noms vſités en l'Algebre, ie les nomme des quarrés ou des cubes, &c.

Il eſt auſſy a remarquer que toutes les parties d'vne meſme ligne, ſe doiuent ordinairement exprimer par autant de dimenſions l'vne que l'autre, lorſque l'vnité n'eſt point déterminée en la queſtion, comme icy a^3 en contient autant qu'abb ou b^3 dont ſe compoſe la ligne que i'ay nommée $\sqrt{C.a^3-b^3+abb}$: mais que ce n'eſt pas de meſme lorſque l'vnité eſt déterminée, a cauſe qu'elle peut eſtre ſouſentendue par tout ou il y a trop ou trop peu de dimenſions: comme s'il faut tirer la racine cubique de $aabb-b$, il faut penſer que la quantité $aabb$ eſt diuiſée vne fois par l'vnité, & que l'autre quantité b eſt multipliée deux fois par la meſme.

Tercera página de la edición de 1637 de *La Geometría* de Descartes donde se introduce el simbolismo de la notación cartesiana.

Vemos, pues, que Descartes asigna una letra a cada segmento, que de hecho designa (y mide) su longitud. Además, introduce los exponentes para escribir las potencias; utiliza a^3, a^4, a^5, a^6, etc., para representar las respectivas potencias de a, pero usa indistintamente aa o a^2 para el cuadrado, lo cual tiene su explicación pues mientras que para escribir tanto aa como a^2 se precisan dos signos en las potencias superiores a a^2 hay una gran economía de lenguaje al escribir a^n. Además, vemos que Descartes designa la raíz cúbica $\sqrt[3]{a^3 - b^3 + abb}$ mediante $\sqrt{C.a^3 - b^3 + abb}$ y la igualdad por medio del símbolo ∞ que pudiera provenir de las dos primeras letras de la palabra latina *aequare*.

El penúltimo párrafo tiene una gran trascendencia. Con anterioridad a Descartes, geométricamente sólo tenían sentido las potencias cuadrática a^2 y cúbica a^3, que representaban respectivamente un cuadrado de lado a y un cubo de arista a. El propio Vieta, con su *ley de los homogéneos*, había permanece fiel al espíritu del pasado geométrico de los griegos, gobernado por la *Teoría de las Proporciones*, que había liberado a los antiguos del trauma de la inconmensurabilidad. Afortunadamente Descartes eliminó esta reminiscencia clásica en la Regla XVI de las *Regulae* (RXVI.AT.X.457):

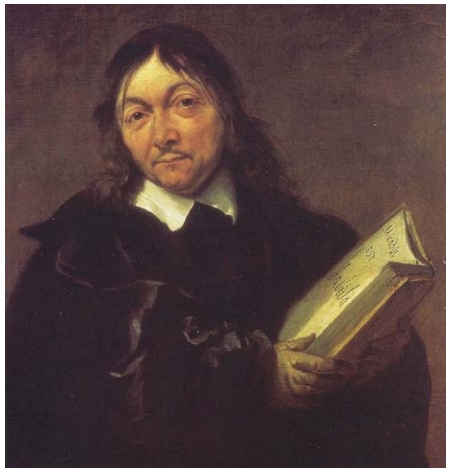
> «*La misma magnitud aunque, aunque sea llamada cubo o bicuadrado, nunca debe ser propuesta a la imaginación* [...] *más que como una línea o como una superficie. Por lo tanto es preciso notar sobre todo que la raíz, el cuadrado, el cubo, etc., no son otra cosa que magnitudes en proporción continua, a la que siempre se supone antepuesta aquella unidad asumida* [...]; *a esta unidad hace referencia inmediatamente la primera proporcional y por medio de una única relación; la segunda, por su parte, por medio de la primera y por lo tanto por medio de dos relaciones; la tercera, mediante la primera y la segunda, y por medio de tres relaciones, etc. Llamaremos, pues, en lo sucesivo, primera proporcional a aquella magnitud que en Álgebra es denominada raíz, segunda proporcional a la que es llamada cuadrado y así las restantes.*»

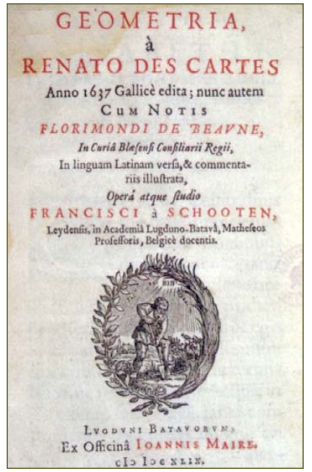
Como en muchas otras cuestiones lo que Descartes aventura en las *Reguale* lo consolida en *El Discurso del Método* o en *La Geometría*, como es el caso. En la notación cartesiana introducida en este tercer epígrafe del Libro I de *La Geometría* hay una clave geométrica que estriba en que un segmento de recta es considerado tanto como magnitud geométrica continua como una medida numérica, pero la potencia de una línea recta sigue siendo una línea recta, así que cuadrado y cubo no indicarán magnitudes planas o espaciales, sino la segunda o tercera potencia de un número, de modo que las operaciones aritméticas quedan incluidas en un terreno estrictamente algebraico. En este punto, Descartes rompe con la tradición griega al abandonar el principio de homogeneidad. *La Geometría* da carta de naturaleza a las potencias superiores a^4, a^5, a^6,..., todas ellas son legítimas líneas. De esta forma se produce una cierta unificación del Álgebra y la Geometría. Descartes habría alcanzado lo que se había propuesto en *El Discurso del Método*, ya aludido anteriormente:

> «*Había estudiado* [...] *el Análisis de los antiguos y el Álgebra de los modernos,* [...]*, el primero está siempre tan constreñido a considerar las figuras, que no puede ejercitar el entendimiento sin fatigar mucho la imaginación, y en la última hay que sujetarse tanto a ciertas reglas y cifras, que se ha hecho un arte confuso y oscuro,* [...]*. Esto fue causa de que pensase que era necesario buscar algún otro método que, reuniendo las ventajas de estos tres, estuviese libre de sus defectos.* [...] *pensé que, para considerarlas mejor particularmente, debía suponerlas en línea* [recta]*, pues nada hallaba más simple ni que más distintamente pudiera representarse a mi imaginación y a mis sentidos. Y que para retenerlas o comprenderlas era necesario explicarlas mediante algunas cifras lo más cortas que fuera posible; de esta manera tomaría lo mejor del Análisis geométrico y del Álgebra y corregiría los defectos del uno por medio de la otra*» (DM.AT,VI, 17-20).

Con sus radicales reformas, Descartes habría superado la esclavitud a la dependencia de las figuras en la Geometría de los antiguos y la falta de trasparencia del Álgebra de los modernos. Por si fuera poco, Descartes eliminaba otra limitación de la Geometría griega y del *Arte Analítica* de Vieta, la de las tres dimensiones.

LA NOTACIÓN MATEMÁTICA CARTESIANA





1. **Retrato de Descartes por Weenix. Museo de Utrecht.**
2. **La edición en latín de 1649 de van Schooten (con notas de F. De Beaune) de *La Geometría* de Descartes. Esta edición es la primera separada de *El Discurso del Método* y contribuyó de forma muy considerable a la difusión de la obra de Descartes.**

La notación que Descartes introduce en la Regla XVI de las *Regulae* y perfecciona al comienzo de *La Geometría* tuvo un papel esencial en su magno proyecto de reforma que alcanzó a una completa reconstrucción de la Matemática sobre premisas muy sencillas, no geométricas como en Euclides, sino algebraicas, y con unos instrumentos muy modestos, sólo el *Teorema de Tales* y el *Teorema Pitágoras*, como confiesa a dos de sus pupilas, la reina Cristina de Suecia y la Princesa Isabel de Bohemia.

Descartes alude una y otra vez en las *Regulae* y en *La Geometría* a la función que debe cumplir una buena notación, simple y clara, formada de «*signos muy breves*»: «*ejercitar el entendimiento sin fatigar mucho la imaginación*» (DM.AT,VI, 17-18), para no distraer el pensamiento en retener cosas, a base de descargar la memoria por medio de la escritura para sólo confiarle lo imprescindible (RXVI.AT.X.458):

> **«*De modo general es preciso observar que jamás debe encomendarse a la memoria ninguna de las cosas que no requieran una continuada atención, si podemos depositarlas en el papel, no sea que un recuerdo superfluo para el conocimiento de un objeto nos prive de alguna parte de nuestro espíritu.*»**

El simbolismo algebraico, que apuntaba a convertirse en el lenguaje universal traería simplificación, generalización, mecanización y unificación en la notación, entrañando economía de pensamiento y difusión rápida. Después de Descartes, el Álgebra es uno de los más potentes lenguajes creados por el hombre, un instrumento para la expresión breve, intuitiva y mecánica de relaciones enormemente complicadas que puedan tener entre sí objetos abstractos cualesquiera, y en su aplicación a la Geometría, el ingenio que exigía la lectura y comprensión de la obra de Euclides quedaría eliminado y reemplazado por procedimientos algorítmicos automáticos.

Aparte de su ingente contribución al nacimiento de la Geometría Analítica, a Descartes le cabe, pues, el mérito de haber dado los pasos más importantes en la introducción de la moderna notación simbólica de las Matemáticas, de modo que el convenio notacional cartesiano se hizo definitivo. La Geometría, es el primer texto matemático en el que un estudiante actual no encontraría dificultades con la notación.

Análisis y Síntesis: planteamiento y resolución de las ecuaciones

Con las nuevas notaciones y símbolos, Descartes realizó una importante simplificación en el lenguaje matemático. Ahora disponía de una Geometría que al poderse expresar de forma algebraica permitía desarrollar procedimientos para resolver problemas geométricos a base de traducirlos al lenguaje algebraico de las ecuaciones, simplificar éstas y finalmente resolverlas (lo que quiere decir construir las soluciones) mediante lo cual Descartes se propondrá rehacer la Geometría.

Cómo se llega a las ecuaciones que sirven para resolver los problemas

«*Así, si se quiere resolver algún problema, debe de antemano considerarse como ya resuelto, y dar nombre a todas las líneas que parecen necesarias para construirlo, tanto a las que son desconocidas como a las otras. Luego, sin considerar ninguna diferencia entre estas líneas conocidas y desconocidas, se debe examinar la dificultad según el orden que se presente como más natural de todos, en la forma como aquellas líneas dependen mutuamente las unas de las otras, hasta que se haya encontrado la manera de expresar una misma cantidad de dos maneras: lo que se denomina una ecuación, pues* [el resultado de] *los términos de una de esas dos formas son iguales a los de la otra.*» (G.AT,VI,372).

He aquí una aplicación directa de los procedimientos del *Análisis* y la *Síntesi*s tal como los había descrito Pappus en *El Tesoro del Análisis* del Libro VII de *la Colección Matemática* y tal como lo había aplicado Vieta con la intervención del Álgebra en su *Arte Analítica*. Todo conduce a determinar la ecuación del problema geométrico, es decir, transitar de la Geometría al Álgebra mediante la metodología cartesiana, siguiendo unas pautas que Descartes ya había insinuado en las Reglas XVII–XXI de las *Regulae* :

a) Suponer el problema resuelto.

b) Dar nombre a todos los segmentos que parecen necesarios.

El propio *Análisis* nos ayudará a determinar quiénes son éstos, tanto los conocidos (datos) como los desconocidos (incógnitas) sin considerar ninguna diferencia entre ellos.

Estos dos primeros pasos corresponden al *Análisis* en sentido de Pappus. Ahora examinando el problema, siguiendo un orden basado en la intuición o en el *Análisis* anterior, estableciendo las relaciones que existen entre las diversas segmentos –los conocidos y los desconocidos– hemos de conseguir expresar un mismo segmento por medio de dos expresiones algebraicas diferentes, lo que permite realizar la *Síntesis*, es decir:

c) Determinar la ecuación entre las longitudes conocidas y las desconocidas.

Finalmente para resolver de forma definitiva el problema quedan dos pasos:

d) Resolver la ecuación resultante.

e) Construir geométricamente la solución.

Al plantearse problemas geométricos en la *Síntesis* se han de obtener soluciones geométricas para cuya construcción el Álgebra será el instrumento analítico esencial.

Así pues, ante un problema geométrico se aplicará todo un protocolo de actuación –el método cartesiano–: se empieza suponiendo el problema resuelto y se consideran las relaciones entre las líneas, lo que lleva al establecimiento de las ecuaciones, es decir, el estudio analítico se complementa con la síntesis algebraica que lleva a la construcción de la solución. El Análisis y el Álgebra que están ordenados al estudio y conocimiento de la figura, permiten traducir los datos geométricos de forma que sean tratables por medio del cálculo algebraico; se concluye el problema de Álgebra planteando y resolviendo las ecuaciones y finalmente los resultados obtenidos deben ser traducidos de nuevo al lenguaje geométrico, operación que nos da por fin la construcción de la solución. El Álgebra es un instrumento

que finalmente nos ha de reconducir a la Geometría.

Continúa el texto de Descartes con la primera manifestación irónica de las diversas presentes en *La Geometría* (G.AT,VI, 374):

> «[...] *Y pueden siempre reducirse así todas las cantidades desconocidas a una sola, cuando el problema puede construirse mediante círculos y líneas rectas, o bien por secciones cónicas o aun por ninguna otra línea que no esté compuesta, sino en uno o dos grados más. Pero no me detengo a explicar esto con más detalle para no privar a cada uno del placer de aprenderlo por sí mismo, ni impedir el cultivo útil del propio espíritu ejercitándolo, que es, a mi parecer, la principal utilidad que puede obtenerse de esta ciencia*, [...]»

Enseguida Descartes realiza una caracterización algebraica de los *Problemas Planos*:

Cuáles son los problemas planos.

«*Si éste puede ser resuelto por la geometría ordinaria, es decir, sin servirse más que de líneas rectas y circulares trazadas sobre una superficie plana, cuando la última ecuación haya sido enteramente desarrollada, no quedará, al fin, más que un cuadrado desconocido, igual a lo que resulta de la adición, o sustracción, de su raíz multiplicada por alguna cantidad conocida* [coeficiente]*, más alguna otra cantidad también conocida* [término independiente]» (G.AT,VI, 374).

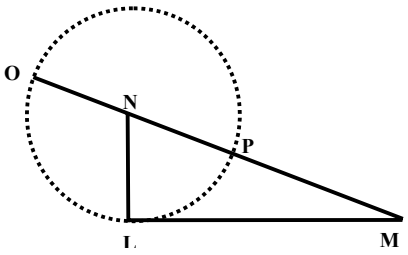
Descartes establece aquí una primera relación entre los problemas de construcción y los de clasificación (que tanta importancia tendrá en el Libro II) relacionando claramente un tipo concreto de expresiones algebraicas con los instrumentos que permiten trazar determinadas construcciones. En una primera trasferencia de la Geometría al Álgebra, si un problema geométrico lleva a una ecuación cuadrática, será resoluble con regla y compás, pero Descartes trasciende esta obviedad, identificando totalmente una cuestión geométrica con una cuestión algebraica al establecer que cualquier problema geométrico resoluble con regla y compás conduce a una última ecuación que necesariamente es cuadrática.

Vamos a ver concretamente cómo resuelve Descartes las ecuaciones cuadráticas que corresponden a los *Problemas Planos* y que son las únicas que Descartes trata en el Libro I.

Cómo se resuelven **[las ecuaciones que resultan de los problemas Planos]**

(G.AT,VI, 374-376)

«[...] *Si se tiene, por ejemplo* $z^2 = az + bb$ *construyo el triángulo rectángulo NLM, cuyo lado LM es igual a b, raíz cuadrada de la cantidad conocida bb, y el otro LN es* $\frac{1}{2}a$*, la mitad de la otra cantidad conocida, que está multiplicada por z, que supongo ser la línea desconocida. Luego, prolongando MN, base de ese triángulo, hasta O, de modo que NO sea igual a NL, la línea total OM es z, la línea buscada; ella se expresa:*



$$z = \frac{1}{2}a + \sqrt{\frac{1}{4}aa + bb}$$

Si se tuviera

yy = –ay + bb

e y fuera la cantidad que debe encontrarse, se construye el mismo triángulo rectángulo NLM y de la base MN se quita NP, igual a NL; el resto PM es y, la raíz buscada. De modo que tengo

$$y = -\frac{1}{2}a + \sqrt{\frac{1}{4}aa + bb}$$

Y lo mismo, si tuviera $x^4 = -ax^2 + b^2$

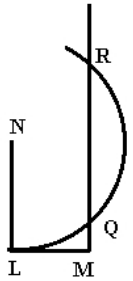
PM sería x^2 *y tendría*

$$x = \sqrt{-\frac{1}{2}a + \sqrt{\frac{1}{4}aa + bb}}$$

y así otros casos.

En fin, si tuviera $z^2 = ax - bb$

se hace NL igual a $\frac{1}{2}a$*, y LM igual a b, como anteriormente; luego, en vez de unir los puntos M y N, se traza MQR paralela a LN y trazando un círculo con centro en N y que pase por L la cortará en los puntos Q y R; la línea buscada z es MQ, o bien MR, pues en este caso ella se expresa de dos maneras, a saber:*



$$z = \frac{1}{2}a + \sqrt{\frac{1}{4}aa - bb}$$

$$z = \frac{1}{2}a - \sqrt{\frac{1}{4}aa - bb}$$

Y si el círculo que tiene su centro en N y pasa por el punto L no corta ni toca la línea recta MQR, no hay ninguna raíz de la ecuación, de manera que puede asegurarse que la construcción del problema propuesto es imposible.

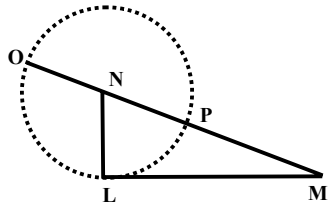
Por otra parte, estas mismas raíces se pueden encontrar por una infinidad de otros medios y he indicado aquí solamente esos muy simples, a fin de mostrar que se pueden construir todos los problemas de la geometría ordinaria, sin hacer más que lo poco que está comprendido en las cuatro figuras que he explicado. No creo que los antiguos lo hayan observado; pues en tal caso ellos no hubieran escrito libros tan voluminosos en que el solo orden de las proposiciones nos muestra que no poseían el verdadero método para resolverlas todas, sino que solamente han recopilado las que habían resuelto.»

Situándose en la tradición, Descartes estudia y resuelve los tres tipos clásicos de ecuaciones z^2=az+bb, yy= –ay+bb, z^2=ax–bb, que, como se sabe, antes de Vieta, eran considerados con coeficientes numéricos concretos y positivos, resultando equivalentes a los tres tipos tradicionales de ecuaciones: z^2=az+b, z^2+az=b, z^2+b=az, con a,b positivos, que en el *Álgebra Geométrica* griega tenían su forma particular de resolución, como tenían, asimismo, en la matemática árabe, cada uno de los tipos, su propio algoritmo de resolución. Pues bien, ahora Descartes, manteniendo por esta vez la homogeneidad, nos brinda la resolución geométrica de cada uno de los casos posibles. Para cada caso, siguiendo el algoritmo algebraico de resolución bien conocido, podríamos obtener el valor del segmento solución z expresado por las correspondientes operaciones con radicales, de donde se advierte que todas las operaciones que intervienen son resolubles, a partir de los segmentos dados, a,b, utilizando sólo la regla y el compás. No obstante, Descartes procede de forma

geométrica construyendo para cada caso el segmento solución, por eso no puede considerar todos los tipos de ecuaciones cuadráticas (por ejemplo no estudia la ecuación z^2+az+b^2=0 porque no tiene raíces positivas), y de las que resuelve sólo tiene en cuenta las soluciones positivas que son las únicas construbles.

Por otra parte, Descartes dice que «*estas mismas raíces se pueden encontrar por una infinidad de otros medios*». Y la verdad es que no especifica exactamente cómo obtiene la construcción geométrica del segmento solución si aplicando el *Teorema de Pitágoras* o la invariancia de la potencia de un punto respecto de la circunferencia. La equivalencia de ambas aplicaciones deja la construcción geométrica cartesiana en cierta ambigüedad.

Así en la primera ecuación z^2=az+bb, Descartes procede geométricamente indicando cómo puede construir el segmento de longitud z. Construye un triángulo rectángulo NLM cuyos catetos están determinados por los coeficientes de la ecuación: LM=b, LN=(1/2)a y con centro en N traza una circunferencia de radio NL=a que es cortada por la prolongación de la hipotenusa MN en el punto O, resultando que el segmento OM es la recta buscada z.

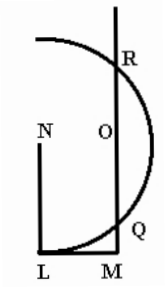


En efecto: MO=MN+NO. Pero por el Teorema de Pitágoras $MN = \sqrt{NL^2 + LM^2}$, de modo que sustituyendo cada recta por su longitud tenemos la expresión algebraica indicada: $z = \frac{1}{2}a + \sqrt{\frac{1}{4}aa + bb}$.

Ahora bien, aplicando la invariancia de la potencia de un punto respecto de la circunferencia (*Euclides,* III.36) se tiene: MO·MP=ML^2, es decir, z·(z–a)=bb, expresión equivalente a la ecuación dada z^2=az+bb.

La construcción geométrica de la solución de la última ecuación z^2=az–bb es un poco más complicada.

Construidos de los elementos geométricos de la figura como indica Descartes, a partir de los segmentos medidos por los coeficientes de la ecuación NL=(1/2)a y LM=b, se aplica la invariancia de la potencia de un punto respecto de la circunferencia: MR·MQ=LM^2.



Si z=MR, tenemos que MQ=a–z, y por tanto: z·(a–z)=b^2, es decir, z^2=az–bb, por tanto el segmento z=MR es una línea solución.

Si z=MQ, tenemos que MR=a–z, y por tanto: z·(a–z)=b^2, es decir, z^2=az–bb, por tanto el segmento z=MR es una línea solución.

Ahora si O es el punto medio de QR tenemos:

z_1=MQ=OM–OQ= $\frac{1}{2}a - \sqrt{\frac{1}{4}aa - bb}$

z_2=MR=MO+OR= $\frac{1}{2}a + \sqrt{\frac{1}{4}aa - bb}$

Descartes construye las dos raíces porque ambas son positivas.

Si MR es tangente al círculo, es decir si b=(1/2)a, las raíces son iguales; mientras que si b>(1/2)a, la línea MR no cortará al círculo y entonces no hay raíces. Descartes expresa esto en un lenguaje tributario todavía de los geómetras griegos:

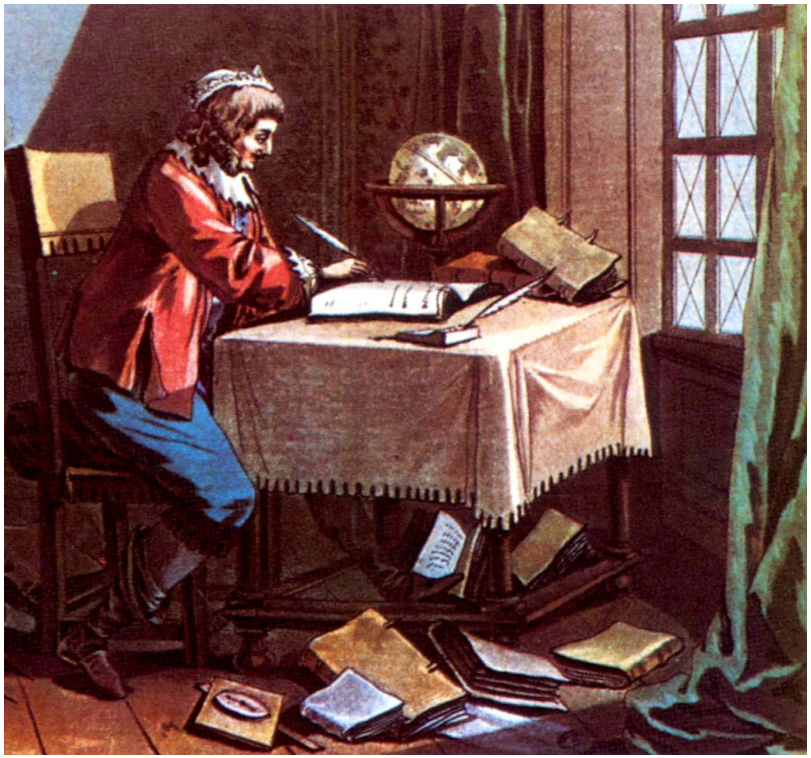
> «*Y si el círculo que tiene su centro en N y pasa por el punto L no corta* [no es secante] *ni toca* [no es tangente] *la línea recta MQR, no hay ninguna raíz de la ecuación, de manera que puede asegurarse que la construcción del problema propuesto es imposible.*»

Vemos cómo Descartes ha vinculado íntimamente el Álgebra con la Geometría, hasta el punto de extraer conclusiones geométricas de un hecho estrictamente algebraico: si la ecuación no tiene solución el problema geométrico no se puede construir, porque «*encontrar la solución*» es «*construir la línea*». Los principios del método cartesiano aplicados a la Geometría inician los problemas geométricos por un proceso intermedio de escritura algebraica que revierte finalmente sobre la geometría del problema conduciendo a la construcción de la línea solución. Es esta intermediación del Álgebra lo que más se echa de menos en la Geometría griega, por eso después de la resolución constructiva de las ecuaciones, Descartes hace un soberbio alarde de la magnificencia de los métodos de su *Geometría* en contraposición con la precariedad de la Geometría de los griegos. Según él, puede construir todos los problemas de la Geometría ordinaria con las escasas cuatro figuras que ha explicado, mientras que el abstruso orden de las complejas proposiciones de los voluminosos libros de la Geometría griega era una prueba palmaria de que los antiguos no disponían de método. Esta afirmación, además de exagerada, es injusta, porque aunque la Geometría griega no pudo disponer de la claridad, simplicidad, flexibilidad, versatilidad y capacidad algorítmica que proporciona el Álgebra simbólica que manejó Descartes, éste estaba al corriente de las grandes obras de la matemática griega, empezando por *Los Elementos* de Euclides, que es una obra indiscutiblemente metódica, con fundamentos en el método axiomático-deductivo muy diferente del de Descartes, pero tanto o más riguroso que lo que se aplica en *La Geometría*. De hecho la *Aplicación de las Áreas* del Álgebra Geométrica griega de los Libros II y VI de *Los Elementos* de Euclides es un procedimiento metódico de resolución de ecuaciones mediante comparación de áreas. Es más, el método cartesiano tiene su inspiración en el método de *Análisis* y *Síntesis*, que instaurado por Hipócrates de Quíos y fundamentado por Platón es explicitado por Pappus, uno de los matemáticos griegos más admirados por Descartes.

SELLOS DE DESCARTES EN EL AÑO 2000 DE LAS MATEMÁTICAS

1. **Emitido en Granada con motivo del año 2000 de las Matemáticas.**
2. **Emitido en Sierra Leona con motivo del año 2000 de las Matemáticas.**

SÍNTESIS DEL PROGRAMA DE REFORMA CARTESIANA DE LA GEOMETRÍA A TRAVÉS DE LOS TEXTOS DE DESCARTES



Retrato de Descartes como escritor (Biblioteca Nacional de París, 1791).

Ejercitar el entendimiento sin fatigar mucho la imaginación (DM.AT,VI, 17)

- **Fundir del Análisis Geométrico de los antiguos («*siempre tan constreñido a considerar las figuras*») [DM.AT,VI.17)] y el Álgebra de los modernos («*que se ha hecho un arte confuso y oscuro*» [DM.AT,VI.17)]; «*para buscar otro método que, reuniendo las ventajas de éstos, estuviese libre de sus defectos*»[DM.AT,VI.18].**
- **Introducir las coordenadas: «*Todos los problemas de Geometría pueden reducirse fácilmente a términos tales, que no es necesario conocer de antemano más que la longitud de algunas líneas rectas para construirlos*» [G.AT,VI, 369].**
- **Reconstruir de forma geométrico-algebraica las operaciones aritméticas, es decir, mostrar «*cómo el cálculo de la aritmética se relaciona con las operaciones de la geometría*» [G.AT,VI, 369].**
- **Introducir una revolucionaria simplificación en la notación: «*explicarlas mediante algunas cifras lo más cortas que fuera posible*» [DM.AT,VI,20].**
- **Indicar «*cómo pueden emplearse letras en geometría*» [G.AT,VI,371].**
- **Enseñar «*cómo se llega a las ecuaciones que sirven para resolver los problemas*» [G.AT,VI,372]; y «*cómo se resuelven*» [G.AT,VI,374] estas ecuaciones (es decir, cómo se construyen las soluciones).**

Mediante estas tareas, Descartes «*tomaría lo mejor del Análisis geométrico y del Álgebra y corregiría los defectos del uno por medio de la otra*» [DM.AT,VI, 20].

Sistemas de referencia. El Problema de Pappus

Hasta aquí, Descartes ha elaborado un potente método analítico-sintético de ataque de los problemas geométricos que utiliza el Álgebra como instrumento algorítmico y con el que se propone no sólo rehacer la Geometría griega sino ir mucho más allá en la resolución de antiguos y nuevos problemas geométricos. Por eso se plantea al final del Libro I el abordaje del famoso *problema de Pappus de las tres o cuatro rectas*, que tan firmemente se había resistido a los geómetras griegos, y que siendo generalizado a 2n–1, 2n rectas campea a lo largo de *La Geometría* de Descartes, de modo que algunos historiadores se atreven a decir que «*Toda La Geometría de Descartes está destinada a la resolución del Problema de Pappus*» o que «*El Problema de Pappus conforma La Geometría de Descartes*».

La profundidad, extensión e inmensa casuística del tratamiento cartesiano del *Problema de Pappus* trasciende de los objetivos de este escrito, por tanto nos ceñiremos a aspectos que incidan sobre los orígenes de la Geometría Analítica, en particular la aparición de los sistemas de referencia y las coordenadas.

Ejemplo tomado de Pappus

(G.AT,VI, 377-380)

> «*Y esto* [la insuficiencia de los métodos de la Geometría griega] *puede verse bien claramente en lo que Pappus ha puesto al principio de su Libro VII, donde después de haberse detenido a citar todo lo que había sido escrito en geometría por los que lo habían precedido, habla finalmente de un problema que, según dice, ni Euclides ni Apolonio habían podido resolver enteramente*; *he aquí sus propias palabras*: [...]»

Descartes transcribe el enunciado del problema en latín y hace una observación:

> «*Os ruego que observéis de paso que el escrúpulo que tenían los antiguos en emplear los términos de la aritmética en la geometría, no podía provenir más que de no ver ellos claramente su relación, lo que producía bastante oscuridad y confusión en la forma como se expresaban*; [sigue el texto latino].»

Descartes no alcanza a interpretar que los escrúpulos de los que habla respondían a la conformación que dio a la Geometría griega la aparición de las magnitudes inconmensurables que impedía asignar a las figuras geométricas números que midieran sus longitudes, áreas y volúmenes y por tanto los griegos tenían que calcular directamente con las figuras, que se trataban como magnitudes. Así se hacía en el *Álgebra Geométrica* del Libro II de *Los Elementos* de Euclides en la que los números se sustituyen por segmentos de recta y las operaciones entre ellos se llevan a cabo mediante construcciones geométricas pero sin el concurso de Álgebra simbólica porque era inexistente, por lo que las ecuaciones se resolvían estableciendo relaciones que comparaban –a partir del Libro V de *Los Elementos* de Euclides mediante razones– áreas dadas con áreas buscadas o deseadas para la resolución de un problema. En la Geometría griega, los segmentos rectilíneos no tenían longitud ante la eventualidad de la inconmensurabilidad y como consecuencia las «*operaciones*» con los segmentos daban rectángulos y paralelepípedos, que eran objetos de naturaleza estrictamente geométrica imposibles de confundir con el producto de las longitudes de sus lados, ya que, hasta Diofanto, estaba ausente el sentido aritmético de las operaciones. De ahí las limitaciones del respeto a la homogeneidad y de la acotación tridimensional.

La gran innovación de Descartes es la asignación de una longitud a los segmentos («*y yo no temeré introducir estos términos de aritmética en la geometría, a fin de hacerme más inteligible*») lo que permite su manipulación algebraica operacional:

> «[...] *es de señalar que para a^2 o b^3 u otras expresiones semejantes, yo no concibo ordinariamente mas que líneas simples, aunque para servirme de los nombres usados en álgebra, los designe por cuadrados, cubos, etc.*» (G.AT,VI, 371),

con base en lo cual Descartes no tiene ningún prejuicio geométrico en hablar en el *Problema de Pappus* del «*producto de cuatro líneas rectas, de cinco o de más*», es decir, el instrumental cartesiano no sólo permitirá resolver de forma brillante el problema clásico, sino que además propicia su más amplia generalización.

Continúa Descartes escribiendo su propio enunciado del *Problema de Pappus*:

> «*Así pues, la cuestión que Euclides había empezado a resolver y que Apolonio había proseguido sin que nadie la hubiera terminado, era ésta: Dadas tres, cuatro o más rectas, se trata de encontrar un punto del que se puedan trazar otras tantas líneas rectas, una sobre cada una de las dadas, y haciendo con ellas ángulos dados, y que el rectángulo formado por dos de esas así trazadas desde el punto, tenga una proporción dada con el cuadrado de la tercera, si no hay más que tres; o bien con el rectángulo de las otras dos, si hubiera cuatro; o bien si hay cinco que el paralelepípedo compuesto por tres tenga la proporción dada con el paralelepípedo formado por las dos que restan y por otra línea dada. O bien si hay seis, que el paralelepípedo formado por tres tenga una proporción dada con el paralelepípedo de las otras tres. O bien si hay siete, que lo que se produce multiplicando cuatro la una por la otra, tenga la razón dada con lo que se produce por la multiplicación de las otras tres y además por otra línea dada. O si hay ocho, que el producto de la multiplicación de cuatro tenga la proporción dada con el producto de las otras cuatro. Y así este problema se puede extender a todo número de líneas. Pero, a causa de que hay siempre una infinidad de diversos puntos que pueden satisfacer lo que aquí se pide, se requiere también conocer y trazar la línea sobre la cual deben todos ellos encontrarse; y Pappus dice que cuando no hay más que tres o cuatro líneas rectas dadas, es en una de las tres secciones cónicas, pero él no trata de determinarla ni describirla; ni explicar la línea en que los puntos deben encontrarse cuando el problema está propuesto para un mayor número de líneas. Solamente agrega que los antiguos habían imaginado una, que mostraban ser útil* [para la resolución del problema], *y aunque parecía la más manifiesta, sin embargo no era la primera. Lo que me ha dado ocasión para ensayar si, por el método de que me valgo, se puede ir tan lejos como ellos fueron*.»

Respuesta al problema de Pappus

(G.AT,VI, 380-382)

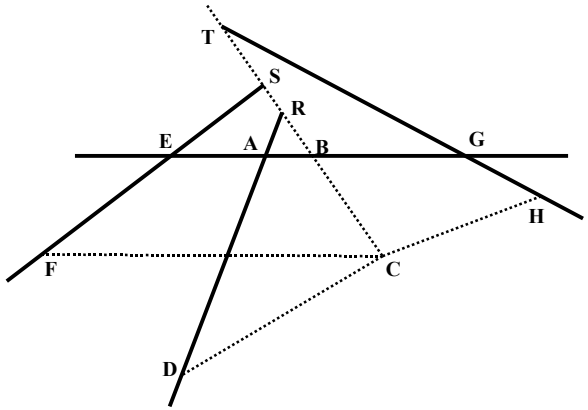
> «*He comprendido ante todo, que planteado el problema para tres, cuatro o cinco líneas, se puede siempre encontrar los puntos buscados, por la geometría simple, es decir sin servirse más que de la regla y el compás, ni hacer otra cosa que lo ya dicho; excepto solamente cuando, siendo cinco las líneas, ellas son todas paralelas. ...*
>
> .....
>
> *Y he encontrado que cuando no hay más que tres o cuatro líneas dadas, los puntos buscados se encuentran todos no solamente en una de las tres cónicas sino a veces en la circunferencia de un círculo o en una línea recta. ...*
>
> .....
>
> *De modo que pienso haber satisfecho enteramente lo que Pappus nos dice haber sido buscado por los antiguos, trataré de dar la demostración en pocas palabras: pues ya me cansa tanto escribir.*

Sean AB, AD, EF, GH, etc., varias líneas dadas y debe encontrarse un punto, como C, del cual trazando otras líneas a las dadas, como CB, CD, CF y CH, de manera que los ángulos CBA, CDA, CFE, CHG, etc. Sean dados, y que el producto de la multiplicación de una parte de estas líneas, sea igual al producto de la multiplicación de las otras; o bien que ellas tengan otra proporción dada, lo que no hace, en modo alguno, más difícil el problema.»

Cómo deben ponerse los términos para llegar a la ecuación de este ejemplo

(G.AT,VI, 382-385)

«*Primeramente yo supongo la cosa como ya hecha y para salir de la confusión de todas esas líneas, considero una de las dadas y una de las que hay que encontrar, por ejemplo AB y CB como las principales y a las cuales trato de referir todas las otras. Sea designado x el segmento de la línea AB comprendido entre los puntos A y B; y CB sea designado y; y todas las demás líneas se prolonguen hasta que corten a estas dos también prolongadas, si es necesario y si no le son paralelas; como se ve cortan la línea AB en los puntos A, E, G y la línea BC en los puntos R, S, T.*»



He aquí uno de los puntos de mayor interés de *La Geometría* de Descartes. Empieza el análisis:

a) se supone el problema resuelto,

b) se da nombre a todos los segmentos necesarios para representarlos, tanto los conocidos como los desconocidos,

c) se reconstruye algebraicamente el problema hasta obtener una ecuación que permitirá alcanzar la síntesis.

Pero para facilitar el proceso *Análisis-Síntesis* Descartes introduce el primer sistema de coordenadas de *La Geometría* (G.AT,VI, 383):

> «[...] *Considero una de las dadas y una de las que hay que encontrar, por ejemplo AB y CB como las principales y a las cuales trato de referir todas las otras.*»

Los segmentos $x = \overline{AB}$, $y = \overline{CD}$ son las «*coordenadas*» del punto C en el sistema de referencia establecido.

A continuación del texto (G.AT,VI, 383-385), Descartes utiliza el sistema de referencia introducido para obtener, mediante una serie de cálculos elementales aunque prolijos, la expresión de cada uno de los segmentos que dan las distancias (CB, CD, CF, CH). Resultan ser todas ellas ser combinaciones afines (diríamos hoy) de «*las coordenadas*» (x,y) del punto C, es decir, de la forma Ax+By+C, donde A, B, C, son cantidades no nulas, salvo cuando hay relaciones de paralelismo:

> *«Se ve así que cualquiera que sea el número de líneas dadas, todas las líneas trazadas desde C, que forman ángulos dados, conforme al enunciado, se pueden siempre expresar, cada una por tres términos de los que uno está compuesto por la cantidad desconocida y multiplicada o dividida por alguna otra conocida, y la otra, de la cantidad desconocida x, también multiplicada o dividida por alguna otra conocida y la tercera, de una cantidad toda conocida.*
>
> *Además se ve que multiplicando varias de estas líneas entre sí, las cantidades x e y que se encuentran en el producto, no pueden tener cada una más que tantas dimensiones como líneas haya.*»

Así pues, aunque Descartes no lo explicita, en el caso de tres o cuatro rectas el problema equivaldría a una ecuación cuadrática de la forma: $Ax^2+Bxy+Cy^2+dx+ey+f=0$. De esta forma aparecerían por primera vez las curvas como lugares geométricos definidos por ecuaciones, en el umbral del Libro II de *La Geometría*.

Utilizando métodos de la Geometría Analítica moderna, en particular la forma normal de la ecuación de la recta, podemos resolver fácilmente el problema de tres líneas, por ejemplo, encontrando que el lugar geométrico es una cónica. Sean $a_1x+b_1y+c_1=0$, $a_2x+b_2y+c_2=0$, $a_3x+b_3y+c_3=0$ las ecuaciones de las rectas, y α_1, α_2, α_3, los ángulos que dan las direcciones sobre las que se deben medir las distancias. El lugar geométrico del punto C(x,y) viene dado por la ecuación:

$$\frac{(a_1x+b_1y+c_1)^2}{(a_1^2+b_1^2)\,\mathrm{sen}^2\alpha_1} = K\,\frac{(a_2x+b_2y+c_2)^2}{\sqrt{(a_2^2+b_2^2)}\,\mathrm{sen}\alpha_2}\cdot\frac{(a_3x+b_3y+c_3)^2}{\sqrt{(a_3^2+b_3^2)}\,\mathrm{sen}\alpha_2}$$

Sigue el texto de Descartes demostrando que si lo que buscamos es un solo punto (para el problema de menos de cinco rectas) el problema es plano, ya casi al final del Libro I:

Cómo se encuentra que este problema es plano cuando no está propuesto para más de cinco líneas

(G.AT,VI, 385-387)

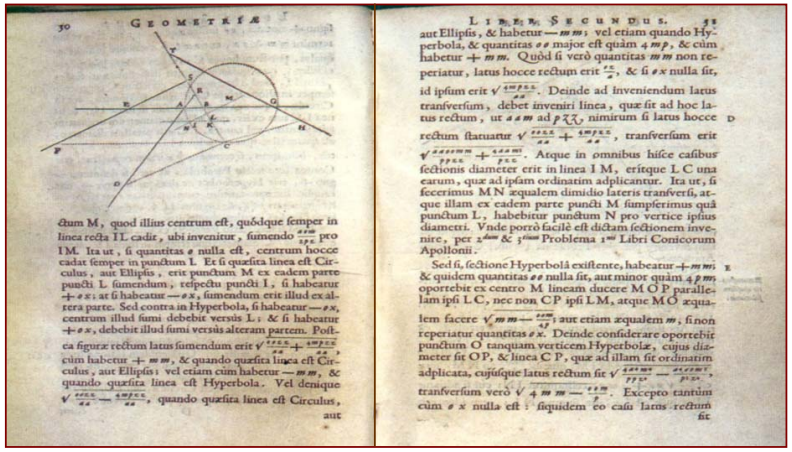
«*Además, a causa de que para determinar el punto C no hay más que una sola condición requerida* [la de la igualdad de las multiplicaciones de líneas]*, puede tomarse a discreción una de estas cantidades desconocidas x o y, y buscar la otra por la ecuación, en la cual es evidente que cuando el problema no está propuesto para más de cinco líneas, la cantidad x que no es utilizada para la expresión de la primera de las líneas nunca puede tener más de dos dimensiones. De modo que tomando y como una cantidad conocida tendremos*

xx = + o – ax + o –bb [$x^2 = \pm ax \pm b^2$]

y así se podrá encontrar la cantidad x con la regla y el compás de la manera ya explicada. Lo mismo tomando sucesivamente infinitos valores para la línea y, podemos hallar otros tantos para la línea x; y así se tendrá una infinidad de diversos puntos tales como el que se ha señalado con C, por medio de los cuales se describirá la línea curva pedida.»

Nuevamente debemos señalar aquí otro de los puntos de mayor interés, para nosotros, de *La Geometría* de Descartes: una aproximación al concepto de función a través de la expresión analítica de una ecuación. Por ejemplo para cuatro rectas la consideración de las coordenadas del punto C=(x,y), nos lleva a una ecuación polinómica de segundo grado F(x,y)=0, que representa una infinidad de pares (x,y) que satisfacen el problema de las cuatro rectas. Tal ecuación es una expresión matemática bien definida que depende de dos variables x, y, de modo que conociendo una de ellas se puede hallar rigurosamente la otra (tras la resolución algebraica de una ecuación de lo que Descartes se ocupará, en el Libro III de *La Geometría*, para grados superiores), lo que equivale a la determinación geométrica del punto C. Así pues, el conocimiento de la ecuación permite conocer los puntos de la curva. Todavía no hay una identificación de la curva con la ecuación, Descartes sólo ha introducido el concepto de curva definida por puntos y esperará al Libro II de *La Geometría* para estudiar ampliamente el problema y establecer cuáles son las razones geométricas que permiten considerar una expresión como una representación de una curva.

EL *PROBLEMA DE PAPPUS*



El *Problema de Pappus*. Edición latina de van Schooten de *La Geometría* de Descartes, 1659.

El *Problema de Pappus* –llamado en su enunciado más sencillo *lugar de tres o cuatro rectas*–, es una de las cuestiones más importantes de toda la Historia de la Geometría, por ser la piedra de toque de aplicación de los diversos métodos y técnicas geométricas. Planteado por los geómetras griegos a partir de Euclides, estudiado por Apolonio y Pappus, su dificultad desbordaba, siglo tras siglo, las posibilidades del Análisis geométrico griego. Campeando a lo largo de *La Geometría*, como si fuera su punto de inspiración, casi como un reto a alcanzar, será Descartes quien lo resuelva de forma brillante y general poniendo de manifiesto la potencia de unos métodos analíticos, que en el curso de los años se convertirán en la esencia de la Geometría Analítica.

Las rectas normales a una curva. El Método del círculo

Descartes desarrolla en el Libro II de La Geometría (G.AT,VI, 412-423) un método para el trazado de las tangentes a las líneas curvas –el llamado «*método del círculo*»–, mediante la construcción previa de la recta normal. Es sin duda uno de los más significativos problemas de aplicación del método cartesiano, con el que, además, Descartes participa e interviene, a través de su celebre polémica con Fermat, en el ámbito matemático de la primera parte del siglo XVII, muy ocupado en la resolución del problema del trazado de las tangentes a las líneas curvas.

Una vez concebida y definida en la primera parte del Libro II, de forma clara y distinta, *la naturaleza geométrica de las líneas curvas*, Descartes introduce uno de los principios básicos de su método ponderando su importancia en la resolución de los problemas sobre curvas:

Para encontrar todas las propiedades de las líneas curvas basta con saber la relación que tienen todos sus puntos con los de las líneas rectas, y la manera de trazar otras líneas que las corten en todos esos puntos en ángulo recto.

(G.AT,VI, 412-413)

«*Luego con sólo saber la relación que tienen todos los puntos de una línea curva con todos los de una línea recta, en la forma que he explicado, es fácil también conocer la relación que ellos tienen con todos los otros puntos y líneas dadas; y, por lo tanto, conocer los diámetros, los ejes, los centros, y otras líneas o puntos que tengan con la línea curva alguna relación particular, o más simple que otros; y, de ahí imaginar diversos modos de describirlas, y elegir los más fáciles. Y también, con sólo esto, se puede aun, encontrar casi todo lo que puede ser determinado respecto a la medida del espacio que abarcan, sin que haya necesidad de que yo me extienda más. Y, por último, en lo que respecta a todas las otras propiedades que pueden atribuirse a las líneas curvas, ellas no dependen más que de la magnitud de los ángulos que ellas forman con otras líneas. Pero, cuando puedan trazarse líneas rectas que las cortan en ángulo recto* [normales], *en los puntos en que se encuentran con aquéllas con las que forman los ángulos que se quieren medir, o, lo que aquí tomo como igual, en que ellas cortan sus contingentes* [tangentes], *la magnitud de esos ángulos no es más difícil de encontrar que si ellos estuvieran comprendidos entre dos líneas rectas. Creo por esto haber dado aquí todo lo que se requiere para los elementos de la líneas curvas, cuando haya expuesto la manera general de trazar líneas rectas que las corten en ángulos rectos en los puntos* [de las curvas] *que de ellas se elijan. Y me atrevo a decir que éste es el problema más útil y más general no sólo que yo conozca, sino aun que yo haya anhelado jamás conocer en Geometría.*»

Descartes determina que «p*ara encontrar todas las propiedades de las líneas curvas basta con saber la relación que tienen todos sus puntos con los de las líneas rectas*», descrita por medio de una expresión –que es la ecuación de la curva–, y establece cómo se puede utilizar esta expresión algebraica para encontrar los elementos geométricos más notables de las curvas (diámetros, ejes, centros, etc.) y, en particular, las normales y tangentes. Con ello, Descartes enuncia uno de los principios fundamentales de la llamada Geometría Analítica: el conocimiento de la relación que liga las coordenadas de los puntos –los segmentos o «*las líneas rectas*» a las que alude– de una curva, es decir, la ecuación de la curva, es un elemento esencial para dilucidar y desentrañar las propiedades y elementos de la curva. La ecuación de la curva realiza un tránsito de la Geometría al Álgebra, que, por su carácter operacional, permite, realizando cálculos y en particular resolviendo ecuaciones, regresar a la Geometría, para encontrar y solucionar cuestiones geométricas, de modo que se fija una correspondencia entre las propiedades algebraicas de la ecuación y las

propiedades geométricas de la curva asociada. Como consecuencia, la tarea de probar un teorema en Geometría se traslada de forma muy eficiente a probarlo en Álgebra y, además, ésta se convierte en un poderoso instrumento de investigación geométrica. Pues bien, en este lugar, Descartes aplica toda esta filosofía geométrico-algebraica a encontrar las rectas normales de las líneas curvas.

De pasada Descartes define el ángulo entre dos curvas en un punto como el ángulo que forman las normales en ese punto, además, añade de soslayo la frase (G.AT,VI, 413):

> «[...] *Con sólo esto, se puede aun, encontrar casi todo lo que puede ser determinado respecto a la medida del espacio que abarcan*»,

de modo que el asunto se podría aplicar al cálculo de áreas determinadas por curvas de las que se conoce la ecuación, es decir, al otro problema candente en los círculos matemáticos de la primera parte del siglo XVII, el cálculo de cuadraturas, problema al que Descartes no presta atención alguna en su obra matemática.

ESCENAS DE LA VIDA DE DESCARTES

Escenas de la vida de Descartes, de C.P.Marillier. Grabado del siglo XVIII.

Este grabado muestra algunos hechos importantes de la vida de Descartes:

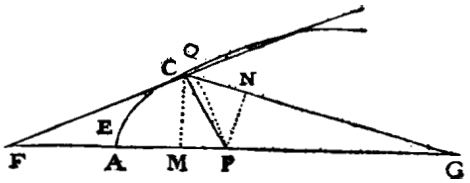
1. **Izquierda-arriba: muerte de la hija ilegítima del filósofo, Francine, el 7 de setiembre de 1640.**
2. **Izquierda-abajo: episodio de la esgrima con los marineros del río Elba que querían desvalijarle (1621).**
3. **Arriba-derecha: Descartes imparte lecciones de Filosofía a la reina Cristina de Suecia (1649).**
4. **Abajo-derecha: Contemplando un cartel con un problema en Breda (Holanda). Encuentro con Beeckman (1618).**

A continuación, Descartes entra directamente en el problema. Curiosamente utiliza la notación anterior xx para indicar x^2, en cambio adopta la notación potencial x^n para el resto de las potencias. Transcribiremos el texto de Descartes utilizando una notación uniforme x^n para todas las potencias:

Manera general de encontrar líneas rectas que corten las curvas dadas o sus tangentes, formando ángulos rectos.

(G.AT,VI, 413-414).

«*Sea CE la línea curva y que deba trazarse una recta por el punto C que forma con ella ángulos rectos.*



Supongamos que la cosa está hecha y que la línea buscada es CP, que prolongo hasta el punto P en que encuentra a la línea recta GA que supongo ser aquella a cuyos puntos se refieren todos los de la línea CE; de manera que haciendo MA o CB=y, y CM o BA=x, hay alguna ecuación que explica la relación que existe entre x e y [el punto B no figura en el dibujo; guiándose por las siguientes figuras, se obtendría como intersección de la perpendicular a AM por A y la paralela a AM por C].

Luego haciendo PC=s, PA=v, o bien PM=v–y, por el triángulo rectángulo PMC obtengo s^2, que es el cuadrado de la base, igual a $x^2+v^2-2vy+y^2$, que son los cuadrados de los dos lados; es decir que tengo

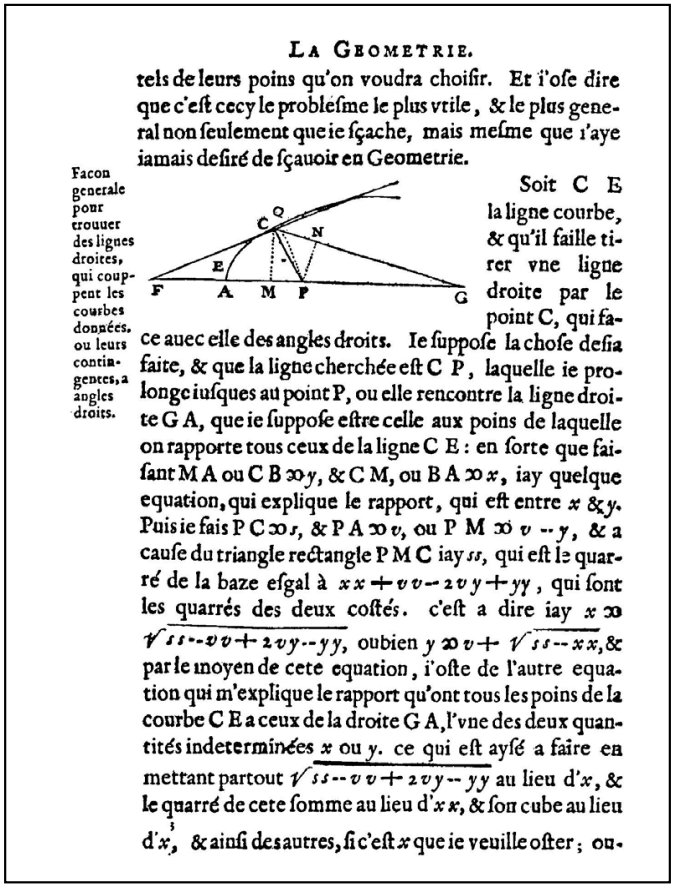
$$x = \sqrt{s^2 - v^2 + 2vy - y^2}$$

o bien

$$y = v + \sqrt{s^2 - x^2}$$

y por medio de esta ecuación, saco de la otra ecuación que da la relación que tienen todos los puntos de la curva CE con los de la recta GA [la ecuación de la curva], *una de las dos cantidades indeterminadas x o y; lo que es fácil de hacer poniendo* $\sqrt{s^2 - v^2 + 2vy - y^2}$ *en lugar de x, y el cuadrado de esta suma en lugar de x^2, y su cubo en lugar de x^3; y así los otros términos si es x que yo deseo sacar; o bien, si es y, poniendo en su lugar* $v + \sqrt{s^2 - x^2}$ *; y el cuadrado o el cubo, etc. De modo que quede siempre según esto, una ecuación en la cual no hay más que una sola cantidad indeterminada x o y.*

EL MÉTODO DEL CÍRCULO EN *LA GEOMETRÍA* DE DESCARTES



LA GEOMETRIE.

tels de leurs poins qu'on voudra choiſir. Et i'oſe dire que c'eſt cecy le probleſme le plus vtile, & le plus general non ſeulement que ie ſçache, mais meſme que i'aye iamais deſiré de ſçauoir en Geometrie.

Facon generale pour trouuer des lignes droites, qui coupent les courbes données, ou leurs contingentes, a angles droits.

Soit C E la ligne courbe, & qu'il faille tirer vne ligne droite par le point C, qui face auec elle des angles droits. Ie ſuppoſe la choſe deſia faite, & que la ligne cherchée eſt C P, laquelle ie prolonge iuſques au point P, ou elle rencontre la ligne droite G A, que ie ſuppoſe eſtre celle aux poins de laquelle on rapporte tous ceux de la ligne C E: en ſorte que faiſant M A ou C B ∞ *y*, & C M, ou B A ∞ *x*, iay quelque equation, qui explique le rapport, qui eſt entre *x* & *y*. Puis ie fais P C ∞ *s*, & P A ∞ *v*, ou P M ∞ *v* − *y*, & a cauſe du triangle rectangle P M C iay *ss*, qui eſt le quarré de la baze eſgal à $xx + vv - 2vy + yy$, qui ſont les quarrés des deux coſtés. c'eſt a dire iay $x \propto \sqrt{ss - vv + 2vy - yy}$, oubien $y \propto v + \sqrt{ss - xx}$, & par le moyen de cete equation, i'oſte de l'autre equation qui m'explique le rapport qu'ont tous les poins de la courbe C E a ceux de la droite G A, l'vne des deux quantités indeterminées *x* ou *y*. ce qui eſt ayſé a faire en mettant partout $\sqrt{ss - vv + 2vy - yy}$ au lieu d'*x*, & le quarré de cete ſomme au lieu d'xx, & ſon cube au lieu d'x^3, & ainſi des autres, ſi c'eſt *x* que ie veuille oſter; ou-

Página de la edición de 1637 de *La Geometría* de Descartes relativa al trazado de rectas normales a las curva –*método del círculo*– donde Descartes aplica uno de los Principios fundamentales de la Geometría Analítica (G.AT,VI, 412):

> «*Para encontrar todas las propiedades de las líneas curvas basta con saber la relación que tienen todos sus puntos con los de las líneas rectas*»

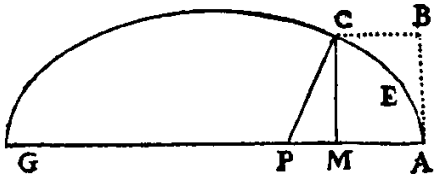
Esta frase contiene uno de los principios más importantes de la Historia de la Matemática, que instaura los fundamentos de la Geometría Analítica. La relación que liga los segmentos o «*las líneas rectas*» que hacen la función de «*coordenadas*» de los puntos de una curva, es decir, la ecuación de la curva, permite conocer las propiedades y los elementos característicos de la curva. La ecuación de la curva establece, pues, una correspondencia entre las propiedades algebraicas de la ecuación y las propiedades geométricas de la curva asociada. De esta forma la resolución de un problema de Geometría se traslada de forma muy eficaz a resolverlo en Álgebra y, además, ésta se convierte en un poderoso instrumento de investigación geométrica.

A continuación, Descartes aplica el método desarrollado a la elipse (G.AT,VI, 414):

Si fuese CE una elipse y MA el segmento de su diámetro [eje] *al cual corresponde CM, y siendo r su lado recto y q el transverso se tiene, por el teorema 13 del Libro I de Apolonio:*

$$x^2 = ry - \frac{r}{q}y^2,$$

[ecuación de la elipse referida a ejes oblicuos, siendo uno de ellos el diámetro y el otro la tangente en su extremo]



de donde sustituyendo x^2, *queda:*

$$s^2 - v^2 + 2vy - y^2 = ry - \frac{r}{q}y^2$$

o bien

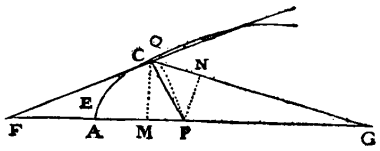
$$y^2 + \frac{qry - 2qvy + qvv - qs^2}{q - r} \text{ igual a cero.}$$

pues mejor, en este lugar, considerar así en conjunto toda la suma que hacer una parte igual a otra.»

Dada la curva CE de eje AG y vértice A, Descartes se plantea trazar la normal en el punto C, para lo cual debe encontrar un punto P sobre el eje AG que al trazar el segmento PC nos de la normal. De acuerdo con la metodología cartesiana, comienza el análisis del problema:

a) se considera resuelto, y b) se da nombre a los todos los segmentos que parecen necesarios: MA =y, CM=x, PC=s, PA=v.

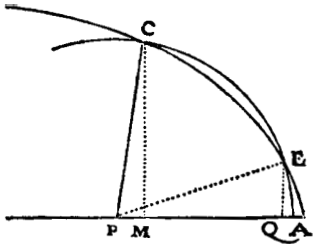
En la síntesis se indica que «[...] *Hay alguna ecuación que explica la relación que existe entre x e y*», es decir, la ecuación de la curva.



Ahora Descartes considera la circunferencia de centro el punto P y radio el segmento PC, que cortará a la curva en algunos puntos según la naturaleza de la curva. Todavía en el ámbito geométrico del problema, Descartes intuye que el segmento PC será la normal si la circunferencia es tangente a la curva en el punto C.

Descartes prosigue realizando para la parábola cálculos análogos a los de la elipse, tras lo cual vuelve al problema general en una forma que justifica que a su regla se le denomine «*método del círculo*» (G.AT,VI, 417):

> «*Ahora, después de encontrada una ecuación así, en lugar de utilizarla para conocer las otras cantidades x o y, que son ya dadas, puesto que el punto C es dado, se la debe emplear para encontrar v o s que determinan el punto P pedido. Y, a este efecto se debe considerar que si ese punto P es el punto que deseamos encontrar, el círculo del cual es centro y que pasa por el punto C, tocará a la línea curva CE sin cortarla; pero si el punto P está ya sea más próximo o más alejado del punto A de lo debido, ese círculo cortará a la curva no sólo en el punto C, sino necesariamente en algún otro. Debe también considerarse que cuando este círculo corta la línea curva CE, la ecuación por la cual se busca la cantidad x o y, o alguna semejante, suponiendo PA y PC conocidas, contiene necesariamente dos raíces, que*



> *son desiguales. Pues por ejemplo, si este círculo corta a la curva en los puntos C y E, trazando EQ paralela a CM, los nombres de las cantidades indeterminadas x e y convendrán igualmente a las líneas EQ y QA que a CM y MA, pues PE es igual a PC por ser del círculo; si bien, buscando las líneas EQ y QA por PE y PA que se suponen dadas, se tendrá la misma ecuación que si se buscara CM y MA por PC y PA, de lo que se deduce, evidentemente, que el valor de x o de y, o de cualquier otra cantidad que se suponga, será doble en esta ecuación, es decir que habrá dos raíces desiguales entre sí, de las que una será CM y la otra EQ, si es x que se busca; o bien una será MA y la otra QA, si es y; y así las otras. Es cierto que si el punto E no se encuentra del mismo lado de la curva que el punto C, no habrá más que una de estas raíces que sea verdadera y la otra será opuesta o menor que cero; pero cuanto más próximos estén estos dos puntos el uno del otro, tanto menor diferencia habrá entre las dos raíces; y, ellas serán enteramente iguales, si ellos están juntos en uno, es decir si el círculo que pasa por C toca la curva CE sin llegar a cortarla.*
>
> *Además, debe considerarse que cuando hay dos raíces iguales en una ecuación, ella tiene necesariamente la misma forma que si se multiplica por si misma la cantidad que se supone ser desconocida menos la cantidad conocida que le es igual; después de lo cual si esta última expresión tiene dimensión inferior a la de la ecuación precedente, se la multiplica por otra suma que tenga tanta dimensión como la que le falta, de modo que pueda haber ecuación separadamente entre cada uno de los términos de la una y cada uno de los de la otra.*»

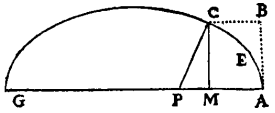
Para cada punto C de la curva hay que determinar el punto P –problema geométrico– que permite trazar el segmento PC, normal a la curva en C, es decir, hay que poner v y s en función de x e y –problema algebraico–. Suponer el problema resuelto permite determinar en

un terreno algebraico la ecuación de un círculo. Regresando a la Geometría, si la recta trazada por P es la normal «*el círculo del cual es centro y que pasa por el punto C, tocará a la línea curva CE sin cortarla*», es decir, será tangente, cortará a la curva en un solo punto, un punto doble (en sentido actual). Ahora, volviendo al Álgebra, para que el círculo «*toque*» (sea tangente) a la curva es preciso que la ecuación resultante tenga una raíz doble. Esta ecuación resultante proviene del sistema de ecuaciones formado por la ecuación de la curva y la ecuación del círculo.

Aquí aparece otro principio fundamental de la Geometría Analítica: «la intersección de curvas –que es un problema geométrico– se reconduce a la resolución de sistemas de ecuaciones – que es un problema algebraico–.

Descartes resolverá el problema algebraico final que se plantea mediante *el método de coeficientes indeterminados*, y en este tema, así como en el asunto paralelo de las raíces dobles, también es un pionero.

Continúa Descartes aplicando el método a la elipse (G.AT,VI, 419):

> *Así, por ejemplo, digo que la primera ecuación encontrada más arriba* [la de la elipse]*, a saber*
>
> $$y^2 + \frac{qry - 2qvy + qv^2 - qs^2}{q - r}$$
>
> 
>
>
> *debe tener la misma forma que la que se obtiene haciendo e igual a y, y multiplicando y–e por sí misma: de lo que resulta*
>
> $y^2 - 2ey + e^2$,
>
> *de manera que se pueden comparar separadamente cada uno de sus términos y decir que, puesto que el primero, que es y^2 es el mismo en la una y en la otra; el segundo que en una expresión es:*
>
> $$\frac{qry - 2qvy}{q - r}$$
>
> *es igual al segundo de la otra que es –2ey.*
>
> *De donde, buscando la cantidad v que es la línea PA, se tiene*
>
> $$v = e - \frac{r}{q}e + \frac{1}{2}r,$$
>
> *o bien, por haber nosotros supuesto e=y, se tiene*
>
> $$v = y - \frac{r}{q}y + \frac{1}{2}r,$$
>
> *y también podría encontrarse s por el tercer término:*
>
> $$e^2 = \frac{qv^2 - qs^2}{q - r}$$
>
> *pero, puesto que la cantidad v determina bien el punto P, que es el único que buscamos, no hay necesidad de proseguir.*»

De forma análoga Descartes realiza la argumentación y el cálculo para la parábola y la Concoide de Nicomedes (G.AT,VI, 420-424).

DESCARTES EN LA PRENSA

Caricatura de Descartes que publicó el 23 de marzo de 1996 la sección TO de la revista LA ESFERA del Diario *EL MUNDO* de Madrid, con motivo del cuarto centenario de su nacimiento.

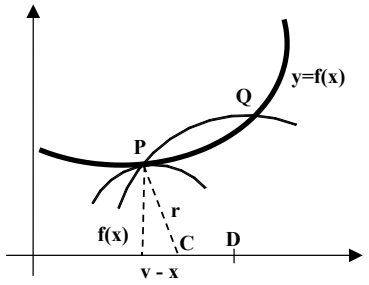
A Descartes se le considera, junto con Fermat, el fundador de la Geometría Analítica. En el propio ámbito de *La Geometría*, Descartes hizo trascendentales contribuciones a la *Teoría de Ecuaciones*, donde vislumbró importantes cuestiones como la *regla de los signos* para descifrar el número de raíces negativas y positivas de cualquier ecuación algebraica, la *Regla de Ruffini* y el *Teorema Fundamental del Álgebra*. Además en sus estudios sobre poliedros, parece ser que Descartes llegó a conocer la conocida *Formula de Euler* que relaciona aristas, caras y vértices de un poliedro.

Al inicio del Libro II de *La Geometría* Descartes introduce los llamados *compases cartesianos*, ingenios que tienen la misma precisión que los instrumentos platónicos y que utilizará para la construcción de diversas curvas geométricas de gran importancia en la resolución de ecuaciones correspondientes a problemas geométricos.

Para mayor comprensión del *método del círculo* de Descartes, interpretemos la técnica cartesiana para una función algebraica general, de forma deliberadamente anacrónica en términos del lenguaje moderno. Tendríamos lo siguiente:

Sea la curva y=f(x), y P un punto cualquiera de ella de abscisa x, donde queremos trazar la normal. Descartes supone como siempre el problema resuelto y la solución dada por la recta CP, siendo C=(v,o) la intersección de la normal con el eje de abscisas.

En general un círculo con centro en un punto D próximo a C y que pase por P, cortará a la curva y=f(x), no sólo en P, sino en otro punto Q, cercano a P, pero si CP es la normal a la curva en el punto P, este punto será un punto doble de la intersección de la curva y=f(x) y el círculo $(x–v)^2 + y^2=r^2$.



Eliminando la y de ambas ecuaciones resulta que la ecuación

$$[f(x)]^2 + (v–x)^2 = r^2 \qquad (1)$$

donde v,r, son fijos, debe tener la abscisa x de P como raíz doble.

Pero una función algebraica con una raíz doble x=e, debe ser de la forma: $(x–e)^2\cdot\Sigma b_n x^n$,

de modo que se puede imponer la condición de raíz doble anterior en la forma:

$$[f(x)]^2 + (v–x)^2 – r^2 = (x–e)^2\cdot\Sigma b_n x^n \qquad (2).$$

Identificando coeficientes se encuentra el valor de v, en términos de la raíz doble e.

En general mediante el método de Descartes lo que se halla es la «*subnormal*» v–x, que permite hallar la pendiente de la normal: –f(x)/(v–x) y de ésta la pendiente de la tangente, es decir, nuestra derivada: (v–x)/f(x).

La condición de raíz doble sobre (1) hoy la impondríamos (utilizando las derivadas formales de una curva algebraica), aplicando que *«toda raíz doble de una función es raíz de su derivada*», por tanto de (1) se deduce:

$$2f(x)\cdot f'(x) – 2(v–x) = 0,$$

y de aquí f'(x)=(v–x)/f(x), obteniéndose el mismo valor que antes para la pendiente de la tangente.

Apliquemos la técnica de Descartes, con lenguaje actual, a diversos casos sencillos.

A_1.- La parábola $y=x^2$.

La ecuación (2) ahora se escribe: $x^4 + (v-x)^2 - r^2 = (x-e)^2\cdot(x^2 + bx + c)$.

Identificando coeficientes se obtiene: $v=2e^3 + e$, sustituyendo e=x, la «*subnormal*» vendrá dada por: $v-x=2x^3$, y la pendiente de la tangente en el punto (x,x^2) de la curva será:

$$(v-x)/f(x) = 2x^3/x^2 = 2x.$$

A_2.- La parábola $y^2=2px$.

La ecuación (2) ahora se escribe: $2px + (v-x)^2 - r^2 = (x-e)^2$.

Identificando coeficientes se obtiene v=e+p, sustituyendo e=x, la «*subnormal*» vendrá dada por v–x=p, y la pendiente de la tangente en el punto $(x, \sqrt{2px})$ de la curva será:

$$\frac{v-x}{f(x)} = \frac{p}{\sqrt{2px}} = \frac{p}{y}\ .$$

La ecuación de la tangente a la parábola $y^2=2px$ en el punto (x_o,y_o) será pues: $y - y_o = \frac{p}{y_o}(x - x_o)$, de donde haciendo operaciones resulta:

$yy_o = p(x+x_o)$, expresión habitual de la tangente a la parábola.

B.- La elipse $\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1.$

La ecuación (2) ahora se escribe: $b^2\cdot\left(1 - \frac{x^2}{a^2}\right) + (v-x)^2 - r^2 = (x-e)^2\cdot c$

Identificando coeficientes resulta: $-(b^2/a^2) + 1 = c$, $-2v = -2ec$.

Despejando v se tiene: $v = e\cdot[1-(b^2/a^2)]$.

Sustituyendo e=x, la «*subnormal*» vendrá dada por $v-x=(-b^2x)/a^2$, de modo que la pendiente de la recta tangente en el punto (x,y) será:

$$\frac{v-x}{f(x)} = -\frac{b^2x}{a^2y}.$$

De aquí resulta que la ecuación de la recta tangente a la elipse en el punto (x_o,y_o) se expresará: $y - y_o = -\frac{b^2x_o}{a^2y_o}(x - x_o)$, de donde haciendo operaciones resulta:

$\frac{xx_o}{a^2} + \frac{yy_o}{b^2} = 1$, expresión habitual de la tangente a la elipse.

Hemos visto que la técnica cartesiana, bajo un punto de vista de Geometría Algebraica (diríamos hoy), encuentra las tangentes a las curvas –vía la normal–, mediante la técnica de considerar el doble contacto del círculo osculador como una característica de la normal. De este modo, Descartes obtiene un método de tratar el problema, que al intuir que será el germen de una ciencia futura, le concede una importancia capital, al reconocer que las tangentes y normales a las curvas son rectas que, de alguna forma, imponen sus leyes a las curvas.

El problema del trazado de las normales a una curva en un punto, es considerado el mayor éxito del método cartesiano, marcando una impronta en la génesis de la Geometría Analítica por la capacidad que desarrolla Descartes de establecer puentes de ida y vuelta entre el Álgebra y la Geometría: análisis geométrico de los problemas, síntesis del análisis en el Álgebra de ecuaciones y traducción geométrica de los resultados algebraicos, un magnífico y poderoso diccionario reversible entre dos lenguajes, el geométrico y el algebraico, con la posibilidad de traducir no sólo en el ámbito gramatical –puntos por coordenadas, curvas por ecuaciones–, sino también en el dominio sintáctico –las relaciones entre los elementos geométricos, por ejemplo intersecciones de curvas, se traducen en relaciones entre los correspondientes elementos algebraicos, por ejemplo mediante sistemas de ecuaciones–.

Con un énfasis inusitado Descartes considera que este problema es el más importante, no sólo de cuantos ha resuelto sino de cuantos anhelara descubrir en Geometría (G.AT,VI,413):

> «[...] *Y me atrevo a decir que éste es el problema mas útil y mas general no sólo que yo conozca, sino aun que yo haya anhelado jamás conocer en Geometría*».

Con estos antecedentes, se comprende la sorprendente acritud con que se desarrolló su polémica con Fermat, a partir de la difusión de los métodos de máximos y mínimos ideados por este «*aficionado*» ya que se aplicaban también al trazado de tangentes. El desarrollo de la controversia, en el que participaron casi todos los matemáticos del círculo de Mersenne, tuvo la feliz virtualidad de ir obligando progresivamente a Fermat a aclarar la naturaleza de sus procedimientos, en el curso de lo cual nuevas curvas nacieron para la Geometría Analítica y para el Cálculo Infinitesimal, que simultáneamente estaba eclosionando gracias a toda la parafernalia analítica que ofrecían los métodos de Fermat y Descartes.

EL CARTESIANISMO

Estatua de Descartes. Palacio de Versalles.

El término latinizado del apellido de Descartes ha dado nombre tanto a su doctrina filosófica: el cartesianismo, basada en el método de la razón matemática, como a las aplicaciones geométricas de *La Geometría:* la Geometría cartesiana, llamada en su forma académica *Geometría* Analítica.

Pocos sabios han dejado su nombre a una doctrina filosófica o a una teoría matemática, pero todavía menos han tenido la gloria de verlo adjetivado en el lenguaje coloquial. Cartesiano ha pasado a ser sinónimo de *racional* y *metódico* en el sentido de analista y riguroso. Así se reconoce cuando se habla, por ejemplo, de una mente cartesiana .

DESCARTES SEGÚN SPINOZA: GEOMETRÍA Y FILOSOFÍA

El método de la Filosofía cartesiana basado en la razón matemática impactó en sus contemporáneos, como los filósofos Malebranche y Spinoza, quien en el prólogo de su obra *Los Principios de la Filosofía cartesiana* escribe el siguiente panegírico de Descartes:

> «*Se alzó al fin este astro, el más destellante de nuestro siglo, René Descartes, quien, en primer lugar, mediante un método nuevo, hizo pasar de las tinieblas a la luz cuanto en las Matemáticas había permanecido inaccesible a los antiguos y todo cuanto los contemporáneos habían sido incapaces de descubrir; luego puso los cimientos inquebrantables de la Filosofía sobre los cuales es posible asentar la mayor parte de las verdades en el orden y con la certidumbre de las Matemáticas, tal como él mismo lo demostró realmente y como aparece más claro que la luz del día a todos cuantos han estudiado atentamente sus escritos, cuya alabanza nunca será tan alta como merece.*»

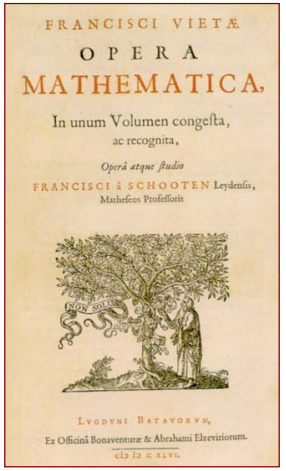
DESCARTES Y LA REINA CRISTINA DE SUECIA

Detalle del cuadro *Cristina de Suecia y su corte*, de P.Dumesnil. Descartes aparece haciendo una demostración geométrica.

Descartes, tomando una decisión excepcionalmente poco cartesiana, acepta la invitación de Cristina de Suecia, que deseaba «*intruirse en Filosofía*», y se desplaza en setiembre de 1649 a Suecia, «*el país de los osos, entre rocas y hielos*». La reina le cita «*en la biblioteca, todas las mañanas, a las cinco*», «*en Suecia se hielan hasta los pensamientos*». Descartes, que tenía la costumbre de levantarse tarde desde la época de estudiante en *La Flèche*, enferma gravemente y muere el 11 de febrero de 1650, oficialmente de una pulmonía, pero tal vez tuvo que ver la añoranza, los caprichos de la reina y los celos de los cortesanos. A pesar de las disputas ente la reina y el embajador de Francia, el cuerpo de Descartes permaneció en Suecia y no fue repatriado hasta 1666. No obstante, el cráneo se quedó en Estocolmo hasta 1822, cuando el químico sueco Berzelius lo regaló a Cuvier. En la actualidad el cuerpo de Descartes descansa en París, en la Iglesia de Saint-Germain-des-Prés, aunque el cráneo se exhibe en el Museo del Hombre.

La Geometría de Descartes y la Geometría Analítica

Descartes participó de los objetivos y propósitos reformadores de Vieta al realizar la construcción geométrica de las raíces de las ecuaciones algebraicas, que empieza, una vez fijada una unidad, por la asignación a cada segmento de una longitud, facilita la asociación implícita del sistema de números reales con los puntos de una línea recta –origen de las coordenadas– y proporciona con esta base un substrato geométrico a las operaciones aritméticas. Después muestra cómo se pueden construir –con instrumentos euclidianos pero con el concurso del Álgebra– las soluciones de las ecuaciones algebraicas, soslayando la necesidad que había en el Álgebra Geométrica griega de conservar la homogeneidad y eliminando la barrera dimensional. Todo ello son ingredientes de lo que mucho después de Descartes se llamaría la Geometría Analítica.


FRANCISCI VIETÆ
OPERA
MATHEMATICA,
In unum Volumen congesta,
ac recognita,
Operâ atque studio
FRANCISCI à SCHOOTEN Leydensis,
Matheseos Professoris
LVGDVNI BATAVORVM,
Ex Officinâ Bonaventuræ & Abrahami Elzeviriorum.

Portada de la primera edición de O*pera mathematica* de Vieta. Fue editada por Van Schooten, en Leyden, en 1646, justo tres años antes de que el propio Van Schooten, y también en Leyden, publicara *La Geometría* de Descartes. Se cree que esta edición de Van Schooten contiene la mayor parte de lo escrito por Vieta, con notaciones nuevas introducidas por el editor.

Con su *Arte Analítica*, Vieta había ya establecido una conexión entre Álgebra y Geometría, al obtener las ecuaciones que corresponden a diversas construcciones geométricas, en el caso de problemas geométricos determinados, es decir manejando sólo ecuaciones determinadas, en las que la variable aunque es una incógnita, es una constante fija a encontrar. Descartes, en *La Geometría* desarrolla esta idea para problemas geométricos indeterminados mediante la consideración de ecuaciones indeterminadas en variables continuas que representan segmentos geométricos. En un sentido general, se puede decir que la invención de la Geometría Analítica por Descartes consiste en la extensión del *Arte Analítica* de Vieta a la construcción geométrica de las soluciones de ecuaciones indeterminadas.

Al partir del rastro de Vieta, Descartes alcanza el principio fundamental de la Geometría Analítica, que expresado en lenguaje moderno, consiste en el descubrimiento de que las ecuaciones indeterminadas en dos incógnitas, f(x,y)=0, se corresponden con lugares geométricos, en general curvas, determinadas por todos los puntos cuyas coordenadas relativas a dos ejes satisfacen la ecuación. Un aspecto de esta idea es anunciado por Descartes en un enunciado básico que viene a decir: «*una ecuación en dos cantidades indeterminadas determina, con respecto a un sistema dado de coordenadas, una curva*», expresado como vimos, en el Libro II de *La Geometría* de la siguiente forma (G.AT,VI, 412):

> «*Para encontrar todas las propiedades de las líneas curvas basta con saber la relación que tienen todos sus puntos con los de las líneas rectas* [la ecuación de la curva]»,

La relación a la que alude Descartes es la ecuación de la curva en un sistema de coordenadas, una expresión algebraica que permite estudiar las propiedades y encontrar los elementos característicos de la curva –diámetros, ejes, centros, normales, tangentes, etc.– como asegura Descartes (G.AT,VI, 412–413):

> «*Luego con sólo saber la relación que tienen todos los puntos de una línea curva con todos los de una línea recta* [la ecuación], *en la forma que he explicado, es fácil también conocer la relación que ellos tienen con todos los otros puntos y líneas dadas; y, por lo tanto, conocer los diámetros, los ejes, los centros,* [...]. *Y también, con sólo*

esto, se puede aun, encontrar casi todo lo que puede ser determinado respecto a la medida del espacio que abarcan, [cuadratura].»

He aquí pues una correspondencia entre las propiedades geométricas de la curva y las propiedades algebraicas de la ecuación asociada que anuncia la esencia de la Geometría Analítica como puente entre el Álgebra y la Geometría y poderoso instrumento de solución de problemas geométricos mediante la intervención del Álgebra, una vez se ha definido un sistema de coordenadas, mediante el que se obtiene la ecuación de la curva como relación algebraica que liga las coordenadas de los puntos de la curva. El carácter algorítmico y operacional del Álgebra convierte a ésta en una potente herramienta no sólo de resolución de problemas geométricos concretos sino también en un magnífico útil de exploración e investigación geométrica, que en esto estriba realmente la eficiencia de la Geometría Analítica.

De esta forma se comprende que a la Geometría Analítica, la «*Geometría de coordenadas*» se le llame Geometría cartesiana. Aunque a tenor de lo que hizo Descartes más bien habría que llamarle «*Geometría de ordenadas*», ya que fijadas las dos incógnitas que componen la ecuación, los segmentos de la primera se miden a partir de un punto inicial –*origen de coordenadas*–, a lo largo de un eje dado y los segmentos de la segunda –que son determinados por la ecuación– se elevan como «*ordenadas*» formando un ángulo con el eje. Así resultan lo que Descartes llama por primera vez en el *Problema de Pappus* «*líneas principales de referencia*» (G.AT,VI,383). No hay, pues, una identidad entre la Geometría Analítica moderna y la de Descartes, incluso son un cierto anacronismo las expresiones actuales como «*sistema de coordenadas cartesianas*». De hecho, Descartes pocas veces utiliza ejes perpendiculares, conocidos como cartesianos, sino que emplea diferentes sistemas de coordenadas, en general oblicuos.

Descartes tenía una opinión muy negativa sobre los métodos sintéticos de los antiguos griegos, por la ocultación del proceso inventivo y la excesiva particularidad, y en consecuencia no participó como hicieron muchos matemáticos coetáneos en el movimiento de restauración de los trabajos perdidos de Apolonio. Descartes era consciente de que su método estaba suplantando a los antiguos. De hecho ése era su propósito desde el principio: no sólo rehacer la Geometría griega, sino crear un nuevo método para la resolución de antiguos y nuevos problemas que rompiera de forma definitiva con la tradición griega y llega incluso a reemplazarla.

Como consecuencia de la aparición de los inconmensurables, el *Álgebra Geométrica* de los griegos estructura casi toda la Matemática griega, con una rigidez que obliga a un tratamiento sintético de los problemas, esclaviza a depender de la naturaleza geométrica intrínseca de las figuras, de modo que cada problema exige un tratamiento local que atomiza la casuística de los casos específicos y precisa de sutiles construcciones geométricas para cada caso particular. Es decir, cada demostración de la Geometría euclídea exigía nuevos e ingeniosos argumentos originales y estaba tan ligada a las figuras que «*que no puede ejercitar el entendimiento sin fatigar mucho la imaginación*», como diría Descartes (DM.AT,VI,17–18). Pero lo más grave era la ocultación del procedimiento y el método de descubrimiento como manifiesta en la Regla IV de las *Regulae* (RIV.AT.X.376-377). Incluso Descartes llega a decir que «[...] *los antiguos no poseían un verdadero método, ...,*» sino «*ellos no hubieran escrito libros tan voluminosos* [para resolver las cuestiones geométricas],» (G.AT,VI,376).

Las Geometría Analítica de Descartes nace, precisamente, de su interés por la metodología. Como escribe Kline (1985, p.51):

> «*Las contribuciones de Descartes a las Matemáticas propiamente dichas no ofrecieron nuevas verdades, sino, más bien, una sólida metodología que ahora llamamos Geometría Analítica.*»

La Geometría de Descartes tienen su anclaje en la Geometría Griega, pero se plantea como

tarea esencial encontrar nuevos métodos más simples, más operativos, más resolutivos, más heurísticos y sobre todo más generales. La intencionalidad de Descartes es palmaria hasta en el propio título de la obra de la que es tributaria *La Geometría*, que con título abreviado se llama *Discours sur la Méthode*, en la que Descartes plasma, de forma clara y distinta, «*el método para dirigir correctamente la razón y buscar la verdad en las ciencias*», es decir, primero el método, después la ciencia que resulta de su fiel aplicación.

Bajo esta filosofía de trabajo Descartes es uno de los principales artífices de la inflexión radical que presenta la Matemática del siglo XVII respecto a la clásica griega, que es ponderada y es la fuente de formación y de inspiración de los matemáticos, pero se abandonan y critican sus métodos porque no son heurísticos. Con un nuevo enfoque se trata ante todo de crear y descubrir, más que de expresar demostrativamente o axiomáticamente. Es más relevante la forma de resolución de los problemas que el estilo de la presentación. Lo que importa es la obtención de métodos que permitan resolver de forma directa y operativa los problemas y escribirlos formalmente siguiendo la línea de la propia investigación geométrica, es decir, métodos que al describir el proceso inventivo enseñen a descubrir y rompan la clásica dualidad helénica *invención-demostración* que tiene lugar en dos estadios de tiempo y espacio diferentes. Se pondera la heurística y se busca afanosamente la fusión, en un solo acto matemático, del descubrimiento y de la demostración. Esto es lo que ante todo persigue Descartes en *El Discurso del Método* la búsqueda de un «*ars inveniendi*», es decir, un método que sirva para descubrir nuevas verdades más que para probar lo que ya se ha hallado. Pues bien, aquí es donde interviene el Álgebra como instrumento inherente a la Geometría Analítica.

En efecto, el Álgebra mecaniza la Matemática de forma que el pensamiento se simplifica y disminuye el esfuerzo de la mente ante la automatización de los procesos. Para Descartes el Álgebra debe preceder a las demás ramas de la Matemática y en cierto modo es una extensión de la Lógica, como motor del razonamiento, en la línea de lo que llamaba Matemática universal (la *Mathesis* de las *Regulae*, RIV.AT.X.376). El Álgebra es la ciencia universal del razonamiento. Y al concretar sobre el ámbito geométrico, el Álgebra es la clave para reconocer los problemas de la Geometría y unificar cuestiones cuya forma geométrica no parece guardar *a priori* relación alguna. Es decir, el Álgebra aporta los principios de clasificación y jerarquía de los problemas y es el instrumento fundamental para discutir con agilidad y generalidad las cuestiones geométricas. El Álgebra simbólica literal, con incógnitas, variables y parámetros, libera de la necesidad de tratar casos específicos y ejemplos concretos y permite formulaciones generales y desarrollos de procedimientos de resolución independientes de la estructura geométrica particular, que posibilita la aplicación de las mismas técnicas a situaciones análogas.

Concretemos aún más qué función cumple el Álgebra en la Geometría Analítica desde el punto de vista del Análisis griego, con el fin de justificar el propio nombre de *Geometría Analítica*, que algunos consideran inapropiado. El término *Análisis* se aplica desde Platón y Pappus para describir el proceso de remontarse desde lo que se desea demostrar hasta llegar a alguna verdad conocida –admitida o probada anteriormente–. En este sentido es lo opuesto a la *Síntesis*, que es la presentación deductiva de lo que se halló mediante el *Análisis*. Es bajo estas concepciones que todavía Vieta y Descartes consideraban el Análisis para describir la aplicación del Álgebra a la Geometría, puesto que el Álgebra era el instrumento adecuado para *analizar* el problema de construcción geométrica. En efecto, Vieta había definido el *Análisis* como «*doctrina bene inveniendi in mathematica*». Para Vieta el *Arte Analítica* constaba de tres partes: *zetetic* donde se determinan las propiedades de los elementos que pide el problema a partir de las propiedades de los datos, *poristic* que es el proceso de verificación y *exegetic* que es la demostración de la proposición. Tal como había sido usado por Platón y Pappus, la palabra *Análisis* hacía referencia al orden de las ideas en una demostración. El *Análisis* es la descomposición en elementos más simples que se hace en el camino de la investigación, la *Síntesis* es la composición o reordenación que se hace en la exposición. Vieta aplicará la palabra *Análisis* en la Geometría algebraica, a la que mira como una nueva forma de Análisis matemático y usa el término bajo el significado de los griegos, pero remarca que en la fase *zetética* de ataque algebraico del problema se procede indirectamente a base de asumir lo que se quiere probar o construir y se maneja y se opera

con las cantidades incógnitas como si fueran conocidas.

Descartes reproducirá, en el segundo epígrafe del Libro I de *La Geometría* titulado «*cómo se llega a las ecuaciones que sirven para resolver los problemas*», estas ideas en torno a la aplicación del *Análisis*: (G.AT,VI,372):

> «*Así, si se quiere resolver algún problema, debe de antemano considerarse como ya resuelto, y dar nombre a todas las líneas que parecen necesarias para construirlo, tanto a las que son desconocidas como a las otras. Luego, sin considerar ninguna diferencia entre estas líneas conocidas y desconocidas, se debe examinar la dificultad según el orden que se presente como más natural de todos, en la forma como aquellas líneas dependen mutuamente las unas de las otras, hasta que se haya encontrado la manera de expresar una misma cantidad de dos maneras: lo que se denomina una ecuación, pues* [el resultado de] *los términos de una de esas dos formas son iguales a los de la otra.*»

Así pues, tanto para Vieta como para Descartes, el Álgebra se convierte en el instrumento adecuado para emprender el camino analítico en Geometría, que de acuerdo con las palabras mencionadas de Descartes, debe seguir un protocolo de actuación. Señalemos, punto por punto, los pasos a seguir, según Descartes, para resolver cualquier problema geométrico, que en esencia son casi los mismos que realizamos de forma académica cuando resolvemos actualmente cualquier problema de Geometría Analítica:

a) se da nombre a todos los segmentos que parecen necesarios;

b) se supone el problema resuelto, es decir, se supone conocida la longitud buscada;

c) se plantea la ecuación entre las longitudes conocidas y desconocidas;

d) se resuelve esta ecuación;

e) se concluye con la construcción geométricamente de la solución.

Éste es el camino que sigue el método cartesiano en el que el estudio analítico se funde con la síntesis algebraica en transición de la Geometría al Álgebra y del Álgebra a la Geometría. El Análisis mediante el Álgebra traduce los datos geométricos de forma que sean tratables por medio del automatismo del cálculo algebraico, esto es el Análisis Algebraico. Se comprende, pues, el nombre de *Geometría Analítica* que en el curso de la Historia se le dio al instrumento desarrollado por Descartes, aunque tal vez hubiera sido mas descriptivo el de Geometría Algebraica –que curiosamente resultaría de la permutación de los términos de Álgebra Geométrica con que se nombra buena parte de la Matemática de *Los Elementos* de Euclides–, aunque este nombre también sería deficiente, toda vez que la Geometría Analítica es mucho más que una mera aplicación del Álgebra a la Geometría, ya que requiere el uso de coordenadas. En suma, la Geometría Analítica sería el Análisis moderno, siendo el Álgebra por su carácter algorítmico el principal instrumento de la aplicación de ese Análisis, por eso también se podría definir con mayor precisión como la aplicación del Análisis Algebraico a la Geometría. Históricamente, hasta muy tarde se han utilizado los términos Álgebra y Análisis como sinónimos. Así aparecen, por ejemplo, en la famosa *Encyclopédie*, donde D'Alembert escribe:

> «*El Análisis es propiamente el método de resolver los problemas matemáticos, reduciéndolos a las ecuaciones. El Análisis para resolver problemas, emplea el recurso del Álgebra, o cálculo de las magnitudes en general: estas dos palabras, Análisis y Álgebra, son a menudo miradas como sinónimas* [...] *Algunos autores definen el Álgebra (como siendo) el arte de resolver los problemas matemáticos: pero ésta es la idea del Análisis o del arte analítico más bien que del Álgebra.*»

Cuestiones nominalistas aparte, volvamos a los orígenes para reiterar que la Geometría Analítica recibe su nombre y sus procedimientos del método de *Análisis* de los griegos y

permite recuperar el Análisis Geométrico de los antiguos mediante la acción del Álgebra, ya que el carácter algorítmico de ésta promociona y acentúa las aptitudes heurísticas del Análisis. Así lo observamos claramente en *La Geometría*, donde uno se convence de la posibilidad cartesiana de reconstruir toda la Geometría con una simplicidad sorprendente como el propio Descartes asegura (G.AT,VI,376):

> «[...] *Se pueden construir todos los problemas de la geometría ordinaria, sin hacer más que lo poco que está comprendido en las cuatro figuras que he explicado*»,

y con unos instrumentos francamente modestos, sólo los Teoremas de Tales y de Pitágoras, como indica a la Princesa Elisabeth en la comunicación epistolar de noviembre de 1643:

> «*Yo no considero otros teoremas que los lados de los triángulos semejantes están en proporción, y que, en los triángulos rectángulos, el cuadrado de la base es igual al cuadrado de los dos lados*».

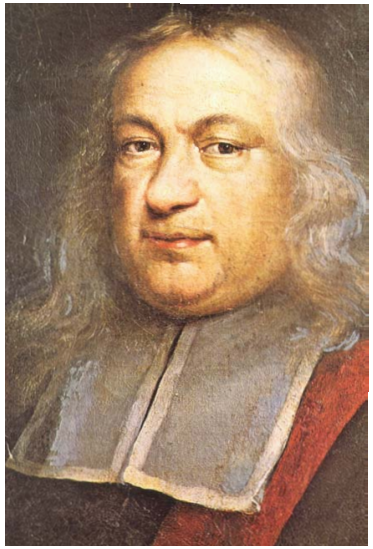
Descartes hace esta observación respecto de la resolución del problema de Apolonio, pero lo mismo se puede decir de *La Geometría* en general.

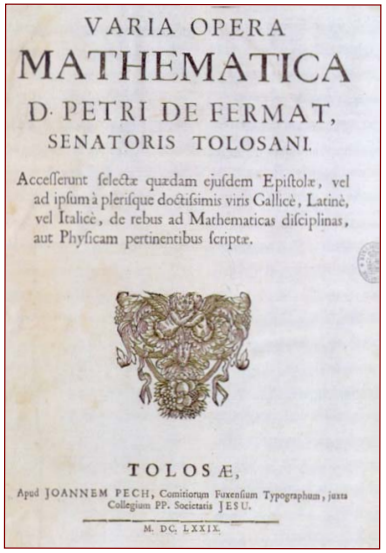
Es asombroso ¡cómo se puede hacer tanto con tan poco! Y es que el enfoque analítico, siempre con el recurso algorítmico del Álgebra simbólica, permite la generalización de los métodos y la aplicación uniforme de los mismos procedimientos a cuestiones similares.

Durante algunos años después de 1637, la Geometría Analítica fue considerada como la invención de un solo hombre –Descartes–, debido a que los trabajos de Fermat sobre Geometría Analítica –la *Introducción a los Lugares Planos y Sólidos*–, conocida como *Isagoge*, no fueron publicados en vida del autor. Por ello es difícil aquilatar el grado de influencia que tuvo sobre sus contemporáneos. No obstante, tanto la *Isagoge* como los trabajos de Fermat sobre máximos y mínimos y su aplicación a las tangentes fueron conocidos –por voluntad de Fermat a través de manuscritos que acompañaban a su correspondencia con Mersenne–, por el círculo de matemáticos de París, incluso antes de la aparición de *La Geometría* de Descartes, pero así como las tangentes de Fermat causaron una gran impresión sobre todo en Descartes, la *Isagoge* parece que fue rápidamente eclipsada por el trabajo de Descartes. Mientras algunos aspectos de los máximos y mínimos y las tangentes de Fermat fueron incorporados a algunas publicaciones de otros matemáticos, la *Isagoge* no aparece en imprenta hasta la publicación de *Varia Opera Mathematica* de Fermat por parte de su hijo Samuel en 1679, catorce años después de la muerte de su autor, cuarenta y dos años después de la publicación de *La Geometría* de Descartes y casi cincuenta años después de ser escrito el tratado, en unos momentos en que la influencia cartesiana se había extendido notablemente, de modo que la memoria de Fermat sobre Geometría Analítica, ya incluso con una notación obsoleta, tenía simplemente un valor histórico para atestiguar (debido a la fecha de composición y a su contenido) la independencia de la Geometría Analítica de Fermat respecto de la de Descartes. Así pues, el nombre de Geometría cartesiana con que se denomina a veces a la Geometría Analítica no hace justicia a ambos fundadores, incluso entre profesionales de la Matemáticas se desconoce, a veces, la copaternidad de Fermat, pero es bien cierto que fue bajo la forma cartesiana como este magnífico instrumento se impuso y echó raíces en la Matemática.

Al contrario que la *Isagoge, La Geometría* de Descartes tuvo una rápida difusión, de modo que por la importancia que se dio a la obra, enseguida aparecieron nuevas ediciones que recibieron infinidad de comentarios por parte de matemáticos coetáneos. Además, no todo el mundo entendía la obra de Descartes, de modo que incluso algunos eruditos solicitaron aclaraciones para poderla seguir. Estas preocupaciones latentes en los ámbitos matemáticos propiciaron el que van Schooten –que había sido el diseñador de las figuras de la primera edición– añadiera, a su traducción latina de 1649, toda una serie de comentarios propios, las *Notas Breves* de F. de Beaune y aportaciones de Witt, de Hudde, de van Heuraet y otros, que contribuyeron a extender su difusión e incrementar su inteligibilidad. Tanto éxito tuvo la publicación de van Schooten que se reeditó en 1659 y 1695.

LA GEOMETRÍA ANALÍTICA DE FERMAT DE LA *INTRODUCCCIÓN A LOS LUGARES PLANOS Y SÓLIDOS*




VARIA OPERA
MATHEMATICA
D. PETRI DE FERMAT,
SENATORIS TOLOSANI.
Accesserunt selectæ quædam ejusdem Epistolæ, vel ad ipsum à plerisque doctissimis viris Gallicè, Latinè, vel Italicè, de rebus ad Mathematicas disciplinas, aut Physicam pertinentibus scriptæ.
TOLOSÆ,
Apud JOANNEM PECH, Comitiorum Fuxensium Typographum, juxta Collegium PP. Societatis JESU.
M. DC. LXXIX.

1. **Retrato de Fermat como consejero del parlamento de Toulouse (atribuido a Antoine Durand). Académie des Sciences et Belles Lettres de Toulouse.**
2. **Edición de Samuel de Fermat de *VARIA OPERA MATHEMATICA* de *D. PETRI DE FERMAT*. Tolosa, 1679.**

La particular forma que tenía Fermat de trabajar en Matemáticas, así como de comunicar sus descubrimientos, unida a la despreocupación por la conservación de sus papeles y la constante reticencia en torno a su eventual publicación –Fermat no escribió grandes tratados, sino apuntes episódicos y notas marginales–, supuso que a su muerte en 1665 gran parte de su trabajo quedara desperdigado en numerosos ambientes científicos de toda Europa. Por esta razón su influencia directa no tuvo la envergadura y la inmediatez de la de Descartes.

Catorce años después de la muerte de su padre, habiendo reunido la mayor parte de los escritos latinos, así como un número suficiente de cartas inéditas, Samuel de Fermat hizo imprimir en 1679 *Varia Opera Mathematica*, que a pesar de las lagunas de importantes desarrollos de Fermat y de las excesivas incorrecciones –Samuel no era matemático–, constituyó hasta finales del siglo XIX –se reimprimió en 1861– la única publicación donde se podían estudiar los trabajos de Fermat

La Geometría Analítica de Fermat tiene su origen en su profundo conocimiento de la Geometría de Apolonio y Pappus y del *Arte Analítica* de Vieta. Fermat se dio cuenta de que las relaciones de áreas, expresadas según el Álgebra Geométrica de los griegos en forma de proporción, mediante las que Apolonio escribía las propiedades intrínsecas de las cónicas se prestaban con gran facilidad a ser traducidas en el lenguaje de ecuaciones del Álgebra simbólica de Vieta. De esta forma el *symptoma* de la curva de *Las Cónicas* de Apolonio, forma retórica de la expresión de la curva en el lenguaje pitagórico de la *Aplicación de las Areas*, evolucionaba hacia la *ecuación característica* de la curva de la *Introduccción a los Lugares Planos y Sólidos* (*Ad Locos Planos et Solidos Isagoge*) de Fermat, memoria que contiene la llamada Geometría Analítica de Fermat .

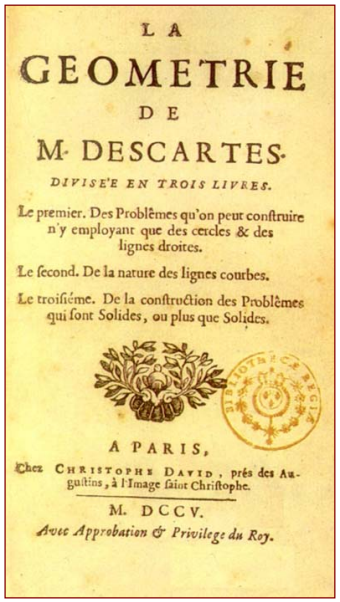
Al vincular los trabajos matemáticos de Vieta y Apolonio, Fermat alumbra su Geometría Analítica que establece un efectivo puente entre la Geometría y el Álgebra, lo que le permitirá la asociación de curvas y ecuaciones, a base de aplicar el Análisis algebraico de Vieta a los problemas de lugares geométricos de Apolonio y Pappus, definidos, en un sistema de coordenadas, por una ecuación indeterminada en dos incógnitas. De este modo Fermat resolverá los problemas del Análisis Geométrico de los antiguos mediante la mecánica operatoria del Álgebra simbólica.

Con la Geometría Analítica de Fermat se alcanzaba el máximo grado de consumación en la aplicación a los problemas geométricos del antiguo método de *Análisis* (de ahí procede el adjetivo Analítica que acompaña al sustantivo Geometría), siendo el Álgebra por su carácter algorítmico el principal instrumento de la aplicación de ese Análisis.

La Geometría Analítica se convierte enseguida, en la mente de Fermat, en una poderosa herramienta heurística de investigación, mediante la cual él mismo resolverá de forma prodigiosa y brillante, numerosos problemas, antiguos y nuevos, en particular numerosas cuestiones de lugares geométricos, máximos y mínimos, tangentes, cuadraturas y cubaturas, centros de gravedad y problemas de rectificación de curvas.

NUEVAS EDICIONES DE *LA GEOMETRÍA* DE DESCARTES

RENATI DES CARTES
GEOMETRIA,
Una cum Notis
FLORIMONDI DE BEAUNE,
In Curia Blesensi Consiliarii Regii, & Commentariis illustrata,
Operâ atque studio
FRANCISCI à SCHOOTEN,
in Acad. Lugd. Batav. Matheseos Professoris.
AB EODEM DUM VIVERET DILIGENTER RECOGNITA,
locupletioribus Commentariis instructa, multaque egregiis accessionibus, tam ad uberiorem explicationem, quàm ad amplificandam hujus Geometriæ excellentiam facientibus exornata.
Nunc verò à Viro Clarissimo denuo revisa, & ab innumeris mendis, quibus priores Editiones scatebant, repurgata, una cum notis quibusdam & animadversionibus cumultuariis in universum Opus, huic quartæ editioni recens adjectis.
Accedit insuper
COMPENDIUM MUSICÆ.
Cum Gratia & Privilegio Sacræ Cæs. Majest.
FRANCOFURTI AD MOENUM,
Sumptibus FRIDERICI KNOCHII, Bibliop.


LA
GEOMETRIE
DE
M. DESCARTES.
DIVISE'E EN TROIS LIVRES.
Le premier. Des Problêmes qu'on peut construire n'y employant que des cercles & des lignes droites.
Le second. De la nature des lignes courbes.
Le troisiéme. De la construction des Problêmes qui sont Solides, ou plus que Solides.
A PARIS,
Chez CHRISTOPHE DAVID, prés des Augustins, à l'Image saint Christophe.
M. DCCV.
Avec Approbation & Privilege du Roy.

1. **Edición latina de 1695 de van Schooten de *La Geometría* de Descartes.**
2. **Edición francesa de 1705 de *La Geometría* de Descartes.**

A diferencia de las obras de Fermat, *La Geometría* de Descartes tuvo numerosas ediciones, tanto en latín como en francés, algunas de ellas con prolijos comentarios para hacerla más inteligible, es decir, que eran auténticas ediciones críticas. Por ello los rudimentos de Geometría Analítica de *La Geometría* de Descartes recibieron una amplia difusión.

La Edición latina de 1695 de van Schooten contiene entre otros elementos los siguientes:

- **Geometria, una cum notis Florimondi De Beaune.**
- **Francisci à Schooten In Geometriam Renati Des Cartes Commentarii.**
- **Johannis Huddenii Epistola prima de Reductione Æquationum.**
- **Johannis Huddenii Epistola secunda de maximis et minimis.**
- **Renati Des Cartes Principia Matheseos Universalis seu Introductio ad Geometriæ Methodum**
- **conscripta ab. Er. Bartholino. De Æquationum Natura, Constitutione, et Limitibus Opuscula Duo.**
- **Incepta à Florimondo De Beaune ab Erasmio Bartholino.**
- **Johannis De Witt Elementa Curvarum Linearum edita operà Francisci à Schooten.**
- **Francisci à Schooten Tractatus de Concinnandis Demonstrationibus Geometricis ex Calculo Algebraico.**

La proyección histórica de la Geometría Analítica cartesiana

Al partir del *Análisis Geométrico* griego como trampolín y punto de arranque y al utilizar el *Arte Analítica* de Vieta como instrumento algorítmico básico, Descartes conduce el *Análisis* a su máximo poder heurístico para la resolución de los problemas geométricos, a base de aunar el estudio analítico con la síntesis algebraica, lo que permitirá, mediante las ecuaciones, transitar de la Geometría al Álgebra y del Álgebra a la Geometría. La Geometría Analítica resultante, dotada del simbolismo literal, con todo el potencial de la mecánica algorítmica operatoria de cálculo, manipulación y simplificación propias de las ecuaciones del Álgebra, reemplaza las ingeniosas construcciones geométricas de la rígida, farragosa y retórica *Álgebra Geométrica* de los griegos por sistemáticas operaciones algebraicas que permiten mediante un proceso analítico-sintético de resolución de los problemas, no sólo reconstruir la Geometría clásica con más claridad, flexibilidad, operatividad y versatilidad, sino crear, además, una potente heurística geométrica mediante la cual Descartes pudo plantear y resolver de forma admirable, brillante y prodigiosa problemas difíciles, algunos clásicos –el *Problema de Pappus* y el *Problema de Apolonio*, entre otros–, y otros nuevos como la determinación de las rectas normales a las curvas.

Descartes creará, además, las herramientas geométrico-algebraicas para resolver con eficacia otros muchos problemas como los de lugares geométricos, el estudio de elementos notables de las curvas y los problemas infinitesimales –extremos, tangentes, cuadraturas y cubaturas, centros de gravedad, rectificación, etc.– de gran interés e importancia en los ambientes científicos de la primera mitad del siglo XVII. En este sentido, la Geometría Analítica cartesiana tuvo una decisiva influencia, según veremos, como instrumento clave de la eclosión de multitud de métodos y técnicas infinitesimales, que convergen en la invención del Cálculo Infinitesimal.

Una de estas herramientas fundamentales es precisamente la nueva notación cartesiana. Descartes aplica a los problemas y a las ecuaciones un nuevo y potente simbolismo simplificador, explicativo y resolutivo, que va mucho más allá de la abreviatura cósica iniciada por Diofanto y desarrollada por los algebristas italianos e incluso allende la escritura simbólica de Vieta. Los símbolos y términos de la matemática son el soporte de sus conceptos y métodos, por tanto tienen una gran importancia, y a pesar de la arbitrariedad en la elección de los signos, conviene adoptar un criterio unificador, que al ser adoptado universalmente, facilita la interpretación y la comprensión, ahorra tiempo y espacio, entraña economía de pensamiento y permite una mayor y más rápida difusión. Esto es precisamente lo que consiguió Descartes con los convenios notacionales fijados en *La Geometría*, que han tenido la virtualidad de convertirse en algo poderosamente definitivo, de modo que *La Geometría*, es el primer texto matemático en el que un lector actual no encontraría dificultades con la notación.

Como escribe E.Colerus, en un lenguaje casi místico, en su *Breve Historia de la Matemática* (Vol. II, pág.17, Doncel, Madrid, 1973):

> «*La Matemática no es sino una obra mágica del pensamiento, y los espíritus aparecen cuando se les invoca con las fórmulas adecuadas*.»

La notación de Descartes fue la fórmula oportuna para su magno proyecto de reforma que alcanzó a una completa reconstrucción de la Matemática sobre premisas muy sencillas, no geométricas como en Euclides, sino algebraicas. La *Geometría* de Descartes elimina toda una serie de limitaciones que encorsetaban a la Geometría griega:

- Limitación pitagórica de la inconmensurabilidad.
- Limitación platónica de los instrumentos geométricos –regla y compás–.
- Limitación euclídea de la homogeneidad dimensional.
- Limitación tridimensional.
- Limitación de la dependencia de las figuras geométricas.
- Limitación de la imposibilidad de asignar números a las figuras geométricas.

A título de ejemplo sobre lo que introduce la Geometría Analítica en el panorama matemático pensemos en el estudio de las curvas, un tema de importancia esencial en la Geometría de toda época y en particular en los estudios infinitesimales. Las limitaciones algebraicas del Álgebra Geométrica de los griegos, consecuencia de los inconmensurables, hizo imposible la introducción en el mundo griego de nuevas curvas por medio de ecuaciones. Las curvas se obtenían constructivamente mediante intersección de superficies y lugares geométricos y también a través de relaciones de áreas o longitudes, que daban la propiedad de definición de la curva, De esta forma, el elenco de curvas que manejaron los griegos hubo de ser necesariamente muy limitado –*las cónicas* de Menecmo y Apolonio, *la espiral* de Arquímedes, *la cuadratriz* de Hipias o Dinóstrato, *la cisoide* de Diocles, *la hipopede* de Eudoxo, *la concoide* de Nicomedes, y pocas más–.

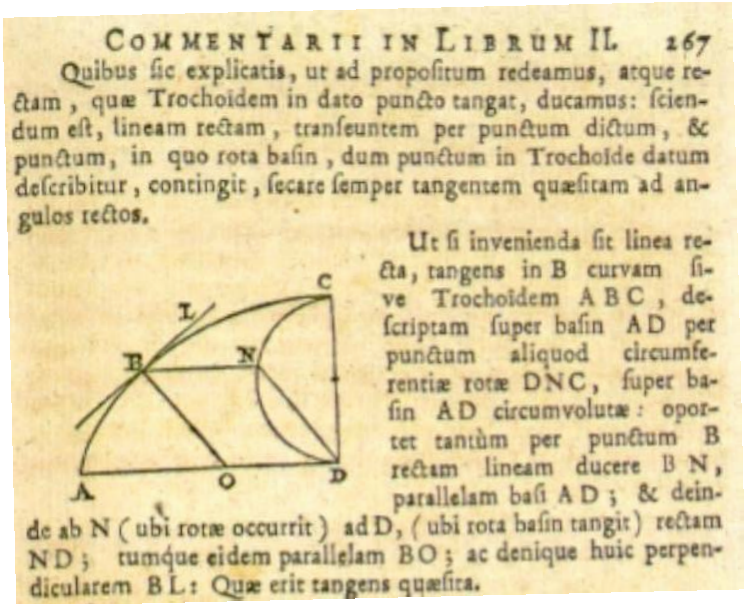
La investigación infinitesimal en los albores del siglo XVII tiene lugar con el planteamiento y resolución de problemas de cuadraturas y tangentes sobre curvas. Cabe decir que pioneros de los métodos y técnicas del Cálculo del siglo XVII, como Kepler e incluso Cavalieri, no tuvieron a su disposición los desarrollos geométricos de Descartes, de modo que el número de curvas que manejaron y a las que podían aplicar las técnicas algorítmicas del Cálculo que iban descubriendo era muy limitado, prácticamente las mismas que conocieron los griegos. Además, todavía manejaron las curvas en el farragoso lenguaje del *Álgebra Geométrica* griega mediante relaciones de áreas y proporciones. El trabajo de Descartes en *La Geometría* abre el camino a la introducción sistemática de nuevas curvas y a un manejo más útil, sencillo y operativo, mediante las ecuaciones de las curvas. En efecto, de acuerdo con el *Principio Fundamental de la Geometría Analítica* las curvas planas están determinadas por la ecuación canónica asociada, y por tanto, por el simple hecho de escribir una ecuación una nueva curva queda definida en el ambiente geométrico para la indagación de problemas infinitesimales vinculados a ella, de modo que aparece en el ambiente matemático de los dos primeros tercios del siglo XVII una ingente cantidad de nuevas curvas que se definen a propósito de la introducción de la Geometría Analítica, entre las que sobresalen: el caracol de Pascal, el folium de Descartes o galande de Barrow, la curva de Lamé, la espiral logarítmica, la kappa-curva, la curva tangentoidal, pero sobre todo las parábolas, hipérbolas y espirales generalizadas o de orden superior –llamadas de Fermat por ser él quien las introdujo– y por encima de todas ellas, en cuanto a importancia, la cicloide, la reina de todas las curvas, llamada la *Helena de la Discordia*, por las polémicas que surgieron sobre cuestiones de prioridad y acusaciones de plagio acerca de la resolución de problemas vinculados a ella. El vasto conjunto de nuevas curvas promueve la aparición de multitud de variadas técnicas algorítmicas infinitesimales al disponer de un amplio material geométrico al que aplicarlas. Se comprende, pues, la amplitud panorámica que sobreviene en el ámbito matemático con la emergencia de la Geometría Analítica.

La propia Geometría Analítica en sí misma era un instrumental algorítmico de primer orden, por eso jugó un papel decisivo en la investigación Infinitesimal. Las Geometría Analítica de Descartes permite utilizar la expresión algebraica de la ecuación de una curva para encontrar sus elementos geométricos más notables –diámetros, ejes, centros, etc.– y, en particular, en el terreno infinitesimal resolver los problemas de cuadraturas y tangentes relacionadas con la curva. Es decir, la ecuación de la curva es un elemento esencial para esclarecer las propiedades y encontrar los elementos relevantes de la curva. La Geometría Analítica traslada los problemas infinitesimales de la Geometría al Álgebra, la cual por su carácter operacional, permite, tras la realización de cálculos y en particular la resolución de ecuaciones, regresar a la geometría del problema, para encontrar y solucionar cuestiones geométricas. Como consecuencia, la tarea de probar un teorema o resolver un problema geométrico de índole infinitesimal se conduce de forma muy eficiente a probarlo o resolverlo mediante el Álgebra, de modo que la aplicación de la Geometría Analítica proporciona una potente técnica de resolución de problemas infinitesimales, y algo que es todavía más importante, un poderoso instrumento de investigación geométrica en el ámbito infinitesimal.

La Geometría Analítica desarrollada por Descartes tuvo, pues, un papel decisivo en todo este proceso de alumbramiento de las técnicas del Cálculo del siglo XVII . El impacto del Álgebra tiene lugar no sólo sobre la Geometría sino también sobre el Cálculo Infinitesimal a través de la propia Geometría Analítica, apareciendo como resultados positivos los intentos

de aritmetización del método de exhaución de los griegos, que conducirá a la utilización incipiente y subrepticia de los límites. Se comprende entonces por qué la aparición de la Geometría Analítica en el horizonte matemático del siglo XVII tuvo una incidencia capital en la aparición de multitud de métodos y técnicas infinitesimales que condujeron al descubrimiento del Análisis Infinitesimal por parte de Newton y Leibniz.

Puede decirse que el Cálculo anterior a Newton y Leibniz es una ingente casuística de métodos heurísticos, aplicados a problemas geométricos específicos, que se resuelven mediante técnicas *ad hoc* vinculadas a las correspondientes figuras geométricas, que desarrollan multitud de resultados particulares que, al traducirlos al lenguaje moderno, muestran los conceptos esenciales del Cálculo, que de alguna manera yacían en ellos, pero de forma tan fragmentaria que sólo se referían a problemas individuales y no a teorías generales, aunque la perspectiva de generalización estaba implícita en esos métodos. Es precisamente la Geometría Analítica de Descartes la que favorece este proceso de búsqueda del algoritmo válido en general e independiente de la estructura geométrica intrínseca de cada problema. La generalidad del Álgebra frente a la especificidad de la Geometría, permite, por ejemplo, que en la traducción geométrico-algebraica en que consiste la Geometría Analítica, cada caso particular del trazado geométrico de la tangente, que es diferente y específico para cada curva, de acuerdo con su naturaleza geométrica, deje de serlo y se pueda aplicar, mediante un proceso analítico, el mismo procedimiento a todas las curvas de las que se conozca su expresión analítica –su ecuación–, es decir, el proceso algorítmico de cálculo de una derivada. He aquí una muestra muy significativa de la trascendencia de la Geometría Analítica como herramienta que simplifica y reduce una extensa tipología de problemas geométricos –el trazado de las tangentes de las diversas curvas– a un único y concreto problema analítico –el cálculo de la derivada–.



COMMENTARII IN LIBRUM II. 267

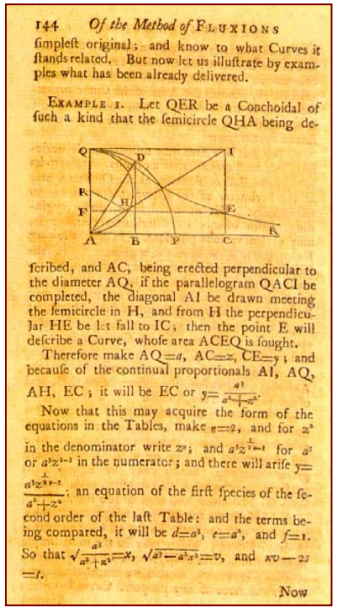
Quibus ſic explicatis, ut ad propoſitum redeamus, atque rectam, quæ Trochoidem in dato puncto tangat, ducamus: ſciendum eſt, lineam rectam, tranſeuntem per punctum dictum, & punctum, in quo rota baſin, dum punctum in Trochoide datum deſcribitur, contingit, ſecare ſemper tangentem quæſitam ad angulos rectos.

Ut ſi invenienda ſit linea recta, tangens in B curvam ſive Trochoidem ABC, deſcriptam ſuper baſin AD per punctum aliquod circumferentiæ rotæ DNC, ſuper baſin AD circumvolutæ: oportet tantùm per punctum B rectam lineam ducere BN, parallelam baſi AD; & deinde ab N (ubi rotæ occurrit) ad D, (ubi rota baſin tangit) rectam ND; tumque eidem parallelam BO; ac denique huic perpendicularem BL: Quæ erit tangens quæſita.

Página de la edición de van Schooten de 1695 de *La Geometría* de Descartes.

Se trata de un apunte de van Schooten donde se explica el cálculo cartesiano de la tangente a la cicloide, sin duda la curva más importante sobre la que se ensayaron los métodos y técnicas infinitesimales que aparecieron a lo largo del siglo XVII como consecuencia del desarrollo de la Geometría Analítica de Descartes.

LA GEOMETRÍA ANALÍTICA Y EL DESCUBRIMIENTO DEL CÁLCULO INFITESIMAL



144 *Of the Method of* FLUXIONS

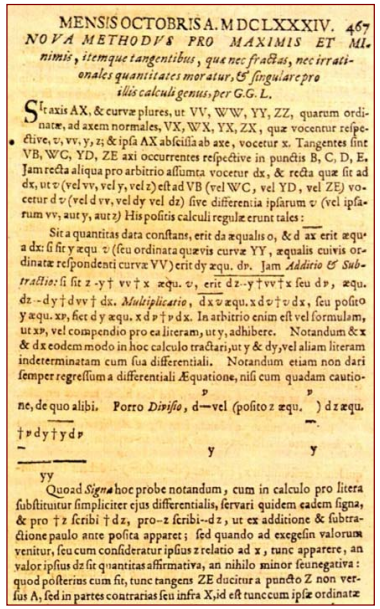
ſimpleſt originals; and know to what Curves it ſtands related. But now let us illuſtrate by examples what has been already delivered.

EXAMPLE 1. Let QER be a Conchoidal of ſuch a kind that the ſemicircle QHA being deſcribed, and AC, being erected perpendicular to the diameter AQ, if the parallelogram QACI be completed, the diagonal AI be drawn meeting the ſemicircle in H, and from H the perpendicular HE be let fall to IC; then the point E will deſcribe a Curve, whoſe area ACEQ is ſought.

Therefore make AQ=*a*, AC=*z*, CE=*y*; and becauſe of the continual proportionals AI, AQ, AH, EC; it will be EC or $y=\frac{a^3}{a^2+z^2}$.

Now that this may acquire the form of the equations in the Tables, make *a*=2, and for z^2 in the denominator write z^n; and $a^1z^{\frac{2}{2}n-1}$ for a^3 or a^1z^{n-1} in the numerator; and there will ariſe $y=\frac{a^1z^{\frac{2}{2}n-1}}{a^1+z^n}$; an equation of the firſt ſpecies of the ſecond order of the laſt Table: and the terms being compared, it will be $d=a^1$, $e=a^1$, and $f=1$. So that $\sqrt{\frac{a^3}{a^2+z^2}}=x$, $\sqrt{a^2-z^2}=v$, and $xv=zs$ $=t$.

Now



MENSIS OCTOBRIS A. MDCLXXXIV. 467

NOVA METHODVS PRO MAXIMIS ET MINIMIS, itemque tangentibus, quæ nec fractas, nec irrationales quantitates moratur, & ſingulare pro illis calculi genus. per G.G.L.

Sit axis AX, & curvæ plures, ut VV, WW, YY, ZZ, quarum ordinatæ, ad axem normales, VX, WX, YX, ZX, quæ vocentur reſpective, *v*, *vv*, *y*, *z*; & ipſa AX abſciſſa ab axe, vocetur *x*. Tangentes ſint VB, WC, YD, ZE axi occurrentes reſpective in punctis B, C, D, E. Jam recta aliqua pro arbitrio aſſumta vocetur *dx*, & recta quæ ſit ad *dx*, ut *v* (vel *vv*, vel *y*, vel *z*) eſt ad VB (vel WC, vel YD, vel ZE) vocetur *dv* (vel *dvv*, vel *dy* vel *dz*) ſive differentia ipſarum *v* (vel ipſarum *vv*, aut *y*, aut *z*) His poſitis calculi regulæ erunt tales:

Sit a quantitas data conſtans, erit *da* æqualis 0, & *d* $\overline{ax}$ erit æqua *dx*: ſi ſit *y* æqu. *v* (ſeu ordinata quævis curvæ YY, æqualis cuivis ordinatæ reſpondenti curvæ VV) erit *dy* æqu. *dv*. Jam *Additio & Subtractio*: ſi ſit *z* -*y*† *vv*† *x* æqu. *v*, erit *dz*-*y*†*vv*†*x* ſeu *dv*, æqu. *dz*-*dy*†*dvv*† *dx*. *Multiplicatio*, *dxv* æqu. *xdv*†*vdx*, ſeu poſito *y* æqu. *xp*, fiet *dy* æqu. *xdp*†*pdx*. In arbitrio enim eſt vel formulam, ut *xp*, vel compendio pro ea literam, ut *y*, adhibere. Notandum & *x* & *dx* eodem modo in hoc calculo tractari, ut *y* & *dy*, vel aliam literam indeterminatam cum ſua differentiali. Notandum etiam non dari ſemper regreſſum a differentiali Æquatione, niſi cum quadam cautione, de quo alibi. Porro *Diviſo*, *d*—vel (poſito *z* æqu.) *dz* æqu. $\frac{\overline{p}\,dy\dagger y\,dp}{yy}$.

Quoad *Signa* hoc probe notandum, cum in calculo pro litera ſubſtituitur ſimpliciter ejus differentialis, ſervari quidem eadem ſigna, & pro †*z* ſcribi †*dz*, pro-*z* ſcribi-*dz*, ut ex additione & ſubtractione paulo ante poſita apparet; ſed quando ad exegeſin valorum venitur, ſeu cum conſideratur ipſius *z* relatio ad *x*, tunc apparere, an valor ipſius *dz* ſit quantitas affirmativa, an nihilo minor ſeunegativa: quod poſterius cum ſit, tunc tangens ZE ducitur a puncto Z non verſus A, ſed in partes contrarias ſeu infra X, id eſt tunc cum ipſæ ordinatæ

1. **Página de *The Metod of Fluxions* de Newton (Londres, 1737), con el trazado de la tangente a la concoide. En esta obra Newton unifica la mayor parte de resultados sobre tangentes y cuadraturas que ocuparon a buena parte de los matemáticos del siglo XVII.**
2. **Página de *Nova Methodus pro Maximis et Minimis* de Leibniz (Leipzig, 1684), donde aparecen las reglas para derivar sumas, productos y cocientes. Esta obra es considerada como la primera publicación sobre Cálculo Infinitesimal de la historia.**

«*Apoyándose en hombros de gigantes*» como Fermat y Descartes, apurando y exprimiendo la capacidad de unificación y generalización que permitían los procedimientos del Álgebra y de la Geometría Analítica, bajo concepciones y métodos infinitesimales diferentes, Newton y Leibniz fueron capaces de separar la ganga geométrica de los resultados de sus antecesores y encontrar el principio general que les permitiría reducir las operaciones fundamentales del Cálculo Infinitesimal a una operativa universalmente válida, concibiendo la idea de sustituir todas las operaciones de carácter geométrico involucradas en el cálculo de tangentes, por una única operación analítica, la derivación del Cálculo Diferencial, que resolvería, además por inversión –cálculo de la antiderivada o primitiva– los problemas de cuadraturas del Cálculo Integral, a través del *Teorema Fundamental del Cálculo*, que vincula ambos problemas y permite la obtención de cuadraturas mediante la resolución del problema inverso de la tangente.

En la brillante operación realizada por Newton y Leibniz, que se ha venido en llamar el descubrimiento del Cálculo Infinitesimal, y que es sin lugar a dudas, uno de los logros más importantes en la Historia del Pensamiento matemático, coadyuvó de forma decisiva la creación y aplicación de un simbolismo que propiciara traducir en fórmulas los resultados y en algoritmos los métodos, a base de utilizar los recursos algebraicos de la Geometría Analítica para independizar el discurso matemático de las figuras geométricas y con todo ello reconocer y aislar los conceptos fundamentales del Cálculo Infinitesimal y crear un cuerpo de doctrina dotado de algoritmos eficaces, es decir, funcionando como un Cálculo operacional que resuelve todos los problemas planteados anteriormente, mediante procedimientos uniformes y con una proyección a nuevos y más complicados problemas, como un potente instrumento de investigación. En palabras del propio Leibniz, se trataba de hacer con las técnicas del Cálculo lo mismo que había hecho Vieta con la Teoría de Ecuaciones y Descartes con la Geometría.

Con la fusión del *Análisis Geométrico* griego y la síntesis algebraica de Vieta, Descartes da a luz la Geometría Analítica, algo auténticamente revolucionario que consigue aniquilar, en gran parte, su raíz primigenia: la propia Geometría griega. Es la primera vez en la Historia moderna que se tiene el convencimiento de haber superado a los antiguos en algún aspecto. A este respecto, pueden ser muy oportuno recordar las reflexiones de O.Spengler, matemático y ensayista de éxito, de los años 20 del siglo pasado, en su libro *La decadencia de Occidente* donde desarrolla su teoría de la Historia como una sucesión de ciclos culturales. Para Spengler, no hay una Matemática que se desarrolle linealmente y cuyo contenido vaya acumulándose a través de los siglos, sino que hay tantas Matemáticas como culturas, como ciclos históricos y cabría distinguir entre otras, «*la Matemática antigua*», de la cultura griega y la «*Matemática moderna*» de la cultura occidental y cristiana, esencialmente distintas y que serían fruto y consecuencia de los componentes culturales de cada época y al mismo tiempo un factor decisivo en la configuración global de las mismas. En el capítulo I de la citada obra, titulado «*El sentido de los números*», Spengler (1998, p.144): escribe:

> «*No hay una Matemática, hay muchas Matemáticas.* [...] *El espíritu antiguo creo su Matemática casi de la nada. El espíritu occidental, histórico, había aprendido la Matemática antigua, y la poseía, aunque sólo exteriormente y sin incorporarla a su intimidad*; *hubo, pues, de crear la suya modificando y mejorando, al parecer, pero en realidad aniquilando la matemática euclidiana, que no le era adecuada. Pitágoras llevó acabo lo primero; Descartes lo segundo. Pero los dos actos son, en lo profundo, idénticos.*»

En efecto, Descartes parte de la Geometría griega para construir algo completamente nuevo, que se convertirá en una Matemática universal, que, en particular apartará a la Geometría del eje central de la Matemática y la destrona de forma definitiva de su rango de reina de la Matemática de modo que la Matemática algebrizada de Descartes desplazará y ocupará el lugar de la Matemática geometrizada de los griegos. Descartes trastoca completamente la jerarquía de las diversas partes de la Matemática, de modo que en su pensamiento la Aritmética y el Álgebra no sólo preceden lógicamente a la Geometría, sino que, además, son superiores en esencia, porque al ser las ciencias de las magnitudes, son mucho más generales y aplicables, entre otros ámbitos al de la Geometría. Es más con Descartes el Álgebra figura en primera línea como técnica, como método de combinación y construcción, de tal modo que es el cálculo algebraico el que legitima los resultados de la nueva Geometría Analítica, que destruye los escrúpulos de los griegos relativos a la definición de las curvas y hace inútil la teoría de la construcción geométrica, que queda sustituida por la síntesis de la construcción algebraica. Así pues, Descartes, con su Geometría Analítica, otorga al Álgebra el gobierno soberano de las Matemáticas, hasta que en el siglo XIX Gauss afirme que es la Aritmética quien debe ocupar el trono de esta ciencia.

La importancia que la posteridad ha concedido a *La Geometría* de Descartes no coincide con los aspectos que interesaban a su autor, porque la idea esencial de futuro de la Geometría Analítica es la tan reiteradamente apuntada de la asociación de ecuaciones y curvas en un sistema de coordenadas, pero como bien señala Kline (1992, vol.1. p.419):

> «*Para Descartes esto* [asociar ecuación y curva] *no era más que un medio para un fin, a saber la resolución de problemas de construcciones geométricas.*»

De hecho las construcciones geométricas que con tanto esmero describe Descartes en *La Geometría* desde el mismo comienzo de la obra han ido perdiendo importancia, porque a diferencia de lo que sucede en la Matemática griega y en la del siglo XVII, la constructibilidad ha dejado de ser una condición necesaria para la existencia. No obstante, más allá del acento en la construcción geométrica de las soluciones de las ecuaciones, por fortuna, Descartes también dio unos usos alternativos a las ecuaciones de las curvas, como en la resolución del *Problema de Pappus* (G.AT,VI,377-387) y en la determinación de las normales a las curvas (G.AT,VI,412-423), donde se sirve de las propiedades geométricas de las curvas para «*construir*» las raíces comunes de las ecuaciones determinando los puntos

de encuentro de las curvas correspondientes; y a la inversa, partir de las ecuaciones y de sus raíces para obtener los puntos de intersección de las curvas correspondientes. Por eso estos dos problemas han ocupado siempre un lugar distinguido en todo estudio de *La Geometría* de Descartes. De hecho aquí emerge el segundo Principio fundamental de la Geometría Analítica, que al sintetizar los desarrollos de Descartes (G.AT,VI, 417), se expresaría en la forma:

> «*El problema geométrico de la intersección de curvas se reconduce al problema algebraico de resolución de sistemas de ecuaciones.*».

La solución de Descartes al problema del trazado de las normales a una curva en un punto se considera uno de los más brillantes logros del método cartesiano, y es, sin duda, uno de los más conocidos y apreciados. Descartes despliega una eficaz alfombra que enlaza la Geometría y el Álgebra –análisis geométrico de los problemas, síntesis algebraica del análisis mediante las ecuaciones y finalmente y traducción geométrica de los resultados algebraicos–. Descartes va escribiendo un verdadero diccionario reversible entre los dos lenguajes, el geométrico y el algebraico, que traduce no sólo lo gramatical –puntos por coordenadas, curvas por ecuaciones–, sino que va mucho más allá al alcanzar el dominio sintáctico –relaciones entre los elementos geométricos, por ejemplo, intersecciones de curvas, se traducen en relaciones entre los correspondientes elementos algebraicos, mediante sistemas de ecuaciones–. No es extraño que, con una retórica altisonante, Descartes considere que el problema del trazado de las normales a una curva en un punto es el más importante, no sólo de cuantos ha resuelto sino de cuantos aspirara a descubrir en Geometría (G.AT,VI, 413). Realmente es una de las muestras más representativas de las raíces cartesianas de la Geometría Analítica por la inmensa capacidad que desarrolla Descartes para establecer caminos reversibles que conectan una y otra vez el Álgebra y la Geometría, mediante los que Descartes aplica toda la potencia algorítmica del Álgebra para resolver problemas geométricos, que en ello consiste la virtualidad de la Geometría Analítica.

A pesar de ciertas reticencias por parte de Pascal, Barrow, Hobbes, e incluso Newton en la aceptación de los nuevos métodos de la Geometría Analítica de Descartes, la extensión de sus aplicaciones a todos los ámbitos de la Matemática fue cada vez más inexorable. A ello contribuyó sobremanera la difusión de las diversas ediciones críticas de van Schooten, plenas de comentarios explicativos, aclaraciones complementarias y apostillas extensivas de los métodos cartesianos del propio editor y de otros matemáticos.

En una de las entradas de la Enciclopedia, la que define el concepto de *Curva*, D'Alembert expresa la idea básica de la asociación de curvas y ecuaciones de la Geometría Analítica de Descartes en relación con los lugares geométricos de los antiguos:

> «*Descartes es el primero que haya pensado en expresar las líneas curvas por medio de ecuaciones. Esta idea sobre la que se funda la aplicación del Álgebra a la Geometría ha sido muy feliz y fecunda. Está claro que al resolver la ecuación de una curva se obtiene uno o varios valores de la ordenada y para una misma abscisa x, y que, en consecuencia, una curva trazada no es otra cosa que la solución geométrica de un problema indeterminado, es decir, que tiene una infinidad de soluciones: es lo que los antiguos llamaban lugar geométrico. Así pues, aunque ellos no pudieron tener la idea de expresar las curvas por medio de ecuaciones, habían visto, sin embargo, que las curvas geométricas no eran otra cosa que el lugar, es decir la sucesión de una infinidad de puntos que satisfacían a la misma cuestión. Por ejemplo, que el círculo era el lugar de todos los puntos que describen los vértices de los ángulos rectos, que se pueden formar sobre una misma base dada tomada como diámetro del círculo, y así para las demás curvas.*»

Aunque la referencia de D'Alembert no hace ningún honor a su compatriota Fermat, tiene el interés de conocer el concepto que se tenía de *La Geometría* de Descartes, ciento treinta años después de su publicación.

En las dos centurias siguientes a la de Fermat y Descartes, matemáticos de la talla de Euler, Monge, Lagrange, Lacroix, etc. imprimirán a la Geometría Analítica un ingente desarrollo hasta situarla en el umbral de la Geometría Analítica moderna –la que se imparte hoy académicamente–, salvo en lo que se refiere al instrumento vectorial, que la convertirá en una de las vetas más fructíferas del pensamiento matemático, en un instrumento responsable de la increíble pujanza y del impresionante progreso que ha desarrollado la Matemática desde entonces. Por ejemplo, más allá del Análisis Matemático, del encuentro de esta materia con la Geometría Analítica aplicada al estudio de curvas y superficies surge, sobre todo tras los trabajos de Euler y Monge, la Geometría Diferencial.

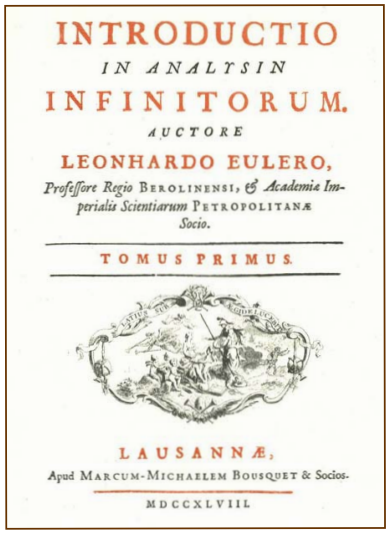
La Geometría Analítica goza de una serie de virtudes que hacen de ella una cómoda y didáctica herramienta matemática para el abordaje de los problemas geométricos. Por una parte permite que las cuestiones geométricas puedan formularse algebraicamente y que los objetivos geométricos puedan alcanzarse por medio del Álgebra, e inversamente, facilita la interpretación geométrica de los enunciados algebraicos, lo que propicia una percepción más intuitiva de su significado, con la posible apertura a la visión de nuevos problemas y conclusiones. Así lo ve Lagrange cuando escribe en sus *Leçons élémentaires de mathématiques* (1795):

> «*Mientras el Álgebra y la Geometría han estado separadas, su progreso ha sido lento y sus aplicaciones limitadas; pero cuando estas dos ciencias han sido vinculadas, se han prestado su fuerza mutuamente y han caminado juntas hacia la perfección.*»

Ilustremos estas ideas de Lagrange mediante las originales motivaciones de Descartes, es decir, la asociación de curvas y ecuaciones. Toda curva construida según una regla geométrica se puede representar mediante su propia ecuación, que caracteriza a la curva y por ello es diferente de la que corresponde a otra curva distinta. De este modo, las propiedades geométricas de una curva pueden ser descubiertas sin más que examinar el comportamiento algebraico de su ecuación. Los vínculos entre curvas, por ejemplo, si se cortan o si son tangentes, se pueden predecir estudiando las relaciones algebraicas que existen entre sus ecuaciones. Por tanto, una vez que de la definición geométrica o cinemática de una curva hayamos derivado la ecuación algebraica que tiene asociada, el establecimiento de las propiedades geométricas restantes de la curva es una cuestión de cálculo algebraico. El poder algorítmico de la máquina simbólica creada por el Álgebra aplicado a la Geometría convierte a la Geometría Analítica en un magnífico instrumento de investigación. Así lo describe de forma magistral el historiador y filósofo de la ciencia Hull (1981, p.268):

> «*Su mérito consiste en que capacita para hallar resultados geométricos mediante un procedimiento sistemático que, si se aplica bien, no puede prácticamente fallar. El descubrimiento de nuevos teoremas particulares que en el caso de los métodos griegos, dependía siempre de la llama genial de la imaginación* [que se fatiga según Descartes] *o bien de la buena suerte* [de la idea feliz], *pasa a la esfera de la competencia profesional ordinaria. El progreso de la Geometría, esencial para el de la ciencia, se hace ahora mucho menos romántico, pero mucho más rápido de lo que fue. La Geometría Analítica ha afectado probablemente a la vida humana más profundamente, aunque menos violentamente, que la máquina de vapor o el aeroplano. La creación de nuevos métodos generales es de mucha mayor importancia que el descubrimiento de conocimientos particulares, por interesantes o útiles que éstos sean.*»

EULER, EL *SHAKESPEARE DE LAS MATEMÁTICAS*


INTRODUCTIO
IN ANALYSIN
INFINITORUM.
AUCTORE
LEONHARDO EULERO,
Professore Regio Berolinensi, & Academiæ Imperialis Scientiarum Petropolitanæ Socio.
TOMUS PRIMUS.
LAUSANNÆ,
Apud Marcum-Michaelem Bousquet & Socios.
MDCCXLVIII.

Portada de *Introductio in Analysin infinitorum* (1748) de Euler.

En la *Introductio* Euler trata sistemáticamente la Geometría con coordenadas Euler da un paso de gigante en la sistematización de la Geometría Analítica de dos y de tres dimensiones y es el introductor como Descartes de nuevas y definitivas notaciones.

De acuerdo con Descartes, Euler reconoce que «*La naturaleza de una curva cualquiera viene dada por una ecuación en dos variables, x,y*». Euler sustituyó el término cartesiano de *construcción* por el de gráfico. La *Introductio* es uno de los primeros tratados donde se dan numerosos gráficos de curvas específicas dadas por sus ecuaciones indicando claramente las unidades utilizadas en el eje de abscisas. Quizá lo más sobresaliente de la *Introductio*, desde el punto de vista del desarrollo de la Geometría Analítica, sea el tratamiento general de los problemas. A partir de Euler surge una de las grandes ventajas de los métodos analíticos modernos frente al enfoque sintético de los antiguos: muchos casos específicos de las cuestiones geométricas pueden ser incluidos en una formulación global. Este aspecto de generalidad que permitía el Álgebra frente a la singularidad de cada problema en la Geometría de los griegos era uno de los rasgos más relevantes señalados por Descartes, pero había sido en parte pasado por alto durante la siguiente centuria, incluso en cuestiones muy básicas como por ejemplo en el estudio de la ecuación de la recta, que se subdividía en numerosos casos diferentes. Euler manejó ya una única forma general: $\alpha x+\beta y-a=0$. Euler realizó un estudio exhaustivo de las cónicas y las cuádricas que alcanzó a su clasificación.

En Geometría elemental el resultado más famoso de Euler es la conocida *Recta de Euler*, sobre la que se sitúan tres de los puntos notables de un triángulo, el Ortocentro, el Baricentro y el Circuncentro, resultado de la más bella Geometría, que ignorado por todas las generaciones anteriores de geómetras, de Euclides a Descartes, de Apolonio a Fermat y de Arquímedes a Newton, fue obtenido por Euler como magistral aplicación de la Geometría Analítica.

La *Introductio* de Euler es una de las tratados más importantes de toda la Historia de la Matemática. C.B.Boyer dice sobre ella en su obra *History of Analytic Geometry* (Scripta Mathematica, New York,1956), p.180:

> **«*La Introductio es probablemente el libro de texto más influyente de los tiempos modernos. Es el trabajo que convirtió el concepto de función en básico para las Matemáticas* [...]. *La Introductio es para el Análisis elemental lo que Los Elementos de Euclides es para la Geometría.*»**

W.Dunham escribe: «*Euler es el Shakespeare de las Matemáticas*» (*El Universo de las Matemáticas*, Pirámide, Madrid, 1995, p.103).

Decía Euler: «*Mejor que de nuestro juicio, debemos fiarnos del cálculo algebraico*».

LOS ARTÍFICES DE LA GEOMETRÍA ANALÍTICA MODERNA

MONGE Y LAGRANGE

Monge y Lagrange son dos de los matemáticos mas importantes de la época de la Revolución francesa. Ambos dieron un impulso inusitado a la Geometría Analítica.

Monge escribe algunas memorias, a modo de libros de texto para los cursos que se imparten en la Escuela Politécnica, que son auténticos manuales de Geometría Analítica. Sobresalen *Feuilles d'analyse apliquée a la géométrie* (1795) y *Application de l'algèbre à la géométrie* (1802), donde da una forma bastante definitiva a la Geometría Analítica. Además de generalizaciones de teoremas elementales como el Teorema de Pitágoras, aparecen las fórmulas de traslación y rotación de ejes para las ecuaciones del cambio de ejes de coordenadas, el tratamiento habitual de rectas y planos, la determinación del plano que pasa por tres puntos mediante coeficientes indeterminados, los cosenos directores, las condiciones de paralelismo y perpendicularidad, los ángulos entre rectas y planos, la determinación de los planos principales de una cuádrica, etc. Monge extendió para el tetraedro ortocéntrico el resultado de la *Recta de Euler* demostrando que el Baricentro está a doble distancia del Ortocentro que del Circuncentro. Ante la impresionante profusión de importantes resultados sobre Geometría Analítica obtenidos por Monge, no es extraño que Lagrange asombrado exclamara:

> «*Con sus aplicaciones del Análisis a la Geometría este demonio de hombre* [Monge] *conseguirá hacerse inmortal*».

Lagrange también realizó importantes contribuciones a la Geometría Analítica, siempre bajo la filosofía de aplicar el carácter algorítmico del Álgebra para superar toda representación concreta. En sus desarrollos analíticos, su formulación, de una brillante elegancia, ya está muy próxima a la escritura del Álgebra Lineal, por ejemplo en cálculos que se asemejan a aspectos matriciales y determinantes. En el artículo *Solutions analytiques de quelques problèmes sur les pyramides triangulaires* (1775), Lagrange resolvió, de forma puramente analítica, diversas cuestiones ya conocidas sobre la geometría del tetraedro: las fórmulas, en función de las coordenadas de los vértices, del área, centro de gravedad y volumen, así como los centros y radios de las esferas inscrita y circunscrita. Al desconocer la ecuación normal del plano, Lagrange obtiene las alturas como problemas de mínima distancia mediante los recursos del Cálculo Infinitesimal. Estos resultados están redactados de tal forma analítica que pueden ser entendidos sin aludir a figura alguna, como el propio Lagrange escribe en el artículo:

> «*Me siento halagado por el hecho de que las soluciones que voy a dar serán ciertamente de interés para los geómetras tanto por los métodos como por los resultados. Estas soluciones son puramente analíticas y pueden entenderse incluso sin figuras.*»

Y efectivamente, no hay ni una figura a lo largo de este trabajo, lo que prefigura la ulterior concepción y estructura de la Geometría Analítica.

Con los trabajos mencionados de Monge y Lagrange, que para algunos historiadores representan una auténtica «*Revolución analítica*», la Geometría Analítica se convirtió en una rama de las Matemáticas independiente y cerrada, muy próxima al enfoque actual en cuanto a los métodos y la notación, salvo en lo que se refiere a las cuestiones vectoriales.

A partir de Descartes habrá dos tipos de tratamiento de los problemas geométricos que darán lugar a dos Geometrías, la Analítica, que aplicará el nuevo lenguaje algebraico y la Sintética, que prescindirá del mismo. Gracias al lenguaje analítico podrán resolverse problemas para los que el lenguaje geométrico puro era impotente, como hallar la normal o la tangente a una curva, calcular el área encerrada por una curva, máximos y mínimos, y demás problemas infinitesimales, cuya resolución se inicia en simultaneidad con el trabajo cartesiano. Pero aunque un problema pueda ser tratado de las dos formas, la analítica dependerá menos de la geometría de la figura y por tanto será más simple y más general. Por ejemplo, para demostrar que las alturas o mediatrices de un triángulo se cortan en un punto, en Geometría Sintética hay que considerar por separado la forma del triángulo según los ángulos, porque ello condiciona si la intersección tiene lugar en el interior o en el exterior del triángulo. En Geometría Analítica, los dos casos se consideran de consuno.

Las maravillosas virtudes de la Geometría Analítica, ponderadas por todos los grandes matemáticos a partir de Descartes, no suponen ni obligan a abandonar la Geometría Sintética, simplemente el profesional sabe que hay dos métodos geométricos y utilizará uno u otro según el objetivo del problema o según el gusto y el sentido estético.

> ¿*Por qué renunciar a las diversas herramientas del taller geométrico*?

Por ejemplo, el gran maestro Euler, en un pequeño artículo de 1747 que lleva el poco original título de *Variae demostrationes geometricae*, aplica Geometría Sintética pura para demostrar, con una elegancia incomparable, la clásica y famosa *Fórmula de Herón* para el área del triángulo en función de los lados (Dunham, 2000, p.215). Pero en otro artículo de 1767, haciendo gala de una increíble intuición geométrica y de una audaz perseverancia algebraica, Euler descubre y demuestra, con la más bella y brillante aplicación de Geometría Analítica, el resultado que ha pasado a los manuales de Geometría con el nombre de *Recta de Euler*:

> «*En cualquier triángulo el Ortocentro, el Baricentro y el Circuncentro están sobre la misma recta. Además, el Baricentro está dos veces más lejos del Ortocentro que del Circuncentro*.»

Conocidos ambos ejemplos, podemos decir que la versatilidad analítica, sintética algebraica, geométrica, teórica y práctica de Euler no tiene límites. Las dos demostraciones eulerianas podrían representar en su propia persona a los dos bandos, el analítico y el sintético, enfrentados en una controversia que se remonta al umbral de la aplicación de los métodos cartesianos. Por fortuna, para los grandes artífices de la Matemática, como el propio Euler o Monge, la polémica es de lo más estéril y debe ceñirse a cuestiones de tipo exclusivamente estético, sin elevarla a juicios de valor acerca de cuál de las dos Geometrías es superior, aunque se esté de acuerdo en que, ciertamente, por el automatismo del Álgebra Simbólica que se aplica en la Geometría Analítica, la Geometría Sintética, como dice W.Dunham (*Euler, el maestro de todos los matemáticos*. Nivola, Madrid, 2000. Cap.7. pp.229-230):

> «*requiere a menudo un punto de intuición, que habitualmente se conoce como inspiración*. [...] ¿*Cómo sabía Euler qué hacer* [en uno y otro problema]. *En última instancia, la respuesta a esta pregunta se halla en el misterioso territorio de la imaginación humana*. [...] *Por supuesto, uno puede preguntarse si la Geometría Analítica es realmente Geometría. Carente de gracia y elegancia, dependiente de lo que Carnot llamó "los jeroglíficos del Análisis", ¿no es una mera aplicación de una fuerza algebraica inexorable*?»

La fuerza incuestionable de la Geometría Analítica y su generalidad e independencia de la «*idea feliz que trae la divina inspiración*», permite entender que, por ejemplo, el discípulo de Monge, Poncelet, uno de los artífices de la Geometría Proyectiva moderna, autor de la importante obra *Traité des propriétés projectives des figures* (1822), y no precisamente un gran admirador de la Geometría Analítica, escribiera:

«*Mientras la Geometría Analítica ofrece su característico método general y uniforme como forma de proceder en la resolución de problemas* [...], *la otra* [la Geometría Sintética clásica] *actúa al azar y depende completamente de la sagacidad de los que la emplean.*»

Por la misma época, Lacroix, el matemático y profesor que más contribuyó a difundir la Geometría Analítica, formula, en su *Traité de calcul* (1810), un punto de vista próximo al actual:

«*Obviando todas las construcciones geométricas se hará ver al lector que existe una manera de considerar la geometría que se podría llamar geometría analítica, y que consiste en deducir las propiedades de la extensión del mínimo número posible de principios por métodos puramente analíticos, de la misma manera que ha hecho Lagrange en su mecánica con respecto a las propiedades del equilibrio y del movimient*o.»

Aun así, Lacroix fue algo reacio a titular sus obras con el nombre de «*Geometría Analítica*». Aunque este nombre había ido apareciendo subrepticiamente a lo largo del siglo XVIII, parece que el primero que lo utiliza como título es Lefrançais en una edición de sus *Essais de géométrie* de 1804 y Biot en la edición de 1805 de sus *Essais de géométrie analytique***.**

La Geometría Analítica se ha convertido en una poderosa herramienta de investigación y exploración científica, en el más útil instrumento para resolver con elegancia, rapidez y plenitud heurística las cuestiones geométricas. Al fundir en un único acto intelectual el descubrimiento y la demostración –el *ars inveniendi* y el *ars disserendi*– la Geometría Analítica permite alcanzar un objetivo básico que se había propuesto su fundador, Descartes:

«*Ejercitar el entendimiento sin fatigar mucho la imaginación*» (DM.AT,VI,17–18).

La Geometría Analítica, de origen remoto en el Análisis Geométrico de los griegos con su incipiente uso retórico de coordenadas en Apolonio y Pappus y su apoyo en la mecánica algorítmica del Álgebra simbólica de Vieta, domina el pensamiento matemático desde la época de Descartes hasta nuestros días. El empleo sistemático de las coordenadas tratadas con el cálculo algebraico, es una potente herramienta algorítmica de resolución de problemas geométricos, un método de un poder y una universalidad tan eficientes en la Matemática, que supera cualquier otro instrumento anterior, y más allá de la Geometría y de la Matemática, la Geometría Analítica ha revolucionado todas las ciencias relacionadas con el tiempo y el espacio, a través del concepto de función, la herramienta más importante para el conocimiento y dominio de la naturaleza. Como escribe Kline (1992, vol.1, p.425):

«*La Geometría Analítica cambió la faz de las Matemáticas*».

La fuerza algebraica inexorable de la Geometría Analítica, su universalidad y su autonomía de la «*fortuna de la inspiración*», democratiza la Geometría y la Matemática en general y pone al servicio de la Humanidad, es decir, de cualquier persona normal, de todo escolar que tenga pequeños rudimentos de Álgebra, un eficaz instrumento que potencia la intuición, facilita la investigación y promueve que no sea imprescindible un gran talento y una gran sagacidad y sutileza intelectual en la resolución de los problemas geométricos. Por eso nos permitimos completar la frase anterior de Kline para sentenciar:

«*La Geometría Analítica cambió la faz de las Matemáticas y de la Educación matemática.*»

Bibliografía

Obras de Descartes

1. DESCARTES,R.: *Vie et Oeuvres de Descartes*. Publiées par C. Adam y P. Tannery. 13 volúmenes. Léopold Cerf, imprimeur París, 1897-1913.
2. DESCARTES,R.: *La Geometría*. Espasa-Calpe. Buenos Aires, 1947.
3. DESCARTES,R.: *The Geometry*. Dover, New York, 1954.
4. DESCARTES,R.: *Oeuvres de Descartes*. Publiées par C. Adam y P.Tannery. 12 volúmenes. Librairie philosophique J.Vrin. París, 1964-74.
5. DESCARTES,R.: *Discurso del Método* / *Reglas para la dirección de la mente*. Orbis. Barcelona,1983.
6. DESCARTES,R.: *Reglas para la dirección del espíritu*. Alianza Editorial, 1034, Madrid, 1989.
7. DESCARTES,R.: *Discurso del Método*. Alianza Editorial,736, Madrid, 1991.
8. DESCARTES,R.: *La Geometria*. Institut d'Estudis Catalans. Barcelona, 1999.
9. DESCARTES,R.: *Discurso del Método / Meditaciones Metafísicas*. Austral. Madrid, 2002.

Obras sobre Descartes

10. ÁLVAREZ,C.: *Descartes y la ciencia del siglo XVII*. Siglo XXI. México, 2000. Caps.1,3,4.
11. CHICA,A.: *Descartes, Geometría y Método*. Nivola. Madrid, 2001.
12. CROMBIE,A.C.: Grandes Matemáticos. *R. Descartes*. Investigación y Ciencia. Cap.3. 18-24. Barcelona, 1995.
13. GARIN,E.: *Descartes*. Crítica, Barcelona ,1989.
14. GÓMEZ PIN,V.: Descartes. Barcanova. Barcelona, 1984.
15. MARITAIN,J.: *Le songe de Descartes*. Buchet/Chastel. París,1922.
16. RODIS-LEWIS,G.: *Descartes et le Rationalisme*. PUF. París, 1996.
17. ROSSELLINI,R.: *Cartesius*. RAI, una produzione Orizzonte 2000. 180'.
18. SIRVEN,J.: *Les années d'apprentissage de Descartes*. Albi. París,1928.
19. SHEA,W.R.: *La magia de los números y el movimiento*: *la carrera científica de Descartes*. Alianza Universidad, 746, Madrid, 1993.
20. TURRÓ,S.: *Descartes. Del Hermetismo a la nueva Ciencia*. Anthropos. Barcelona,1985.
21. VALENSIN,A.: *Imágenes de Descartes*. Cuadernos Taurus. Madrid, 1963.
22. VUILLEMIN,J.: *Mathématiques et métaphisyque chez Descartes*. PUF. París, 1987.
23. WILLIAMS,B.: *Descartes*. Catedra, Teorema, Madrid, 1996.

Obras originales sobre La Matemática Griega

24. EUCLIDES: *Elementos*. Traducción y notas de M.L.Puertas. Gredos, Madrid, 1996.
25. HEAT,T.L.: *The thirteen books of The Elements*. 3 Vols. Dover, New York, 1956.
26. VER EECKE,P.: *Les Coniques d'Apolonius de Pergue*. Blanchard, París, 1959.
27. VER EECKE,P.: *Diophante d'Alexandrie. Les six livres arithmétiques*. Blanchard, París, 1959.
28. VER EECKE,P.: *Les Oeuvres complètes d'Archimède*. Vaillant-Carmanne, Liège, 1960.
29. VER EECKE,P: *Pappus d'Alexandrie. La Collection Mathémathique*. Blanchard. París, 1982.
30. VERA,F.: *Científicos griegos* (Ediciones en español de *Los Elementos* de Euclides, *Las Cónicas* de Apolonio, *La Aritmética* de Diofanto y *La Colección Matemática* de Pappus). Aguilar, Madrid, 1970.

Obras originales de otros autores

31. VIETA: *Introduction a l'Art Analytique*. Traduit par M.F.Ritter. 1867.
32. VIETA: *The Analytic Art*. Traslated by T.R.Witmer. The Kent University Press, Ohio, 1983.
33. VIETA: *Oeuvres Mathématiques*. Traduction du latin in français par J.Peyroux. Blanchard, París, 1991-92.
34. EULER,L: *Introducción al Análisis de los infinitos*. Edición crítica con facsímil de A.Durán de la obra de Euler *Introductio in Analysin Infinitorum*.. SAEM Thales, RSME, Sevilla, 2003.
35. FERMAT,P.: *Oeuvres*. 4 vols. C.Henry,P.Tannery,eds. Gauthier-Villars**.** París,1891-1912.

Obras generales de Historia de las Matemáticas

36. BELL,E.T.: *Les grands mathématiciens*. Payot. París, 1950. Caps.3,4.
37. BELL,E.T.: *Historia de las Matemáticas*. Fondo de Cultura Económica. México, D.F., 1985. Caps.3,6,7.
38. BOYER,C.B.: *History of Analytic Geometry*. Scripta Mathematica. Yeshiva Univ. New York, 1956.
39. BOYER,C.B.: *Historia de las Matemáticas*. Alianza Universidad. Madrid. 1986. Caps. 4,7,9,11,16,17,21,22.

40. CAJORI,F.: *A History of Greek Mathematics*.The MacMillan Company. Londres, 1919. Cap.11.
41. CAJORI, F.: *A History of mathematical notations*. Dover, New York, 1983.
42. COLERUS,E.: *Breve historia de las Matemáticas*, Vol.2, Cap.1. Doncel. Madrid, 1972.
43. DAINVILLE,F.: *L'educatió des jésuites (XVI-XVII siècles)*. Editions de minuit. París, 1978. Cap.3
44. D'ALEMBERT,J.: *Discurso preliminar de la Enciclopedia*. Losada, Buenos Aires, 1954.
45. D'ALEMBERT,J.: *Discurso preliminar de la Enciclopedia*. Orbis, Barcelona, 1984.
46. DEDRON,P.: *Mathématiques et Mathématiciens*. Magnard, París, 1959. Cap.X.3.
47. DUNHAM,W.: *Euler, el maestro de todos los matemáticos*. Nivola, Madrid, 2000. Cap.7.
48. EVES,H.: *An Introduction to the History of Mathematics*. CBS College Publishing,New York, 1983. Caps. 5,6,10.
49. EVES,H.: *Great Moments in Mathematics*. The math association of America, 1977. Vol.1, Cap.23.
50. GELFAND,E.: *EL método de las coordenadas*. MIR, Moscú, 1981
51. GONZALEZ URBANEJA,P.M.: *Las raíces del Cálculo Infinitesimal en el siglo* XVII. Alianza. Madrid, 1992. Cap.1.2.
52. GONZALEZ URBANEJA,P.M.: *El Pensament geomètric en el món grec*. ICE. Universitat de Barcelona, 1996. Caps. 8,10,12.
53. GONZALEZ URBANEJA,P.M.: *La Matemática de la Revolución Francesa*. Seminario *Orotava* de Historia de la Ciencia, Actas, año II, Las Palmas , 1997.
54. GONZALEZ URBANEJA,P.M.: *Matemáticas y matemáticos en el mundo griego* (en *El legado de las Matemáticas. De Euclides a Newton. Los genios a través de sus libros*). Consejería de Cultura. Junta de Andalucía. Sevilla, 2000.
55. GONZALEZ URBANEJA,P.M.: *Los orígenes de la Geometría Analítica*. Fundación Canaria *Orotava* de Historia de la Ciencia. Tenerife, 2003.
56. HEAT,T.L.: *A History of Greek Mathematics*. 2 Vols. Dover, New York, 1981. Caps.5,6,9,10,11,14,19,20.
57. KLEIN,J.: *Greek mathematical thought and the origin of algebra*. Dover, New York, 1992. Caps.10,11.
58. KLINE,M.: *El pensamiento matemático de la Antigüedad a nuestros días*. 3 Vols. Alianza Univ. Madrid, 1992. Caps.3,4,5,15,23.
59. LABRADOR,C.: La *ratio studiorum de los Jesuitas*. Universidad Pontificia de Comillas, Madrid,1986.
60. LORIA,G.: *Histoire des sciences mathématiques dans l'antiquité hellénique*. Gauthiers-Villars, París, 1929. Caps.2,3.
61. MALET,A.: *El Álgebra en el período renacentista. La recuperación de los clásicos griegos*. PPU. Barcelona, 1989. Caps.1,2, Ap.1,3,4.
62. MONTESINOS,J (Compilador): *Historia de la Geometría griega*. Seminario Orotava de Historia de la Ciencia. Tenerife, 1992.
63. MONTESINOS,J.: *Historia de las Matemáticas en la Enseñanza Secundaria*. Síntesis. Madrid.2000. Caps.1,6.
64. MONTUCLA,J.F.: *Histoire des Mathématiques*. Tomo II, Libro II, Cap.V. Blanchard. París, 1968.
65. NICOLAU,F.: *La Matemàtica i els matemàtics*. Claret, Barcelona, 2000. Cap.29.
66. REY PASTOR,J.; BABINI,J: *Historia de la Matemática* .Espasa-Calpe, Buenos Aires, 1951. Caps. III.19, III.22, VI.33.
67. REY PASTOR,J.; BABINI,J.: *Historia de la Matemática*. Gedisa, 2 Vols. Barcelona, 1984. Caps. 3,4,7,8,9.
68. RÍBNIKOV,K.: *Historia de las Matemáticas*. MIR, Moscú, 1974. Cap. 5.2.
69. ROUSE BALL,W: *Histoire des Mathématiques*. Libr. scientifique A.Hermann, París, 1906. Cap.15.
70. SCOTT,J.F.: *A History of Mathematics*. Taylor and Francis, New York, 1975. Cap.7.
71. SMITH,D.E.: *History of Mathematics*. Dover. New York, 1959.Vol.II, Cap.5.8.
72. SMITH,D.E.: *A Source Book in Mathematics*. Dover. 2 Vols. New York. Caps. III.14,III.15.
73. STRUIK.D.J.: *A Source Book in Mathematics, 12OO-18OO*. Harvard University Press, Cambridge, 1969. Caps.2.5,3.1,3.3,3.4,3.5.
74. UNIVERSIDAD COMPLUTENSE: *Libros antiguos de Matemáticas*. Madrid, 2000.
75. VAN DER WAERDEN,B.L.: *Geometry and Algebra in Ancients civilizations*. Springer-V, Berlín, 1983. Caps.3,4.
76. VAN DER WAERDEN,B.L.: *A History of Algebra*. Springer-Verlag. Berlín, 1985. Cap.3.
77. VARIOS AUTORES: *Sigma, el mundo de las Matemáticas*. Grijalbo. Barcelona,1969. Vol.I, Cap.7.
78. VARIOS AUTORES: *Mathématiques a fil des âges*. Gauthiers-Villars, París, 1990. Cap.6.
79. VERA,F.: *Breve Historia de la Geometría*. Losada. Buenos Aires. 1963. Caps. 2,3,4,5,7.
80. VERA,F.: *Veinte matemáticos célebres*. Mirasol. Buenos Aires, 1961. Cap.5.
81. WUSSING,H.: *Lecciones de Historia de las Matemáticas*. Siglo XXI. Madrid, 1989. Cap.7.2.

Obras de Filosofía de la Ciencia y de las Matemáticas

82. BRUNSCHVIC,G,L.: *Les étapes de la Philosophie Mathématique*. Blanchard, París, 1972. Cap. 7.
83. GARCÍA FONT,J.: *Historia de la Ciencia*. Barcelona, 1980. Caps. 5,7,14.
84. HULL,L,W,H.: *Historia y Filosofía de la Ciencia*. Ariel, Barcelona, 1981. Cap.8.
85. KLINE,M.: *Matemáticas, la pérdida de la certidumbre*. Siglo XXI, Madrid, 1985. Caps.2,3.
86. LORENZO,J.: *Introducción al estilo matemático*. Tecnos, Madrid, 1971. Cap.4.
87. MAGEE,B.: *Historia de la Filosofía*. Blume. Barcelona, 1999. Caps.1,4.
88. RUNES,D.: *Historia ilustrada de la Filosofía*. Grijalbo. México, 1967.
89. RUSSELL,B.: *Historia de la Filosofía occidental*. Austral.Madrid,1995. Vol.2. Libro 3.Cap.9.
90. SPENGLER,O.: *La decadencia de Occidente*. Austral, Madrid, 1998.Cap.I.1.
91. VARIOS AUTORES: *Historia del Pensamiento*. Orbis, Vol.2. Cap.3. Barcelona, 1983.

Artículos de revistas científicas

92. BERTHET,J.: *La méthode de Descartes avant le discours*. Rev. de Métaphisique, 1896, 388-415.
93. BOS,H.J.M.: *On the Representation of Curves in Descartes' Géométrie*. Archiv of History of Exactes Sciences, 24, 295-338, 1981.
94. BOYER,C.B.: *Early contributions to Analytic Geometry*. Scripta Mathematica, XIX, 1953, 97-108, 230-238..
95. BOYER,C.B.: *Analytic Geometry in the alexandrian age*. Scripta Mathematica, XX, 1954, 30-36, 143-154.
96. BOYER,C.B.: *Fermat and Descartes*. Scripta Mathematica, XX, 1955. 189-217.
97. BOYER,C.B.: *Post-Cartesian Analytic Geometry*. Scripta Mathematica, XXI, 1956, 101-135.
98. BOYER,C.B.: *Mathematiciens of the French Revolution*. Scripta Mathematica. XXV, 1960, 11-31.
99. CAÑÓN,C.: *El alcance de un sueño. La Mathesis Universales* (en *Descartes, el sueño y la razón*). Revista Quark. UPF. Barcelona, abril-junio, 1996. 22-28.
100. COOLIDGE,J.L.: *The Origin of Analytic Geometry*. Osiris,I,1936, 231-25O.
101. DAINVILLE,F.: *Lénseignement des mathématiques dand les Collèges Jésuites de France du XVIe au XVIIIe siècle*. Revue d'Histoire des Sciences, 1954, 6-21.
102. GALUZZI,M.: *Il Problema della Tangenti nella "Géométrie*" di Descartes. Archiv of History of Exactes Sciences, 22, 1980, 37-51.
103. GIBSON,B.: *La Géometrie de Descartes au point de vue de sa méthode*. Revue de Métaphisique, 1896, 387-398.
104. GONZALEZ URBANEJA,P.M.: *La aparición de los inconmensurables*. Mundo científico, nº 220, 2/2001.
105. MAHONEY,M.S.: *Another Look at Greek* Geometrical Analysis. Archiv of History of Exactes Sciences, 5, 1968-69, 318-348.
106. MOLLAND,G.: *Shifting the foundations: Descartes transformation of ancient gemeotry*. Historia Mathematica, 3, 1976, 21-49.
107. RITTER,F.: *Francois Viète, inventeur de l'algebre moderne, 1540-1603, essai sur sa vie et son oeuvre*. Revue occidentale philosophique sociale et politique, 10 (1895) 234-74.
108. VARIOS AUTORES. *Descartes, les nouvelles lectures* (Artículos sobre Descartes en el cuarto centenario de su nacimiento). Magazine littéraire,342. París, abril,1996.
109. VARIOS AUTORES. Artículos en la Prensa española sobre Descartes en el cuarto centenario de su nacimiento. LA VANGUARDIA, 26/3/96; EL PAÍS, 3/1/96, 2/3/96, 25/3/96, 30/3/96, 19/7/96; EL MUNDO, 23/3/96; AVUI 27/3/96, ABC, 29/3/96, 20/4/96, 4/5/96; HISTORIA Y VIDA, 341, 8/96.
110. WHITESIDE,D.T.: *Patterns of mathematical thought in de later 17th century*. Archiv of History of Exactes Sciences, 1, 1960-62, Cap.7, págs. 290-311.

La Historia de la Matemática como recurso didáctico e instrumento de integración cultural de la Matemática

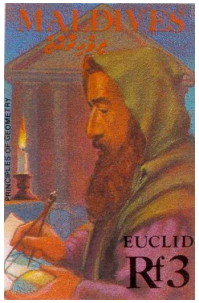
HISTORIA DE LA MATEMÁTICA PARA LA ENSEÑANZA SECUNDARIA

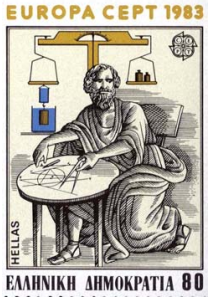
ESTUDIO CRÍTICO DE TRES OBRAS CUMBRES DE LA LITERATURA MATEMÁTICA:

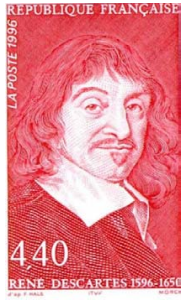
LOS ELEMENTOS DE EUCLIDES

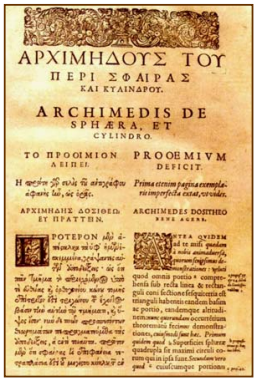
EL MÉTODO DE ARQUÍMEDES

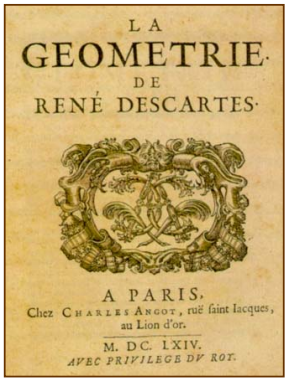
LA GEOMETRÍA DE DESCARTES











Pedro Miguel González Urbaneja
pgonzale@pie.xtec.es

BIBLIOTECA CLÁSICA GREDOS, 155

EUCLIDES

ELEMENTOS

LIBROS I-IV

INTRODUCCIÓN DE
LUIS VEGA

TRADUCCIÓN Y NOTAS DE
MARÍA LUISA PUERTAS CASTAÑOS

EDITORIAL GREDOS

Asesor para la sección griega: CARLOS GARCÍA GUAL.

Según las normas de la B. C. G., la traducción de este volumen ha sido revisada por PALOMA ORTIZ.

© EDITORIAL GREDOS, S. A.

Sánchez Pacheco, 81, Madrid, 1991.

Depósito Legal: M. 28297-1991.

ISBN 84-249-1463-5. Obra completa.
ISBN 84-249-1464-3. Tomo I.

Impreso en España. Printed in Spain.
Gráficas Cóndor, S. A., Sánchez Pacheco, 81, Madrid, 1991. — 6422.

INTRODUCCIÓN GENERAL

I. «EUCLIDES», UN NOMBRE PARA LA GEOMETRÍA

A mediados del siglo pasado, Augustus de Morgan se hacía eco de una creencia común cuando aseguraba que no hay un sistema de geometría digno de tal nombre que se aparte sustancialmente del plan trazado en los *Elementos*. Hoy ya no podemos compartir esta opinión. Pero tenemos tan buenos motivos como los que había entonces para maravillarnos. No suele ocurrir que un solo tratado funde de una vez por todas una disciplina científica; aún es más extraño que además represente por más de veinte siglos el espejo y la norma del rigor en ésa y otras ciencias de la misma familia. Se diría que la historia de las fundaciones, pródiga de suyo en gestas y milagros, sólo conoce como mucho un caso así: la fundación de la geometría y del llamado «método axiomático» con los *Elementos* de Euclides. Por ello, quizás, los empeñados en buscarle algún término de comparación acuden a la mitología: el tópico más recordado a este propósito es el nacimiento de Atenea, en la plenitud de sus armas y atributos, de la cabeza de Zeus. Pero ni siquiera la retórica de un parto tan singular haría plena justicia al acontecimiento, pues, para

colmo, el autor de los *Elementos* estuvo lejos de ser un dios resplandeciente; Euclides fue —a juzgar por lo poco que sabemos— un oscuro profesor.

Los visos de milagro de los *Elementos* tienen mucho que ver, desde luego, con la escasez de noticias sobre los precedentes matemáticos del tratado y con la absoluta falta de información sobre las circunstancias que rodearon su composición por parte de Euclides. Sólo sabemos que, en el curso del siglo III a. C., los *Elementos* vuelven ociosa cualquier obra anterior de análoga factura en geometría; representan el tratado «elemental» por antonomasia. Luego, con el paso del tiempo, las señas de identidad del propio Euclides se van reduciendo a ésta: «*ho stoikheiotés* = el autor de los *Elementos,* el "elementador"» *(vid.,* por ej., Elías, *Comentarios a las Categorías de Aristóteles* 251, 18). Su nombre no tardará en convertirse en el epónimo que hoy nos es familiar: un nombre para la geometría. Eliano, un polígrafo de los siglos II-III, ya afirmaba en un tratado sobre los animales [*Perì dsóon idiótetos*] que «las arañas pueden trazar un círculo sin necesitar nada de Euclides», *i.e.* sin necesidad de conocimientos geométricos. Aún es más explícita la confesión de E. M. Forster a la hora de presentar a Euclides en su célebre guía de Alejandría: «Nada sabemos de él: a decir verdad, hoy lo consideramos como una rama del saber más que como un hombre» *(Alejandría. Historia y Guía,* 1961[3], Sec. I E, i).

Cabe pensar, entonces, que la fortuna institucional misma del tratado no sólo hizo que se perdiera el rastro de otros *Elementos* anteriores, sino que veló para siempre el rostro de su autor. La cuestión de si los *Elementos* de Euclides ocultan su pasado, es un asunto polémico y delicado. Lo que nadie discute es el desvanecimiento de quien los escribió.

1. *Oscuro autor, incierto personaje*

Sobre la vida —no ya los milagros— del propio Euclides, sólo parece haber dos referencias relativamente dignas de crédito. Una: fue más joven que los discípulos de Platón (muerto en el 347), mayor que Arquímedes (nacido hacia el 287), coetáneo del instaurador de la dinastía tolemaica, Tolemeo Sóter (367/6-283). Y dos: enseñó o formó escuela en Alejandría.

Una indicación sobre el momento que toca vivir a Euclides se encuentra al final del resumen histórico de Proclo *(Comentarios al libro I de los Elementos de Euclides* 68, 6-23). Según Proclo: «vivió en tiempos del primer Tolemeo, pues Arquímedes, que vino inmediatamente después, menciona a Euclides» (68, 12-14). Este juicio no se justifica por la alusión a los *Elementos* I, 2, que aparece en *Sobre la esfera y el cilindro* de Arquímedes, pues seguramente se trata de una interpolación; pero casa perfectamente con la observación de un autor bastante próximo, Ateneo —siglo II a. C.—, quien apunta en el *Banquete de los sofistas* (VI 250) que Euclides había sido un «comensal [*parásitos*]» de Tolemeo I; cuadra también con las distancias que el tratado guarda a veces, en la demostración de resultados anteriormente conocidos, respecto de ciertas pruebas reseñadas por Aristóteles y respecto de algunas contribuciones debidas a los matemáticos anteriores más o menos relacionados con la Academia de Platón (Teeteto, Eudoxo, Menecmo) [1]. Precisamente, en el pasaje antes citado, Proclo introduce a Euclides con estos términos:

[1] Cf. las pruebas recordadas por Aristóteles que se encuentran en T. L. HEATH, *Mathematics in Aristotle,* Nueva York-Londres, 1949, reimp. 1980. Sobre las contribuciones de la tradición matemática anterior, *vid.*

No mucho más joven (que Hermótimo de Colofón y Filipo de Medma, discípulos de Platón) es Euclides, quien compiló los elementos poniendo en orden varios teoremas de Eudoxo, perfeccionando muchos resultados de Teeteto y dando así mismo pruebas incontestables de aquello que sus predecesores sólo habían probado con escaso rigor *(Com.* 68, 6-12).

Esta apreciación viene corroborada por otras fuentes. A Tecteto se le reconocían ciertos progresos en el estudio sistemático de las magnitudes inconmensurables iniciado por Teodoro (Platón, *Teeteto* 147c-148b); pues bien, en un comentario sobre el libro X de los *Elementos* atribuido a Pappo, Euclides es alabado por dar criterios seguros para la conmensurabilidad y la inconmensurabilidad en general, precisar algunas nociones y distinciones en este sentido, establecer varias clases de magnitudes irracionales y mostrar con claridad su campo de aplicación. Eudoxo había sido celebrado por su contribución a una teoría general de la proporción; pues bien, el escolio 3.° al libro V de los *Elementos,* que sienta las bases de esa teoría, declara:

Algunos dicen que este libro es descubrimiento de Eudoxo, coetáneo de Platón... Pero la disposición del libro en punto a los elementos y al orden de consecuencia de los teoremas es reconocida por todos como obra de Euclides (Stamatis, ed. cit., V (1977), pág. 213).

W. R. KNORR, *The Evolution of the Euclidean Elements,* Dordrecht-Boston, 1975; D. H. FOWLER, *The Mathematics of Plato's Academy,* Oxford, 1987. En I. THOMAS, ed., *Selections of Greek Mathematics,* Londres-Cambridge (Mass.), 1939 (1967[3]) hay algunos textos significativos en versión bilingüe (griego-inglés). Las referencias al *Comentario* de Proclo se atienen a G. FRIEDLEIN, ed., *Procli Diadochi in primum Euclidis Elementorum librum commentarii,* Leipzig, 1873 (reimp. 1967); suele ser más accesible la versión de G. MORROW, *Proclus. A Commentary on the First Book of Euclid's Elements,* Princeton, Nueva Jersey, 1970.

Por otra parte, sabemos que Euclides toma en sus *Fenómenos* ciertas proposiciones de *Sobre la esfera en movimiento* de Autólico —sin mencionar la fuente—; Autólico fue maestro de Arcesilao, nacido hacia 315; aunque no disponemos de más datos, es probable que Autólico, a su vez, se moviera entre la 2.ª mitad del siglo IV y el 1.er tercio del siglo III. En fin, las fechas del reinado de Tolemeo (304-285/3) son las mejor documentadas; tal vez sea un privilegio del que funda una dinastía frente a quien se limita a fundar un cuerpo de conocimientos.

La segunda referencia, acerca de la enseñanza de Euclides en Alejandría, procede de Pappo *(Colección* VII 35): cuenta que Apolonio de Perge había frecuentado durante algún tiempo —en torno a 250— a los discípulos de Euclides en Alejandría. La existencia y el peso de esta escuela alejandrina están fuera de dudas. Hay señales del celo de algunos miembros por asegurar una especie de ortodoxia euclídea tanto en el cultivo de la matemática elemental como en la investigación avanzada; ni los talentos mejor dotados y más innovadores de la época (Conón, Arquímedes, Apolonio) dejaron de plegarse a la pronta institucionalización alejandrina de ciertos supuestos y métodos de la prueba matemática consagrados por los *Elementos.*

En suma: si redondeamos las fechas de que disponemos, podemos estimar que Euclides alcanzó su madurez en torno al año 300 a. C. Es hacia entonces cuando, por convención, se suelen fechar los *Elementos* [2].

[2] Cf., por ej., F. O. HULTSCH, «Eukleides», en la PAULY-WISSOWA, *Realencyclopädie der klassischen Altertumswissenschaft,* Stuttgart, 1909 (reimp. 1970), VI 1, col. 1004; I. BULMER-THOMAS, «Euclid. Life and Works», en CH. C. GILLISPIE, dir., *Dictionary of Scientific Biography,* IV, Nueva York, 1970-1980, (reimp. 1981), pág. 415. En P. M. FRAZER, *Ptolemaic Alexandria,* Oxford, 1972, I, págs. 386-388; II, nota 82, puede verse una discusión al respecto.

Con el curso del tiempo, Euclides deviene un personaje de historias y leyendas, a veces presa de malentendidos. Las historias de los polígrafos griegos tienden a moralizar en un severo tono neoplatónico. Según Estobeo *(Extractos, preceptos, consejos* II 31, 114), en cierta ocasión uno de sus oyentes, nada más escuchar la demostración del primer teorema, se había apresurado a preguntar por la ganancia que podía obtener de cosas de ese género; Euclides, volviéndose hacia un sirviente, habría ordenado: «Dale 3 óbolos, pues necesita sacar provecho de lo que aprende». También se decía que, en otra ocasión, Tolemeo Sóter —a quien podemos suponer harto ocupado en asuntos de otro tipo— se había interesado, cortésmente quizá, por una vía de acceso al conocimiento geométrico menos fatigosa, más rápida y llana que la de los *Elementos;* la respuesta atribuida a Euclides habría sido algo seca y cortante: «No hay camino de reyes en geometría», réplica que el propio Estobeo prefiere poner en boca de Menecmo y ante Alejandro Magno *(ibid.* II 31, 115). Algunos matemáticos de ayer (G. H. Hardy) y de hoy (E. C. Zeeman) todavía consideran a Euclides como un digno y envarado colega de otra universidad. Esta impresión no se compadece con la que sugiere Pappo cuando alaba su talante comprensivo y afable [*eumenés*] *(Col.* VII 35). Pero un viejo profesor de matemáticas, al que pintan celoso de su rigor y su pureza, puede tener achaques de impaciencia como los referidos en esas dos anécdotas más edificantes que ciertas. Bien: sea así, si así os parece.

Los polígrafos árabes imaginaron leyendas más audaces. Euclides había sido hijo de Naucrates, nieto de Zenarco —y tal vez de Berenice—; había nacido en Tiro y residido en Damasco, pero sin renegar de su linaje y sus hábitos griegos; por lo demás, en los *Elementos* no había hecho

sino refundir el trabajo anterior de un tal Apolonio, carpintero por más señas (Casiri, *Bibliotheca Arabico-Hispana Esculiarensis* I 339).

La hora de los malentendidos llegó, en fin, con los comentadores y los editores postmedievales y renacentistas. Algunos dieron al autor de los *Elementos* la falsa identidad de Euclides de Mégara, un filósofo socrático coetáneo de Platón; la confusión, tal vez nacida de Valerio Máximo *(Sobre dichos y hechos memorables* VIII 12) en tiempos de Tiberio, se generalizó a través de las ediciones impresas de los *Elementos* entre 1482 y 1566, hasta ser disipada por Commandino (1572) y por Clavio (1574). Otra especie común en el siglo XVI fue considerar únicamente los enunciados de problemas y teoremas como propios de Euclides, mientras se adjudicaban las pruebas —la *expositio* o el *commentum* de estas proposiciones— a Teón, el editor alejandrino de los *Elementos* en el siglo IV. Este desmembramiento del cuerpo de los *Elementos* prosperó con ediciones influyentes, como la de Zamberti (bajo el elocuente título: *Euclidis megarensis philosophi... elementorum libros XIII cum expositione Theonis insignis mathematici,* Venecia, 1505) o la de Lefèvre *(Euclidis Megarensis ... libri XV, Campani ... in eosdem commentariorum libri XV. Theonis ... in tredecim priores commentationum libri XIII,* París, 1516); al final, dio lugar a la impresión de manuales que sólo recogían los enunciados de las proposiciones euclídeas (e. g., Roma, 1545; París, 1557). Por otro lado, la adaptación escolar y la divulgación de una base geométrica y aritmética que la cultura de la época hacía necesaria, invitaban a una actitud análoga de selección: a entresacar de los *Elementos* ciertos principios, en particular definiciones, y determinadas proposiciones a cuyo enunciado seguían las pruebas más sumarias e intuitivas disponibles. De hecho,

amparándose en el valor cultural o en los servicios instructivos de la geometría (en nuestro país, por ejemplo, podían comprender desde «la educación de nobles» hasta «la fortificación militar»), no pocos tratados de matemática aplicada de los siglos XVII y XVIII seguirán practicando esta suerte de selección o despojo de Euclides con premeditación y alevosía.

Por lo demás, la neblina que envuelve a nuestro personaje sigue dando que hablar en nuestros días. Hay quien ha imaginado que la única salida a la incertidumbre creada por su cronología imprecisa y por la índole un tanto heterogénea de los *Elementos*, es plantear una especie de «cuestión euclídea». ¿No serán «Euclides» un nombre colectivo y los *Elementos* obra de una escuela? [3] Imposible no es. Pero tampoco parece una salida razonable: no tiene la menor base documental y nos metería en dificultades mayores que las que pretende salvar.

2. *Obras*

Euclides escribió por lo menos una decena de obras. Hoy disponemos de dos: los *Elementos* [Στοιχεῖα] y los *Datos* [Δεδομένα]. La segunda bien podría ser un texto auxiliar, complementario de la primera; tiene que ver con los libros I-VI y con la práctica del análisis geométrico como vía de resolución de problemas: en el supuesto de que ciertas partes de una figura estén dadas (por lo que se refiere a su magnitud, posición, etc.), muestra la manera de determinar otras partes de la figura en el mismo respecto. Tenemos recensiones de otras dos. Una es *Fenómenos*

[3] *Vid.* J. ITARD, *Les livres arithmétiques d'Euclid*, París, 1961, pág. 11; «Euclid», en *Encyclopaedia Universalis. Corpus.* 7, París, 1985, pág. 518. Cf. G. KAYAS, ed., *Euclide. Les Éléments*, París, 1978, Introd., págs. xi-xvi.

[Φαινόμενα], ampliamente comentada por Pappo *(Col.* VI 104-130); se trata de una obra de astronomía teórica compuesta sobre la base de la geometría esférica —cuyo estudio tenía menos arraigo tradicional que el de la geometría plana de los *Elementos—;* introduce una noción absoluta de horizonte frente al sentido relativo que antes había tenido *horídson* en Aristóteles y Autólico. La otra es *Óptica* [Ὀπτικά] (Pappo, *Col.* VI 80-103; Proclo, *Com.* 69, 2), un tratado acerca de la perspectiva y la visión directa —por oposición a la geometría de los rayos reflejados, la catóptrica, o de los rayos refractados, la dióptrica—, marginado seguramente por las mayores luces de la *Óptica* de Tolemeo (aparecida a mediados del siglo II); conservamos, con todo, una recensión del editor alejandrino Teón y otra versión anterior, editadas por Heiberg-Menge *(Euclidis opera omnia*, VII, 1895); formula por vez primera el principio de la propagación rectilínea de la luz. Para las demás obras de Euclides contamos con referencias más o menos cumplidas. De *Sobre divisiones de figuras* [Περὶ διαιρέσεων βιβλίον] (Proclo, *Com.* 69, 4) hay una versión árabe; la reconstrucción moderna de R. C. Archibald (1915) se beneficia también de una sección recogida en *Practica geometriae* de Leonardo de Pisa (1220). Pero no quedan restos de los *Porismas* [Πορίσματα] (Pappo, *Col.* VII 14; Proclo, *Com.* 301, 22-26); ésta ha sido una grave pérdida, pues allí Euclides abordaba al parecer cuestiones de matemática superior de un tipo y un alcance que hoy no estamos en condiciones de precisar [4]. *Sobre cónicas* debió de

[4] *Vid.* T. L. HEATH, ed., *The Thirteen Books of Euclid's Elements*, Londres, 1926², (reimp. Nueva York, 1956) 1, págs. 10-15. BULMER-THOMAS, art. cit., 1981, pág. 42, opina, como Heath, que esta obra podría haber sido el primer tratado sobre geometría proyectiva de que hubiera noticia. C. B. BOYER, *Historia de la matemática*, Madrid, 1986,

ser una obra en cuatro libros sobre las secciones cónicas de la que sólo indirectamente da noticia Pappo *(Col.* VII 30); Apolonio, en el prefacio del libro I de sus *Cónicas,* se limita a aludir al trabajo de Euclides en este campo sin mayores referencias; es posible que en tiempos de Pappo, entre finales del siglo III y principios del IV, ya se hubiera perdido sustituida por el gran tratado de Apolonio. Los dos libros de *Sobre superficies* [Τόποι πρὸς επιφανείᾳ] (Pappo, *Col.* VII 3) estudiaban conos, cilindros, esferas y posiblemente otras construcciones sobre superficies de sólidos en revolución, tal vez en la línea desarrollada luego en *Sobre conoides y esferoides* de Arquímedes. Un tratado de paternidad euclídea harto dudosa discurriría sobre *Catóptrica* [Κατοπτρικά] (está mencionado en Proclo: *Com.* 69, 2); naturalmente se ocuparía de espejos; la compilación hoy conservada bajo este título se considera obra de Teón, nuestro activo y casi inevitable mediador con la obra de Euclides. Algo más fiables son los indicios de unos *Elementos de música* [Αἱ κατὰ μυσικὴν στοιχειώσεις] (Proclo, *Com.* 69, 3; Marino, prefacio de los *Comentarios sobre los Datos de Euclides,* recogidos en Heiberg-Menge, VIII 454, 19); pero la llamada *Sectio Canonis,* que ha solido endosarse al propio Euclides, no es al parecer sino un estudio de intervalos compuesto a partir de extractos de algún original euclídeo; y la otra parte que se suponía integrante de esos *Elementos,* una *Introducción a la armonía* [εισαγωγὴ ἁρμονική], hoy suele reconocerse como contribución de Cleónides, discípulo de Aristóxeno. Por último, Proclo *(Com.* 69, 27-70, 18) se hace eco de un escrito *Sobre paralogismos* [Ψευδάρια] que abundaba en casos y ejercicios prácticos dirigidos a formar y depurar el razonamiento

pág. 142, considera que podría haber representado una aproximación a un tipo de geometría analítica.

del alevín de matemático; Miguel de Éfeso, en sus comentarios a las *Refutaciones Sofísticas* de Aristóteles, lo titula *Pseudo-graphemata;* no llegó a ser conocido por los comentadores árabes de Euclides.

Las fuentes árabes atribuyen a Euclides algunos tratados mecánicos *(Sobre lo ligero y lo pesado* [Περὶ κούφων καὶ βαρέων], *Sobre la palanca* [Περὶ Ζυγοῦ]). Creo que lo más que cabe aventurar en este sentido es la posible existencia de una mecánica geométrica, inspirada en consideraciones ajenas a la dinámica de Aristóteles y anterior o, al menos, relativamente primitiva en comparación con la estática de Arquímedes. Dicho en otras palabras: *Sobre la palanca* podría ser quizás una muestra de que, ya antes de Arquímedes, la mecánica aristotélica no era la única posible. Pero las referencias de que disponemos sobre todo esto se limitan a ser compatibles con lo que hubiera podido hacer Euclides de haber hecho efectivamente algo al respecto [5].

3. *Los «Elementos»*

La fama de Euclides descansa, según es bien sabido, en la rara fortuna de su tratado de elementos matemáticos. «Los *Elementos* contienen una guía incontestable y perfecta de la exposición científica misma en materia de geometría», sentencia Proclo al final de su breve catálogo de las

[5] Para más detalles sobre las obras de Euclides, aparte de los *Elementos, vid.* la ed. capital de J. L. HEIBERG - H. MENGE, *Euclidis opera omnia,* Leipzig, 1883-1916, vols. VI-VIII. Hay referencias en la ed. cit. de HEATH, 1926², 1, págs. 7-18, así como en su *A History of Greek Mathematics,* Londres, 1921 (reimp. Nueva York, 1981), vol. I, págs. 421-446. También tienen interés las breves indicaciones de I. BULMER-THOMAS, art. cit., 1981, págs. 425-431.

obras euclídeas *(Com.* 70, 16-18). Este juicio resume la opinión del mundo cultural helenístico y, con ligeras modulaciones, se ha mantenido prácticamente vigente hasta bien avanzado el siglo XIX. Algunas opiniones modernas son incluso mucho más exageradas: Peyrard, en 1819, aún recuerda a propósito de esta obra: «M. Legendre ... me repetía a menudo que la geometría era una lengua muerta después de Euclides» (ed. cit. de J. Itard (1966), prefacio, pág. 1).

Los *Elementos,* en la presente versión, son un conjunto de 132 definiciones, 5 postulados, 5 nociones comunes o axiomas y unas 465 proposiciones distribuidas en 13 libros. Entre los comentadores árabes se extendió la creencia de que el tratado incluía además otros dos libros, el XIV y el XV, que venían a complementar el estudio de los sólidos regulares del libro XIII. Pero el supuesto libro XIV es seguramente obra de Hypsicles, un matemático alejandrino, quizás discípulo del propio Euclides. Y el presunto libro XV es un producto mucho más tardío y de inferior calidad aún que el anterior; hay quienes lo atribuyen a Damaskios (siglo VI), discípulo de Isidoro de Mileto —uno de los arquitectos de la Basílica de Santa Sofía—.

A pesar de su común identificación con la geometría, el tratado comprende diversos campos temáticos de la matemática elemental y emplea métodos variados. Entre esos cuerpos figuran la teoría de la geometría plana (libros I-IV) y la geometría del espacio (XI-XIII), la teoría generalizada de la proporción (V-VI), la teoría aritmética (VII-IX); el libro X es un tanto singular y, en sustancia, ofrece una conceptualización precisa de la inconmensurabilidad y una clasificación prolija de las variedades de rectas irracionales. Dentro de estas teorías, puede haber así mismo algunos núcleos de singular relieve, e.g., la concepción de

las paralelas —en el marco de la geometría— o el estudio de los primos relativos —en el marco de la aritmética—. Por lo que concierne a los métodos utilizados, los más relevantes podrían ser el procedimiento elemental de construcción por «reglas y compás», el procedimiento de aplicación de áreas —conocido entre algunos intérpretes modernos por la denominación equívoca de «álgebra geométrica»—, y un proceder demostrativo fundado en el principio de continuidad (libro V, definición 4) y en el lema de bisección (libro X, Prop. 1) —que, a su vez, se ha visto tildado de «método de exhausción», por obra y gracia de la matemática renacentista—. Pero, por encima de todo, lo que tradicionalmente ha llamado más la atención de los comentadores y observadores *methodologically minded* es la portada «axiomática» de los *Elementos* —el conjunto de definiciones, postulados y nociones comunes—, con la que se abre el libro I y a la que se supone el pórtico de la gloria de la geometría euclídea.

Más adelante, volveremos sobre todos estos ingredientes teóricos y metódicos del contenido de los *Elementos.* Ahora no estará de más detenerse a considerar brevemente el lugar y el sentido de la composición de Euclides dentro de la antigua tradición matemática griega de este tipo de tratados.

II. Los «Elementos», una escritura que hizo historia

Los griegos fueron conscientes de la importancia que revestía la composición de *Elementos* matemáticos. «En geometría es bueno ejercitarse en lo que se refiere a los elementos», aseguraba Aristóteles *(Tópicos* VII 14, 163b23). Una razón puede ser la apuntada por Proclo *(Com.* 71,

9-12): si se cuenta con unos elementos de geometría, cabe entender el resto de esta ciencia, mientras que sin ellos no será posible comprender su complejidad y las demás partes resultarán inalcanzables. Esta consideración se extiende a las matemáticas en general. Según un testimonio del que se hace eco Herón *(Definiciones* 160, 8 ss.), los peripatéticos decían que la retórica, la poesía y la música popular eran accesibles y podían entenderse sin un curso de iniciación, pero nadie era capaz de adquirir conocimiento de las cosas que se denominaban «matemáticas», a menos que tuviera una instrucción específica; «y por este motivo — explica Herón— el estudio de tales materias fue llamado "matemática"» *(i.e. mathematiké,* término relacionado con el verbo *manthánо:* «aprender», «instruirse», «llegar a conocer»).

1. *La tradición de los tratados de elementos*

Los *Elementos* de Euclides coronan una tradición de tratados elementales hoy desaparecidos. A juzgar por el resumen historiográfico de Eudemo que nos facilita Proclo *(Com.* 65 ss.), Hipócrates de Quíos, hacia 470-400 a. C., fue el primero en componer un libro de elementos. Posteriormente, un tal León, algo mayor que Eudoxo y más joven que Platón, compuso otro prestando mayor atención al número de los elementos y a su empleo en las pruebas; también tiene interés anotar que se aplico a la investigación de diorismos [*diorismoí*], *i.e.* a la investigación de condiciones que determinan cuándo un problema es soluble o no. Teudio de Magnesia, en fin, confeccionó unos *Elementos* que mejoraban el orden deductivo de exposición y daban forma general a resultados que sólo habían tenido hasta entonces un alcance particular; fue seguramente un

manual con cierta circulación dentro de la academia platónica; algunas referencias matemáticas sumarias de los *Analíticos* de Aristóteles revelan por lo menos la familiaridad con un manual de este género. Proclo no menciona ningún otro tratado anterior a los *Elementos* de Euclides —aunque hace alusión a Hermótimo de Colofón, que continúa la línea de trabajo de Teeteto y Eudoxo, y es más joven que Euclides, como descubridor de «muchos de los elementos»—.

El término «elemento [*stoikheîon*]» podía emplearse con múltiples usos en diferentes contextos (gramatical, cosmológico, astronómico, matemático) [6]. Su equivocidad es manifiesta sin salir siquiera del contexto geométrico. Por un lado, recibían el nombre de *Elementos* las compilaciones que reunían ciertos conocimientos primordiales y básicos. Por otro lado, se llamaban «elementos» las proposiciones que desempeñaban un cometido capital en la obtención o en la organización deductiva de otros resultados (Proclo, *Com.* 72, 5-7). Según esto, los *Elementos* venían a ser tratados que exponían con mayor o menor acierto los «elementos» —en este último sentido— de algún dominio de las matemáticas. Pero también se distinguía, desde Menecmo (siglo IV a. C.) al menos, entre una significación más amplia y otra más restringida de estos «elementos», de estas proposiciones capitales. En un sentido genérico, todo cuanto sirve de medio para obtener o establecer alguna otra cosa puede considerarse «elemento» de tal resultado; así entendidos, son «elementos» los lemas asumidos, los teoremas demostrados y los problemas resueltos cuando se utilizan en la prueba de nuevos teoremas o en la

[6] *Vid.*, por ej., W. BURKERT, «Stoicheion. Eine semasiologische Studie», *Philologus* CIII (1959), 167-197.

solución de problemas ulteriores. Por ejemplo, en los *Elementos* de Euclides la construcción de un triángulo equilátero (libro I, Prop. 1) puede considerarse elemento de la obtención de una recta igual a otra recta dada (I, Prop. 2). En cambio, el título de «elemento» en un sentido propio y restringido se reserva para designar un grupo selecto de asunciones y proposiciones: las que tienen el estatuto de principios [*arkhaí*] dentro de una disciplina y constituyen la base de partida sobre la que se teje la trama deductiva de las demás proposiciones como un cuerpo más o menos sistemático de conocimientos. Esas proposiciones primordiales —definiciones, postulados— son así elementos, propiamente dichos, de los teoremas derivados de ellas. Proclo señala que este significado más estricto es el que mejor cuadra a los elementos compilados en el tratado de Euclides. Y después de Euclides, probablemente bajo su influjo, es el que prevalece en algunos otros tratados básicos o elementales [*stoikheióseis*] sobre otras ramas de la matemática.

De todo ello se desprende que la composición de *Elementos* era un empeño importante en varios aspectos, tanto teóricos y metodológicos como didácticos o disciplinarios. Pero no implicaba una estrategia única, uniforme, y por lo demás nada garantizaba en principio el éxito de la empresa o la significación del plan concebido: podía conducir a reducciones triviales o excesivamente limitadas de la materia objeto de estudio, podía descansar en unas asunciones primeras desprovistas de la calidad o de la capacidad de los auténticos principios —Aristóteles ya había reparado incidentalmente en que la captación [*noûs*] de los *arkhaí* peculiares de una ciencia no siempre estaba asegurada de antemano. En su larga digresión sobre la noción de *stoikheîa*, Proclo también advierte:

> en todas las ciencias es difícil seleccionar los elementos y disponerlos en su debido orden... Y de quienes han abordado esta empresa, unos han sido capaces de compilar más proposiciones y otros menos; unos han dado pruebas muy sucintas y otros han dilatado su exposición con interminables minucias; algunos han evitado el empleo de la reducción al absurdo, otros han descartado el uso de las proporciones, otros han ideado unos preliminares dirigidos contra quienes desechan los principios; y, en suma, muy diferentes métodos han sido los inventados por los diversos autores de *Elementos* (*Com.* 73, 16-25).

A la luz de esto podemos apreciar la excelencia del tratado de Euclides sobre los demás tratados, anteriores o posteriores, de este género. Contamos con razones de dos tipos para acreditarla: unas parecen ser más bien de carácter disciplinario o didáctico, hacen referencia a las virtudes expositivas e instructivas de la composición euclídea; las otras son más bien de carácter metodológico y guardan relación con la sistematización deductiva de las teorías o cuerpos de conocimiento que contiene. Tanto unas como otras son criterios coligados y aducidos por el mismo Proclo —aunque seguramente la distinción que he apuntado apenas fuera significativa en el medio helénico que da origen a esta suerte de tratados [7]—. De los valores metodológicos, Proclo deja constancia en un pasaje donde trata de

[7] Puede que los griegos no advirtieran tanta diferencia como nosotros entre un tipo y otro de criterios. PLATÓN ya aludía al «logos que se escribe con la ciencia en la mente del que aprende y es capaz de defenderse a sí mismo» (*Fedro* 276a); más aún, según ARISTÓTELES, los silogismos didácticos [*didaskalikoí*] de los *Tópicos* no son otros que los demostrativos [*apodeiktikoí*] de los *Analíticos*, vid. *Refutaciones sofísticas* 2 165b1-4 y 8-9. En lo sucesivo y por lo regular, mencionaré los títulos de Aristóteles por sus abreviaturas habituales, e. g., *SE (Refutaciones sofísticas)*, *APr. (Primeros Analíticos)*, *APo. (Segundos Analíticos)*, etc.

resumir las virtudes que realzan los *Elementos* euclídeos *(Com.* 69, 4 ss.). La primera es el acierto y la discriminación que gobiernan la selección de los teoremas y problemas considerados, pues Euclides «no ha incluido todo lo que podría haber dicho sino sólo lo pertinente para la construcción de los elementos». La segunda virtud es la riqueza de métodos empleados:

> Ha utilizado toda suerte de deducciones concluyentes, unas que obtienen su fuerza de los principios mismos, otras que se derivan de pruebas demostrativas, pero todas ellas incontestables, precisas y congruentes con el conocimiento científico. A mayor abundamiento se ha servido de todos los métodos dialécticos...;

por si esto fuera poco, «hace un uso hábil y certero de diversas formas de conversión geométrica». La tercera y última virtud mencionada por Proclo guarda una relación más estrecha con la sistematización deductiva: consiste en la continuidad de las pruebas y, más especialmente, en el hecho de proceder conforme al orden de consecuencia debido.

Los valores didácticos o disciplinarios aparecen en el contexto antes citado de la digresión en torno a la noción de elementos y la confección de tratados elementales.

> Un tratado así —asegura Proclo— ha de verse libre de todo cuanto sea superfluo, pues eso obstaculiza el aprendizaje; debe cribar todo lo comprendido por el objeto de estudio de forma coherente y conducente al fin propuesto, en orden a ser de mayor utilidad para el conocimiento; ha de poner sumo cuidado tanto en la claridad como en la concisión, pues lo contrario entorpece la comprensión; debe proponerse la formulación de los teoremas en términos generales, pues parcelar la instrucción de la materia en casos particulares dificulta la consecución del conocimiento. De acuerdo con todos estos criterios —concluye nuestro comentador—, el sistema de los elementos de Euclides supera a los demás *(Com.* 73, 26-74, 11).

Esta significación didáctica y disciplinaria del tratado de Euclides reaparece entre los objetivos que Proclo atribuye expresamente al propio autor *(Com.* 70, 19 ss.). Los propósitos de Euclides al componer los *Elementos* han sido, según nuestro comentarista del libro I, dos: uno hace referencia al objeto de la investigación; el otro se refiere justamente al valor instructivo del tratado.

Por lo que toca al primero, Proclo, en su vena platonizante habitual, supone que la geometría globalmente considerada tiene que ver con las figuras «cósmicas», con los cinco sólidos regulares de que trata el libro XIII, y asegura que el propósito final de la contribución de Euclides es precisamente la construcción de estos poliedros, sumamente importantes para una cosmología platónica. La verdad es que el tratado, aunque se cierre con el broche de estas figuras tendenciosamente llamadas «cósmicas» o «platónicas», no deja entrever ninguna intención de Euclides en tal sentido. El silencio de Euclides sobre todo cuanto se refiera a un propósito filosófico o, en general, extramatemático de su tratado, es denso y tan perceptible como el fervor neoplatónico de Proclo y de otros muchos comentadores helenísticos. Si, por nuestra parte, llevados de una pasión hermenéutica similar, nos empeñáramos en extraer o identificar «una filosofía» a partir de ciertas asunciones u orientaciones implícitas en esos textos matemáticos clásicos, nos encontraríamos con posiciones irremediablemente equívocas. En concreto, la presunta «filosofía de la matemática» de los *Elementos* nos podría recordar alguna vez a Platón, otras veces a Aristóteles, pero sus cauces de pensamiento —digamos— son generalmente los abiertos por

la tradición matemática más antigua (en el caso de la geometría plana y de la aritmética) o más moderna (en el caso de la teoría de la proporción y del estudio de rectas racionales e irracionales) —tradiciones también conocidas por Platón y Aristóteles de modo que en ocasiones no es fácil distinguir las voces de los ecos—. En general, ni antes ni después del divino Platón, los motivos filosóficos que no venían tamizados por el desarrollo relativamente autónomo de la tradición matemática misma, alcanzaron a tener un influjo crucial o decisivo sobre el curso teórico o sistemático de la geometría griega [8]. Por otro lado, si bien es

[8] Ni siquiera es perceptible un efecto directo de los problemas de la infinitud y del continuo suscitados por la dialéctica eleática y por la discusión cosmológica posterior. *Vid.* W. R. KNORR, «Infinity and continuity: the interaction of mathematics and philosophy in antiquity», en N. KRETZMANN, ed., *Infinity and Continuity in Ancient and Medieval Thought*, Ithaca, Nueva Jersey, 1982, págs. 112-145. Cabe apreciar algún eco de las aporías cosmológicas en algunas nociones explicitadas por Euclides en la cabecera «axiomática» de los *Elementos*, vid., e. g., V. VITA, «L'infinito matematico in Aristotele e nel suo tempo», *Boll. di Storia delle Scienze Matematiche* VI - 2 (1986), 109-132. Pero no hay indicios de que tales aporías filosóficas o dialécticas determinaran o cuestionaran el curso de la geometría de modo parecido a como influyeron en ciertas posiciones filosóficas acerca de la matemática o en el curso de la propia filosofía natural; antes bien, las nociones y proposiciones geométricas solían considerarse como un fondo de datos o resultados disponibles en orden a la resolución de ciertos problemas cosmológicos, *vid.*, por ej., ARISTÓTELES, *Física* VI 2, 233a32-b15. También es errónea, amén de simplista, la interpretación de los *Elementos* sugerida por Popper: se habrían concebido justamente como respuesta a *un* problema y tal problema no habría sido sino de orden *cosmológico (vid.* K. R. POPPER, «The cosmological origins of Euclidean geometry», en I. LAKATOS, ed., *Problems in the Philosophy of Mathematics*, Amsterdam, 1967, págs. 18-20). En L. VEGA, *La trama de la demostración (Los griegos y la razón tejedora de pruebas)*, Madrid, 1990, I, §§ 2-4, se discuten diversos aspectos críticos de esa presunta influencia dialéctica o filosófica sobre el desarrollo

cierto que la geometría euclídea se hace eco de —al mismo tiempo que facilita— diversas aplicaciones (astronómicas, armónicas, ópticas, geográficas), ninguna de ellas merece un excurso o una mención siquiera en los *Elementos*.

En cambio, el segundo objetivo —relativo al aprendizaje de la disciplina de la geometría—, aunque tampoco tiene una mención expresa en el tratado, resulta francamente visible: consiste, dice Proclo, en «perfeccionar la intelección del conjunto de la geometría por parte del estudiante». Se echa de ver, sin ir más lejos, en el interés tanto teórico como práctico de las pruebas euclídeas: Euclides se interesa no sólo por demostrar que algo es el caso (que la figura tal o cual tiene tales o cuales propiedades), sino por demostrar cómo se pueden obtener este y otros tipos de resultados (e.g., en I, 1-26; I, 45; II, 14; IV, 2-16); el antiguo profesor alejandrino no sólo enseña una ciencia, sino que, en cierto modo, parece empeñado en enseñar a aprenderla y construirla (y, para colmo de felicidad, sin el menor alarde de pedagogía).

2. *El lugar del tratado de Euclides en la tradición de «Elementos»*

Para tener mayores luces sobre la significación histórica de la contribución de Euclides, conviene volver a su dimensión metodológica y recordar en esta perspectiva la distinción entre el sentido amplio o genérico y el sentido estricto de «elemento».

de las matemáticas. En fin, sobre el interés de Proclo en presentar a Euclides como un platónico convencido, *vid.* D. J. O'MEARA, *Pythagoras revived. Mathematics and Philosophy in Late Antiquity*, Oxford, 1989, págs. 170-176 en particular.

Supongamos un armazón deductivo de la forma siguiente:

$$\langle A_1, A_2 \ldots \rightarrow \langle P_j \rightarrow P_k \rightarrow \ldots \rightarrow Q_n \rangle \rangle,$$

donde las proposiciones de tipo A son principios o asertos primordiales en un campo de conocimiento, las proposiciones de tipo P son teoremas o resultados conocidos, y la proposición de tipo Q es la conclusión que se acaba de obtener; la flecha '→' señala el orden de deducción.

En la medida en que P_j, P_k ... funcionan como lemas o pasos deductivos que llevan a la obtención de Q_n, cumplen la función de elementos en sentido amplio o genérico. Ahora bien, si P_j, P_k ... han sido a su vez derivados o probados a partir de unos primeros principios A_1, A_2 ..., podemos decir que también hay elementos en el otro sentido, más propio y restringido: pero éstos no son las proposiciones de tipo P, sino las de tipo A, de las que así mismo se deriva indirectamente, a través de P_j, P_k ..., el último resultado, Q_n. En el primer caso, cuando nos atenemos al fragmento $\langle P_j, P_k \ldots \rightarrow Q_n \rangle$ y tomamos P_j, P_k ... como puntos de partida para la obtención de Q_n —al margen de si tienen o no el carácter de pasos intermedios, derivables de unas proposiciones básicas del tipo A—, nos movemos en el estrecho ámbito de un «núcleo» deductivo. En el segundo caso, en cambio, designamos ciertas proposiciones A_1, A_2 ... como puntos absolutamente iniciales de ésta o de cualquier otra cadena deductiva vinculable a ellas y, así, damos una disposición «axiomatiforme» a un cuerpo de conocimientos o de proposiciones: pasamos a movernos en el ámbito de una «teoría» deductiva.

Los dos casos evidencian el proceder eslabonado de la deducción, la transitividad de los encadenamientos deductivos (si vale una deducción '$\alpha \rightarrow \beta$' y otra '$\beta \rightarrow \gamma$', vale así mismo la deducción '$\alpha \rightarrow \gamma$'), y su capacidad para ir acumulando nuevos resultados Q_n, Q_{n+1}... sobre la base de los ya conocidos. Pero el segundo caso, la introducción expresa de unos primeros principios A_1, A_2 ..., nos permite subsumir el núcleo particular $\langle P_j, P_k \ldots \rightarrow Q_n \rangle$ o cualquier otro fragmento parecido dentro de un contexto mucho más comprensivo y nos autoriza a armar este contexto como un cuerpo sistemático de conocimientos. En el marco de una ciencia deductiva, en matemáticas, un tratado de *elementos* en el primer sentido cobra visos de ser una organización parcial o un ensayo provisional cuando es contemplado en la perspectiva de una sistematización más cabal y comprensiva como la representada por un tratado de *elementos* en el sentido segundo, más estricto. He ahí, a los ojos de Proclo y de otros comentadores antiguos, uno de los motivos de la excelencia metódica del tratado de Euclides dentro de la tradición de los *Elementos* de geometría. Por lo que hoy sabemos, el de Euclides es el primer tratado que distingue expresamente un conjunto determinado de primeros principios, con un alcance general en este área de conocimiento, y el único que los subdistingue en definiciones [*hóroi*], postulados [*aitémata*] y nociones comunes [*koinaì énnoiai*].

Actualmente también nos tienta la conversión de esta clave metodológica en una clave de interpretación histórica: así, podríamos ver el desarrollo de la tradición de los *Elementos* como una linea de sistematización progresiva desde sus inicios con Hipócrates hasta su culminación en Euclides. Por regla general no debemos caer en tentaciones de este tipo: la historia es mucho menos simple y lineal que nuestras metodologías, y nuestras reconstrucciones racionales del progreso científico suelen producir bellas his-

torias, relatos edificantes, antes que una historia verosímil. Sin embargo, en el presente caso, no es inverosímil una evolución de este signo: desde unos *Elementos* que más bien consistían en principios instrumentales, asunciones capaces de aglutinar un núcleo de resultados o de proposiciones conocidas en torno a una cuestión determinada, hasta unos *Elementos* fundados en principios axiomatiformes, asunciones capaces de tejer un cuerpo de conocimientos como una teoría deductiva. Desde luego, la desaparición de los tratados anteriores al de Euclides impide pronunciarse sobre esta conjetura con pruebas documentales; nos contentaremos con indicios.

Los primeros *Elementos* se remontan a Hipócrates de Quíos. De este Hipócrates apenas sabemos algo. Por un lado, según Proclo *(Com.* 212, 24-213, 11), fue el primero en reparar en que el problema de la duplicación del cubo era reducible al problema de hallar dos medias proporcionales en proporción continua (tal vez guiado por una analogía con el problema más sencillo de la duplicación del cuadrado, para cuya solución bastaba hallar una media proporcional [9]). Por otro lado, Simplicio *(Comentarios a la Física de Aristóteles* 60, 22-68, 32), refundiendo al parecer un extracto de la *Historia* de Eudemo, recoge la solución de Hipócrates al problema de la cuadratura de la lúnula: es el fragmento de cierta entidad más antiguo que nos ha

[9] De creer a Platón, en su tiempo podían resolver este problema hasta los esclavos que no habían recibido la menor instrucción en geometría con tal de que fueran griegos o hablaran griego —esto al menos pretende demostrar *Menón* 82b-85b—. Así pues, no tendrán las bendiciones de Platón quienes excluyan de un *curriculum* la opción matemáticas + griego, como combinación disparatada, o mantengan el hábito de las dos (in)culturas: ciencias/letras *(—Graeca sunt? Non leguntur. —Mathematica? Non intelliguntur).*

llegado de los matemáticos griegos y muestra con cierto detalle el proceder deductivo de Hipócrates en la solución de este problema. Tanto la reducción seguida en el primer caso como el procedimiento analítico de prueba adoptado en el segundo caso, permiten sospechar que el método empleado por Hipócrates se atiene a la pauta siguiente:

El punto de partida es una proposición que sienta un conocimiento capital o un resultado primordial para la cuestión planteada —en el fragmento citado se trata de la proposición: los segmentos semejantes de círculos guardan entre sí la misma razón que los cuadrados de sus bases—. Esta proposición se reduce o prueba por referencia a otra más simple y general —a saber: los cuadrados de los diámetros guardan entre sí la misma razón que sus círculos respectivos—, que, a su vez, pasa a ser probada o establecida —remitiéndose a que la razón que guardan los círculos entre sí es la misma que la que tienen sus segmentos, pues los segmentos que son semejantes comprenden las mismas partes en sus círculos respectivos—. Por último, se distinguen los diversos casos comprendidos por el problema (e.g.: la cuadratura de una lúnula cuya circunferencia exterior sea igual, mayor o menor que un semicírculo) y se procede a resolverlos sobre la base de la proposición de partida; es interesante advertir que algunos de los casos examinados requieren la consideración de «diorismos», condiciones determinantes de su solución efectiva. En suma: Hipócrates se ocupa ante todo de resolver problemas; los trata de forma retroductiva (retrotrayendo la prueba de unas proposiciones iniciales a la prueba de otras más básicas), de un modo similar al que más adelante se denominará «método de análisis» y se supondrá característico de la investigación geométrica griega; a este fin selecciona las proposiciones pertinentes no en calidad de principios geo-

métricos generales, sino en función de las condiciones y supuestos oportunos para el problema particular que debe resolver. El aspecto general de esta trama no es otro que el de un «núcleo» deductivo. Este procedimiento también resulta afín al señalado por Platón en *Menón* 86e-87a, cuando recomienda el proceder «a partir de hipótesis [*ex hypothéseos*]» que siguen los geómetras: aquí una hipótesis viene a representar un medio de prueba en el sentido de constituir un supuesto o una condición a los que se retrotrae la solubilidad de la cuestión considerada.

Si unimos estos cabos sueltos, nos inclinaremos a pensar que los *Elementos* atribuidos a Hipócrates respondían más bien al sentido amplio o genérico de *stoikheîa* y a su empleo como principios instrumentales, como proposiciones capitales dentro de núcleos deductivos ordenados a la solución de problemas. En la misma línea cabría situar los *Elementos* de León, habida cuenta de que León también es recordado por su contribución al descubrimiento de «diorismos, cuya finalidad es determinar cuándo el problema investigado es susceptible de solución y cuándo no» (Proclo, *Com.* 67, 1-2). Ahí se detienen nuestras pistas y nada podemos decir sobre la índole del tratado de Teudio. Pero Platón, en la *República,* da a entender que el método de hipótesis de los geómetras ya no sólo consiste en la heurística *analítica* de problemas, sino que ha iniciado una conversión *sintética* hacia la deducción de teoremas: la geometría ha empezado a instituir ciertas aserciones como principios de los que ya no cabe dar razón ni considerar posibles alternativas, pues los geómetras los toman como si fueran obvios para cualquiera y, partiendo de ellos, deducen el resto hasta concluir en el objeto final de la prueba. Esta inmovilización axiomatiforme descalifica ahora, a los ojos de Platón, el conocimiento discursivo de los geómetras frente al método dialéctico del filósofo que continúa ascendiendo en pos de otros principios superiores hasta la idea suprema de Bien *(República* VI 510c-511b, VII 533c). Aristóteles será así mismo testigo de las ambigüedades de la situación: declara que son elementos «las proposiciones geométricas cuyas demostraciones están contenidas en las pruebas de otras proposiciones, de todas o de la mayoría» *(Metafís.* B 3, 998a25-26) y que «las demostraciones primeras e implícitas en otras demostraciones se llaman elementos» *(ibid.* Δ 3, 1014a35-b2). Un postrer testimonio de esta época de transición podría ser la distinción misma de Menecmo entre los dos sentidos, el amplio o genérico y el restringido, del término *elemento* en geometría.

Todavía cabe añadir a estos indicios algunas pistas procedentes de otra tradición ambigua: la de los problemas y teoremas. Un problema representaba ante todo un objeto geométrico (e.g., la construcción de una figura) a hacer. Un teorema era, en cambio, una proposición a establecer acerca de alguna característica —propiedad, relación— esencial de objetos matemáticos construidos o dados. (En esta línea cabe recordar que el primer teorema de los *Elementos,* la proposición I, 4, viene a continuación de la prueba de tres problemas previos.) Proclo glosa además otra diferencia metodológica de interés *(Com.* 79, 12-80, 16): en la solución de problemas se puede contar con la lógica relativamente abierta de ciertas condiciones de posibilidad; pero en la demostración de teoremas sólo cuenta la lógica estricta de la necesidad o de la imposibilidad, capaz de sentar conclusiones absolutas y generales. Esta diversificación tiene que ver con ciertos motivos internos (conceptuales, teóricos) de desarrollo del rigor de la prueba deductiva en matemáticas —motivos relacionados sobre todo con el estudio de las magnitudes inconmensurables y con el tra-

bajo sobre el concepto de proporción—; por otra parte, confirma la congruencia de la solución de problemas con los medios reductivos y heurísticos del *análisis* geométrico, y la afinidad de la prueba de teoremas con la sistematización deductiva de la *síntesis*. El «mecánico» Carpo, recordado por Proclo, apuntaba además que al vérselas con un problema cabía acudir a una vía relativamente eficaz de descubrimiento de la solución —el análisis— sin que, por el contrario, se conociera ningún método heurístico general que se aplicara con igual fortuna a los teoremas. Todos éstos son aspectos en los que no puedo entrar aquí [10]. Baste mencionar, por último, el hecho de que esta doble orientación de la investigación en geometría había conducido a controversias acerca de la índole de las proposiciones geométricas dentro del círculo platónico o entre gente más o menos relacionada con la Academia: Espeusipo y Anfínomo defendían su calidad de teoremas, pues no en vano versaban sobre objetos eternos de una ciencia teórica, contemplativa; en cambio, Menecmo vindicaba su condición primigenia de problemas. Lo cierto, en cualquier caso, es que los ecos de esa diversificación y de esta polémica apenas se dejan notar en la sistematización deductiva uniforme

[10] *Vid.* P. TANNERY, *La géométrie grecque. 1ère Partie: Histoire de la géométrie élémentaire*, París, 1887, Hildesheim-Zurich-Nueva York, (reimp. 1988); G. A. ALLMAN, *Greek Geometry from Thales to Euclid*, Nueva York, (reimp. 1976); G. CAMBIANO, «Il metodo ipotetico e la origine della sistemazione euclidea della geometria», *Rivista di Filosofia* 58 (1967), 115-149; W. R. KNORR, *The Evolution...*, op. cit., 1975; *The Ancient Tradition of Geometric Problems*, Boston, 1986; L. VEGA, *La trama de la demostración...*, op. cit., 1990, I, § 3; IV, § 1. En el ya citado I. THOMAS, ed., *Selections of Greek Mathematics*, 1939, 1967[3], pueden verse textos y testimonios al respecto; de esta compilación han partido las versiones y comentarios, a veces un tanto *sui generis*, de J. D. GARCÍA-BACCA, *Textos clásicos para la historia de las ciencias*, I, Caracas, 1961.

de los *Elementos* de Euclides. Tanto los problemas como los teoremas se presentan como proposiciones con un patrón de prueba común en el tratado euclídeo [11].

La consideración de este patrón general de las pruebas de Euclides muestra otro aspecto del papel normalizador que desempeñan sus *Elementos*. Si su disposición axiomatiforme marca la norma de exposición de una disciplina deductiva como la geometría, el tipo de prueba seguido se convierte parejamente en el canon a seguir por toda demostración geométrica digna de tal nombre. Proclo, una vez más, puede servir de guía en este terreno; en el contexto de su larga glosa de la proposición 1 del libro I, señala una especie de pauta general de prueba de las proposiciones de los *Elementos* (*Com.* 203, 1 ss.). Cuando Euclides la sigue de modo cabal —e.g., en el caso citado I, 1—, comprende los pasos:

a) Enunciado [*prótasis*]: proposición del objeto a construir si se trata de un problema, o del aserto a establecer si se trata de un teorema; su formulación perfecta declara por una parte lo que está o se considera dado y, por otra

[11] No obstante, cabría reparar en cierta afinidad entre la prueba de los problemas y el uso de los postulados, así como entre la demostración de los teoremas y el uso de las definiciones. Pero su significación no se puede exagerar (por ejemplo, en la línea de una relación entre los postulados y las cuestiones de «existencia» matemática como la sugerida por H. G. ZEUTHEN, «Die geometrische Construction als "Existenzbeweis" in der antiken Mathematik», *Mathematische Annalen* 47 (1896), 222-228; cf. W. R. KNORR: «Construction as existence proof in ancient Geometry», *Ancient Philosophy* 3 (1983), 125-148. Por lo demás, la distinción entre los problemas y los teoremas tampoco representa una distribución cabal del conjunto de las proposiciones de los *Elementos:* hay proposiciones que envuelven rasgos característicos de ambas «categorías» y las hay que parecen discurrir con independencia de una y otra.

parte, lo que se busca probar —siempre en términos generales—.

b) Exposición [*ékthesis*]: presentación de lo dado o introducción de un caso determinado de aplicación del enunciado mediante la cláusula «sea...» y el uso de letras como abreviaturas que designan los elementos del caso (puntos, líneas, figuras, magnitudes, números).

c) Determinación ò delimitación [*diorismós*]: especificación del objeto de la prueba por referencia al caso expuesto; en los problemas se concreta la tarea con la fórmula; «lo que se requiere es...», en los teoremas se concreta la aserción con la fórmula «digo que...». Por *diorismós* también se entiende a veces —como ya sabemos— una delimitación en el sentido más preciso de fijar las condiciones de posibilidad de la prueba: si lo buscado es imposible o es posible y, entonces, cómo se puede conseguir efectivamente; cuando tiene este significado de condición o límite de la prueba, suele seguir inmediatamente a la *prótasis* a manera de apéndice del enunciado del problema *(vid.* e. g., I, 22).

d) Preparación [*kataskeué*]: urdimbre o disposición de construcciones y relaciones, a partir de lo dado y en orden a la obtención del resultado propuesto.

e) Demostración [*apódeixis*]: proceso demostrativo propiamente dicho que consiste en la derivación de consecuencias sobre la base de los conocimientos previos, ya sean proposiciones primordiales (definiciones, postulados, nociones comunes), ya sean proposiciones sentadas en pruebas anteriores.

f) Conclusión [*sympérasma*]: aserción de que se ha satisfecho el *diorismós* en el caso de un problema o reiteración de la *prótasis* en el caso de un teorema como confirmación de que el objeto de la prueba ha sido establecido.

Viene marcada indefectiblemente por una partícula consecutiva fuerte ('por consiguiente [*ára*]'). La conclusión de los problemas se remata con la cláusula final: «que era lo que había que hacer [*hóper édei poiésai*]»; la de los teoremas, con la fórmula: «que era lo que había que demostrar [*hóper édei deîxai*]».

Proclo reconoce que no siempre se dan todos estos pasos en las pruebas de los *Elementos*, pero llama la atención sobre tres que, a su juicio, nunca han de faltar: a) el enunciado, b) la demostración, c) la conclusión. «Porque es igualmente necesario conocer de antemano el objeto de la prueba, demostrarlo por los medios debidos y concluir lo que se ha probado» *(Com.* 203, 17-19).

Hay indicios de que antes de Euclides ya existía en medios matemáticos un patrón tradicional de prueba parecido al que consagran los *Elementos*. Aparte de ciertas huellas incidentales que se traslucen en el lenguaje metasilogístico que Aristóteles emplea en algunos pasajes de los *Primeros Analíticos* (e. g., I 5, 27a2-14), esas semejanzas pueden vislumbrarse en la práctica seguida por Autólico en los primeros tratados matemáticos completos que hoy se conservan *(Sobre la esfera en movimiento, Sobre ortos y ocasos)*. El procedimiento tradicional seguía los pasos de la enunciación, una especie de fusión de la exposición o *ékthesis* con la preparación, y finalmente la demostración [12]. Así

[12] También serán luego destacados por los traductores medievales de los *Elementos (vid.* más adelante, IV, § 3.2 B). En realidad, este proceder (enunciado, construcción sobre un caso particular cualquiera, prueba) y su trasfondo inferencial intuitivo (referido a un diagrama o un caso-muestra) podrían considerarse típicos de la práctica matemática más común a lo largo de la historia. No estaría de más replantear a esta luz las discusiones y contraposiciones habituales, en punto a rigor, entre esta especie de «ingeniería de la prueba» que suele practicar la gente cuando hace

a) La composición euclídea fue, para empezar, un repertorio básico de los resultados probados y las proposiciones demostradas; un archivo tan cumplido que hizo superfluo cualquier otro tratado matemático del mismo alcance y género, y devino una referencia común en las investigaciones subsiguientes: siempre que hacía falta un lema elemental bastaba, por lo regular, mencionar su presencia en los *Elementos* sin que fuera preciso detenerse a probarlo. Los *Elementos* eran fuente de autoridad: cualquier asunción hecha en su nombre quedaba por ello mismo autorizada. Hay raras excepciones, e. g., la independencia de Arquímedes o incluso su renuencia a citar a Euclides, que parecen confirmar la regla [13]. Este papel singular, por otro lado, facilita el crecimiento normal de la geometría por adiciones sucesivas que a partir de esta base pueden ir completando lo que falta —esta forma de progreso de una disciplina, *epídosis*, era la prevista ya por Aristóteles en la *Ética Nicomáquea* I 7, 1098a23-25—. El largo rosario de comentarios y escolios de los *Elementos* es buena muestra de la existencia de un desarrollo parecido.

b) Los *Elementos* fijaron una especie de estándar metodológico o nivel básico de exigencia tanto en lo referente a la sistematización deductiva de un cuerpo de conocimientos como en lo referente al rigor informal de la prueba matemática. También representaron, por otro lado, una

[13] Sobre las curiosas relaciones entre Arquímedes y los *Elementos* de Euclides, *vid.* W. R. KNORR, «Archimedes and the *Elements:* Proposal for a revised chronological ordering of the Archimedian Corpus», *Archive for the History of Exact Sciences* XIX (1978), 211-290; cf. algunas acotaciones críticas oportunas en T. SATO, «A reconstruction of *The Method* Proposition 17 and the development of Archimedes' thought on quadratura», *Historia scientiarum* 31 (1986), 61-86, y 32 (1987), 75-142.

normalización de la exposición demostrativa de las proposiciones geométricas. Estos dos aspectos de lo que luego se llamaría *ars disserendi,* el metodológico o sistemático y el disciplinario o expositivo (solidarios, como ya he sugerido, en la tradición griega de los *Elementos),* determinaron, mediante su ejemplar materialización en el tratado de Euclides, la instauración alejandrina de la geometría como la disciplina matemática —*i.e.* demostrativa— por excelencia. Más aún: fuera del dominio matemático, los *Elementos* también ejercieron de modelo de la ciencia sistemática y precisa, según muestran los deseos expresos de Galeno de incorporar esa estructuración deductiva a la teoría médica con el fin de liberarla de disquisiciones filosóficas y discusiones dialécticas. Ni que decir tiene que la proyección general de los *Elementos* como directriz regulativa, arquetipo de ciencia rigurosa e incontrovertible, es más bien programática o corre a veces el riesgo de resultar prematura, cuando no de volverse contraproducente [14].

c) La impronta institucional de los *Elementos* se echa de ver, en fin, no sólo en sus repercusiones metódicas y disciplinarias, sino en consecuencias de otro género, más sustantivas. Influyeron decisivamente en ciertas formas de

[14] En la misma *Ética Nicomáquea* —a continuación del pasaje antes citado—, ARISTÓTELES había advertido: «Pero ... no hemos de buscar del mismo modo el rigor en todas las cuestiones, sino, en cada una según la materia que subyazga a ella y en el grado pertinente para la investigación particular en curso» (1098a25-28). G. E. R. LLOYD ha llamado la atención sobre los efectos perniciosos de esta proyección axiomatiforme tanto para el desarrollo de la propia matemática como para otros desarrollos científicos fuera de ese ámbito; *vid.* «Greek notions and practice of proof», 1987, incluido en su *Demystifying Mentalities* (Cambridge, 1990, c. 3, págs. 87-90 en particular). Proclo, sin embargo, todavía intentará dar visos de respetabilidad a su teología neoplatónica por medio de una *stoikheiosis* «axiomática».

conceptualización científica de su tiempo, especialmente las deudoras de la geometría. Aunque esta orientación geométrica ya viniera avalada por algunas tradiciones preeuclídeas, los *Elementos* marcaron un hito decisivo en la geometrización de las matemáticas y de sus dominios de aplicación (desde la astronomía o la estática hasta la geografía). También sancionaron la suerte de otras tradiciones marginales (la proporción numérica, el cálculo logístico) [15].

Con todo y con esto, las indicaciones anteriores no quieren dar a entender en absoluto que la matemática griega, desde el siglo III a. C. en adelante, sólo sea un conjunto de notas a pie de página de los *Elementos*. De hecho, el tratado aparece sobre un transfondo cultural de discusiones y tradiciones dispares que militarán en contra de su implantación dogmática y absoluta. Según recuerda Proclo (*Com.* 199, 3-14), son varios los autores helenísticos que oponen objeciones a la geometría y, en particular, a la sistematización euclídea. La mayor parte de ellos alimentan dudas sobre los principios y tratan de mostrar que las pro-

[15] A pesar de algunos estudios recientes, e.g. de D. H. FOWLER, la suerte de estas tradiciones marginales, tanto antes como después de Euclides (en Herón, en Diofanto), sigue sin recibir la atención que merece. Por lo demás, es obvio que la rara fortuna institucional de los *Elementos* de Euclides no se debe únicamente a sus características propias o a sus virtudes internas. También contribuyeron otras circunstancias de muy diverso orden, cuya consideración nos llevaría ahora demasiado lejos. *Vid.*, por ej., G. GIANNANTONI, «Su alcuni problemi circa i rapporti tra scienza e filosofia nell'età ellenistica», en G. GIANNANTONI y M. VEGETTI, eds., *La scienza ellenistica*, Nápoles, 1984, págs. 41-71; M. VEGETTI: «La scienza ellenistica», ibid., págs. 431-470; G. CAMBIANO, «La démonstration géométrique», en M. DETIENNE, dir., *Les savoirs de l'écriture en Grèce ancienne*, París, 1988, págs. 251-272; L. VEGA, *La trama de la demostración...*, op. cit., IV, § 4.

posiciones derivadas resultan infundadas: unos, como los escépticos, porque no admiten los principios ni, en general, la existencia de una ciencia demostrativa; otros, como los epicúreos, porque discrepan de ciertos principios geométricos. Pero también hay quienes, reconociendo los principios, niegan que sus consecuencias se hallen demostradas a menos que se añadan o declaren otros supuestos no explicitados en los *Elementos;* entre éstos últimos, quizás se halle algún filósofo (e. g., el epicúreo Zenón de Sidón, al que responde Posidonio en defensa de Euclides), pero su actitud parece ser típica de los comentadores y editores más o menos matemáticos del tratado. Todas estas noticias y otras ocasionales, incluso las procedentes de una fuente crítica tan puntillosa como Sexto Empírico, no son demasiado precisas; desde luego, no documentan ni avalan en absoluto la presunción de una especie de geometría rival de los *Elementos*, una geometría —digamos— «no euclídea» [16]. Pero son elocuentes: hacen ver que el estatuto pri-

[16] Aparte de las referencias de Proclo, tenemos noticias de una oposición frontal a la geometría deductiva en general, en particular a la euclídea, por parte de algunos filósofos escépticos y también quizás epicúreos; en el primer caso, contamos con las objeciones de Sexto Empírico en *Pròs Mathematikoús* III; en el segundo caso, no es tan fácil identificar al crítico o el alcance de la crítica (cf., por ejemplo, la discutida *Carta a Heródoto* de EPICURO, §§ 56-59, trad. en C. GARCÍA GUAL, *Epicuro*, Madrid, 1981, págs. 98-99). Pero en ambos casos se plantean por lo regular cuestiones conceptuales y filosóficas (e. g., dificultades relacionadas con las nociones de punto, línea, superficie, sólido, con la idea de divisibilidad...), antes que cuestiones técnicas. *Vid.* M. GIGANTE, *Scetticismo e epicureismo*, Nápoles, 1981, págs. 209-214 en especial; I. MUELLER, «Geometry and scepticism», en J. BARNES y otros, *Science and Speculation. Studies in Hellenistic Theory and Practice*, Cambridge-París, 1982, págs. 69-95. De la filosofía epicúrea precisamente se ha querido extraer una especie de geometría «no euclídea», una alternativa atomista que descartaría la bisecabilidad o divisibilidad indefinida de los *Elementos*

vilegiado de los *Elementos* no cierra la discusión o acaba con las discrepancias, ni siquiera en el seno de la tradición

(e. g.: I, 10; X, 1); *vid.* e. g., H. VON ARNIM, «Epikurs Lehre vom Minimum», *Almanach der Kaiserlichen Akad. der Wissenschaften,* Viena, 1907, págs. 383-399; S. LURIA, «Die Infinitesimaltheorie der antiken Atomisten», *Quellen und Studien zur Geschichte der Mathematik* 2 (1933), 106-172. Pero los que luego han abrigado esta presunción, como J. MAU, terminan por reconocer que no está documentada, *vid.* su «Was there a special Epicurean mathematics?», *Phronesis,* Sup. 1 (1979), 421-430; también es instructiva la crítica de G. VLASTOS: «Minimal parts in Epicurean atomism», *Isis* 56 (1965), 121-147, y «Zeno of Sidon as a critic of Euclid», en L. WALLACH, ed., *The Classical Tradition,* Ithaca, Nueva Jersey, 1966, págs. 148-159. La asunción de indivisibles tiene, desde luego, una larga tradición en el pensamiento griego y, según PROCLO, *Com.* 279, 5, Jenócrates, discípulo de Platón, ya se había encargado de introducirla en matemáticas; pero no hay constancia de que se produjera una geometría atomista peculiar y rival de la establecida.

Otra suerte de geometría «no euclídea» es la que se ha querido adivinar en algunas referencias de ARISTÓTELES a la teoría de las paralelas; por ejemplo, en *APo.* II 17, 66a13-15, para mostrar que una conclusión falsa puede seguirse de premisas diferentes aduce este caso: la intersección de dos líneas paralelas podría seguirse bien de la hipótesis de que el ángulo interno es mayor que el externo del mismo lado, bien de la hipótesis de que la suma de los ángulos del triángulo vale más de dos rectos (esas hipótesis niegan respectivamente las proposiciones I, 27, y I, 28, dos teoremas característicos de la teoría euclídea de las paralelas). A juicio de un intérprete como I. TÓTH, «Das Parallelenproblem in Corpus Aristotelicum», *Archive for History of Exact Sciences* 3 (1967), 249-322, ésta y otras alusiones por el estilo significan la existencia de geometrías basadas en la negación de la teoría de las paralelas de los *Elementos,* antes del propio Euclides. Pero los textos aducidos y sus contextos quedan lejos de apoyar una interpretación tan aventurada. Cf. el detenido examen crítico de G. J. KAYAS, «Aristote et les géométries non-euclidiennes avant et après Euclide», *Revue des Questions Scientifiques* 147 (1976), 175-194, 281-301, 475-465. Por lo demás, los matemáticos griegos conocían resultados de este tipo en geometría esférica, resultados que tomaban no como una alternativa sino como un desarrollo parejo al de la geometría euclídea, trazado a imagen y semejanza de los *Elementos.* Así:

matemática, sobre puntos particulares —alguno tan capital como la identificación de las nociones comunes necesarias y suficientes, o como la índole de un postulado—; hacen ver que esa consagración tampoco descarta el empleo de otras definiciones alternativas o la revisión crítica de varias pruebas euclídeas. Por ejemplo, Apolonio parece arrogarse el derecho de probar alguna de las nociones comunes, supuestamente indemostrables; Herón sólo admite tres nociones comunes, corrige demostraciones y evita el uso de la reducción al absurdo; Pappo explicita axiomas suplementarios y otros supuestos implícitos en las definiciones; Teón, el editor principal de los *Elementos,* añade alguna demostración de propia cosecha; el mismo Proclo se suma a la tradición, presidida quizás por Tolemeo, de los que intentan probar como teorema el postulado euclídeo de las paralelas. La verdad es que ni la matemática de más estricta observancia euclídea se privó del placer de agregar lemas y puntualizaciones a Euclides. Así pues, los *Elementos,* a pesar de la transfiguración institucional que les sobreviene, no dejan de ser una obra estimulante dentro de una tradición viva. Constituyen una fuente de glosas y comentarios a la vez que conviven con otras formas de hacer

las proposiciones 10 y 11 del tratado *Perì sphaíras* de Menelao —astrónomo alejandrino de finales del siglo I— establecen que la suma de los ángulos de un triángulo esférico excede de dos rectos, y este exceso es proporcional al área; pues bien, el tratado está compuesto por analogía con la geometría plana de Euclides, y ni a Menelao ni a ningún autor antiguo conocido (incluidos los comentadores árabes que traducen el tratado de Menelao y están interesados en la teoría de las paralelas, como Nasir al-Din al-Tūsī) se les ocurre pensar en algo parecido a una «geometría alternativa». Sobre esta tradición geométrica esférica, puede verse B. A. ROSENFELD, *A History of Non-Euclidean Geometry. Evolution of the Concept of a Geometric Space,* Nueva York-Berlín-Heidelberg, 1988; cap. 1, págs. 1-22 en particular.

matemáticas, e. g., con la construcción geométrica por medio de la *neusis* —la interposición de una recta que se inclina hacia un punto dado y podría suponer el uso de una especie de regla marcada móvil— o por medios que se decían «mecánicos»; con el cálculo numérico, «logístico»; con ciertas primicias algebraicas como la avanzadas por Diofanto. Por consiguiente, hay que distinguir entre, por un lado, la significación metodológica y disciplinaria que el tratado de Euclides pudo adquirir en el medio helenístico y, por otro lado, la escolarización que luego hubo de padecer al transformarse en el manual de geometría por excelencia.

Es cierto que esta normalización escolar ya se insinúa en la edición [*ékdosis*] estándar de los *Elementos* que hace Teón de Alejandría en el siglo IV. Pero no se torna claramente perceptible hasta su conversión, primero árabe y después escolástica, en una especie de libro de texto susceptible de simplificación a la medida de los usos que la geometría alcanzaba a tener en una y otra cultura —usos al principio mucho menores, y más triviales por cierto, en la cristiana—. Luego asistimos a una suerte de bifurcación. Por una parte, esta misma condición escolar es la que prevalece en muchos usos del texto desde el Renacimiento e, incluso, se va acentuando hasta finales del siglo pasado cuando la ascensión del cálculo y del álgebra termina por desalojar los *Elementos* de su posición hegemónica en la enseñanza media y universitaria de Francia y del Reino Unido. Pero, por otra parte, sobre todo en medios filosóficos sensibles al desarrollo de la ciencia moderna, los *Elementos* continúan ejerciendo un papel de paradigma del rigor informal y de la demostración geométrica. Así: bajo el lema del «more geometrico» o del «esprit de la géométrie» del siglo XVII, Euclides se convierte en el epónimo no tanto

de una disciplina matemática como de un método de axiomatización. Algunos escolásticos tardíos ya se habían aplicado en París o en Padua a examinar la peculiar lógica deductiva de la prueba euclídea. Sin embargo, es a partir de los programas «racionalistas» del siglo XVII (alentados por Descartes, Pascal, Leibniz) cuando adquiere pleno sentido esta observación de Brunschvicg: «Euclides, para las numerosas generaciones que se han nutrido de su sustancia, puede que haya sido menos un profesor de geometría que un profesor de lógica» *(Les étapes de la philosophie mathématique,* París, 1947², § 49, pág. 84).

De ahí no se sigue que los *Elementos* sean el acta inaugural del método axiomático clásico de los siglo XVII-XIX, aunque sí constituyen una especie de preludio y ofician como un tópico casi obligado de referencia. Más adelante volveré sobre este punto, donde conviene hacer otra distinción de importancia: entre la «axiomatización» euclídea, i.e. la trama deductiva de los *Elementos,* y la axiomatización «euclidiana», i.e. el método axiomático desarrollado en los tiempos modernos por diversas contribuciones a la geometría clásica.

La fortuna institucional de los *Elementos* puede inducir también a otro malentendido que conviene despejar: consiste en creer que forman un tratado homogéneo y compacto, tan autosuficiente y clausurado en sí mismo que ha borrado todo rastro de las tradiciones matemáticas anteriores. La ausencia de *Elementos* preuclídeos favorece esta impresión, pero sólo puede ser una impresión aparente y superficial. Más justo es concluir que lejos de condenar todas esas tradiciones a la extinción, la síntesis de Euclides contribuyó en cierto modo a la supervivencia y a la integración de algunas de ellas. Como Knorr ha apuntado *(The Evolution of Euclidean Elements,* op. cit., 1975, pág. 312),

lo deseable sería que las otras ramas de la matemática griega —la aritmética, en particular— hubieran encontrado compiladores de una inteligencia y de una capacidad semejante a las mostradas por Euclides es el caso de la geometría. Por lo demás, la composición de los *Elementos* tampoco es la de una obra marmórea, esculpida en un solo bloque y perfectamente acabada. Si la contemplamos en una visión global y panorámica, nos recuerda más bien una vieja catedral en cuya construcción, aunque esté presidida por un plan arquitectónico deliberadamente sistemático e integrador, se han estremezclado ya desde un principio materiales procedentes de diversas épocas y formaciones teóricas con distinto grado de desarrollo, a todo lo cual más tarde —a partir de los comentadores alejandrinos y de manos de sus sucesivos editores— se han ido añadiendo ciertos arreglos e, incluso, alguna que otra restauración moderna. Para no quedarnos en una simple impresión, veamos ahora de cerca la constitución interna de los *Elementos*.

III. La constitución de los «Elementos»

1. *El «pórtico axiomático»*

El libro I empieza sin el menor miramiento con una serie de definiciones [*hóroi*] como las siguientes:

1. Un punto es lo que no tiene partes.
2. Una línea es longitud sin anchura.
3. Los extremos de una línea son puntos.
4. Una línea recta es aquella que yace por igual respecto de los puntos que están en ella.
5. Una superficie es lo que sólo tiene longitud y anchura. < ... >

15. Un círculo es una figura plana comprendida por una línea tal que todas las rectas que caen sobre ella desde un punto interior son iguales entre sí.
16. Y el punto se llama centro del círculo. < ... >
23. Son rectas paralelas las que estando en el mismo plano y siendo prolongadas indefinidamente en ambos sentidos, no se encuentran una a otra en ninguno de los dos sentidos.

Aristóteles ya había recogido antiguas nociones del punto como «unidad con posición» (*Metafís.* Δ 7, 1016b26) y «extremo de una línea» (*Tópicos* VI 4, 141b21). Además, había precisado que el punto es algo sustancialmente indivisible y provisto de posición, pero no es parte de la línea ni la línea se compone de puntos —aunque, por movimiento, un punto pueda generar una línea y resultar así origen de una magnitud—. El punto se asemeja al *ahora*, instante indivisible que no forma parte del tiempo y se limita a marcar el comienzo, el final o una división en el tiempo (*Fís.* IV 11, 22a1-21). El paradigma de esta concepción de las unidades de magnitud sería la idea de unidad que transmite la def. 1 del libro VII, donde Euclides sienta la bases tradicionales de la aritmética: «la unidad es aquello en virtud de lo cual cada una de las cosas que hay es una». Esta noción puede haber surgido como una respuesta filosófica a las aporías y problemas suscitados por la amenaza eleática de una divisibilidad infinita, según se desprende de algunas referencias de Platón (e. g., *Rep.* VII 524e-566a) y del propio Aristóteles, aunque tampoco es una idea extraña a la tradición pitagórica. Euclides se hace eco de esta noción, pero no emplea la denominación usual del punto en contextos filosóficos, *stigmé* (punción, marca producida por un objeto puntiagudo), sino el término *semeîon* (signo, señal convencional) que, aun estando presente ya en

el contexto matemático, con Euclides deviene característico de este contexto hasta que los comentadores vuelven al uso indiscriminado de *stigmé* y *semeîon* para referirse al punto. Las definiciones 2 y 5 muestran, con mayor claridad aún, la conformidad de Euclides con nociones y formulaciones anteriores. Se mueven dentro del marco general que Aristóteles evoca en la *Metafísica:* «lo completamente indivisible según la cantidad y carente de posición se llama *unidad,* y cuando es completamente indivisible con posición se llama *punto;* lo divisible de una sola manera se llama *línea;* lo divisible de dos maneras, *superficie,* y lo divisible de las tres maneras *cuerpo*» (Δ 7, 1016a24-28). Así se introducen las dimensiones. Pero este planteamiento y en particular las definiciones citadas 1-2, 5, merecen más atención por sus resonancias intuitivas que por su rendimiento geométrico —en otras palabras: son más útiles a Edwin A. Abbot para imaginar *Planilandia* que a Euclides para sistematizar los *Elementos*—. La Def. 3 relaciona los puntos con las líneas y su orden de aparición evita una objeción aristotélica contra una definición que introdujera el punto en estos mismos términos: si ésta fuera la definición inicial, equivaldría a explicar lo anterior (el punto) por lo posterior (la línea). La Def. 4 es probablemente original de Euclides: elabora una noción de línea recta a partir de su concepción como un rayo óptico o visual, familiar quizás en medios relacionados con la observación astronómica. La imagen subyacente podría ser la que Platón emplea en *Parménides* (137e): en una recta, el medio se interpone delante de uno y otro extremos de manera que eclipsa (encubre, impide ver) los extremos. Ésta no es la única imagen que los griegos se formaron de la línea recta. En todo caso, la versión de Euclides tiene un tono abstracto que soslaya cualquier referencia visual o concreta. La

Def. 15, en cambio, responde fielmente al criterio de equidistancia que Platón utiliza en el mismo pasaje del *Parménides* para determinar la noción de lo redondo o lo circular; Euclides añade la precisión de que el punto equidistante, el centro, es un punto interior, *i.e.* un punto que se encuentra en el mismo plano que el círculo. Al igual que la Def. 2, esta Def. 15 es estática, no es genética: no envuelve unas condiciones de movimiento y generación como las que había considerado una tradición anterior *(vid.* Aristóteles, *Acerca del alma* I 4, 409a4-6), o como las que luego supondrá Herón (e. g.: un círculo es la figura descrita cuando una línea recta, manteniéndose en el mismo plano, se mueve sobre uno de sus extremos tomado como punto fijo hasta retornar a su posición de partida. *Definiciones,* Def. 27). La Def. 23 declara una noción de paralela no ignorada por Aristóteles, quien conocía también ciertas implicaciones de la teoría de las paralelas, como la de que el ángulo interno formado por una transversal no es mayor que el externo del mismo lado, o la de que los ángulos de un triángulo suman dos ángulos rectos *(APr.* II 17, 66a11-15). Con todo, es significativa la opción de Euclides por un criterio primordial de no intersección, no encuentro, para definir las rectas paralelas, frente a otras caracterizaciones posibles, e. g.: la fundada en el criterio de equidistancia (preferido por Posidonio, Gémino, Simplicio), la derivada del criterio de igual dirección (tal vez sugerido por Filopón), o la asociada a una condición aristotélica de finitud (explotada más tarde por Proclo). De hecho, el criterio euclídeo de no intersección y su empleo efectivo a través de un postulado congruente —el (v), véase *infra*— no incurren en ciertas peticiones de principio, ya denunciadas por Aristóteles *(APr.* II 16, 65a4-5), y que no dejarán de reproducirse después de Euclides entre quienes tratan

de ofrecer una concepción más obvia o mejor fundada —aparentemente— que la propuesta en los *Elementos*.

Euclides presenta a continuación cinco postulados [*aitémata*]:

(i) el de trazar una línea de cualquier punto a cualquier punto;
(ii) el de prolongar una recta finita continuamente en línea recta;
(iii) el de describir un círculo con cualquier centro y distancia;
(iv) el de que todos los ángulos rectos son iguales entre sí;
(v) el de que de si una línea recta al caer sobre dos rectas hace los ángulos interiores de un mismo lado menores que dos ángulos rectos, entonces las dos rectas, si son prolongadas indefinidamente, se encontrarán por el lado en el que están los ángulos menores que dos rectos.

El postulado (i) demanda la construcción efectiva de la línea recta una vez conocida su definición. Por contra, la existencia o disponibilidad del punto se da por descontada. Euclides también omite la implicación de que la línea trazada entre dos puntos es única —si dos rectas coincidieran en los mismos extremos, coincidirían cabalmente en toda su longitud—, aunque luego venga a suponerla en la prueba de la proposición I, 4, que establece las condiciones de la igualdad entre triángulos. Varios editores corregirán luego esta omisión interpolando como un nuevo postulado o como otra noción común, una asunción equivalente: dos rectas no encierran un espacio. El postulado (ii) puede entenderse de modo parejo en el sentido de que un segmento rectilíneo sólo puede prolongarse de una única manera por cada extremo y, por ende, dos líneas rectas no podrán tener un segmento común. Una de las críticas del epicúreo Zenón de Sidón a Euclides fue precisamente la de no haber explicitado este supuesto en la prueba de I, 1; con mayor razón cabría denunciar su ausencia en la

prueba de I, 4; mucho más adelante, en la prueba de X, 1, aparecerá tal supuesto de modo expreso e inopinado. El post. (iii) es a la definición del círculo lo que el post. (i) era a la definición de la recta: demanda su construcción efectiva una vez que se cuenta con la noción correspondiente. Se ha pensado a menudo que este postulado, así como el (ii), envuelve la idea de un espacio continuo e infinito: al no indicarse restricción alguna del tamaño del círculo descrito, éste puede ser indefinidamente pequeño en un espacio continuo —con una distancia mínima entre los puntos contiguos— o puede ser indefinidamente grande, infinito; esta suposición de infinitud parece necesaria para establecer la verdad universal del teorema I, 16, cuya prueba es criticable en diversos aspectos [17]. Pero creo que resulta muy aventurado atribuir a Euclides la intención de fijar, con tales postulados, las propiedades definitorias o estructurales del espacio que hoy calificaríamos de «euclidiano». Euclides, según todos los visos, trabaja con objetos geométricos determinados (segmentos, figuras, etc.), no con la idea abstracta o formal de espacio que hoy nos es familiar. En todo caso, lo que sí hacen los postulados (i)-(iii) es sentar las bases operativas de un procedimiento de construcción «por regla y compás»: con la regla se trazan (o «se tiran [*ekbállontai*]») rectas, con el compás se describen círculos; el compás se colapsa cuando se levantan sus dos pies, de modo que no sirve para transportar segmentos; pero la solución de los problemas I, 1-3 muestra que este recurso es superfluo y que los postulados se bastan para trasladar segmentos determinados —esta suficiencia

[17] *Vid.* T. L. HEATH, ed., *The Thirteen Books of Euclid's Elements*, ed. cit., 1926², 1, págs. 280-281; A. DOU, «Los paralogismos de Euclides y Saccheri en la teoría de las paralelas», *Revista de la Real Academia de Ciencias* 61 (1967), 155-174.

da un toque de elegancia al método euclídeo—. En los postulados (i)-(iii) se nota el peso de una tradición entrenada en el estudio de problemas geométricos. A Enópides, un matemático del siglo v a. C. algo mayor que su paisano Hipócrates de Quíos, se le atribuyen la investigación de dos problemas recogidos en los *Elementos* (I, 12, y I, 23), y su solución por medio de este procedimiento de regla y compás (Proclo, *Com.* 283, 7-10, y 333, 4 ss.). La contribución de Euclides bien pudo ser entonces la explicitación cabal de los supuestos de este antiguo método.

El planteamiento euclídeo reviste importancia por diversos motivos. Por una parte, es sintomático del proceder constructivo de Euclides en este ámbito de la geometría plana. Este proceder nos recuerda la directriz aristotélica (e. g., *APo.* I 10, 76a31-36) de que, en relación con determinados objetos, se han de asumir como principios la noción de lo que son y el hecho de que son o existen realmente, mientras que en otros casos cabe suponer su concepto, pero habrá que demostrar su realidad efectiva. Así, Euclides define, por ejemplo, la recta, el círculo, el triángulo equilátero, el cuadrado; postula la construcción de la recta y del círculo; prueba la construcción del triángulo equilátero (I, 1), del cuadrado (I, 46). Pero no deberíamos ver aquí un prurito aristotélico ni, menos aún, una clave para adivinar la noción de construcción o de «existencia matemática» propia de los *Elementos*. Pues Euclides también emplea otros medios de construcción de objetos geométricos que no están cubiertos ni por estos ni por otros postulados explícitos: uno es la generación de una figura plana o sólida mediante el movimiento de otra (e. g., la generación de una esfera por rotación completa de un semicírculo sobre su diámetro); otro es la producción de figuras planas pasando un plano a través de un sólido (según acontece con las secciones cónicas); tales procedimientos se fundan, si acaso, en una definición como la de esfera (XI, Def. 14) o la de cono (XI, Def. 18). Más aún, en el ámbito de las magnitudes en general del libro V o en el de la aritmética del libro VII, Euclides nunca se considera obligado a explicitar postulado alguno de «existencia». Hay, en fin, proposiciones que se dejan leer en términos que hoy diríamos «existenciales» (e. g., VII, 31: «todo número compuesto es medido por algún número primo»; IX, 20: «los números primos son más que cualquier multitud acotada de números primos»), cuya demostración discurre por reducción al absurdo y desde supuestos que no se podrían considerar «constructivos».

Por otra parte, en los postulados (i)-(iii) late una tradición consciente no sólo de la efectividad de un método elemental de construcción, sino de sus limitaciones. Según Pappo (*Col.* III 54-56; IV 270-272), los geómetras habían llegado a distinguir tres tipos de problemas: los *planos*, para cuya solución bastaban líneas rectas y círculos; los *sólidos*, que suponían el uso de secciones cónicas; los *lineales*, que pedían curvas más complicadas, como las «cuadratrices» de Hipias o las espirales de Arquímedes. Quizás llevara su tiempo reconocer que no todo problema de la geometría plana es un problema *plano*. Problemas como la bisección del ángulo, la duplicación del cuadrado, la trisección del ángulo, la duplicación del cubo, la cuadratura del círculo son problemas bien conocidos en la tradición matemática anterior a Euclides. Todos ellos pueden plantearse en geometría plana; pero sólo los dos primeros admiten una solución efectiva con el simple uso de unas figuras elementales (la recta, el círculo) y de las configuraciones compuestas por ella. Euclides ofrece un criterio de discriminación al respecto: sea A un problema planteado en

los términos de la geometría plana; A es un problema *plano* sólo si es efectivamente soluble por el procedimiento de regla y compás; luego, si A es un problema plano, A tiene una construcción efectiva sobre la base sentada en los *Elementos*. Puede que los matemáticos alejandrinos de estricta observancia euclídea trataran de imponer además una directriz de parsimonia en este sentido: si A es un problema de la geometría plana, ha de abordarse su solución con esos medios *elementales* antes de proceder de modo expeditivo a su solución por otros procedimientos más complicados (como las curvas empleadas en los problemas lineales) o de dudoso origen (como los recursos de carácter «mécanico»), cuyo uso no está sancionado por la práctica seguida en los *Elementos*.

El postulado (iv) parece pertenecer de entrada a una clase de asunciones diferentes de la compuesta por (i)-(iii). Proclo recuerda que Gémino situaba este aserto entre las nociones comunes o axiomas, pues no hace referencia a unas posibilidades de construcción, sino que establece una propiedad esencial de los ángulos rectos: la de ser una magnitud determinada capaz de representar un patrón invariable para medir los demás tipos de ángulos. Proclo añade por su cuenta y riesgo que no es un postulado en un sentido aristotélico; luego da a entender que posee un carácter axiomático; y por último parece estar bien dispuesto a ofrecer una prueba *(Com.* 188, 1 ss.). Lo cierto es que los mss. que se derivan de la edición de Teón suelen situarlo entre las nociones comunes. Su aplicación comporta la asunción de unas figuras invariables y esto significa, a nuestros ojos educados en la geometría moderna, estipular que el espacio en general es homogéneo; de ahí que este postulado (iv) se considere hoy equivalente a una caracterización axiomática del espacio en términos de uniformidad o isotropía. Pero Euclides, una vez más, parece estar pensando en una propiedad básica de un objeto geométrico antes que en una característica estructural del espacio de la geometría euclidiana.

El postulado (v) también resulta un tanto peculiar. Por una parte, se asemeja a los postulados operativos (i)-(iii) al demandar, aparentemente, la existencia de puntos de intersección o encuentro de rectas con rectas. Por otra parte, completa el criterio de paralelismo avanzado en la definición de rectas paralelas —y, quizás en consonancia con este sentido asociado a una definición, no faltarán editores de los *Elementos* en los siglos XVI y XVII que lo incluyan entre las nociones comunes—. En tercer lugar y sobre todo, no goza de la evidencia inmediata que aureola los demás principios geométricos. «Debe ser borrado por completo de la lista de los postulados porque se trata de un teorema henchido de dificultades, que Tolemeo se propuso resolver en un libro, y su demostración requiere varias definiciones y teoremas», afirma Proclo *(Com.* 191, 21 ss.); dos motivos para dudar de su condición primordial de postulado son: a) la existencia de asíntotas, líneas que se aproximan más y más a medida que se prolongan sin llegar a encontrarse —¿por qué no ocurre esto mismo en el caso de las líneas rectas?—, y b) la prueba en los *Elementos* de una proposición conversa de este postulado, I, 17: la suma de dos ángulos cualesquiera de un triángulo es menor que dos rectos —¿por qué hemos de suponer entonces que (v) es un postulado en vez de ser, como I, 17, un teorema?—.

La larga historia de los intentos fallidos de apear de su pedestal el postulado euclídeo de las paralelas es la más popular de todas cuantas se refieren al desarrollo de las matemáticas. Comprende intentos de deducción directa en

el cuerpo de la geometría plana y ensayos de demostración indirecta mediante la reducción al absurdo de proposiciones equivalentes a la negación del postulado. Los fracasos sucesivos han ido arrojando una serie de formulaciones parejas al asediado pero incólume postulado (v), es decir: han ido explicitando supuestos deductivamente equivalentes a este postulado en el seno de los *Elementos* o, más en general, en el marco de la geometría euclidiana. Esta historia es sumamente instructiva, no sólo por lo que toca a la suerte de la geometría clásica al desembocar en la existencia de otras geometrías no euclidianas, sino por la repercusión de este desenlace en la 2.ª mitad del siglo XIX. Sus secuelas se hicieron sentir tanto en la filosofía de la matemática y en la filosofía de la ciencia en general —e. g., al replantear las relaciones entre las construcciones geométricas y el mundo físico o el espacio natural—, como en la metodología de las ciencias deductivas en particular [18].

[18] La literatura generada por el largo proceso de discusión del postulado (v) y su desembocadura en las variantes hoy conocidas (geometrías neutrales, euclidianas, no euclidianas) ha sido tan abundante que al cumplirse la 1.ª década del presente siglo ya merecía un catálogo especializado como el de D. SOMMERVILLE, *Bibliography of Non-Euclidean Geometry*, Londres, 1911. Hay una selección de los textos que tejen esta historia en J. FAUVEL y J. GRAY, *The History of Mathematics. A Reader*, Londres-Basingstoke, 1987; § 6C, págs. 235-239, §§ 16A-16D, págs. 508-540. Si dejamos aparte las primeras escaramuzas helénicas en torno al postulado de las paralelas (*vid.* T. L. HEATH (1926²), ed. cit., 1, págs. 202-208 especialmente), los momentos álgidos de su historia podrían considerarse éstos:

1.°) Viene marcado por los comentarios árabes que siguen empeñados en su deducción directa a partir de alguna otra proposición congruente con la teoría; *vid.* K. JAOUICHE, *La théorie des parallèles en pays d'Islam*, París, 1986.

2.°) Un período de reelaboración moderna de la teoría euclidiana de las paralelas desde Saccheri (1733), *vid.* la edic. de G. B. HALSTED, *G. Sacheri's Euclides Vindicatus*, Nueva York, 1986²; Saccheri toma una vía indirecta para establecer el postulado: éste se seguiría de su propia negación dentro de la geometría euclidiana, pues la adición de una proposición contradictoria con el postulado al resto de los axiomas o principios euclidianos arrojaría unos resultados completamente absurdos —a los ojos del vindicador—. La contribución principal de estas dos primeras fases es la explicitación de proposiciones equivalentes al postulado o derivadas de él.

3.°) El tercer momento corresponde a la formulación de las geometrías no euclidianas de Lobatchevski (1829-30), J. Bolyai (1832), B. Riemann (1854), que justamente parten de la negación o el abandono del postulado de las paralelas para obtener otras geometrías tan consistentes como lo pudiera ser la geometría euclidiana; *vid.* el estudio clásico de R. BONOLA (1912): *Geometrías no euclidianas*, Madrid, 1923 (aunque preferible a esta edición española es la inglesa de H. S. Carslaw: BONOLA, *Non-Euclidean Geometry*, Nueva York, 1955, al incluir como suplemento textos básicos de Lobatchevski y Bolyai). Las revisiones históricas más comprensibles son las de B. A. ROSENFELD, *A History of Non-Euclidean Geometry...*, op. cit., 1976, 1988, y J. C. PONT, *L'aventure des parallèles*, Berna, 1986; también tiene interés en este sentido J. GRAY, *Ideas of Space*, Oxford-Nueva York, 1989². Sobre ciertos puntos conexos o derivados, pueden verse P. PARRINI, *Fisica e geometria dall'Ottocento a oggi*, Turín, 1979, y R. J. TRUDEAU, *The Non-Euclidean Revolution*, Boston-Basilea-Stuttgart, 1987. Otros aspectos metodológicos aparecen en E. AGAZZI y D. PALADINO, *Le geometrie non euclidee ed i fondamenti della geometria*, Milán, 1978; P. FREGUGLIA, *Fondamenti storici della geometria*, Milán, 1982 (en particular, caps. 5-6, págs. 233-327).

El pórtico de los *Elementos* se remata con una selección de nociones comunes [*koinai énnoiai*] de este tenor:

(i*) Las cosas iguales a una misma cosa son iguales entre sí.
(ii*) Si cosas iguales se añaden a cosas iguales, los totales son iguales.
(iii*) Si cosas iguales se sustraen de cosas iguales, los restos son iguales.
(iv*) Las cosas que coinciden entre sí, son iguales.
(v*) El todo es mayor que la parte.

Suele considerarse que (i*)-(iii*) no sólo provienen de una tradición fielmente seguida por Euclides, sino que cumplen a plena satisfacción las condiciones que el programa aristotélico de la ciencia demostrativa había impuesto en los *Segundos Analíticos* a los principios de este tipo, e. g., la de ser verdaderamente primordiales e incontestables, la de tener un alcance más general que el limitado por el campo temático de una ciencia determinada. La noción común (iii*) es, en concreto, la ilustración favorita de Aristóteles cuando se refiere a esta clase de principios indemostrables (*APo.* I 10, 76a41; 76b20; 11 77a31). Pero es posible que los matemáticos helenísticos tuvieran alguna diferencia al respecto: al menos, hay noticias de un intento de Apolonio de demostrar alguna de estas nociones comunes típicamente indemostrables (Proclo, *Com.* 194, 10 ss.). Puestos en el otro extremo, varios comentadores y editores de los *Elementos* fueron añadiendo luego a (i*)-(iii*) nuevos axiomas acerca de las relaciones de igualdad y desigualdad. En todo caso, las nociones comunes (i*)-(iii*) no plantean los problemas de autenticidad de las dos siguientes.

La noción común (iv*) representa un axioma de congruencia que no parece tener la generalidad de las anteriores. Herón, de hecho, no la reconoció como tal y Proclo la acepta alegando que ni se deben multiplicar excesivamente los axiomas innecesarios ni se deben reducir drásticamente al mínimo los axiomas imprescindibles (*Com.* 196, 15-21). Es posible que se trate de una interpolación temprana, con la intención de que la coincidencia intuitiva (o la congruencia geométrica) se viera como una aplicación de la igualdad y se beneficiara de las virtudes señaladas en (i*)-(iii*), la transitividad y la regularidad con respecto a la adición y la sustracción. Sea como fuere, su utilización en las proposiciones I, 4, o I, 8 de los *Elementos* da a entender que responde al procedimiento tradicional de superposición de figuras por desplazamiento de una y su colocación sobre la otra. A juicio de Platón, este recurso era uno de los que descalificaban a los geómetras de su tiempo por contaminar el pensamiento geométrico con la manipulación y el movimiento de objetos, pero ni antes ni después de Euclides dejó de aplicarse. Se suponía tácticamente que el movimiento no deforma los objetos así tratados.

La noción común (v*) también se nos presenta en Proclo con la cara de un axioma legítimo y la cruz de un principio discutido —e. g., por Herón—. En la prueba de I, 6, Euclides ante dos triángulos determinados, arguye que si uno fuera igual al otro, «el menor <sería> igual al mayor; lo cual es imposible». Esta misma fórmula se encuentra en una obra —anterior a los *Elementos*— de Autólico: «pues igual es el segmento ΔH al segmento ΔZ, el menor al mayor, lo cual es imposible [*íse ára estìn he* ΔH *tê* ΔZ, *he elásson tê meídsoni, hóper estìn adynaton*]» (*Sobre la esfera en movimiento,* Prop. 3). Cabe pensar entonces, como conjetura Heath, que (v*) viene a ser una generalización de los casos de este tipo que adquiere el aspecto de una condición sobre la relación de magnitud entre un todo, o el total, y una cualquiera de sus partes. O se puede pensar, por el contrario, que la noción común precede a las aplicaciones particulares como la de I, 6, y responde más bien a una opción intuitiva por los conjuntos finitos que trata de sortear algunas paradojas eleáticas sobre el infinito, e. g., que el doble resulta igual a la mitad, como ocurre en el argumento del «estadio» de Zenón de Elea; esto es lo que se imagina Szabó. También es posible, en fin, recordar el uso euclídeo de las nociones de «mayor» y «menor» (por ejemplo, en la línea de las definiciones

11: «Un ángulo obtuso es un ángulo mayor que un ángulo recto», 12: «Un ángulo agudo es un ángulo menor que un ángulo recto»), y entender que el todo y una de sus partes se relacionan como figuras compuestas de modo que la mayor contiene a la menor como una parte propia (e. g., si de trata de ángulos, de modo que ambos comparten el vértice y uno de los lados) [19]. Por otra parte, cabe reconocer que la noción común (v*) cuadra perfectamente con el punto de vista finito que Euclides adopta en otras proposiciones primordiales —e. g., en la Def. 23 o en los postulados (ii) y (v), donde su idea de línea recta se contrae a la idea de un segmento rectilíneo con la capacidad de ser indefinidamente extensible—. Pero así mismo hay que lamentar que la selección de (v*), dentro del contexto de las relaciones de orden de magnitud —mayor/menor—, sea mucho menos afortunada que la de (i*)-(iii*) en el contexto de las relaciones de igualdad; por ejemplo, (v*) no facilita la expresión de la transitividad y de otras propiedades de un orden de magnitud. De hecho, como ya he sugerido, (i*)-(iii*) no dejaron de inspirar nuevos axiomas emparentados (e. g.: «si se añaden iguales a desiguales, los totales resultan desiguales», «si se sustraen iguales de desiguales, los totales resultan desiguales»); la noción común (v*), por el contrario, permaneció solitaria y absorta en su propia (y aparente) evidencia.

Esta visión sumaria de la portada axiomática de los *Elementos* nos permite adelantar algunas características generales de la composición de Euclides. En primer lugar, su compleja relación con la tradición matemática anterior (los principios constituyen un lugar privilegiado para observar esta relación; recordemos, por ejemplo, la índole de varias de sus definiciones: explicitan o explican nociones de raigambre geométrica, en vez de constituir definiciones propiamente dichas —son en ciertos casos nociones tan simples y básicas que de suyo parecen indefinibles; pueden resultar a veces un tanto superfluas para la organización sistemática de los *Elementos*—). Desde este punto de vista, los principios euclídeos se pueden clasificar, creo, en tres tipos:

a) Las nociones o asunciones que han sido tomadas directamente del legado tradicional, e. g.: las definiciones 2, 5, 15; la noción común (iii*).

b) Las nociones o asunciones de importación indirecta, pues, al parecer, incluyen cierta elaboración o alguna precisión por parte de Euclides, e. g.: las definiciones 1, 3; los postulados (i)-(iii); las nociones comunes (iv*) y (v*).

c) Las nociones o asunciones originales —a lo que se me alcanza— del propio Euclides, dado que muestran una postura o un compromiso relativamente característicos, e. g.: las definiciones 4, 23; el postulado (v).

En segundo lugar, lo que hemos visto ya deja traslucir la naturaleza sustancialmente intuitiva de la «axiomatización» euclídea. Volvamos, por ejemplo, a la elucidación de los supuestos que obran en el método de la regla y el compás. No es una explicitación cabal; no es una reconstrucción suficiente para prescindir del sustrato intuitivo y del soporte diagramático sobre los que venía descansando la práctica tradicional de las pruebas geométricas. Se echa en falta, especialmente, un postulado de continuidad

[19] Cf. T. L. HEATH, *infra*, ed. cit., 1926², 1, pág. 232; A. SZABÓ, *The Beginnings...*, op. cit., 1978, págs. 291-298; R. J. TRUDEAU, *The Non-Euclidean...*, op. cit., 1987, págs. 35-36. Por lo demás, según L. BRUNSCHVICG, no representaría sino un axioma de desigualdad en la línea de los añadidos a (i*-iii*) por los comentadores y editores euclídeos, *vid.*, *Les étapes de la philosophie mathématique*, París, 1947³, § 52, pág. 89.

que asegure la existencia de puntos de intersección de las figuras elementales, las rectas y los círculos. Euclides confía en poder contar con un punto siempre que lo necesite y donde le convenga. El postulado (v) permite disponer en cierto modo de puntos de intersección entre rectas. Pero nada se dice sobre los puntos de intersección de rectas con circunferencias o de circunferencias con circunferencias. Serán las configuraciones gráficas quienes se encarguen de proveerlos. Así pues, las representaciones diagramáticas siguen desempeñando un papel sustancial en la efectividad constructiva de los métodos elementales de solución de los problemas planos (en la prueba de I, 1, sin ir más lejos). Estos y otros signos inducen a pensar que la esmerada organización deductiva de los *Elementos* no deja de atenerse al estudio tradicional de objetos geométricos determinados. Guarda relación con una nueva perspectiva sobre la constitución de cuerpos de conocimiento y con la preocupación por una exposición coherente y ordenada de sus *elementos,* que procede desde las asunciones más básicas y las pruebas más simples hasta los resultados derivados y las pruebas más complejas. Pero todo esto poco tiene que ver con una presunta conciencia axiomática que empieza declarando el espacio geométrico de referencia, así como el conjunto de condiciones y operaciones que determinan los objetos posibles dentro de este universo, sus propiedades pertinentes y sus relaciones mutuas. Si tal impresión fuera correcta, los defectos de la «axiomatización» euclídea no se deberían tanto a unas imperfecciones más o menos eventuales —e. g., a la existencia de supuestos tácitos— como, sobre todo, a la falta de una perspectiva axiomática propiamente dicha. Para recabar más información sobre este y otros conceptos de la constitución de los *Elementos,* pasemos a considerar los cuerpos

temáticos desarrollados en los libros que componen el tratado.

2. *Teoría elemental de la geometría plana. (Libros I-IV)*

El libro I contiene, además de las definiciones, postulados y nociones comunes que hemos visto, 48 proposiciones de las que 14 son problemas y 34 son teoremas. Las proposiciones I, 1-26 versan principalmente sobre triángulos; representan una especie de entrenamiento en el método de la regla (no marcada) y del compás (colapsable). Las proposiciones I, 27-32 desarrollan la teoría euclídea de las paralelas: I, 32 establece precisamente que la suma de los tres ángulos de un triángulo es igual a dos rectos. Pero antes de I, 27, donde se recurre por vez primera al postulado (v), ya se han dado pasos en esta dirección: I, 12 sienta la posibilidad de trazar una perpendicular a una recta dada desde un punto exterior a ella; I, 16 dice que, en todo triángulo, si se prolonga (infinitamente) uno de los lados, el ángulo exterior será mayor que cualquiera de los ángulos internos y opuestos; I, 17 dice que, en todo triángulo, la suma de dos ángulos cualesquiera es menor que dos rectos. De todo esto resulta la unicidad de la perpendicular: por un punto exterior a una recta sólo cabe trazar una perpendicular a ella —es un corolario explicitado por Proclo—. Este camino conduce a una versión popular del propio postulado (v): por un punto exterior a una recta sólo cabe trazar una paralela —sobre el supuesto de la prolongabilidad infinita que ya estaba involucrado en I, 16—. I, 28 expone uno de los criterios derivados del paralelismo euclídeo: si una recta al caer sobre dos rectas hace los ángulos interiores del mismo lado iguales a dos rectos, dichas

rectas serán paralelas. I, 29 da un criterio converso: si una recta cae sobre dos rectas paralelas, hace los ángulos interiores del mismo lado iguales a dos rectos.

Las proposiciones restantes se aplican a la determinación de áreas de paralelogramos, triángulos y cuadrados; preparan una métrica de la geometría plana que culmina en los resultados I, 45; I, 47-48 y II, 14. Importa señalar que en I, 35, donde se establecen las condiciones de igualdad entre paralelogramos, los *Elementos* introducen subrepticiamente una noción nueva de igualdad. Hasta entonces, la igualdad se venía entendiendo como coincidencia de formas y de figuras, como *congruencia geométrica*. A partir del I, 35, también se entenderá como igualdad de contenido o áreas, como *equivalencia geométrica*. En I, 44 («Aplicar a una recta dada, en un ángulo rectilíneo dado, un paralelogramo igual a un triángulo dado»), Euclides empieza a emplear el procedimiento geométrico de aplicación y transformación de áreas. Era un recurso antiguo; según Eudemo, citado por Proclo *(Com.* 419, 15-16), se debía a la Musa de los pitagóricos. I, 45 prueba la construcción de un paralelogramo equivalente al área de una figura rectilínea dada y establece así la posibilidad de representar cualquier área rectilínea como un rectángulo. Más adelante, en II, 14, sentará esta misma posibilidad para el otro tipo de paralelogramo que se consideraba básico, el cuadrado. I, 47 prepara el camino mostrando la expresión de la suma de dos cuadrados como un cuadrado. I, 47 es justamente una versión elemental del celebrado «teorema de Pitágoras» —resultado no exclusivamente original del padre de la secta pitagórica, ni seguramente demostrado por él como un teorema estricto—. La prueba de esta versión, a todas luces propia de Euclides, discurre al margen de la teoría de la proporción y del recurso a la semejanza de

triángulos. El carácter elemental de I, 47 y su converso, I, 48, convierte ambos teoremas en un broche de oro del libro I desde un punto de vista disciplinario; son una viva muestra de la capacidad de Euclides para dominar por medios relativamente sencillos buena parte de la tradición geométrica antigua. Desde el punto de vista teórico tienen menos importancia en la medida en que la teoría de la proporción del libro V permite establecer otra versión del teorema de alcance más general (V, 31); pero esto mismo también contribuye a hacer significativa la opción de Euclides, en el contexto del libro I, por un desarrollo inicial y sistemático en los términos elementales de la aplicación y transformación de áreas. El libro I, en su conjunto, es uno de los más logrados tanto en la perspectiva «axiomatiforme» de su organización deductiva, como en la perspectiva disciplinaria de una introducción progresiva en los métodos elementales de la geometría plana. Revela además la capacidad de Euclides para reconstruir de modo sistemático un legado antiguo y forjar nuevas pruebas que se adecuen a esta reconstrucción.

El libro II consta de 14 proposiciones, 2 de ellas problemas y las otras 12 teoremas. En realidad, parece recoger un núcleo formado por antiguos problemas aunque sean tratados a modo de teoremas, y esto da al libro un aire de miscelánea dirigida a ilustrar el uso y alcance del desarrollo elemental del método de aplicación de áreas. Por ejemplo, II, 11 da una solución necesaria para la inscripción de un pentágono regular en un círculo —operación que no tendrá lugar hasta IV, 10-11), y es un problema parejo al de cortar un segmento en razón media y extrema, cuestión que habrá de resolverse en el marco de la teoría de la proporción (VI, 30). II, 12 y 13 suplementan I, 47, y completan la teoría de las relaciones entre los cuadrados

de los lados de cualquier triángulo, sean rectangulares o no. II, 14, la construcción de un cuadrado equivalente a una figura rectilínea dada, es una especie de culminación del método elemental de la aplicación de áreas. Por lo demás, de esta antigua tradición de la aplicación de áreas Apolonio toma prestados algunos términos empleados en su investigación de las secciones cónicas [*parabolé:* «aplicación» / «parábola»; *hyperbolé:* «exceso» / «hipérbola»; *éleipsis:* «defecto» / «elipse»], aunque las diferencias conceptuales entre uno y otro campo sean notorias.

La aplicación de áreas suele recibir en nuestro tiempo —a partir de H. G. Zeuthen (1886): *Die Lehre von den Kegelschnitten im Altertum*— la denominación de «álgebra geométrica» de los griegos. Un rectángulo corresponde al producto de dos cantidades en álgebra, de modo que en la aplicación a una recta dada de un rectángulo o de un cuadrado equivalentes a un área determinada podemos ver la contrapartida geométrica de la división algebraica de un producto de dos cantidades por una tercera. La adición y sustracción de productos se traduce así mismo en la adición o sustracción de rectángulos o cuadrados. La extracción de una raíz cuadrada se vierte, en fin, por el hallazgo de un cuadrado equivalente a un rectángulo dado, *i. e.* por lo logrado en II, 14 con la ayuda de I, 47. En suma, cabe leer varios resultados de II en términos algebraicos (véanse más adelante estas versiones en las notas correspondientes a este libro). Con todo, la denominación de «álgebra geométrica» no deja de ser bastante ambigua y harto problemática. De hecho, ha significado tesis diversas entre las que destacan dos: 1) la tesis *histórica* de que el «álgebra geométrica» de los griegos es una geometrización de los métodos algebraicos practicados por la matemática prehelénica, en particular babilonia; 2) la tesis *hermenéutica* de que el «álgebra geométrica» consiste en una geometría algebraica, cuyas magnitudes (líneas, áreas) representan cantidades cualesquiera y cuyos teoremas traslucen relaciones o estructuras algebraicas [20].

La tesis histórica (1) puede tener distinto alcance según el grado de influencia que se atribuya a los precedentes babilonios, y según se considere el caso de la tradición geométrica griega en general o ciertos casos particulares como el de Herón de Alejandría. Herón emplea un procedimiento que se diría «algebraico» en las pruebas de II, 2-10 como consecuencias de II, 1: sus pruebas no requieren el uso de figuras, sino simplemente el trazo de una línea y, por

[20] Esta tesis 2) procede de H. G. ZEUTHEN, *Die Lehre von den Kegelschnitten in Altertum* (Copenhague, 1896) y luego se ha mantenido como en estado de solución en algunas historias notables de las matemáticas *(vid.*, por ejemplo, B. L. VAN DER WAERDEN (1950, 1954[2]), *Science Awakening*, Nueva York, 1963). La tesis 1) se debe a una sostenida línea de investigación de O. NEUGEBAUER: desde su clásico «Zur geometrischen Algebra» *(Quellen und Studien zur Geschichte der Mathem., Astron. u. Phys.*, Abt. B 3 (1936), 245-259), hasta revisiones como «The survival of Babylonian methods in the exact sciences of Antiquity and the Middle Ages», *Proc. Amer. Philos. Society* 107 (1963), 528-535. La discusión se ha reavivado posteriormente merced a la intervención crítica de S. UNGURU, «On the need to rewrite the history of greek mathematics», *Archive for the History of Exact Sciences* 15 (1975), 67-114; «History of ancient mathematics. Some reflections on the state of the art», *Isis* 70 (1979), 555-565. También gracias a que el testigo de Neugebauer ha sido recogido por otros historiadores, en especial por VAN DER WAERDEN, *vid.* sus «Defense of a "schocking" point of view», *Archive for History of Exact Sciences* 15-3 (1976), págs. 199-210; *Geometry and Algebra in Ancient Civilizations*, Berlin-Heidelberg-Nueva York, 1983. Creo que hay observaciones juiciosas en I. MUELLER, *Philosophy of Mathematics and Deductive Structure in Euclid's Elements*, Cambridge (Mass.)-Londres, 1981, págs. 42-44, 50-52; J. L. BERGGREN, «History of Greek mathematics: a survey of recent research», *Historia Mathematica* 11 (1984), págs. 397-398 en especial.

otro lado, sus planteamientos bien podrían ser, en general, una forma helenística de asumir el legado de la matemática mesopotámica, con la que coinciden en la manera de abordar los problemas y de resolverlos mediante ejemplos numéricos. Pero el caso de la tradición geométrica griega ya es otro cantar. Por un lado, esta tradición comporta un sentido de la demostración ajeno a las pruebas de los babilonios. Por otro lado, la tesis de un influjo específicamente babilonio sobre ella presupone una difusión de la matemática oriental en Grecia desde finales del siglo v a. C. que no es fácil precisar y, menos aún, documentar; por ejemplo, ¿cómo determinaremos si tal influjo tuvo lugar antes de Euclides (como la transmisión del gnomon) o después (como la asimilación tolemaica de la medición de ángulos en grados junto con el sistema sexagesimal)? En fin, la conjetura de una influencia babilonia directa habrá de justificarse frente a otras hipótesis posibles. Por ejemplo, la hipótesis de una invención griega independiente —como ya he dicho, la aplicación de áreas fue, según Eudemo, un «descubrimiento de la Musa de los pitagóricos»—. O la hipótesis de un *poo!* y un trasiego de conocimientos matemáticos relativamente esparcidos por esta transitada zona del Mediterráneo oriental, donde ya se habían relacionado desde antiguo los sabios jonios (Tales, Pitágoras, Enópides de Quíos) y los egipcios, y puede que se vinieran compartiendo ciertos resultados: la difusión de los triplos pitagóricos y del propio «teorema de Pitágoras» bien puede ser el reflejo concreto de un acervo común o, cuando menos, de una suerte de ósmosis. Ante la escasez de datos al respecto, todas estas conjeturas son igualmente verosímiles. Por eso, el mayor mérito de la tesis (1), al margen de su esfuerzo por identificar una raíz definida de la tradición matemática griega, quizás consista

en llamar la atención sobre los motivos operativos y el interés por el cálculo que subyacen en ciertos desarrollos geométricos griegos —bien que esos motivos e interés sólo levanten cabeza de vez en cuando: con la versión pitagórica de la proporción, con Herón, con Diofanto—.

La tesis (2) es, a mi juicio, mucho menos aceptable. Se parece demasiado a una anacrónica «operación rescate»: a la traducción de los *Elementos* a un lenguaje más familiar para el lector de hoy, el algebraico. Pero, en este caso, el cambio de lenguaje equivale no a decir lo mismo de otro modo sino a hablar de otras cosas. En principio, una interpretación algebraica propiamente dicha incluye la abstracción estructural de unas leyes o condiciones operacionales que pueden cumplirse en muy diversos dominios de objetos. Pero esta concepción estructural no es la de Euclides: él no generaliza, más bien duplica o desdobla la aplicación de áreas en los libros II y VI, por razones metódicas o disciplinarias; y en otros casos que también se prestarían a una generalización análoga (e. g., a propósito de los conceptos de magnitud y número) tampoco da un paso en esa dirección estructural. Más aún: en II no hay ecuación alguna ni trasuntos algebraicos, sino la exposición de unos resultados deductivamente inconexos entre sí; baste reparar en que una vez probada la distributividad de la multiplicación (II, 1), dicha propiedad desaparece de escena y las pruebas de otras propiedades básicas como la conmutatividad o la asociatividad aparecen distanciadas en otros libros de los *Elementos*. Resumiendo, la impresión que produce el libro II se asemeja a ésta: el propósito de Euclides es ilustrar el método de aplicación de áreas y, en esta perspectiva, su proceder es más instructivo que el «algebraico» u otro de formato similar, pues en vez de remitir al aprendiz a un conjunto de fórmulas abstractas

o sistemáticas, lo pone en condiciones de probar desde un principio, por un método elemental y ejemplificado en diversos casos, otras proposiciones del mismo género [21].

El libro III parte de 11 definiciones y contiene 37 proposiciones, 5 de ellas problemas y las otras teoremas. Presenta la geometría del círculo e incluye el estudio de círculos, sus segmentos, intersecciones y tangencias. Tiene especial relieve la construcción de III, 1 para determinar el centro del círculo, aunque su prueba supone que una recta y un círculo no pueden tener más de dos puntos en común y ésta es una condición que cabe establecer a partir de III, 2. La suposición tácita más general es la característica de la geometría plana euclídea, la de continuidad. El libro no está muy cuidado desde un punto de vista sistemáticamente deductivo. Apenas guarda relación con los anteriores y tampoco tiene mucha cohesión interna: dejando al margen III, 1, hay 12 proposiciones (2, 3, 5-8, 12, 14, 16, 18, 20, 23) que no dependen de resultados presentes en III, y hay 13 proposiciones (4, 7, 8, 12-13, 15, 25, 29-30, 33-35) que nada contribuyen a la obtención de otros resultados dentro del mismo libro. Tal vez por esta razón cundan las revisiones y mejoras posteriores de varias de sus pruebas y abunden las alternativas propuestas por los comentadores y editores.

El libro IV parte de 7 definiciones y consta de 16 proposiciones, todas ellas problemas. Estudia inscripciones y circunscripciones de figuras regulares rectilíneas y círculos, y ofrece la construcción de polígonos regulares, como el pentágono o el hexágono, por la duplicación de los lados. El libro está formado en cierto modo por núcleos y quizás el más notable sea el dispuesto en torno a la inscripción

[21] *Vid.* T. L. HEATH, ed. cit., 1926², 1, pág. 377.

del pentágono regular, IV, 11. En este resultado concurren IV, 1-2, 5, 10, amén de material procedente del libro anterior (e.g., III, 26-27, 29), y de él parten las pruebas de IV, 12, la circunscripción del pentágono, y IV, 16, el problema final de la inscripción de un polígono regular de quince lados. Según el escolio IV, 2.°, «este libro es descubrimiento de los pitagóricos». Puede que estos orígenes y su afinidad con la tradición de los problemas lo hicieran un tanto refractario a un tratamiento sistemático. En todo caso, su calidad desde el punto de vista instructivo y disciplinario es muy superior a su cohesión deductiva interna o a su congruencia teórica.

3. *La teoría generalizada de la proporción. (Libros V-VI)*

Con el libro V pasamos al campo de la teoría de la proporción y a un legado matemático relativamente reciente. La teoría se refiere a las magnitudes como términos de la relación de proporcionalidad. La forma de aparición de la teoría —en sustitución de nociones anteriores más limitadas y concretas de razón y proporción, en la primera mitad del siglo IV—, el hecho de que estas magnitudes hayan de cumplir ciertas condiciones —la de ser homogéneas y «arquimedianas»—, el punto oscuro de las relaciones entre las ideas de magnitud y de número, son motivos de que la califique de teoría «generalizada», en vez de referirme a ella como teoría sencillamente «general», de la proporción.

El libro comprende 18 definiciones, algunas de ellas sustanciales, y 25 proposiciones, todas ellas teoremas. Según la definición 1, «una magnitud [*mégethos*] es una parte de una magnitud, la menor de la mayor, cuando mide [*ka-*

tametrê] a la mayor». Cabe entender que las magnitudes son abstracciones o idealizaciones de objetos geométricos que únicamente consideran la cantidad, *i.e.* la longitud en el caso de las líneas, el área en el caso de las figuras planas, el volumen en el caso de los sólidos. Según la definición 2, «la mayor es un múltiplo de la menor cuando es medida por la menor». Así pues, son magnitudes susceptibles de multiplicación. Si x es un m-múltiplo de y, x mide m veces y; entonces: $x = m.y = ((y+y)_1 + y)_2 + \ldots + y)_m$, de modo que la multiplicación $m.y$ de una magnitud y equivale a una adición reiterada m veces. Con ello se está suponiendo la existencia, para una magnitud dada, de un número indefinido de magnitudes iguales a ella —un supuesto tácito en V, pero tal vez declarado en otra obra de Euclides, *Datos*, cuya definición 12 explica que áreas, rectas, ángulos y razones «están dados en magnitud cuando podemos hacer otros iguales a ellos»—.

Hay además otras nociones y suposiciones nunca declaradas: por un lado, el concepto de medición y las relaciones «medir a/ser medido por»; por otr lado, el supuesto de que siempre es posible considerar la m-parte de una magnitud, de manera que para toda magnitud x hay una magnitud y tal que $x = m.y$. Esta suposición obra tácitamente en la prueba de V, 8 sobre la proporcionalidad de las composiciones de magnitudes proporcionales; conviene anotar de paso que tal suposición no vale para los números estudiados en los libros VII-IX. Según la Def. 3, «una razón [*lógos*] es un tipo de relación en lo que se refiere al tamaño entre dos magnitudes homogéneas». De aquí podría sacarse la impresión de que una razón, a imagen y semejanza de un espíritu elemental, es algo que se deja sentir con más facilidad que definir. La definición siguiente viene a introducir ciertas precisiones; además confirma

la impresión de que la noción de razón no tiene mucho sentido fuera del contexto de la teoría de las proporciones [22]. La Def. 4 es una de las consideradas básicas en la teoría: «Se dice que tienen una razón entre sí las magnitudes que, al ser multiplicadas, una de ellas puede exceder a la otra». Aparte de conocer una suerte histórica un tanto accidentada, su interpretación también ha sido variopinta. Hay escoliastas que ven en ella una caracterización de la homogeneidad de las magnitudes involucradas; los hay que entienden la definición en el sentido de excluir las magnitudes infinitas, tanto infinitamente grandes como las infinitamente pequeñas —más justo es entender que excluye la relación de razón entre una magnitud finita y otra infinita del mismo género—; hay quienes leen la definición como si implicara magnitudes inconmensurables —aunque la distinción entre las magnitudes conmensurables y las inconmensurables carece de lugar en este contexto—. Además cabe relacionar la Def. 4 con un supuesto que Arquímedes recuerda o explicita en la *Cuadratura de la parábola* y formula de modo más general en *Sobre la esfera y el cilindro:* «de líneas desiguales, superficies desiguales y sólidos desiguales, el mayor excede al menor por una magnitud que añadida a sí misma puede exceder a cualquier otra

[22] Sobre la suerte de la idea de razón antes y después de esta teoría generalizada de la proporción, *vid.* D. H. Fowler, «Ratio in early Greek mathematics», *Bull. Amer. Mathem. Society* 1-6 (1979), 807-846; «Anthyphairetic ratio and Eudoxian proportion», *Archive for History of Exact Sciences* 24 (1981), 69-72; «*Logos* (ratio) and *Analogon* (proportion) in Plato, Aristotle, and Euclid», en J. Petitot, dir., *Logos et Théorie des Catastrophes* (Colloque de Cerisy), Ginebra, 1988, págs. 444-472. También cabe considerar las relaciones entre la concepción inicial de razones y de proporciones y la teoría musical, a la luz del análisis filológico de A. Szabó, *The Beginnings of Greek Mathematics*, Dordrecht-Boston-Budapest, 1978; § 2.8, págs. 134-137.

dada del mismo género» (I, asunción 5). De este supuesto se sigue la existencia de los múltiplos de una magnitud dada: si $x < y$, hay una magnitud m tal que $m.x > y$. Suele recibir el nombre de «condición arquimediana» —o el de «postulado de continuidad»—. Dentro de los *Elementos,* contribuye a la prueba de V, 8 y mucho más adelante, a través precisamente de la Def. V, 4, a la prueba de X, 1. Vistas así las cosas, la asunción de Arquímedes extiende el criterio euclídeo de razón postulando que si dos magnitudes guardan entre sí una razón conforme a la Def. 4, su diferencia guardará una razón en el mismo sentido con cualquier otra magnitud homogénea, *i.e.* si $x > y$, hay una magnitud n tal que $n.(x—y) > z$. Pero, en fin, también cabe entender que la Def. 4 de Euclides y la condición arquimediana son dos versiones independientes de un supuesto que obraba en las pruebas por «exhausción» que había avanzado Eudoxo: Euclides define la relación «A guarda razón con B» entre magnitudes, dentro de su teoría de la proporción, y hace referencia a la multiplicación; Arquímedes postula en cambio una condición precisa para que ciertas clases de magnitudes homogéneas (líneas, áreas, sólidos) se atengan efectivamente a la definición euclídea, y hace referencia a la adición de diferencias. A esta luz, la versión euclídea parece una reconstrucción tan elaborada como lejana de la base deductiva eudoxiana en la que se apoyan el lema de bisección (el teorema X, 1) y sus aplicaciones (e. g., XII, 2, 7, 10); mientras que la versión de Arquímedes es una elucidación certera y más próxima del supuesto que late en las pruebas correspondientes a esos resultados. Según esto, el interés de Euclides no estriba de momento en seguir los pasos dados por Eudoxo, sino más bien en sentar las bases propias de su teoría abstracta de la proporcionalidad.

La Def. 5 es la piedra angular de la teoría: «Se dice que están en la misma razón unas magnitudes, la primera con respecto a la segunda y la segunda con respecto a la cuarta, cuando si se toman unos equimúltiplos cualesquiera de la primera y la tercera, y unos equimúltiplos cualesquiera de la segunda y la cuarta, los primeros equimúltiplos exceden a la par, o son parejamente iguales, o resultan parejamente deficientes que los últimos equimúltiplos tomados unos y otros en el orden correspondiente». Si suena a cosa abstrusa, la culpa es del lenguaje común. Su contenido conceptual es claro: suministra un criterio suficiente y necesario de proporcionalidad. La Def. 6 estipula que las magnitudes que guarden la misma razón se llamen proporcionales [en proporción, *análogon*]. Diremos entonces que los términos x, y, z, w son proporcionales *(«x* es a *y* como *z* es a *w»)* si y sólo si se da lo siguiente: si $m.x > n.y$ entonces $m.z > m.w$, o si $m.x = n.y$ entonces $m.z = n.w$, o si $m.x < n.y$ entonces $m.z < n.w$, para todo m, n. Conviene caer en la cuenta de que una proporción no es una igualdad entre dos objetos, una relación binaria o una identidad entre razones, sino más bien una relación cuaternaria, como revela el hecho de que Euclides se crea obligado a demostrar la proposición V, 11 («las razones que son las mismas que una misma razón también son las mismas la una que la otra «—*i.e.* si A:B::C:D y E:F::C:D entonces A:B::E;F—), en vez de considerar que se trata de una aplicación obvia o trivial de la noción común (i*); así pues, la notación usual de una proporción euclídea en los términos *«(x, y) = (z, w)»* resulta inadecuada.

La def. 7 representa a su vez un criterio de no proporcionalidad. Reza así: «Cuando, de los equimúltiplos, el múltiplo de la primera magnitud excede al de la segunda pero el múltiplo de la tercera no excede al de la cuarta, entonces

se dice que la primera está en una razón mayor con la segunda que la tercera con la cuarta». Es decir: x es a y más que z a w si y sólo si se da $m.x > n.y$ pero no se da $m.z > n.w$, para algún m, n. Suplementariamente se supone la existencia de un cuarto término proporcional: se halla implícito en la prueba de V, 18, aunque más tarde Euclides viene a demostrar un caso particular, VI, 12: «dadas tres líneas, hallar una cuarta proporcional». Consecuencia importante de este supuesto es la condición de tricotomía (dadas dos magnitudes o razones cualesquiera, una es mayor o igual o menor que la otra), coherente con el sistema ordenado de objetos que hoy podríamos tomar como el dominio de referencia de la teoría. Todo ello le permite a Euclides el recurso adicional de las pruebas indirectas de proporcionalidad *(i. e.* que x es a y como z es a w*)* por reducción al absurdo de las alternativas de desproporción *(i. e.* que la razón de x a y sea mayor o sea menor que la de z a w*)*. Si el criterio de equimúltiplos que da lugar a la Def. 5 se inspira en Eudoxo —según suele reconocerse—, esta Def. 7 bien puede ser una contribución enteramente original de Euclides.

La teoría de la proporción del libro V viene siendo en nuestros días un objeto predilecto de reconstrucciones y reestructuraciones axiomáticas [23]. Al margen de ellas, el libro constituye por sí mismo una presentación sistemática de ideas y resultados que proceden de investigaciones relativamente recientes en la época de la confección de los *Elementos*. El escolio V 1.° recoge la opinión de quienes dicen que el libro es un descubrimiento de Eudoxo (miembro

[23] Cabe destacar las formuladas por F. BECKMAN, «Neue Gesichtspunkte zum 5. Buch Euklids», *Archive for History of Exact Sciences* 4 (1967), 1-144; I. MUELLER, *Philosophy of Mathematics and Deductive Structure...*, op. cit., 1981, cap. 3, págs. 145-148 en especial.

distinguido de la Academia de Platón hacia 370-360); el escolio V 3.° precisa que el libro, aunque se considere como una contribución de Eudoxo, debe su organización a Euclides; esta organización puede haber incluido variantes euclídeas equivalentes, pero no idénticas, a los supuestos y lemas empleados por Eudoxo [24]. En cualquier caso, su notable cohesión interna y el poder de sistematización de esta teoría de la proporción le han valido al libro cierto crédito «axiomático». H. Hasse y H. Scholz, en su influyente (pero no muy afortunado) artículo «Die Grundlagenkrisis der griechischen Mathematik» *(Kantstudien* 33 (1928), 4-34), lo describen como el primer intento de una axiomatización completa. Desde luego, es más justo ver en él una reelaboración de las nociones recibidas en orden a conseguir unos conceptos y unos criterios precisos como los formulados en las definiciones 4, 5 y 7; también es notable el grado de abstracción de la teoría. Pero algo que hoy echamos de menos es justamente la base axiomática que le deparan sus actuales reconstrucciones: la explicitación de ciertos postulados existenciales, la expresión de unas leyes combinatorias y la acotación de las condiciones que determinan los objetos que pueden pertenecer a su dominio de referencia. Por otro lado, aún nos aguarda un punto oscuro pero singularmente significativo: el de las rela-

[24] *Vid.* W. R. KNORR, *The Evolution...*, op. cit., 1975, págs. 261-263; «Archimedes and the pre-Euclidean proportion theory», *Archives Intern. d'Histoire des Sciences* 28-103 (1978), 183-244; cf. también J. L. BERGGREN, «History of Greek mathematics: a survey...», art. cit., 1984, pág. 399 especialmente. J. L. GARDIES, por otra parte, ha intentado reconstruir la teoría presuntamente original de Eudoxo y ha destacado los problemas que supone su inmersión en el contexto de los *Elementos:* vid. *L'héritage épistémologique d'Eudoxe de Cnide (Un essai de reconstitution)*, París, 1988.

ciones entre las magnitudes del libro V y los números del libro VII. Esta oscuridad en una cuestión capital para la fundamentación de unas investigaciones matemáticas cercanas a Euclides, según muestra el libro X, no es propia de una materia que ya está suficientemente trabajada y dispuesta a recibir la bendición deductiva de su coronación axiomática; menos excusable es todavía si se trata de la fundación axiomática misma de una teoría general. La oscuridad en ese respecto más algunas otras ambigüedades y lagunas indican que la reflexión de Euclides sigue en parte empeñada en el duro trabajo de análisis conceptual y de depuración deductiva que habían iniciado algunos pioneros como Teeteto y Eudoxo.

El libro VI empieza con 4 definiciones y contiene 33 proposiciones, de las que 8 son problemas y las restantes teoremas. Aplica la teoría de la proporción a la geometría plana desarrollando una teoría de los polígonos semejantes y generalizando el procedimiento de la aplicación de áreas. No es raro entonces que algunos de sus problemas vengan a ser réplicas generalizadas de resultados obtenidos por medios más elementales en los dos libros primeros (e. g., VI, 31 es una generalización de I, 47; VI, 28, de II, 5; VI, 30, de II, 11; VI, 13, de II, 14). El libro está compuesto por diversos núcleos: VI, 4-7 sientan las condiciones de semejanza entre dos triángulos (a partir del criterio establecido en la Def. 1: «figuras rectilíneas semejantes son las que tienen sus ángulos rigurosamente iguales y los lados en torno a los ángulos iguales proporcionales», noción ya conocida por Aristóteles *(APo.* II 17, 99a13-14) y no muy precisada aquí, pues omite la necesidad de tomar los ángulos iguales en el mismo orden). VI, 9-13 se ocupan de problemas de cortar rectas en proporciones dadas o determinar rectas que las satisfagan (VI, 13 prueba la manera de hallar una media proporcional entre dos rectas dadas). Otro núcleo, 18-26, apunta hacia 28 y 29, donde hoy suelen verse contrapartidas geométricas de ecuaciones cuadráticas. Hay, en fin, proposiciones como VI, 32-33 que miran hacia las construcciones del libro XIII. En suma, el libro carece de cohesión interna e incluso presenta alguna que otra anomalía desde un punto de vista deductivo como, por ejemplo, la prueba de un caso particular VI, 14 sin reducción al caso general VI, 23, o la separación entre una proposición (VI, 24) y su conversa (VI, 26) por la interposición de otra, VI, 25, con cierto interés histórico —recoge un resultado atribuido a Pitágoras (la construcción de una figura semejante a una figura rectilínea dada e igual a otra figura rectilínea dada)—, pero sin un cometido teórico o metódico de relieve en este contexto.

4. *Teoría de la aritmética. (Libros VII-IX)*

Los libros VII-IX presentan y desarrollan la aritmética de los *Elementos*. En conjunto, comprenden 102 proposiciones presididas por las 23 definiciones que Euclides introduce en el libro VII. Puede ser un signo de unidad esta agrupación de las definiciones a la cabeza del VII, que contrasta con el hábito anterior de irlas distribuyendo por cada uno de los libros I-VI. También suelen llamar la atención la ausencia de postulados y el hecho de que en VII-IX no se reconozca ningún problema como tal. Las pocas proposiciones que no constituyen teoremas expresos, e. g., VII, 2-3, o IX, 18-19, son investigaciones más teóricas que prácticas: se marcan el propósito de hallar algo (VII, 2: la medida común máxima de dos números no primos entre sí; VII, 3: la medida común máxima de tres: IX, 18: un

tercer término proporcional a unos números dados; IX, 10: un cuarto término proporcional), en vez de hacerlo o construirlo —aunque los mss. de la tradición teonina rematan la prueba de VII, 3 con la fórmula equívoca «que es lo que había hacer»—. La aritmética de los *Elementos* considera que los números son unos objetos susceptibles de hallazgo, no de generación o producción, y en VII-IX desaparece la habitual terminología diagramática que describe o prescribe acciones como las de construir, levantar, prolongar, cortar, etc. Los objetos aritméticos parecen ser entidades u objetos de otro tipo que los geométricos.

Algunas de las definiciones dignas de consideración son las siguientes:

Según la Def. 1, «la unidad [*monás*] es aquello en virtud de lo cual cada una de las cosas que hay es llamada una». Jámblico *(Introducción a la aritmética de Nicómaco* 11, 5) dice que ésta es la definición de unidad de los ⟨pitagóricos⟩ modernos; los pitagóricos antiguos la habían definido en ocasiones como una «divisoria entre número y partes» o como una «cantidad delimitante». Con la definición moderna ya parece familiarizado Platón (e. g.: «hombres asombrosos, ¿acerca de qué números discurrís en los cuales la unidad se halla tal como la consideráis, siendo en todo igual a cualquier otra unidad sin diferir en lo más mínimo ni contener en sí misma parte alguna?», *República* VII 526a). Tiene visos de ser una definición dirigida a separar la unidad de la multiplicidad y de la divisibilidad. Su sentido resulta algo menos impreciso a contraluz de la noción subsiguiente de número. Según la Def. 2, «un número es una pluralidad compuesta de unidades». Las relaciones entre la unidad y el número recuerdan la concepción que Aristóteles manifiesta en diversos lugares de la *Metafísica* (e. g.: «uno» significa «medida de una pluralidad», y «número» significa «pluralidad medida» y «pluralidad de medidas», y por eso es razonable que la unidad no sea un número, pues tampoco la medida es medidas, mientras que tanto la medida como la unidad son principios», *Metafís.* N 1, 1088a4-8). En el contexto de los *Elementos,* la unidad no es un número ni hay una cosa tal como la pluralidad formada por una sola unidad; de hecho, Euclides da pruebas separadas para las unidades (VII, 15) y para los números (VII, 9), aunque no siempre sea coherente con esta distinción. Por otro lado, a tenor de la prueba de VII, 31 («todo número compuesto es medido por algún número primo»), la pluralidad numérica es una pluralidad finita —la prueba discurre por una reducción al absurdo de la suposición opuesta cuyo desenlace sería la asunción de una divisibilidad infinita de los números, pero esto, declara Euclides, «es imposible en los números [*adynaton én arithmoîs*]»—. Así pues, el primer supuesto de la aritmética euclídea viene a ser que hay una cantidad indefinida de unidades y que sus colecciones finitas constituyen números. A esto se añade otro rasgo característico de la concepción aritmética de los *Elementos,* ya apuntado antes: los números no se producen, se encuentran; los números no se generan dentro de una serie mediante la idea de sucesión, ni se determinan por esta formación genealógica —según acostumbramos a pensar hoy en la estela de un Peano o de un Frege—. La verdad es que no faltan a veces en la aritmética de los *Elementos* razonamientos «inductiviformes» sobre colecciones y secuencias de números naturales dados (e. g., en la línea VII, 2, 3, 33); pero esta teoría no tiene una idea estructural de la serie infinita de los números enteros ni cuenta con un principio general de inducción matemática; por ejemplo, en VII, 16, Euclides prueba la conmutatividad de la multiplicación no con

arreglo a este principio, sino mostrando su aplicación en unos números dados (que representan, podemos suponer, números cualesquiera).

Según la Def. 3, «un número es parte de un número, el menor del mayor, cuando mide al mayor»; ahora bien, como añade la Def. 4, un número no es parte sino «partes, cuando no lo mide». (Por «parte [*méros*]» se entiende una parte alícuota o submúltiplo; por «partes [*mére*]», un número de partes alícuotas o una fracción propiamente dicha.) Recíprocamente, «el número mayor es un múltiplo del menor cuando es medido por el menor», dice la Def. 5. El uso posterior de estas definiciones envuelve nuevas suposiciones tácitas sobre la noción de medida y la relación de medir una cantidad un número determinado de veces, amén de los axiomas pertinentes —e. g.: si A mide a B y B mide a C, A mide a C; si A mide a B y a C, A mide a B+C—. Podemos colegir que «X mide a Y» significa «X es un factor o divisor exacto de Y», pues «Y es medido por X» monta tanto como «Y es un múltiplo de X». Entonces, X es *parte* de Y si y sólo si $X < Y$ y X es, en efecto, un factor o divisor de Y —siendo Y a su vez múltiplo de X—. Por ejemplo, el número 2 es parte del número 6. Pero X es *partes* de Y si y sólo si $X < Y$, sin que X sea un factor o divisor exacto de Y —ahora nos hallamos ante una fracción propiamente dicha—. En este sentido, el 4 no es *parte* sino *partes* del número 6.

Euclides recoge a continuación, en las definiciones 6-10, un conjunto de nociones de clara estirpe pitagórica: las de número par (Def. 6: «número par es el divisible en dos partes iguales»), número impar (Def. 7: «número impar es el no divisible en partes iguales, o el que difiere de un número par por una unidad»), y otras nociones derivadas. Según la Def. 11, «un número primo es el que sólo es

medido por una unidad»; Aristóteles ya había precisado que el número primo no era medido por ningún número (*APo.* II 13, 96a36), en razón de que la unidad no es un número. Según la Def. 12, «números primos entre sí son aquellos que sólo son medidos por una unidad como medida común». La importancia que cobran en el libro VII los números primos relativos es considerable. Quizás se derive de la efectividad del algoritmo euclídeo para hallar la medida común máxima (el máximo común divisor) de dos números, que se presenta en VII, 1 como un «test» de esa propiedad de ser primos entre sí. Puede que el interés un tanto absorbente de Euclides por estos primos relativos le haga descuidar los primos «absolutos» y contribuya a que su prueba del hoy denominado «teorema fundamental de la aritmética» —a saber: la factorización unívoca de los números primos—, no sea una demostración completa ni en el caso del teorema que más se le aproxima, IX, 14, ni en otros teoremas análogos, VII, 30-31.

Según la Def. 15, «se dice que un número multiplica a un número cuando el multiplicado es añadido a sí mismo tantas veces cuantas unidades hay en el otro, y así se produce [*génetai*] algún número». Se trata de una versión clásica de la multiplicación como abreviatura de un proceso reiterativo de adición. La «generación» que se menciona no es sino la constitución de un número en tales términos y no cabe pensar en una suerte de construcción o producción recursiva.

Conforme a la Def. 20, «los números son proporcionales cuando el primero es el mismo múltiplo, o la misma parte, o las mismas partes, del segundo que el tercero del cuarto». Euclides no define la noción aritmética de razón. Su idea de la proporción numérica envuelve de hecho aquellos supuestos de la teoría generalizada de la proporción

que pueden aplicarse en el presente caso. De seguir por esta línea, todo número sería una magnitud, aunque no toda magnitud fuera un número, pues los números del libro VII no cumplen ciertas condiciones de las magnitudes del libro V, como la posibilidad de tomar en todo caso la *m*-parte de una magnitud o como la existencia de un cuarto término proporcional. Pero ésta no es precisamente la perspectiva de los *Elementos;* tampoco se dice en lugar alguno que los números de VII-IX son una clase de magnitudes conmensurables. De manera que si bien Euclides emplea tácitamente algunos presupuestos de la teoría generalizada de la proporción en el desarrollo de la aritmética, lo cierto es que demuestra las propiedades de la proporción numérica por separado y no transfiere a los números las propiedades de la proporción generalizada que podrían convenirles. A lo largo de VII-IX todo parece indicar que nos las habemos con unos objetos matemáticos relativamente autónomos e independientes de las magnitudes de V-VI; al menos por el momento, los *Elementos* no traslucen ningún motivo para entretejer la caracterización de los números con la antes prevista para las magnitudes. Nos aguarda una sorpresa más adelante, en X, 5.

Desde un punto de vista sistemático, el libro VII es una esmerada reconstrucción del antiguo legado aritmético de raíces pitagóricas [25]. Su cohesión deductiva interna, aun-

[25] Las reconstrucciones modernas de la aritmética pitagórica parten de O. BECKER, «Die Lehre vom Geraden und Ungeraden im neunten Buch der euklidischen *Elemente*», *Quellen u. Studien z. Geschichte der Mathem., Astron. u. Phys.*, Abt. B 3 (1936), 533-553. Buen exponente del camino recorrido en esta línea de reconstrucción es M. MALDENDIER, «Eine axiomatik zum 7. Buch der *Elemente* von Euklid», *Mathem.-Phys. Semesterberichte* 22 (1976), 240-254. Una exposición clásica del legado pitagórico ha sido la de W. BURKERT (1962): *Lore and Science in Ancient Pythagoreanism*, Cambridge (Mass.), 1972, págs. 427-447. También ha ejercido cierta influencia B. L. VAN DER WAERDEN, *Die Pythagoreer: Religiöse Bruderschaft und Schule der Wissenschaft*, Zurich, 1979. W. R. KNORR, *The Evolution...*, op. cit., 1975, págs. 131-169, ofrece una espléndida visión panorámica de la aritmética y otros aspectos de interés; *vid.* también su «Problems in the interpretation of Greek number theory», *Studies in History and Philosophy of Science* 7 (1976), 353-368. Más reciente —y, por lo regular, demasiado entusiástica— es la revisión de L. ZHMUD, «Pythagoras as a mathematician», *Historia Mathematica* 16 (1989), 249-268.

que no tenga una disposición tan lineal y bien dispuesta como la del libro I, resulta parejamente estimable. Algo parecido cabría decir de su calidad y eficacia expositivas como introducción a ciertos métodos peculiares de la aritmética griega. VII, 1-3 presentan el algoritmo euclídeo para la determinación de números primos y el hallazgo de la medida común máxima entre dos o tres números no primos entre sí (en VII, 33, el procedimiento se generaliza «a tantos números cuantos se quiera»). He ahí una de las joyas de la corona de la tradición matemática griega. VII, 4-20 sientan las bases de la teoría de la proporción numérica que se aplica en este y en los otros dos libros de la aritmética. VII, 21-32 desarrollan el importante tema euclídeo de los primos relativos y VII, 33-39 estudian la expresión de razones en sus menores términos y el hallazgo de mínimos comunes múltiplos. Por lo demás, la agrupación de las definiciones en la cabecera del libro y el material fundacional recogido en su primera parte hacen de él una fuente sistemática de ulteriores resultados como los obtenidos en los dos libros siguientes.

El libro VIII se ocupa de series de números en proporción continua, *i. e.* en progresión geométrica, noción que permanece indefinida.

El libro IX es ya una especie de miscelánea aritmética. Incluye el resultado IX, 4, la primicia de la moderna resolución unívoca de un número en sus factores primos, y el teorema IX, 20, que establece la cantidad infinita de los números primos: «los números primos son más que cualquier multitud determinada de números primos». En IX, 21-34, Euclides da la impresión de transcribir un manual anterior seguramente de origen pitagórico que versaría sobre pares, impares y sus relaciones —en opinión de algunos, como Heath, no faltan indicios de la existencia de unos *Elementos* de aritmética en la época de Arquitas, entre el último tercio del siglo v a. C. y la primera mitad del siglo IV; por desgracia, no hay mención ni referencia alguna al respecto [26]—. El libro y esta parte aritmética de los *Elementos* se cierran con la prueba de IX, 36, que da la fórmula para obtener los llamados «números perfectos» —según la Def. 22 de VII, «números iguales a la suma de sus factores» (e. g., el número 6, el 28, el 296)—.

5. *El libro X: «la cruz de los matemáticos»*

El libro X ha venido a conocerse como «la croix des mathématiciens» desde que Simon Stev' 1 empleara esta denominación para recordar que muchos sólo veían en este libro dificultades sin provecho [27]. Creo que el apelativo podría extenderse: si fue entonces una «cruz» para los matemáticos renacentistas, todavía representa hoy una especie de cruz para los historiadores de las matemáticas. Pero creo además que no merecería esa denominación únicamente por los problemas de interpretación que ha suscitado, sino por su condición de encrucijada dentro de los *Elementos*. En él confluyen desarrollos de la teoría generalizada de la proporción y motivos aritméticos; de él parten nuevas vías de construcción de figuras regulares tanto planas como sólidas —tema posterior de los libros XI-XIII sobre la geometría del espacio—. Consta de 16 definiciones repartidas en tres grupos a lo largo del libro; incluye 115 proposiciones, todas ellas probadas como teoremas, aunque varias (unas 24) revistan la forma ya conocida de las investigaciones que se proponen hallar algo con independencia de sus condiciones de generación o de construcción. El libro acusa en parte un laborioso esfuerzo por reconstruir de forma sintética, como si se tratara de hacer un ejercicio metódico y académico, varias series de problemas, susceptibles quizás de soluciones analíticas más simples aunque sean relativamente avanzados. En sustancia, está dedicado al estudio de tipos y criterios de conmensurabilidad e inconmensurabilidad y a la clasificación de rectas irracionales. De ahí la importancia del primer grupo de definiciones que encabeza el libro.

Según la Def. 1, «se dicen conmensurables [*sýmmetroi*] las magnitudes que son medidas por la misma medida, e

[26] *Vid.* T. L. HEATH, ed. cit., 1926², pág. 295. L. ZHMUD, art. cit., 1989, págs. 253-254, va más lejos aún y acoge, sin la menor reserva, la existencia de un manual pitagórico de aritmética anterior a los *Elementos* de Hipócrates —se trata de una conjetura avanzada por P. TANNERY (1887, reimp. 1988), *La géométrie grecque*, op. cit., págs. 87, 95, de la que luego se han hecho eco algunos otros como A. Rey o B. L. van der Waerden—.

[27] *Vid.* C. M. TAISBAK, *Coloured Quadrangles. A Guide to the Tenth Book of Euclid's Elements*, Copenhague, 1982. W. R. KNORR, «La croix des mathématiciens: the Euclidean theory of irrational lines», *Bull. Amer. Mathem. Society* 9 (1983), 41-69; «Euclid's tenth book: An analytic survey», *Historia Scientiarum* 29 (1985), 17-35. D. H. FOWLER, *The Mathematics of Plato's Academy*, Oxford, 1987, cap. 5 en especial, y «An invitation to read Book X of Euclid's *Elements*», ensayo reciente (enero de 1991) todavía inédito.

inconmensurables [*asýmmetroi*] las que no tienen ninguna medida común». De acuerdo con la Def. 2, «las líneas rectas son conmensurables en cuadrado [*dynámei*] cuando los cuadrados ⟨levantados⟩ sobre ellas son medidos por una misma área, e inconmensurables en cuadrado cuando los cuadrados sobre ellas no pueden tener ninguna área como medida común». (El término «*dýnamis* (potencia)» tiene en matemáticas una referencia equivalente a lo que nosotros significamos con «elevar a la segunda potencia» —o incluso con «raíz cuadrada» a tenor de Platón, *Teeteto* 147d [28]—; en contextos geométricos esta potencia particular es obviamente el cuadrado.) A tenor de la Def. 3: «con estos supuestos se prueba que hay una cantidad ilimitada de líneas rectas que son conmensurables e inconmensurables respectivamente, unas sólo en longitud [*mékei*] y otras en cuadrado también, con una línea recta designada. Llámense entonces racional [*rheté*] la línea designada y racionales las líneas que son conmensurables con ella, bien en longitud y en cuadrado, bien en cuadrado sólo, pero irracionales [*alógoi*] las inconmensurables con ella». Y en fin, como estipula la Def. 4, las áreas se llamarán racionales o irracionales según que sean conmensurables o inconmensurables con un cuadrado racional. Sentada la distinción

[28] Este pasaje y su contexto (*Teeteto* 147c-148a), donde Teeteto da cuenta de su investigación de inconmensurables en la perspectiva abierta por las contribuciones de Teodoro, es uno de los testimonios de Platón más oscuros y debatidos hasta el punto de constituir por sí solo un foco principal de discusión de la historia de la matemática griega, *vid.* J. L. BERGGREN, «History of Greek mathematics: a survey...», art. cit., 1984, págs. 395, 401-402. Cf., por ejemplo, W. R. KNORR, *The Evolution...*, op. cit., 1975, págs. 62-108; M. F. BURNYEAT, «The philosophical sense of *Theaetetus*' mathematics», *Isis* 69-249 (1978), 489-513; W. R. KNORR: «Methodology, philology and philosophy», *Isis* 70-254 (1979), 565-570.

capital entre las magnitudes conmensurables e inconmensurables, tiene especial relieve la subdivisión de las rectas conmensurables/inconmensurables en otras dos clases: las que lo son en cuadrado y las que lo son en longitud. Las primeras son el objeto de la Def. 2. Las segundas aparecen determinadas en el contexto de X, 9: las rectas son conmensurables o inconmensurables en longitud según que sus cuadrados tengan entre sí o no la razón que un cuadrado guarda con un número cuadrado —el escolio 62 de X asegura que éste fue un hallazgo de Teeteto; recordemos, por lo demás, que un número cuadrado era ya en la tradición pitagórica el producto de dos factores iguales—. Un porisma de la prueba de X, 9 agrega: si A es conmensurable en longitud, es conmensurable en cuadrado, pero no vale la conversa —hay líneas que siendo conmensurables en cuadrado, como los catetos y la hipotenusa de un rectángulo, no lo son en longitud—; y por ende, si A es inconmensurable en cuadrado, es inconmensurable en longitud, pero tampoco vale a la inversa. La Def. 3 introduce un nuevo motivo de clasificación: la distinción entre rectas racionales e irracionales. El escolio X, 1.°, da a entender que si la primera distinción entre magnitudes conmensurables e inconmensurables tiene lugar de modo natural [*phýsei*], ésta segunda entre líneas o áreas racionales e irracionales tiene lugar por convención [*thései*] y dice relación a la recta designada como racional. Por otra parte, este uso de «racional» en los *Elementos* parece peculiar de Euclides y tiene un sentido más amplio que otros anteriores o posteriores de la misma calificación; en justa correspondencia, su noción de irracionalidad resulta más restringida, pues sólo son líneas irracionales las que resultan ser inconmensurables tanto en longitud como en cuadrado con la recta designada como racional.

Son muchos y diversos los resultados dignos de mención que se encuentran en el libro X. Pero me contentaré con referirme a algunos de los expuestos al principio del libro. X, 1 es la base del método de «exhausción»: «dadas dos magnitudes desiguales, si de la mayor se sustrae una magnitud menor que su mitad, y del resto una magnitud mayor que su mitad y este proceso es reiterado constantemente, quedará una magnitud que será más pequeña que la menor de las magnitudes dadas». Apenas rinde servicio alguno hasta el libro XII. Este teorema, junto con la Def. 4 del libro V, han sido objeto de discusiones historiográficas y matemáticas —como ya he apuntado a propósito de esa definición—. Por un lado, se les ha atribuido la misión de excluir la existencia de infinitésimos con el fin de prevenir las paradojas eleáticas sobre el infinito. No hay datos que avalen esta proyección matemática de un problema de origen cosmológico o físico; si acaso, Aristóteles testimonia que las cuestiones acerca de la divisibilidad infinita de una magnitud extensa no preocupan a los geómetras de su tiempo: a ellos les basta con postular la existencia de líneas finitas del tamaño que se quiera y la posibilidad de seccionar un segmento menor siempre que sea preciso *(Física* III 7, 207b27-34). Por otro lado, se ha planteado la relación de X, 1 con un lema anterior de Eudoxo, asumido luego y declarado por Arquímedes en el prefacio de *La cuadratura de la parábola* y como *lambánon* 5 de *Sobre la esfera y el cilindro* I. Y, en fin, desde un punto de vista más conceptual que histórico, también cabe considerar su significación en el contexto de la «exhausción» y en la perspectiva del cálculo infinitesimal. No puedo detenerme ahora en el método de «exhausción» [29], aunque sí conviene

[29] Una exposición clásica es la de E. J. DIJKSTERHUIS (1956): *Archimedes*, Princeton, Nueva Jersey, reimp. 1987, págs. 130-133 en particular. También pueden verse B. RODRÍGUEZ-SALINAS, «Arquímedes», en el colectivo *Historia de la matemática hasta el s. XVII*, Madrid, 1986, págs. 84-86 en especial; y L. VEGA, *La trama de la demostración...*, op. cit., 1990, IV, § 3.1, para el funcionamiento del método en Euclides *(Elementos* XII, 2).

destacar las distancias y las diferencias entre este procedimiento griego de acotación métrica proporcional de áreas y el moderno cálculo infinitesimal. El método de «exhausción» griego conoce algunas variantes (de aproximación, de compresión), incluye esencialmente la reducción al absurdo y, sobre todo, carece de la noción de límite: por consiguiente, no necesita un principio de continuidad para establecer la existencia del límite de convergencia de una serie. En la práctica griega, el «límite» —por así decir— ya está dado por el objeto geométrico considerado y el método sólo tiene que aplicarse a la determinación de secuencias convergentes. En definitiva, Euclides —al igual que otros matemáticos griegos clásicos, como Eudoxo o Arquímedes— trabaja con objetos geométricos y en una perspectiva métrica de raíz cualitativa, en vez de operar con conceptos o estructuras generales de cálculo. Por lo demás, Euclides no presenta X, 1 como un lema envuelto en el uso de la «exhausción», según cabría esperar de una línea de trabajo eudoxo-arquimediana; antes de una aplicación de este tipo (e. g., en XII, 2), X, 1 aparece en el contexto antifairético [29bis] de las proposiciones X, 3-4, formulado en términos más débiles que los del lema de bisección exacta empleado por Arquímedes —pero un corolario de X, 1 no deja de señalar que «el teorema puede probarse de modo similar aun cuando las partes sustraídas

[29bis] «Anthyphaireîn [ἀνθυφαιρεῖν]» es la denominación técnica de Euclides para su procedimiento de sustracción recíproca, por ejemp. en VII, 1-2, y X, 3-4. Sobre su posible significación preeuclídea, *vid.* W. R. KNORR (1975), o. c., págs. 255-261 en particular.

sean iguales», y así recoge la bisección que luego será habitual en las pruebas arquimedianas—.

La proposición X, 2 se hace eco del criterio antifairético de inconmensurabilidad: dos magnitudes son inconmensurables si la operación metódica de búsqueda de su medida común máxima nunca tiene fin. Las proposiciones X, 3-4 ofrecen el algoritmo euclídeo para la determinación de la medida común máxima de dos o tres magnitudes conmensurables. X, 5 sienta que «las magnitudes conmensurables tienen entre sí la razón que un número guarda con otro número». Parte de la prueba de esta proposición descansa en la Def. 20 del libro VII, *i. e.* en la noción de proporción prevista para términos que sean todos ellos números, y en el supuesto tácito de que los términos que sean proporcionales en el sentido de la Def. 20 de VII, también lo son en el sentido generalizado de la Def. 5 de V. En otras palabras: después de haber dado dos caracterizaciones autónomas y separadas de la proporcionalidad, Euclides viene a suponer que la segunda puede considerarse un caso particular de la primera, o que los números se comportan —hasta cierto punto— de forma pareja a las magnitudes. Esta congruencia ya era familiar en tiempos de Aristóteles; Aristóteles muestra una clara conciencia de la generalidad de las pruebas derivables del concepto eudoxiano de proporción —podían aplicarse a números, tiempos, longitudes, sólidos *(APo.* I 5, 74a17)—; incluso hace la observación de que el método aritmético utilizado para establecer ciertas propiedades de las magnitudes no sería adecuado si las magnitudes no fueran de algún modo como los números *(APo.* I 7, 75b4-5). Sin embargo, no es una congruencia prevista expresamente en los *Elementos* —como pudimos notar a propósito de la def. 20 del libro VII—. Las proposiciones X, 6-8 mantienen y desarrollan

la confluencia introducida en X, 5 para desembocar en el resultado antes citado, X, 9, atribuido a Teeteto.

El libro X recoge investigaciones relativamente recientes —de Teeteto, de Eudoxo y, posiblemente, de Hermótimo (el último autor anterior a Euclides que cita Proclo como «descubridor de muchos de los elementos»)—; se mueve en una línea de elucidación y de catalogación progresiva de unos resultados acerca de diversos tipos de conmensurabilidad/inconmensurabilidad y de clases posibles de rectas racionales/irracionales. Pappo, en un comentario sobre el libro que sólo se ha conservado a través de una versión árabe, reconoce el papel clarificador que corresponde a Euclides [30]. No obstante, el libro X mantiene cierta oscuridad y algunas ambigüedades de principio. Su estructura deductiva también adolece de una cohesión interna deficiente; por ejemplo, las proposiciones 27-35, 42-47, 66-70, 79-84, 85-90, 103-107 representan núcleos aislados en el conjunto del libro, y algunas otras, como 2-4, 24-25, 112-115, no cumplen ningún cometido en el conjunto de los *Elementos*. Del libro X se ha dicho incluso que es un desastre pedagógico (vid., por ej., W. R. Knorr, art. cit., 1983, pág. 59). Con todo y con esto, aún cabe reconocer al libro ciertos valores metódicos y disciplinarios: no sólo contiene el lema del procedimiento de «exhausción» y pre-

[30] Se trata de una versión de Abu Uzman AL-DIMASHQI (de principios del siglo X) hallada en París (ms. 952, 2, Biblioteca Imperial) por F. WOEPCKE, *Mémoires présentés à l'Académie des sciences* XIV (1856), págs. 658-719. Según PAPPO: «Por lo que se refiere a Euclides, ha dado pruebas rigurosas que ha sentado en relación con la conmensurabilidad y la inconmensurabilidad en general; ha precisado las nociones y las distinciones entre magnitudes racionales e irracionales, ha expuesto un gran número de tipos de magnitudes irracionales y por último ha clarificado todo su ámbito», *vid.* T. H. HEATH, *A History of Greek Mathematics*, op. cit. (1921, reimp. 1981), pág. 403.

senta una nueva aplicación del algoritmo euclídeo, sino que logra una acabada clasificación de líneas irracionales que llega hasta catalogar 13 géneros distintos —enumerados al final de X, 111—; también enseña a construir muestras de cada uno de ellos, amén de demostrar que otras subdivisiones de alguno de estos géneros no son especies vacías.

6. *Geometría del espacio. (Libros XI-XIII)*

Los últimos libros, XI-XIII, contienen la geometría del espacio de los *Elementos*. Al igual que sucedía con la aritmética, esta parte descansa en 28 definiciones avanzadas al principio de su primer libro y carece de postulados propios. Consta de 75 proposiciones, 63 de ellas teoremas y las otras 12 problemas —aunque entre éstos aparezcan unas proposiciones mixtas, como XIII, 13-17, cuyo enunciado señala expresamente la construcción de un objeto y la demostración de alguna característica métrica esencial—. Un rasgo interesante de algunas definiciones (la Def. 14 de la esfera, la Def. 18 del cono) es una introducción del movimiento que les da un aire de descripciones genéticas. Por ejemplo, la Def. 14 dice: «cuando, permaneciendo fijo el diámetro de un semicírculo, se hace girar el semicírculo y se lo vuelve de nuevo a la misma posición de la que había partido su puesta en movimiento, la figura así comprendida es una esfera». Esta noción de esfera no era, desde luego, la única que podía conocer Euclides, pues Aristóteles ya estaba al tanto de que el criterio de equidistancia era extensible a esta figura sólida (vid. *De Coel.* II 14, 297a24) y, más tarde, Herón la definirá en esta línea de modo parejo a como Euclides había definido el círculo («una esfera es una figura sólida rodeada por una superficie tal que todas las líneas que caen sobre ella a partir de un punto interior de la figura son iguales entre sí», *Definiciones*, Def. 76). La opción de Euclides puede justificarse por su rendimiento en la prueba de las últimas proposiciones del libro XIII. Otra posibilidad que Euclides parece ignorar por completo es la de servirse en este caso de un postulado análogo al postulado (iii) de la geometría plana. En general, la ausencia de postulados específicos es digna de tenerse en cuenta. Un procedimiento que Euclides emplea habitualmente en su geometría sólida es la reducción de una cuestión tridimensional a otra bidimensional a la que aplica luego resultados o lemas previamente obtenidos. Pero esta estrategia reductiva supone relaciones —no tratadas antes en los *Elementos*— entre planos y puntos, planos y líneas, planos y planos, relaciones que en última instancia quedan todas ellas sin especificar. En fin, el empleo de la geometría plana en el contexto de XI-XIII también discurre al margen de un planteamiento expreso de la posibilidad de transferir los resultados acerca de un plano a la solución de problemas que envuelven más de un plano.

La celebridad de esta última parte de los *Elementos*, dejando aparte sus eventuales servicios a la astronomía helenística, ha residido en el estudio de los cinco poliedros regulares, los renombrados «cuerpos platónicos», en el libro XIII. (Esa fama, como ya he sugerido, no es una prueba de la filosofía platónica de Euclides; antes bien, sí es señal de que las matemáticas no estaban llamadas a librarse del fervor neoplatónico de los comentadores y a'baceas posteriores de este legado clásico.) Euclides prueba la construcción dentro de una esfera de la pirámide o tetraedro (XIII, 13), el octaedro (XIII, 14), el cubo (XIII, 15), el icosaedro (XIII, 16), el dodecaedro (XIII, 17), y remata libro y tratado con un famoso corolario a la proposición XIII, 18, en

el que asegura que no existe ningún otro poliedro regular. En el libro XIII concurren materiales y resultados de casi todos los demás libros, de modo que presenta un alto grado de cohesión externa con el resto de los *Elementos*. Esto, a primera vista, concuerda con uno de los objetivos que Proclo atribuye al tratado: la construcción de los cinco sólidos regulares, los cuerpos cósmicos del *Timeo*. Pero este propósito no está respaldado por ninguna indicación del propio Euclides o del texto mismo. Cabe pensar, por otro lado, en una filiación de Euclides con la tradición matemática de la geometría sólida: tres de los poliedros regulares (el cubo, la pirámide, el dodecaedro) ya habían atraído la atención de los pitagóricos, y los otros dos (el icosaedro, el octaedro) habían sido estudiados por Teeteto, según recuerda el escolio XIII, 1.°. El hecho de que Teeteto pasara por ser «el primero en escribir sobre los cinco sólidos» cuadra bien además con su contribución al análisis de los irracionales y es, seguramente, una congruencia similar la que manifiesta y desarrolla Euclides en el libro XIII [31]. Así pues, podríamos alinear a Euclides no tanto en el ámbito cosmológico de influencia del *Timeo* —cuyo peso se hará sentir más bien pasado algún tiempo—, como en la estela matemática de unas tradiciones e investigaciones anteriores a este mismo diálogo de Platón. Esta línea de trabajo alcanza luego la madurez suficiente para dar lugar —pocos años antes de los *Elementos*— a un tratado

[31] Un tratado de Aristeo sobre los cinco sólidos regulares (compuestos a finales del siglo IV a. C.) podría haber sido la fuente más próxima de inspiración de Euclides, según han apuntado ALLMAN (1889, reimp. 1976): *Greek Geometry...*, op. cit., págs. 201-202, y T. L. HEATH, ed. cit., 1926², 3, págs. 438-439. De este tratado de ARISTEO («el viejo»), *Tôn pénte shkemáton sýnkrisis*, hay noticias por Hipsicles —el presunto autor del libro XIV añadido a los *Elementos*—.

de Aristeo dedicado expresamente a un estudio comparado de los cinco sólidos regulares: de este tratado pudo haberse beneficiado Euclides al componer el libro XIII, al igual que, efectivamente, lo haría más tarde el autor (Hypsicles) del apócrifo libro XIV.

El libro XIII recoge resultados de las otras partes de los *Elementos* salvo de la aritmética y —curiosamente— del libro XII. Además tiene una estructura deductiva interna sumamente cuidada. XIII, 1-5 constituyen lemas básicos para el desarrollo del libro que culmina en el análisis comparativo final de los sólidos regulares en XIII, 18 y en su corolario. Por contra, la cohesión y la capacidad sistemática de los dos libros precedentes, el XI y el XII, es mucho menor. El papel de XI es preparar el camino hacia XIII y lo más significativo de XII es seguramente el empleo del método de «exhausción» en las pruebas de XII, 2, 5, 10-12 y 18 sobre la base del lema X, 1 —se trata de una base parecida pero no igual a la que Eudoxo habría empleado en la obtención de XII, 2, 7, 10 y 18, según da a entender Arquímedes en el prefacio de *La cuadratura de la parábola*—.

Creo que todas estas observaciones de los apartados 1-6 permiten hacerse una idea global de la constitución del tratado a partir de los distintos libros que componen los *Elementos*. La conclusión más obvia y general es el contraste entre la fachada «axiomática» del libro I y la composición no poco heterogénea e irregular del tratado en su conjunto. Aunque la composición euclídea está lejos de justificar un énfasis sobre la informalidad de la prueba matemática como el expresado por Wittgenstein («Me gustaría decir: la matemática es una abigarrada *mezcla* de técnicas demostrativas», *Observaciones sobre los fundamentos de la matemática* (1939-1940) III, § 46), también dista de

ser el sistema axiomático constituido por nuestra geometría euclidiana moderna o contemporánea. Los *Elementos* dan la impresión de hallarse a medio camino entre un extremo y otro. Esta impresión no sólo se debe a la variedad de los cuerpos teóricos involucrados —cuyos nexos sistemáticos no siempre se declaran — o a la diversidad de los métodos de solución y de prueba, sino también a la dispar conformación deductiva de los distintos libros. Por lo que refiere a este último aspecto, cabe sintetizar esas diferencias e irregularidades al hilo de unas nociones genéricas que ya he venido empleando y sólo aspiran a ser vías de aproximación.

Para empezar, propongo una distinción entre una cohesión deductiva interna y una cohesión externa. La cohesión interna será mayor o menor según sea el grado de sistematización deductiva existente entre las proposiciones de un determinado libro; la cohesión externa será mayor o menor según sea el grado de articulación deductiva que se dé entre el libro en cuestión y otros pertenecientes al mismo campo teórico. También cabe distinguir entre la coherencia teórica de un libro y su disposición disciplinaria o didáctica. La coherencia teórica será mayor o menor en razón de si el orden de desarrollo seguido casa mejor o peor con las líneas o focos de interés temático o sustantivo de los resultados presentados; la disposición disciplinaria será mejor o peor en razón de la claridad de la exposición y de su eficacia didáctica —aquí pueden tomarse en consideración la transición expositiva desde unas cuestiones relativamente simples a las otras más complejas, o las facilidades dadas para el aprendizaje y el dominio de unas técnicas de solución o de prueba—. Según esto, el libro X, por ejemplo, tiene cierto grado de cohesión externa, ciertos visos de encrucijada, pero ni su cohesión interna ni

su plan de exposición desciplinaria son muy estimables. La aplicación de áreas, por su parte, cuenta con indudables méritos expositivos y didácticos tanto en el libro II como el libro VI, en especial por lo que se refiere al aprendizaje de esta técnica mediante su aplicación a varios casos; pero, desde luego, ni II ni VI son muestras brillantes de coherencia teórica o de cohesión interna. Aún es más notoria la distancia que media entre la eficaz exposición de la técnica de inscripciones y circunscripciones emparejadas en el libro IV y el bajo grado de cohesión interna y de coherencia teórica del libro en su conjunto. Algo parecido cabe pensar de fragmentos como IX, 21-34, cuyas virtudes didácticas y disciplinarias contrastan con el aire de miscelánea (temática y deductiva) que envuelve todo el libro. Por último, no estará de más recordar que los materiales y resultados recogidos a lo largo del tratado no sólo pertenecen a cuerpos teóricos con distinto grado de desarrollo, sino que provienen de tradiciones e investigaciones de mayor o menor antigüedad: unas separadas de Euclides por cerca de dos siglos y otras —como mucho— por apenas dos o tres generaciones. Así cabe llegar a un esquema general de los casos que representan algunos libros de los *Elementos,* por lo que refiere a su tipo o grado de sistematización deductiva, como el esquema siguiente:

		material recogido	
		más antiguo	más reciente
cohesión interna	mayor:	I, VII	V, XIII
	menor:	III, IV, IX	VI, X
cohesión externa	mayor:	I, VII	V, X, XIII
	menor:	III, IX	XII
coherencia teórica	mayor:	I	XIII
	menor:	IV, IX	VI, XII

De este esquema se desprende que no hay una correlación significativa entre el grado de sistematización de estos libros y la antigüedad relativa del material que exponen o elaboran. Tanto los resultados más antiguos como los desarrollos más próximos al autor de los *Elementos* pueden presentarse bajo reconstrucciones más afortunadas (e. g.: I, VII; V, XIII, respectivamente) o menos afortunadas (e. g.: III, IV; VI, X). Por otra parte, ante esos altibajos, no es difícil sacar la impresión de que los *Elementos*, por regla general, cumplen mejor con el propósito de constituir una introducción eficaz y progresiva a las teorías y métodos de la matemática elemental —la geometría, más en particular—, que con las exigencias de lo que habría de ser una reconstrucción deductiva cabal de todo un cuerpo, complejo y articulado, de conocimientos; por lo menos, su rendimiento suele ser más regular y efectivo en relación con el primero que en relación con éstas últimas. Llegados a este punto, ¿qué significación podemos reconocer hoy a la sistematización euclídea?

IV. Sobre la significación axiomática de los «Elementos»

Si miramos a nuestro alrededor para conocer la suerte actual de la obra de Euclides, lo que más nos llamará la atención será lo siguiente: el lenguaje matemático de los *Elementos* es desde hace tiempo una lengua muerta; ha pasado a la historia [32]. Pero, al margen de su lugar de

[32] *Pace* N. Cuesta Dutari —cf. «Tres notabilísimos pasajes de Euclides», *Llull* 4, 6-7 (1981), 35-41—.

reposo bajo el sol en la historia de la matemática, Euclides sigue siendo un personaje vivo o reciclado en muchos libros de filosofía de la ciencia o de metodología deductiva. Se trata, por supuesto, de la supervivencia vicaria que otorgan la fama y la pedagogía. Los *Elementos* sobreviven gracias a la fama en calidad de arquetipo del método axiomático; sobreviven gracias a la pedagogía como la ilustración cabal de una axiomática determinada, a saber: una axiomática «material», «ingenua», «informal», por contraposición a la axiomática «abstracta», «postulacional», «formal» que la geometría empieza a conocer en el curso del siglo XIX a través de Pasch, Peano, Veronese y Hilbert. Así, en la literatura metodológica al uso, no es infrecuente dar con definiciones como ésta:

Σ es una *axiomática euclídea* si y sólo si: *(i)* Σ es un conjunto posiblemente infinito de proposiciones cerrado con respecto a la relación de consecuencia lógica; *(ii)* hay un subconjunto finito Δ cuyos miembros son verdades primeras inmediatas y autovidentes —llamados «axiomas» de Σ—; *(iii)* cualquier otra proposición de Σ que no pertenezca a Δ, es una verdad demostrable a partir de Δ. Condiciones como la verdad primordial y la autoevidencia de los axiomas son las que harían de Euclides el presunto prototipo de esta axiomática.

Pero, por lo regular, las descripciones de este género no pretenden ser una interpretación histórica de Euclides, sino más bien una estipulación didáctica. Se limitan a aducir un tópico metodológico de referencia o de contraste, el primer ejemplar de *homo axiomaticus*, para dar una imagen del progreso de esta especie o para hacer que resalte a contraluz la estructura de la axiomática contemporánea —que, en definitiva, es la que interesa—. Por eso tampoco importa mucho que la fama y la pedagogía crean hablar

de Euclides, cuando en realidad se están refiriendo al programa clásico de axiomatización nacido en el siglo XVII, bajo lemas como el «more geometrico» o «l'esprit de la géométrie»; y crean hablar de los *Elementos* cuando se refieren a su reconstrucción axiomática en los términos de la geometría *euclidiana* del siglo XIX.

Las confusiones y los malentendidos surgen cuando esta retroproyección va más allá de sus fines iniciales, didácticos o analíticos, y pasa a convertirse en una especie de guía o clave interpretativa del sentido de los *Elementos* [33]. Entonces, aun evitando otros puntos discutibles, hemos de abordar una cuestión de fondo: ¿cuál es el significado del término «axiomatización» aplicado a los *Elementos?* Pues, los *Elementos,* ¿acaso no contienen alguna muestra efectiva de axiomatización? Además, su organización deductiva de ciertos cuerpos de conocimiento y la práctica de una forma característica de la demostración matemática ¿no han sido justamente los motivos que han hecho de los *Elementos* el paradigma clásico de la geometría, y han hecho de la geometría misma la ciencia demostrativa por antonomasia? En suma: la «cuestión axiomática» es una de las que determinan nuestra comprensión y valoración histórica de la contribución de Euclides. Plantearé esta cuestión en el marco configurado por tres hipótesis que parecen agrupar las opiniones más frecuentes o características.

[33] Cf., por ejemplo, D. LACOMBE, «L'axiomatization des mathématiques au III siècle avant J. C.», *Thalès* 3 (1949-1950), 37-58; S. S. DEMIVOD, «Évolution, extension et limites de la méthode axiomatique dans les sciences sur l'exemple de la géométrie», *Archives intern. d'histoire des sciences* 90-91 (1970), 3-30; M. GUILLAUME, «Axiomatique et logique», en J. DIEUDONNÉ, dir., *Abrégé d'histoire des mathématiques, 1700-1900,* Paris, 1978; vol. II, págs. 315-430; P. FREGUGLIA, *Fondamenti storici della geometria,* op. cit., 1982.

1.

Podemos empezar considerando dos hipótesis relativamente descentradas: una hipótesis mínima y una hipótesis máxima.

De acuerdo con la hipótesis mínima, Euclides fue un mero recopilador [34]. Su aportación consiste en seleccionar y reunir diversos legados de conocimiento matemático —algunos de ellos organizados deductivamente (la geometría plana, la teoría de la proporción, la aritmética)— en un buen manual, con una exposición más o menos sistemática según el desarrollo alcanzado por la materia tratada.

La hipótesis carece de base documental. La deuda de los *Elementos* con la tradición matemática es indudable. Pero, como hemos podido ver, no es menos cierta su reelaboración tanto del material más lejano como del más próximo, especialmente en lo que se refiere a su organización deductiva. En efecto, son raros los pasajes que parecen reflejar una mera transcripción —quizás IV, 1-15, sobre inscripciones y circunscripciones, IX, 21-34, sobre la aritmética del número par e impar, y muy pocos más—. Ahora bien, por regla general, hemos de reconocer las innovaciones introducidas por Euclides en la forma y disposición del legado matemático anterior —aparte de alguna que otra

[34] E. C. ZEEMAN, en su (1974): «Research, Ancient and Modern» (recogido en *Catastrophe Theory: Selected Papers, 1972-1977,* Reading, Mass., 1979, págs. 605-614), esboza una divertida caricatura de Euclides como adusto compilador y administrador del Departamento de Matemáticas en Alejandría, por contraposición al talento y la originalidad creadora de Eudoxo; puede que Arquímedes tuviera una impresión parecida, no del propio Euclides, pero sí desde luego de algunos de sus discípulos alejandrinos.

contribución original según todos los visos, e. g., la teoría de las paralelas—. Como ya hemos visto, en la atribución de esas innovaciones metodológicas a Euclides coinciden varios testimonios de Pappo, Proclo y los escoliastas. Y no faltan otros elementos de juicio que corroboran esta apreciación, por ejemplo: las muestras de pruebas anteriores que nos han llegado conservadas por Platón y Aristóteles o transcritas por algunos comentadores tardíos (e. g. Simplicio), cuyas diferencias con el tratamiento euclídeo de esos mismos resultados son notorias e inequívocas.

De acuerdo con la hipótesis máxima, los *Elementos* fundan la ciencia del espacio geométrico y son la expresión cabal del método axiomático de prueba que caracteriza la madurez de la demostración matemática griega. En ambos casos, dan así respuesta a problemas cosmológicos y estímulos dialécticos que provienen, en último término, de Elea (Parménides, Zenón, Meliso). La fundación de la geometría como ciencia sistemática y coherente del espacio responde al reto y a las paradojas de los eleatas en torno a la idea de magnitud extensa y otras nociones asociadas, en particular la idea misma de espacio. E. g.: si lo que hay no tuviera extensión, no existiría; ahora bien, de haber una magnitud extensa, ésta sería infinitamente grande e infinitamente pequeña, como consecuencia de su divisibilidad indefinida (Zenón, 29 B 1); así pues, el espacio como medio o marco extenso es una suposición aporética o absurda (29 B 5; cf. *Teeteto* 180c). La expresión del método axiomático en los *Elementos* es, a su vez, la culminación de una tendencia hacia la abstracción conceptual y la autonomía racional de la demostración, tendencia que, asimismo, arranca de la dialéctica eleática —en particular, de su empleo de la reducción al absurdo— y contrasta con las pruebas empíricas, visuales o «materiales», que utilizaban los primeros pitagóricos. Esta interpretación, propuesta y desarrollada por Arpad Szabó desde los años 1950, ha tenido considerable eco entre los historiadores de la filosofía y ha provocado una reacción crítica no menos considerable entre muchos estudiosos de la matemática griega [35]. Los aspectos más discutidos son los que hacen referencia a una deuda crucial de la tradición matemática con la cosmología y la dialéctica de Elea. No hay pruebas documentales que avalen inequívocamente esta conjetura; hay, en cambio, datos elocuentes en su contra (e. g., el testimonio expreso de Aristóteles sobre la despreocupación de los geómetras por las aporías acerca del infinito y del continuo, *Fís.* III 7, 207b27-34). Pero los aspectos que ahora tienen interés son los relativos a la significación de los *Elementos* como fundación de la ciencia del espacio y como espejo del método axiomático.

Para empezar, en los *Elementos* no se encuentran ni una idea general del espacio geométrico —que será, en última instancia, una aportación moderna de la geometría *euclidiana* a la teoría de las paralelas de la geometría de

---

[35] *Vid.* A. SZABÓ, «Wie ist die Mathematik zu einer deduktiven Wissenschaft geworden?», *Acta Ant. Acad. Scient. Hung.* IV (1956), Budapest, págs. 109-152; «The transformation of mathematics into deductive science and the beginnings of its foundation on definitions and axioms», *Scripta Mathematica* XXVII 1 y 2 (1964), 27-48 y 113-139; «Greek dialectic and Euclid's axiomatics», en I. LAKATOS, ed., *Problems in the Philosophy of Mathematics*, Amsterdam, 1967, págs. 1-8, Discussion, págs. 20-27; (1969) *The Beginnings of Greek Mathematics*, Dordrecht-Budapest, 1978. Por contra, cf. K. BERKA, «Was there an Eleatic background to pre-Euclidean mathematics?», en J. HINTIKKA, D. GRUENDER y E. AGAZZI, eds., *Theory Change, Ancient Axiomatics, and Galileo's Methodology*, Dordrecht-Boston, 1980, págs. 125-131; W. R. KNORR, «On the early history of axiomatics: the interaction of mathematics and philosophy in Greek antiquity», *ibid.*, págs. 145-186.

Euclides—, ni una idea general de magnitud. Éste es un punto claro de contraste de los *Elementos* con la práctica hoy normal del método axiomático. Los primeros pasos de la axiomatización de un campo de conocimientos son la determinación de un dominio de discurso y la especificación de ciertas relaciones estructurales entre los elementos, dentro de ese dominio, bajo la forma de condiciones o cláusulas axiomáticas. Luego cabe considerar los sistemas de objetos capaces de satisfacer tales condiciones, los posibles modelos de esa estructura. Así, podemos entender el espacio euclidiano como un lugar de puntos que cumple unas condiciones axiomáticas como las de infinitud, isotropía, continuidad, etc., aparte de atenerse al postulado (v) o a una condición equivalente. Pero éste no es el medio en el que se mueven los *Elementos*. Los *Elementos* operan de entrada sobre objetos geométricos o matemáticos, unos dados y conceptualizados (en el sentido sugerido por Aristóteles: *APo.* I 10, 76b39-77a2), otros susceptibles de construcción o de hallazgo. El triángulo, el cuadrado, el círculo, etc. son la base de partida para el estudio de las propiedades y las relaciones geométricas (siendo geométricas por ser inherentes a unos objetos geométricos), en vez de ser meras objetivaciones de una red estructural de relaciones o de una idea de espacio como lugar o conjunto de puntos. Por eso las definiciones euclídeas suelen ser expresiones de nociones intuitivas, caracterizadas por separado y más pendientes a veces de su carga tradicional que de su cometido como nudos de un entramado teórico. En suma, los *Elementos* están lejos de significar una ciencia del espacio —más aún: el antiguo pensamiento griego, tan dado a discutir sobre el espacio cosmológico y el lugar físico, no llega a plantearse expresamente un concepto general y preciso de espacio geométrico—.

Las dificultades para ver en los *Elementos* un espécimen axiomático crecen si recordamos su composición teórica y metódica relativamente heterogénea. Un supuesto clásico de la axiomatización de un cuerpo de conocimientos es su homogeneización como un conjunto de enunciados cerrado bajo la relación de consecuencia lógica. Pero en la composición de los *Elementos* no sólo cuenta esta alineación deductiva de los sucesivos problemas y teoremas; no sólo se trata de *hacer saber* tales o cuales resultados. También tienen particular importancia las construcciones diagramáticas y el entrenamiento en ciertos métodos; también se trata de *saber hacer,* de verificar con los objetos y las figuras geométricas una especie de «experimentos mentales» cuyo éxito viene garantizado por unas condiciones de construcción determinadas. Por lo demás, en el capítulo anterior ya hemos podido advertir la existencia de algunas otras desviaciones de carácter didáctico o de carácter conceptual que impiden al orden de exposición discurrir exactamente emparejado con un orden de deducción puramente lógico.

Aún hay más razones para dudar de la sistematización cabalmente axiomática del tratado de Euclides. La verdad es que el rendimiento de los principios euclídeos (definiciones, postulados, nociones comunes) deja bastante que desear en este sentido. Ni cada uno de estos principios es un supuesto necesario ni todos ellos forman un conjunto suficiente para la deducción de los cuerpos teóricos que toman a su cargo. Por ejemplo, las Deff. I, 1 de *punto,* y VII, 1 de *unidad* son «parásitos» filosóficos dentro de la tradición matemática; las Deff. I, 2-7; I, 19, o V, 3, también carecen de empleo propiamente dicho; hay incluso términos que tras ser objeto de definición, nunca más vuelven a mencionarse (e. g.: «oblongo [*heterómekes*]», «rom-

bo [*rhómbos*]», «romboide [*rhomboides*]», en la Def. I, 22; «sector [*tomeús*]», en la Def. III, 10).

En principio, cabría entender estos defectos como fallos o descuidos circunstanciales: también Euclides se nos duerme a veces. Pero si se toman en cuenta los aspectos antes considerados, estas deficiencias acusan más bien limitaciones del tipo de «axiomatización» *sui generis* que presentan los cuerpos más sistemáticos del tratado. No parece casual, en el marco de la estrategia deductiva de los *Elementos,* el hecho de pasar por alto un postulado que prevea la existencia de puntos de intersección de rectas con círculos, y de círculos con círculos, cuando es una existencia dada en su misma construcción diagramática y esta disposición diagramática no sólo es un medio de representación sino que además puede funcionar, llegado el caso, como un medio tácito de inferencia. Tampoco será casual la ausencia de unos supuestos de orden en la teoría de la proporción y de unos supuestos operacionales en la aritmética, si no hay un concepto general de magnitud, ni hay un concepto general de número, ni hay una conceptualización expresa de ciertas relaciones básicas como, sin ir más lejos, la de medir-a o ser-medido-por. En suma, estas y otras deficiencias por el estilo no parecen revelar tanto la inconsciencia o falta de lucidez del bueno de Euclides como una limitación interna de los propios *Elementos:* falla en ciertos momentos el nivel requerido de abstracción y de conceptualización, y falta en gener l una perspectiva estructural axiomática. Así pues, no creo que los *Elementos* constituyan una muestra cabal del método axiomático [36]. Aunque más tarde fueran tomados como pretexto

[36] En un sentido análogo se han pronunciado diversos estudios. Unos más bien parciales, como los de A. SEIDENBERG, «Did Euclid's *Elements*

y aun como emblema de la sistematización axiomática, la verdad es que sólo significan una especie de «axiomatización» en ciernes, un preludio de lo que luego irá logrando en tal sentido la tradición moderna, desde las visiones programáticas del «more geometrico» del siglo XVII hasta las refundiciones axiomáticas efectivas de la geometría euclidiana en el siglo XIX. Es posible que esta tradición moderna se imagine a veces estar explicitando la teoría y la estructura que los *Elementos* ya tenían originariamente, sólo que en un estado implícito y práctico. Pero en realidad está reconstruyendo ciertas partes del legado euclídeo de los *Elementos* en un nuevo contexto: sobre un nuevo terreno filosófico, con una nueva perspectiva metódica e incluso con otros materiales científicos y otros medios matemáticos.

2.

Descartadas las hipótesis anteriores sobre la significación axiomática de los *Elementos,* todo parece devolvernos a una hipótesis tradicional: no debe extrañarnos que la sistematización del tratado de Euclides sea una «axiomatiza-

book I develop geometry axiomatically?», *Archive for History of Exact Sciences* 14 (1974-1975), 263-295; G. G. GRANGER, *Essai d'une philosophie du style,* París, 1988², cap. II, págs. 24-42 en particular. Otros, en cambio, referidos al conjunto de los *Elementos,* como I. MUELLER, «Euclid's *Elements* and the axiomatic method», *British Journal for the Philosophy of Science* 20-3 (1969), 289-309, y sobre todo *Philosophy of Mathematics and Deductive Structure in Euclid's Elements,* op. cit., 1981. Una valoración similar del alcance de la axiomatización entre los antiguos griegos (Euclides —*Óptica*—, Arquímedes y Tolemeo en particular), se encuentra en P. SUPPES, «Limitations of the axiomatic method in ancient Greek mathematical sciences», en J. HINTIKKA y otros, *Theory Change, Ancient Axiomatics...,* op. cit., 1980, págs. 197-213.

ción» *sui generis,* pues consiste justamente en la materialización práctica de la teoría de la ciencia demostrativa propuesta por Aristóteles en los *Segundos Analíticos.* Esta interpretación puede sostenerse con distintos grados de entusiasmo. En plena euforia se llega a sugerir que Euclides aplica la teoría de la deducción de Aristóteles, *i.e.* su lógica silogística [37]. Menor es el entusiasmo y mayor la discreción de quienes se limitan a asumir una tesis familiar para sacar conclusiones. La tesis asegura: 1) en el desarrollo de la matemática hasta Eudoxo, los griegos crearon el primer ejemplo de una ciencia que estudia «axiomáticamente» sus elementos; 2) Aristótcles, precisamente, fue el que dio la primera descripción general de esta ciencia; 3) tan bien lo hizo Aristóteles que su teoría de la ciencia demostrativa representa «*la* concepción griega» de la demostración y «*la* concepción griega» de la sistematización axiomática [38]. La

[37] *Vid.* I. M. Bocheński, «The general sense and character of modern logic», en E. Agazzi, ed., *Modern Logic. A Survey,* Dordrecht-Boston, 1981, pág. 11. Los intentos expresos de calzar las pruebas euclídeas en la horma del silogismo de Aristóteles se remontan a su comentador Alejandro de Afrodisia (e. g., *In Top. Comm.* 14, 21 ss.). Mucho más tarde, entre los aristotélicos de Padua del Cinquecento, volvió a prender la discusión de la reducibilidad de la prueba geométrica al silogismo. Una muestra del éxito que cabe esperar en la empresa es la reducción que perpetra Clavio de la proposición I, 1 en su edic. (1574): *Euclidis Elementorum libri XV,* Roma, 1589², I 1 Scholium, I, págs. 76-77; para colmo de ingenuidad, remata el intento con esta expresión de fe y de esperanza; «Non aliter resolvi poterunt omnes aliae propositiones non solum Euclidis, verum etiam caeterorum Mathematicorum» (loc. cit., pág. 77).

[38] La exposición clásica de esta tesis es el artículo de H. Scholz (1930), «Die Axiomatik der Alten », reimpreso en su *Mathesis Universalis* (edic. a cargo de H. Hermes, F. Kambartel, J. Ritter), Basilea, 1969², págs. 27-44. Puede que represente el tópico más extendido sobre la ciencia demostrativa griega.

conclusión de todo esto se sigue fácilmente: los *Elementos* de Euclides son entonces la muestra más lograda y reconocida de esta concepción greco-aristotélica. Sea como fuere, el caso es que aun los intérpretes menos entusiastas se ven obligados a comparar el programa metodológico de los *Analíticos* con el «pórtico axiomático» de los *Elementos.* Tal comparación es un hábito tradicional que se remonta a Proclo [39].

A primera vista, la correspondencia entre los diversos tipos de principios de la ciencia demostrativa de los *Analíticos* y las definiciones, postulados y nociones comunes de los *Elementos,* resulta sumamente tentadora. Cabe imaginar una distribución correlativa semejante a ésta:

Analíticos		*Elementos*
principios comunes [*protá, koiná, axiómata*]		nociones comunes
principios específicos	definiciones [*hóroi*]	definiciones
	hipótesis [*hypothéseis*]	postulados.

[39] La literatura comparativa que han suscitado los *arkhaí* aristotélicos y los euclídeos es tan abundante como desigual en valor. Aparte de la prolija información recogida en T. L. Heath, *The Thirteen Books...,* ed. cit., 1962², 1, págs. 117-124, hay observaciones de interés en los trabajos de H. D. P. Lee, «Geometrical method and Aristotle's account of first principles», *Classical Quarterly* 29 (1935), 113-124; K. Von Fritz, «Die Arxai in der griechischen Mathematik», *Archiv für Begriffsgeschichte* 1 (1955), págs. 13-103 (luego incluido en *Grundprobleme der Geschichte der antiken Wissenschaft,* Berlín, 1971); A. Gómez-Lobo, «Aristotle's hypotheses and the Euclidean postulates», *Metaphysics* 30 (1977), 111-123; J. Hintikka, «Aristotelian axiomatics and geometrical axiomatics», en J. Hintikka y otros, eds., *Theory Change, Ancient Axiomatics...,* op. cit., 1980, págs. 133-144; W. R. Knorr, «On the early history of axiomatics...», art. cit., 1980; W. Leszl, «Mathematics, axiomatization, and hypotheses», en E. Berti, ed., *Aristotle on Science. The Posterior Analytics,* Padua, 1981, págs. 271-328.

Una correlación de este tipo supone cuando menos que, para empezar, la distribución aristotélica es una clasificación clara e inequívoca; que, en segundo lugar, algo parecido cabe pensar de la tripartición de Euclides; que, en fin, ambas guardan entre sí las correspondencias debidas. Por desgracia, todas estas suposiciones resultan discutibles. Aquí bastará considerar el caso de la distribución euclídea. Se trata de un simple listado de supuestos repartidos bajo los consabidos rótulos: *hóroi, aitémata, koinaì énnoiai.* No voy a entrar ahora en las licencias terminológicas que los matemáticos griegos se permitían a propósito de estas y otras maneras de llamar o indicar esos supuestos básicos [40]. La cuestión es: ¿hay algún criterio que justifique esta distribución tripartita? Como sólo aparece así en los *Elementos,* habremos de atenernos ante todo a su uso efectivo dentro de este contexto.

Proclo, en sus *Comentarios al libro I,* se hace eco de varios criterios al respecto. Uno, de clara impronta aristo-

[40] Podían utilizar, antes de Platón, denominaciones tan genéricas como *arkhaí* o *hypothéseis;* Aristóteles da fe asimismo del uso de *axiómata* y, lo que es más importante, de la diversidad de contextos de empleo de todos estos términos (contextos filosóficos, dialécticos, matemáticos). Para una visión general de esta variedad de usos y contextos, cf. los estudios clásicos de B. EINARSON, «On certain mathematical terms in Aristotle's logic», *American Journal of Philology* 57 (1936), 35-54, 151-172; K. VON FRITZ, «Die ARXAI in der griechischen Mathematik», *Archiv für Begriffsgeschichte* 1 (1955), págs. 13-103. Después de Euclides, hay cierta tendencia a servirse de *hypokeístho* para introducir definiciones y otras premisas primordiales, y de *etéstho* para introducir postulados, pero esta costumbre no constituye una norma; Arquímedes, por ejemplo, vuelve al uso genérico de *lambanómena* (asunciones) para separar estos supuestos básicos del cuerpo de resultados derivados de ellos. En todo caso, la terminología «axiomática» no llega a estar fijada en una época tan tardía como la de Proclo.

télica, distingue entre los principios comunes del conocimiento científico en general o de una familia de ciencias, e. g., las nociones acerca de las relaciones de igualdad y desigualdad, y los principios propios o específicos de un cuerpo determinado de conocimientos, es decir: los postulados en el caso de la geometría *(Com.* 182, 6-14). Un problema planteado por este criterio es qué hacemos con la noción común (iv*) que, al margen de su formulación general, se remite a una relación geométrica de coincidencia y, a veces, suple un postulado de superposición (e. g., en I, 4, 8) o representa una definición de congruencia. Otro problema es cómo explicamos el uso de las definiciones a las que suele confiarse la especificación de los campos de conocimiento cubiertos por los *Elementos* (no hay postulados para las magnitudes, ni para los números; de ahí que en ocasiones una definición haya de cumplir el cometido que hoy asignaríamos a un postulado matemático, e. g. la Def. 4 del libro V en la prueba del lema X, 1).

Otro criterio, seguramente desgranado de la tradición matemática, insiste en la distinción entre nociones comunes o axiomas y postulados: los axiomas están más indicados en la prueba de teoremas, pues allí se trata de establecer ciertos atributos o relaciones inherentes a los objetos geométricos; en cambio, los postulados están más indicados en la solución de problemas, habida cuenta de su mayor afinidad con los procedimientos de construcción *(Com.* 178, 12-179, 8, 181, 5-10). Este criterio no resulta muy eficaz por la relativa indiferenciación, al menos en la fase capital de la demostración [*apódeixis*], entre los teoremas y los problemas; por la existencia de investigaciones híbridas que se plantean una construcción para demostrar las propiedades o relaciones subsiguientes (e. g., la contrucción de sólidos regulares dentro de una esfera en la línea

de XIII, 13-17); y en fin por la existencia de resultados que no responden ni a una ni a otra condición, como los «porismas» en el sentido de cuestiones que se proponen hallar algo que no descansa en una construcción ni está simplemente sujeto a una determinación teórica (e. g., el hallazgo del máximo común divisor en VII, 2-3 o el hallazgo de la medida común máxima en X, 3-4).

Un tercer criterio, quizás elaborado a partir de Aristóteles y de Espeusipo, entiende que los axiomas son principios absolutamente dignos de crédito por sí mismos y no susceptibles de prueba, ciertos e indiscutibles, mientras que los postulados son asunciones corroborables por alguna suerte de demostración o de derivación *(Com.* 182, 21-183, 13). Pero este criterio tampoco funciona con la limpieza que cabría esperar: esa consideración de los postulados parece desmentida por el empeño en demostrar el postulado (v), no precisamente para convalidar su estatuto de postulado, sino para revocarlo y mostrar su calidad de teorema. Por otro lado, ¿cuál sería entonces el papel que toca desempeñar a las definiciones?

De las definiciones, y de su identificación frente a los postulados podría ocuparse un nuevo criterio quizás relacionado con algunas discusiones en el ámbito de la Academia platónica: las definiciones expresan propiedades esenciales de los objetos matemáticos y se mantienen al margen de los compromisos genéticos o constructivos, más propios de los postulados. Pero lo cierto es que hay definiciones genéticas (e. g., las definiciones XI, 14, 18 y 21, que describen respectivamente la esfera, el cono y el cilindro como sólidos generados por figuras planas que rotan alrededor de un eje fijo); hay definiciones que cumplen la tarea de un postulado; y hay, por contra, postulados que revisten la forma de condiciones definitorias (e. g., el postulado

(iv), que decreta la igualdad de todos los ángulos rectos, o el postulado (v), que sienta el criterio euclídeo de las paralelas).

En general, no hay criterios que justifiquen la presencia efectiva de cada principio en uno de los tres grupos distinguidos por Euclides, o criterios que permitan una redistribución cabal de los casos más problemáticos. Por ende, la selección de los principios y supuestos «axiomáticos» de los *Elementos* no parece abrir una perspectiva de comparación muy prometedora. Así pues, no es extraño que las comparaciones entre la base «axiomática» de la teoría de los *Analíticos* y el «pórtico axiomático» de los *Elementos* vengan a resultar tanto menos decisivas y concluyentes cuanto más finas y ponderadas sean.

Pero, al margen de este punto concreto de comparación, cabe intentar un balance global de las similitudes y diferencias que median entre la teoría aristotélica y la práctica euclídea de la demostración axiomatiforme. Las coincidencias más salientes se pueden cifrar en tres:

a) Hay una distinción neta entre las asunciones primeras y las proposiciones probadas, exigida por la noción técnica de demostración [*apódeixis*] que comparten Aristóteles y Euclides: toda demostración a partir de unas premisas debe remitir en última instancia a unos principios primeros e indemostrables, salvo que el proceso de deducción sea circular o se retrotraiga al infinito y la serie de las pruebas resulte, en su conjunto, infundada (Aristóteles, *APo*: I 3, 72b5-23; Proclo, *Com.* 75, 10-76, 4).

b) Esas proposiciones primordiales constituyen además principios de organización deductiva de cuerpos de conocimiento.

c) Se da por supuesta la autonomía teórica y disciplinaria de cada uno de esos cuerpos que cuente con unos

principios propios y específicos de conceptualización y de organización, aunque también haya de compartir ciertas nociones o axiomas comunes que tienen vigencia en cualquier ciencia demostrativa o en una familia de ciencias, e. g., la de las ciencias matemáticas de la magnitud, el número, la proporción. Si dentro de esta familia una ciencia tiene principio específicos, es una ciencia básica (e. g., la geometría, la aritmética); si ha de tomarlos en préstamo de otra, es una ciencia derivada (e. g., la óptica, la teoría musical).

Las diferencias más acusadas pueden contraerse parejamente a otras tres:

a') Si el interés aristotélico por los principios tiene claras connotaciones filosóficas, epistemológicas e incluso dialécticas, los supuestos básicos de los *Elementos* representan más bien un precipitado de la tradición matemática (a veces impregnada de motivos filosóficos como la distinción entre la unidad y el número o la distinción entre los objetos geométricos, ligados a postulados de construcción, y los objetos aritméticos, sólo susceptibles —al parecer— de definición), o son una reelaboración de este legado matemático; responden a una selección determinada por el curso de la investigación en ciertos ámbitos matemáticos y por la forma de exposición de ciertos resultados básicos en esas áreas; tal vez este compromiso con una tradición un tanto dispar y heterogénea haya contribuido a que los principios euclídeos rehúyan un criterio unívoco y uniforme de identificación y clasificación.

b') Si el programa aristotélico se centra en el análisis de las condiciones de la demostración científica que consiste en un *hacer-saber-que* algo es así y no puede ser de otra manera, las pruebas euclidianas también consisten en ocasiones en un *hacer-saber-cómo* se obtiene determinado tipo

de resultados o se resuelve determinada clase de problemas. Señal de este talante no sólo doctrinal y expositivo, sino operativo y práctico es la diferente función que pueden cumplir las definiciones y las hipótesis de los *Analíticos*, por un lado, y las definiciones y los postulados de los *Elementos*, por otro lado: en la teoría aristotélica no pasarán de ser premisas de una argumentación concluyente, mientras que en la práctica euclídea desempeñan un papel importante en la fase de preparación y conformación metódica de la prueba (en la *kataskeué)*, y las premisas de la fase demostrativa propiamente dicha [*apódeixis*] suelen ser más bien las nociones comunes o las proposiciones previamente establecidas.

c') Si la idea de demostración y de ciencia demostrativa no deja de estar marcada en Aristóteles por una lógica subyacente silogística, las pruebas euclídeas proceden con un rigor informal irreducible a los sistemas de análisis lógico disponibles en ese medio antiguo como el aristotélico o el estoico. No es extraño que, con el curso del tiempo, Posidonio, y tras él Galeno, reconozcan, junto a las demostraciones «categóricas» de corte lógico aristotélico o las demostraciones «hipotéticas» cortadas por el patrón lógico estoico, otro tipo de demostración, peculiar de los geómetras, que discurre en términos de relación [*katà tò prós ti*] y se funda en ciertos axiomas [*ék tinon axiomáton tèn aitìan ískhein systáseos*] [41].

Estas coincidencias y diferencias fácilmente apreciables, así como las relaciones un tanto complejas que el programa aristotélico de la demostración pudiera tener, y que

[41] *Vid.* GALENO, *Eysagogè dialektiké* xvi 1, 5; cf. también xvii, xviii. Hay una edición bilingüe (en griego y español) a cargo de M. H. OTERO y A. RAMÍREZ, *Iniciación a la dialéctica*, México, 1982.

la práctica euclídea de la deducción tiene de hecho, con la tradición matemática griega —sin ir más lejos, con la confección de *Elementos* matemáticos—, sitúan la comprensión histórica de los *Elementos* de Euclides en una perspectiva más amplia y más fecunda que la contemplada de ordinario por la hipótesis tradicional. En definitiva, si las hipótesis anteriores, la mínima y la máxima, pecaban de ceguera o miraban en una dirección equivocada, esta hipótesis tradicional de una correspondencia entre las ideas de Aristóteles y los usos de Euclides adolece más bien de miopía relativa.

3.

La comparación de la práctica euclídea con el programa aristotélico de la ciencia demostrativa puede rendir, al fin, un último servicio: el de darnos una clave sustancial del significado de la sistematización de Euclides. Esta clave consiste en la diferencia de «espíritu» o de intención que distingue la teoría de los *Analíticos* de la práctica de los *Elementos,* como si constituyeran dos fases de desarrollo de la idea griega de demostración.

La motivación capital de la teoría de la ciencia demostrativa preconizada por los *Analíticos* consistía en declarar qué es lo que hay que conocer o asumir en orden a forjar o entender una demostración concluyente y reconocerla como tal. La idea determinante en este contexto era la idea misma de demostración. Y una demostración venía a ser sustancialmente aquel «discurso racional [*lógos*] que se inscribe con el conocimiento en la mente del que aprende y es capaz de defenderse a sí mismo» (por decirlo con palabras de Platón, aunque procedan de otro contexto: *Fedro* 276a). Este concepto de demostración es el que alienta en

los análisis ulteriores (lógicos, epistemológicos, metodológicos) que conducen a Aristóteles a lo que podríamos llamar «un punto de vista axiomático», e. g., a la distinción entre principios indemostrables y proposicionales demostrables a partir o a través de ellos, a la subdistinción entre los principios comunes y las tesis propias o específicas de un campo de conocimientos, etc. [42]. En cambio, el propósito distintivo de la sistematización deductiva de los *Elementos* es, a mi juicio, la elucidación y la organización de ciertos campos matemáticos básicos —merced a una oportuna selección y disposición de los *Elementos*—, hasta su conversión en cuerpos autónomos y concluyentes de conocimientos. Así pues, no cabe negar un talante axiomatiforme a los *Elementos* de Euclides: la puesta en limpio y la reconstrucción sistemática de un conjunto de conocimientos, con miras a su constitución como una teoría científica autosuficiente, es, desde luego, uno de los motivos que se juzgan característicos de la axiomatización. El problema de si este talante aparece con Euclides o había surgido anteriormente (e. g., en los tratados de *Elementos* que circulaban por la Academia platónica, o en otro tipo de contribuciones, como la teoría de la proporción de Eudoxo), es una cuestión ahora irrelevante. Lo cierto en cualquier caso es que la idea de demostración propuesta por los *Analíticos* se extiende en los *Elementos* a todo un cuerpo de conocimientos: el «*lógos* que se inscribe en la mente del que aprende y es capaz de defenderse a sí mismo» no es ya una prueba concluyente o una demostración en particular, sino toda una teoría o un cuerpo de conocimientos.

[42] *Vid.* L. Vega, *La trama de la demostración...*, op. cit., 1990, II, § 5.3; en II, § 5.4 y IV § 3.3, se consideran las relaciones entre esta idea aristotélica de demostración y la práctica euclídea de la prueba matemática.

En este sentido, bien podemos ver en Euclides una primicia o un anuncio paradigmático de la demostración axiomática clásica. Claro está que su tratado no responde con igual fidelidad en todos sus extremos a este planteamiento, de modo que no podremos atribuir un talante axiomatiforme a su conjunto como si representara un sistema compacto y homogéneo de elementos matemáticos. Pero cabe reconocer este sentido a algunas de sus partes, en particular a sus bases geométricas. Por lo demás, también es sintomático que la «escuela alejandrina» de Euclides se dedique a consagrar y perfeccionar la geometría de los *Elementos* en esa misma línea, en la línea de una disciplina autónoma, prioritaria y compuesta por una teorías deductivas informalmente autosuficientes.

En conclusión, es aventurado hablar de unas deudas directas y concretas de Euclides con los *Analíticos*. Cierto es que la noción técnica de *apódeixis* tiene un cuño aristotélico; o que la doble dimensión metodológica y disciplinaria de la exposición racional de una materia científica también está prevista en la teoría aristotélica de la ciencia demostrativa. Pero no es menos cierto que el rigor demostrativo de las pruebas geométricas era, asimismo, una demanda del propio desarrollo matemático a principios del siglo IV (según testifica Platón a propósito de Teodoro y Teeteto); o que esa doble motivación ya inspiraba la confección de los tratados de *Elementos* matemáticos (como sugiere Proclo). Y, en fin, si Aristóteles, fundado en estos precedentes y en otros motivos y desarrollos complementarios (filosóficos, dialécticos) de la cultura ateniense de la 1.ª mitad del siglo IV, supo ofrecer un programa de lo que llamaríamos «el punto de vista axiomático», Euclides, a su vez, bien pudo aprovecharse de todas esas fuentes de inspiración, incluidos ciertos puntos del programa aristotélico (con resonancia en Alejandría), para dar un paso más en esta dirección hasta lograr de hecho la organización deductiva autónoma —o, si se quiere, una «axiomatización» *sui generis*— de ciertos cuerpos de conocimiento matemático.

V. El texto de los «Elementos». Versiones y ediciones

El examen detallado de las versiones y ediciones de los *Elementos* es una tarea que no entra en mis propósitos y, desde luego, está más allá de mis limitaciones. John Murdoch ha llegado a decir que ninguna otra obra científica, filosófica o literaria, en su largo caminar desde la Antigüedad hasta el presente, ha caído bajo la pluma de un editor con tanta frecuencia [43]. Por lo menos, cabe asegurar que ninguna otra fuente clásica de conocimiento ha tenido una historia tan rica y variopinta de versiones, ediciones, epítomes y comentarios. Recordemos además que los *Elemen-*

[43] *Vid.* J. Murdoch, «Euclid: Transmission of the *Elements*», en C. Gillispie, dir, *Dictionary of Scientific Biography*, ed. cit., 4, pág. 437. Cf. la voluminosa compilación bibliográfica sobre Euclides de P. Riccardi, *Saggio di una bibliografía euclidea* (Mem. R. Acad. de l'Instituto di Bologna; en 5 partes), Bolonia, 1887-1893. También cabe hacerse una idea a través de C. Thomas-Sanford, *Early Editions of Euclid's Elements*, Londres, 1926 (84 títulos desde 1842 hasta 1600); M. Steck, «Die geistige Tradition der frühen Euklid-Ausgaben», *Forschungen und Fortschritte* 31 (1957), 113-117 (60 títulos más hasta 1600); F. J. Duarte, *Bibliografía: Euclides. Arquímedes. Newton*, Caracas, 1967 (recoge sólo ediciones impresas hasta 1961: 123 corresponden a Euclides y cubren las págs. 19-101 de las 163 que hacen el total del libro). En fin, otros títulos pueden verse en G. Kayas, *Vingt-trois siècles de tradition euclidienne. (Essai bibliographique)*, École Polyt. L.P.N.H.E. X-77, París, 1977, y en M. Steck, *Bibliographia Euclideana*, Hildesheim, 1981.

tos, con todas sus virtudes, no fueron recibidos como un texto sagrado, sino como un patrimonio perteneciente a una tradición viva, dispuesta a mejorar y añadir lo que al parecer hiciera falta. Las interpolaciones empiezan a circular como moneda corriente desde el siglo III d. C. A juzgar por sus referencias, Sexto Empírico a finales del siglo II aún disponía de una versión correcta, mientras que Jámblico (hacia 250-325) tiene que vérselas con una versión corrupta. Sin embargo, durante algún tiempo debieron de seguir existiendo versiones relativamente fieles, pues nuestros mejores manuscritos, como ha observado Heiberg, se ven libres de algunas interpolaciones presentes en el texto utilizado por Jámblico o en versiones empleadas por otros comentadores tardíos. Sea como fuere, en el siglo IV se acentúan las modificaciones inspiradas en motivos de orden didáctico: por ejemplo, hay una tendencia a normalizar el lenguaje matemático del tratado, a insertar escolios, a suplementar o explicar diversos pasajes; en suma, se empiezan a desarrollar los aspectos disciplinarios y escolares que harán de él un manual básico de geometría. Así quedó sellada su suerte histórica.

La forma principal de transmisión de los *Elementos* se aviene a la pauta que parecen seguir los textos científicos más afortunados del legado griego: versiones griegas manuscritas en unciales o en minúsculas; traducciones del griego al árabe en el siglo IX; traducciones latinas de las versiones árabes en el XII; impresión de versiones y exposiciones en latín a finales del siglo XV; poco después, edición de versiones latinas a partir del griego y aparición de la *editio princeps* del propio texto griego; más tarde, versiones en lenguas vernáculas y, por fin, edición crítica del texto en las últimas décadas del siglo XIX. Pero aun dentro de este marco general, la suerte de los *Elementos* merece

una consideración más concreta. Esta consideración, aquí, no pasará de ser una visión panorámica y superficial [44]. Por motivos prácticos, seguiré la convención de distinguir cuatro fases características de la historia del texto: el Euclides griego, el árabe, el latino y el vertido en las lenguas vernáculas modernas.

1. *El Euclides griego*

La normalización primera y básica del texto de los *Elementos* es la edición de Teón de Alejandría, hacia el año 370 d. C. Nuestro Teón no es precisamente un gramático alejandrino, sino un matemático: está mucho menos interesado en la recuperación o depuración del texto que en la significación disciplinaria del tratado. De la edición de Teón procede la tradición central de las versiones, exposiciones y comentarios de Euclides. Los manuscritos y copias de esta tradición suelen hacer referencia expresa a su procedencia de Teón [*ek tês Théonos ekdóseos, apò synoustiôn toû Théonos*] y tienen algunas adiciones características; por ejemplo, la segunda parte de VI, 32 (añadido confesado por el propio Teón en su comentario al *Almagesto* de Tolemeo, I, pág. 201), VII, 22 y probablemente VII, 20, amén de ciertos lemas y de varios porismas. Los ejemplares teoninos son, naturalmente, de muy diversa calidad.

---

[44] *Vid.* J. L. HEIBERG, «Prolegomena critica» en la ed. cit. de HEIBERG-MENGE, *Euclidis opera omnia* V, 1888, págs. xxiii-cxiii (en la edic. STAMATIS (1977), V, págs. xvi-lxxxix); HEIBERG, «Paralipomena zu Euklid», *Hermes* 38 (1903), 46-74, 161-201, 321-356; T. H. HEATH, ed. cit, 1926², 1, págs. 46-63, 75-113; J. MURDOCH, «Euclid: Transmission...», loc. cit., 1981, págs. 437-459; D. H. FOWLER, *The Mathematics in Plato's Academy*, op. cit., 1987, págs. 202-20.

En dos que no podrían contarse entre los mejores, los mss. Veneciano 301 y Parisino Gr. 2343, —y en una cuidada versión italiana de Zamberti—, se funda precisamente la *editio princeps* de los *Elementos* de Simon Grynaeus (Basilea, 1633; aunque, según Stamatis, ésta sería la fecha de la 2.ª impresión y la 1.ª correspondería al año 1630). Sobre la edición de Grynaeus, complementada con una influyente versión latina de Commandino y con el uso de otros mss. en ciertos pasajes concretos, D. Gregory realiza la primera edición completa de las obras de Euclides en griego y en latín (Oxford, 1703). En 1808, François Peyrad advierte que un ms. que Napoleón se había traído de Roma a París como parte del botín, el Vaticano Gr. 190, no contenía las referencias y adiciones teoninas además de presentar otras diferencias notables con las muestras de la tradición (e. g., la presencia de XIII, 6). Peyrad empezó a tener en cuenta este ms. para corregir las ediciones teoninas en su edición trilingüe (griego, latín, francés) de los *Elementos* (París, 1814-1818); pero se mantuvo sustancialmente dentro de la línea marcada por las ediciones de Grynaeus y Gregory. Sólo con la edición de E. F. August (Berlín, 1826-1829), este ms. independiente se convierte en una baza principal para vindicar y establecer el «texto original» —suponía August— de Euclides.

La recuperación del Euclides «prístino» es, desde luego, una ilusión excesiva. El texto de los *Elementos* que hoy cabe considerar como el más aproximado al original es el establecido por la edición crítica de Heiberg (en la ed. cit. de J. L. Heiberg y H. Menge, *Euclidis opera omnia*, I-IV, Leipzig, 1883-1886). La edición de Heiberg descansa en el cotejo de 7 manuscritos principales y en el uso adicional de un palimpsesto del British Museum (Add. 17211) que contiene varios fragmentos del libro X y alguno del libro XIII. También toma en cuenta las noticias existentes en comentarios y referencias de Pappo, Sexto Empírico, Proclo, Simplicio, así como abundantes escolios (recogidos en el vol. V, Leipzig, 1888). Los mss. son los siguientes:

Dentro de la tradición teonina:

F: ms. Florentino XXVIII, 3, Biblioteca Laurentina; procede originariamente del siglo x, aunque ha sido revisado por una mano posterior (del siglo xvi).
B: ms. Bodleiano, D'Orville X; procedente del siglo ix.
V: ms. Vienés, Philos. Gr. 103; procede probablemente del siglo xii.
b: ms. Boloñés, 18-19, Biblioteca Comunale; procedente del siglo xi.
p: ms. Parisino, Bibliothèque Nationale 2466; procedente del siglo xii.
q: ms. Parisino, BN 2344; procedente del siglo xii.

Independiente de la edición de Teón:

P: ms. Vaticano, Gr. 190; procedente del siglo x.

Heiberg abordó la colación de este material con todos los medios puestos a su disposición por la crítica textual del siglo xix, sin ahorrar la discusión de diversos criterios en orden a la complementación, coordinación y conciliación de estos mss. En general, suele considerar mejores o más fiables P y F, pero esta confianza nunca es completa, menos aún absoluta. Por otra parte, suele dar prioridad a estas fuentes sobre otras correspondientes a las versiones y los comentarios de la tradición árabe [45].

[45] La preeminencia de la tradición árabe fue vindicada por M. Klamroth, «Ueber den arabischen Euklid», *Zeitschrift der Deutschen morgenländischen Gesellschaft* 35 (1881), pág. 310, en particular. Cf. J. L. Hei-

Heiberg examinó posteriormente («Paralipomena zu Euklid», art. cit., 1903), otro material; en especial, los comentarios de Herón, transmitidos por Al-Nayrīzī, y algunos fragmentos de antiguos papiros (Herculanensis, 1061; Oxyrhinchus I, pap. 58, xxix; Fayûm, pap. 96, ix). Este examen contribuyó a precisar los rasgos de la tradición central de los mss. derivados de la edición de Teón: según todos los visos, responden a un arquetipo diferente del representado por P y, quizás, también diferente del texto utilizado por algunos comentadores próximos a Euclides, como Herón; cabe sospechar, así mismo, que el arquetipo de P es anterior al teonino. F es el ms. teonino que más se acerca a P; por esta circunstancia, amén de otros detalles, como el cuidado puesto en su confección, F podría considerarse el ms. más importante dentro de esa tradición. Pero no estamos en condiciones de fijar un orden jerárquico, una red de relaciones o un árbol genealógico de los mss. teoninos.

La edición de Heiberg ha sido revisada finalmente por E. S. Stamatis, *Euclidis Elementa* (Leipzig, Teubneriana, 1969-1973, vols. I-IV; vol. V. *Prolegomena critica,* Libros XIV, XV *Scholia,* Leipzig, 1977). Del examen de Stamatis, según confesión propia, no resulta sino la confirmación del texto establecido por Heiberg *(vid.* su «Additamentum Praefationis», I, págs. x-xi). Las mejoras de la edición revisada consisten sustancialmente en la inclusión de las referencias griegas a Euclides («Testimonia», I, págs. xii-xxx) y en la adición de un apéndice II, dentro del mismo vol. I, que recoge fragmentos de papiros (Michigan iii 143, Fayûm 9, Oxyrh. 29). La mayor diferencia apreciable es el carácter más conservador de la edición de Stamatis. Por ejemplo, Stamatis admite sin reservas la noción común: «Y dos rectas no encierran un espacio», que Heiberg considera una interpolación; Proclo la había descartado expresamente alegando que se debía poner coto al desmedido afán de los comentadores por multiplicar los axiomas sin necesidad *(Com.* 184, 8; 196, 23).

BERG, «Die arabische Tradition der *Elemente* Euklid's», *Zeitschrift für Mathematik und Physik* 29 (1884), 1-22; HEIBERG, «Prolegomena critica», loc. cit., 1888, págs. xcvi-xcvii (en la edic. de STAMATIS (1977), págs. lxxiv-lxxvi); T. L. HEATH, ed. cit., 1926², I, págs. 81-82.

Según esto, nuestra base documental acredita la existencia de dos arquetipos distintos: el correspondiente a la edición de Teón y el correspondiente al ms. P —del que también hay huellas, aunque ocasionalmente, en algún otro ms., como el ms. V—. A estos dos arquetipos quizás cabría sumar otros tres, cuya existencia sólo cabe conjeturar por algunos indicios. Uno podría desprenderse del comentario de Herón recogido en las glosas de Al-Nayrīzī. Otro podría inferirse de ciertos fragmentos latinos antiguos, como los atribuidos a Censorino o los transcritos en el palimpsesto de Verona (Biblioteca Capitolare), y de los llamados *excerpta boetiana* conservados en manuales de los agrimensores romanos y en dos «Geometrías» atribuidas a Boecio. El tercero, en fin, correspondería a la tradición de las versiones árabes, habida cuenta de algunas divergencias con los mss. citados (incluido el ms. b, que desde XI, 36, hasta el final del libro XII se distancia significativamente de sus congéneres para aproximarse a fuentes árabes).

2. *El Euclides árabe*

La fase árabe de la transmisión de los *Elementos* parece haber sido no sólo variopinta sino, desde el punto de

vista matemático incluso, sumamente fecunda y creadora. Los *Elementos* no dejan de sugerir excursos filosóficos a autores con escaso interés por la matemática misma (e. g., Avicena, Al-Razi, Averroes). Pero, sobre todo, constituyen un foco de estudios especializados y dan lugar a una rica literatura euclidiana cuyos géneros son: traducciones propiamente dichas [*naql*], epítomes o compendios [*ikhtisār, mukhtasar*], recensiones [*tahrīr*], enmiendas [*islah*], comentarios [*tafsīr, shurūh*]. Una recensión capital fue, por ejemplo, la de Al-Tūsī —en 1248 había completado una *Tahrīr Usūl Uqlīdis* sobre los 15 libros—. Mayor importancia tuvo aún el comentario de Al-Nayrīzī (muerto hacia 922); difundido en la versión latina de Gerardo de Cremona, no sólo fue una preciosa fuente de información sobre los comentarios helénicos, sino que propició ciertos sesgos típicos de la *commentatio* medieval de Euclides (e. g., la tendencia a distinguir y probar por separado los casos particulares de una proposición general, la tendencia al uso de ilustraciones numéricas). También son memorables los comentarios de Al-Jayyānī (cuya comprensión de la definición V, 5 de la proporcionalidad carece de parangón en el Occidente cristiano hasta Isaac Barrow, siglo XVII), Al-Kindī (muerto hacia 873), Al-Fārābī (870-950), Alhazen (965-1039), Omar Jayyām (muerto hacia 1123). Destacan dos núcleos temáticos en torno a los cuales suelen girar los comentarios: por un lado, la fachada «axiomática» del libro I, donde recibe especial atención la prueba del controvertido postulado (v); por otro lado, la teoría de la proporción del libro V y las magnitudes irracionales del libro X.

Pero importan más las traducciones propiamente dichas. Hay noticias de una versión inicial de Al-Hajjāj (que vive hacia 786-833) —quizás la primera traducción árabe—, seguida de otra del mismo autor en la que se puede ver el precedente de los compendios que ofrecen los *Elementos* como un manual matemático claro, convincente y bien dispuesto para el uso escolar. La versión inicial de Al-Hajjāj fue denominada «Hārūni» en honor del califa Hārūn al-Rashīd (cuyo nombre no es menos honrado por los *Cuentos de las mil y una noche),* y se ha perdido; de la segunda versión sólo se conserva un ms. incompleto (Leiden, Or. 399, 1) que comprende los libros I-VI y algunas definiciones del VII, pero no faltan referencias a ella en otros mss. (e. g. en el ms. árabe 907 de la Real Biblioteca de El Escorial). Ishāq b. Hunain (muerto en 910) hizo otra traducción, hoy también desaparecida, revisada luego por Thābit b. Qurra (826-901). Esta versión Ishāq-Thābit se conserva en unos 20 mss. y se considera más fiel a una fuente griega. Por último, al traductor del comentario de Pappo al libro X, Abū Uzman al-Dimashqī (siglo x), se le atribuye otra versión parcial de los *Elementos;* hoy no queda de ella más noticia que esta referencia del *Fihrist* o catálogo bio-bibliográfico de Al-Nadīm [46].

[46] El *Fihrist* («Índice»), compuesto hacia 987, es una preciosa fuente de información sobre la cultura árabe coetánea; *vid.* H. SUTER, *Abhandlugen zur Geschichte der mathematischen Wissenschaften* VI, Leipzig, 1892. La versión AL-HAJJĀJ ha sido editada junto con los manuscritos de AL-NAYRĪZĪ por J. L. HEIBERG, R. O. BESTHORN *et. al., Codex Leidensis 399, 1: Euclidis Elementa ex interpretatione al-Hadschdschadschii cum commentariis al-Nairīzī,* Copenhague, 1893-1932. En 1937, C. BAUDOUX anunciaba una edición de la versión ISHĀQ-THĀBIT que no llegó a aparecer. Hay, sin embargo, una edición rusa (a cargo de J. AL-DABBAG, L. M. KARPOVA, B. A. ROSENFELD *et al.)* de las obras matemáticas de THABIT IBN QURRA, *Nauchnoe nasledstvo,* Moscú, 1984. Sobre la transmisión árabe, pueden verse los arts. antes citados de KLAMROTH (1881) y HEIBERG (1884); pero más accesibles y actualizados son los informes de G. DE YOUNG, «The Arabic textual traditions of Euclid's *Elements*», *Histo-*

La versión de Al-Hajjāj de los XV libros consta de 468 proposiciones (I-XIII: 452; una ausencia característica es, por ejemplo, la de I, 45) y presenta una marcada orientación didáctica (pruebas de casos particulares, ejemplos numéricos). La versión Ishāq-Thābit comprende a su vez 478 proposiciones (I-XIII: 462), dispuestas en ocasiones en un orden de prueba diferente al de al-Hajjāj; también acusa cierta normalización escolar del lenguaje y del uso de figuras. A pesar de sus diferencias, hay indicios de alguna relación entre ellas: puede que la revisión de Thābit tuviera en cuenta además las versiones de Al-Hajjāj. Se aproximan en las definiciones y en los enunciados de los teoremas, así como en los libros XI-XIII en general; por otro lado, abundan las contaminaciones en las muestras existentes de una y otra tradición, e incluso cabría pensar que ciertos mss. (Leiden, 399, 1; Leningrado, C 2145; El Escorial, 907) representan una especie de versión mixta. Con todo, la suerte histórica de una y otra línea de transmisión hacia el Occidente latino fue un tanto dispar. La tradición Ishāq-Thābit sólo tuvo al parecer una incidencia marginal, a través principalmente de la traducción latina de Gerardo de Cremona; en cambio, la versión de Al-Hajjāj alcanzó a tener mayor repercusión merced al poderoso influjo del comentario de Al-Nayrīzī, a la versión latina de Hermann de Carinthia y, sobre todo, gracias a sus afinidades con

ria Mathematica 11 (1984), 147-160; P. KUNITZ, «Findings in some texts of Euclid's *Elements*», en M. FOLKERTS, U. LINDGREN, eds., *Mathemata. Festschrift für Helmuth Gericke*, Stuttgart, 1985, págs. 115-128; R. LORCH, «Some remarks on the Arabic-Latin Euclid», en C. BURNETT, ed., *Adelard of Bath*, Londres, 1987, págs. 45-54. Para hacerse una idea del marco general de estas contribuciones, *vid.* A. P. YOUSCHKEVICH, (1961): *Les mathématiques arabes (VIII^e-XV^e siècles)*, Paris, 1976.

la tradición adelardiana, el cauce central por el que discurrieron las versiones medievales en latín [47].

3. *El Euclides latino*

La suerte de los *Elementos* en latín conoce tres etapas principales: una corresponde a la cultura romana; otra, a la cultura medieval escolástica; la tercera, en fin, a la cultura renacentista y académica posterior. Comprende desde versiones y traducciones hasta adaptaciones y compendios, pasando por diversas formas de exposición y comentario.

3.1. *Tu regere imperio populos, Romane, memento / Hae tibi erunt artes...* La verdad es que tan poco hicieron los romanos por las matemáticas, que suele considerarse digno de memoria el hecho de que Cicerón se permitiera una alusión a Euclides *(De oratore* III 132), mucho antes de que a alguien se le ocurriera traducir su tratado. También es probable que nadie llegara a poner mucho empeño en la tarea, pues hoy sólo nos quedan versiones muy fragmentarias (por lo demás, los romanos ilustrados sabían griego o podían contar con un lector de griego). Según Casio-

[47] Las afinidades de *Adelardo I* con la tradición Al-Hajjāj aparecen resaltadas en M. CLAGETT, «The Medieval Latin translations from the Arabic of the *Elements* of Euclid, with special emphasis on the versions of Adelard of Bath», *Isis* 44 (1953), pág. 19, y en H. L. BUSARD, ed., *The First Latin Translation of Euclid's Elements Commonly Adscribed to Adelard of Bath*, Toronto-Ontario, 1983, págs. 5-6. Las afinidades de *Adelardo I* con la tradición Isāhq-Thābit son destacadas, en cambio, por P. KUNITZ, «Findings in some texts...», art. cit., 1985, págs. 118-119. Puede que esta duplicidad sea un motivo más para pensar en una pronta compenetración o contaminación entre ambas líneas de transmisión de los *Elementos*.

doro (*Institutiones* II 6, 3), Boecio llevó a cabo una traducción de Euclides —el propio Boecio lo había declarado en carta a Teodorico (*Variae* I 45, 4)—. Sin embargo, los restos actuales de la versión de Boecio no pasan de ser extractos procedentes de otras fuentes: una recensión en ese libro II de las *Institutiones* de Casiodoro; una colección de materiales de geometría aplicada a la agrimensura; algunos fragmentos de dos *Geometrías* atribuidas a Boecio (una compuesta en dos libros y la otra en cinco). Comprenden en suma las definiciones, postulados y axiomas de los libros I-V; los enunciados de casi todas las proposiciones de I-IV; las pruebas de I, 1-3. A estos *excerpta* boecianos cabe añadir un fragmento con las definiciones, postulados y axiomas, atribuido al astrólogo Censorino y recogido en el *De nuptiis Philologiae et Mercurii* de Marciano Capella (VI 708 ss.), así como algunos fragmentos de los libros XII-XIII hallados en un palimpsesto de la Biblioteca Capitolare de Verona [48].

3.2. En el latín medieval, la situación cambia espectacularmente gracias a la atención que Euclides va ganando en el *Quadrivium* y en algunas facultades de Artes pero, sobre todo, gracias a la providencial mediación de la cultura árabe. En los saberes griegos, Alá fue durante algunos siglos el más sabio.

A. Hay, no obstante, alguna versión directa del griego. Contamos, por un lado, con dos minúsculos fragmen-

[48] *Vid.* los «excerpta boetiana» recogidos en N. BUBNOV, ed. (1899), *Gerberti opera mathematica*, Hildesheim, 1963; App. II, págs. 161-166 (otros falsamente atribuidos en págs. 166-174; referencias de los siglos VI-XV a opúculos geométricos bajo el nombre de Boecio, App. III, págs. 180-196). *Vid.* también M. FOLKERTS, ed., *«Boethius» Geometrie II: Ein mathematisches Lehrbuch des Mittelalters*, Wiesbaden, 1970; M. GEYMONAT, ed., *Euclidis latine facti fragmenta Veronensia*, Milán, 1964.

tos (I, 37-38; II, 8-9) en un ms. de Munich (Univ. 2. 757) procedente quizás del siglo x y atribuido erróneamente a Boecio. Pero, por fortuna, disponemos además de una traducción anónima de los *Elementos* I-XV, compuesta en torno a 1170 a partir de un ms. teonino que hoy no está identificado: si por un lado haría recordar B, por otro lado, acusaría contaminaciones no teoninas (a este tipo de recensiones mixtas pertenece por ejemplo, según Busard, el ms. S, iii - 5, de El Escorial). La traducción tiene en ocasiones una sintaxis más griega que latina y sigue la norma de versión literal *de verbo ad verbum* (palabra por palabra) propuesta por Burgundio de Pisa a mediados del siglo XII. Cabe suponer que su autor, o al menos uno de sus autores, es la misma persona que el traductor anónimo siciliano del *Almagesto* de Tolemeo; al menos da pruebas de cierta competencia matemática y agudeza lógica (e. g., en el uso de la llamada *consequentia mirabilis* —si de la presunta negación de una proposición se sigue esta proposición misma, tal proposición queda sin más establecida—, para la demostración de IX, 12) [49].

B. El conocimiento de Euclides en el medioevo escolástico descansa en las traducciones latinas hechas durante el siglo XII a partir de versiones árabes. El cauce central de difusión de los *Elementos* fue la tradición adelardiana. Para empezar, se prodigó en un número de copias manuscritas relativamente sorprendente (los mss. conocidos hoy

[49] *Vid.* J. MURDOCH, «Euclides Graeco-Latinuis. A hitherto unknown Medieval Latin translation of the *Elements* made directly from the Greek», *Harvard Studies in classical Philology* 71 (1966), 249-302. Cf. la edición de H. L. BUSARD, *The Mediaeval Latin Translation of Euclid's Elements Made Directly from the Greek*, Stuttgart, 1987; Busard defiende las hipótesis de un ms. griego contaminado y de dos traductores distintos, *vid.* su «Introduction», págs. 2-4 y 9-10, respectivamente.

pasan con creces del medío centenar); por añadidura, constituyó la base de los comentarios de los siglos XIII y XIV que dieron la medida del conocimiento de Euclides en Occidente hasta el siglo XVI (e. g., el comentario atribuido a Alberto Magno y, sobre todo, el influyente comentario de Campano de Novara) [50].

Desde el trabajo clásico de Claggett (1953), las versiones pertenecientes a la tradición adelardiana se han clasificado en tres grupos: *Adelardo I, Adelardo II, Adelardo III*, que aglutinan las principales clases de variantes.

Adelardo I es, según todos los visos, la primera versión latina de los *Elementos* (incluidos los libros XIV y XV) [51]. Su atribución al conocido traductor medieval Adelardo de Bath (cuya madurez se sitúa hacia 1142-1146) no deja de

[50] *Vid.* el trabajo clásico de M. CLAGETT, «The Medieval Latin translations...», art. cit. (1953), págs. 16-44 (recogido en sus *Studies in Medieval Physics and Mathematics*, Londres, 1979, viii); así como J. MURDOCH, «Euclides Graeco-Latinus...», art. cit., 1966, y «The Medieval Euclid: Salient aspects of the translations of the *Elements* by Adelard of Bath and Campanus of Novara», *Revue de Synthèse* 89 (1968), 67-94. Otras indicaciones se encuentran en P. KUNITZ, «Findings in some texts...», art. cit., 1985; H. L. BUSARD, «Some early adaptations of Euclid's *Elements* and the use of its Latin translations», en M. FOLKERTS, U. LINDGREN, eds., *Mathemata...*, op. cit., 1985, págs. 129-164; R. LORCH, «Some remarks on the Arabic-Latin Euclid», art. cit., 1987; M. FOLKERTS, «Adelard's versions of Euclid's *Elements*», en C. BURNETT, ed., *Adelard of Bath*, op. cit., 1987, págs. 55-68. El art. antes citado de Murdoch, «The Medieval Euclid...», 1968, contiene importantes observaciones sobre el comentario de Campano; por lo que se refiere al atribuido a Alberto Magno, *vid.* P. M. TUMMERS, «The Commentary of Albert on Euclid's *Elements of Geometry*», en J. A. WEISHEIPL, ed., *Albertus Magnus and the Sciences. Commemorative Essays 1980*, Toronto, 1980, págs. 479-499.

[51] *Vid.* H. L. BUSARD, ed., *The First Latin Translation of Euclid's Elements Commoly Adscribed to Adelard of Bath*, Toronto-Ontario, 1983.

ser dudosa. Pero, en cualquier caso, parece haber sido compuesta en la 1.ª mitad del siglo XII, en torno a 1130-1140, y es una versión anterior a las caracterizadas como *Adelardo II* y *Adelardo III*. También se discute el peso relativo de las fuentes árabes más notorias, la tradición Al-Hajjāj y la tradición Ishāq-Thābit, en su composición; desde luego, comparte la omisión de I, 45, característica de la versión Al-Hajjāj y de la recensión de Al-Tūsī; sea como fuere, saltan a la vista los frecuentes arabismos (e. g.: «elmuain»/rombo, «elmukaab»/cubo, «alkora»/esfera, «elaalem»/gnomon) y las vacilaciones de un latín en proceso de adaptación a las necesidades de una terminología geométrica (e. g. líneas «inconiunctive» o «equistantes» por «equidistantes»; proporción «repetitione tripla», «ter repetita», «tri repetita» por «triplicata») [52]. Otro rasgo interesante es el que presenta uno de los mss. de este grupo (Bibliothèque Nationale, París, Lat. 16201): anuncia la distribución medieval común entre las definiciones, postulados, nociones comunes y enunciados por un lado, como originales del propio Euclides, y por otro lado las pruebas de las proposiciones, como un *commentum* adicional —que podía ser atribuido a Teón o a Al-Nayrīzī o, incluso, reconocido más tarde como una contribución de Campano montada sobre una versión de tipo *Adelardo II*—. Cabe mencionar, en fin, otro rasgo de la versión *Adelardo I* que resultará típico de la actitud medieval ante los *Elementos:* consiste en marcar expresamente las partes constituyentes de la prueba. Las fórmulas distintivas de *Adelardo I* vie-

[52] En la ed. cit. de H. L. BUSARD, *The First Latin Translation...*, Addendum I, 1983, págs. 391-396, puede verse una tabla comparativa de los arabismos presentes en *Adelardo I* y *II*, y en la versión de Hermann de Carinthia; en el Add. II, págs. 396-399, ciertas peculiaridades de su terminología latina.

nen a ser: *nunc demonstrandum est* (a veces *nunc signandum est*), para señalár la proposición en cuestión; *exempli gratia*, para hacer referencia al caso particular considerado; *rationis causa* para la demostración propiamente dicha; *hoc est quod demonstrare intendimus*, como cláusula final.

La versión *Adelardo II* no es una traducción como *Adelardo I*, sino más bien una versión abreviada, un compendio, que puede atribuirse casi con seguridad al propio Adelardo de Bath: la habría compuesto antes del período 1142-1146 en el que al parecer redacta su obra sobre el astrolabio, pues allí se refiere a esta versión de los *Elementos*. Este sumario o compendio de los *Elementos* pudo haberse elaborado bien a partir de *Adelardo I*, bien a partir de algún texto árabe (contiene arabismos ajenos a *Adelardo I*, e. g., «helmuaerife»/trapecio); en todo caso, muestra cierta familiaridad con los *excerpta* de Boecio e incorpora algunos grecismos (e. g., «orthogonius»/rectángulo, «ambligonius»/obtusángulo, «oxigonius»/acutángulo, «gnomon», «parallelogrammum», «ysosceles»). *Adelardo II* no recoge las demostraciones propiamente dichas de los *Elementos*, sino que se limita a dar indicaciones sumarias sobre el camino que debería seguir el que deseara hacerlas. Muestra así un comportamiento medieval característico ante el legado matemático: la asimilación de algunos resultados especialmente relevantes a través de glosas y digresiones donde las consideraciones analíticas y metodológicas sobre los supuestos y el curso de la prueba se conjugan con las directrices didácticas.

En *Adelardo III* no encontramos una traducción ni una versión en sentido propio, sino una paráfrasis o una edición —*editio specialis* la denomina Roger Bacon—. Se trata de una reelaboración que parte de los enunciados de

Adelardo II para desarrollar a su aire una demostración relativamente cumplida de las proposiciones correspondientes. Sin embargo, no renuncia a la expresión de intereses metodológicos (e. g., a la distinción de las partes de la prueba) ni a una perspectiva didáctica. Antes bien, incluye una introducción con material medieval sobre las artes liberales y la clasificación de las ciencias, además de comentar algunas definiciones. Sus fuentes parecen ser nociones procedentes de Proclo —halladas en los comentarios árabes—, ideas de los siglos XI-XII en torno a las artes liberales y otro material sobre unidades métricas —quizás del tratado *De disciplina geometriae* de Gerberto—.

La tradición adelardiana, en especial *Adelardo II*, y la reelaboración de Campano constituyeron la vía principal de incorporación de los *Elementos* a la cultura escolástica y determinaron su suerte en este marco histórico. Con ellos el texto de los *Elementos* adquirió un perfil peculiar. Hay algunas omisiones notorias, e. g., no sólo la de I, 45, sino la de VI, 12, donde se sienta la existencia de un cuarto término proporcional (Campano la incorpora como una *additio* a la proposición VI, 11, mientras que la tradición adelardiana en general la introduce como postulado en el libro I). Hay confusiones graves, como la suscitada en torno al concepto de proporcionalidad: la definición 5 del libro V se desplaza para dejar paso a una noción espuria de proporcionalidad continua [53]. Tampoco faltan, en fin, las ilustraciones numéricas y aun extramatemáticas de las proposiciones geométricas, sea con fines didácticos, sea con

[53] *Vid.* la ed. cit. de H. L. BUSARD, *The First Latin...*, 1983, pág. 145. Cf. MURDOCH, «The Medieval language of proportions: elements of the interaction with Greek foundations and the development of new Mathematical techniques», en A. C. CROMBIE, ed., *Scientific Change*, Londres, 1963, págs. 237-271.

otros propósitos especulativos o prácticos (e. g., musicales, filosóficos, astronómicos, métricos). Con todo, el rasgo más llamativo y determinante es la suplementación del contenido de los *Elementos* a través de glosas que, naturalmente, se hacen eco de los intereses y las preocupaciones del mundo escolástico medieval: entre esos intereses figuran ciertas formas de exposición didáctica y de análisis lógico; entre estas preocupaciones, los problemas generados por la concepción del infinito y del continuo [54].

C. Al margen de la tradición adelardiana, disponemos de otra versión relativamente libre de un texto árabe en la línea Al-Hajjāj. Se atribuye a Hermann de Carinthia

[54] Sobre la índole de los comentarios medievales, *vid.* J. MURDOCH, «The Medieval Euclid...», art. cit., 1968. Una muestra singular son las *Quaestiones* de ORESME, *vid.* H. L. BUSARD, ed., *Quaestiones super geometriam Euclidis (Janus*, Suppl. III), Leiden, 1961; cf. la recensión de J. MURDOCH en *Scripta Mathematica* 27 (1964), 67-91, así como V. P. ZUBOV, «Autour des *Quaestiones super geometriam Euclidis* de Nicole Oresme», *Medieval Renaissance Studies* 6 (1968), págs. 150-172. En E. GRANT, *A Source Book of Medieval Science*, Cambridge (Mass.), 1974, pueden verse interesantes extractos de Campano, § 27, págs. 136-150, y de Oresme, § 28, págs. 150-159. Los comentarios medievales del legado matemático griego tuvieron, a mi juicio, secuelas un tanto divergentes: por un lado, minaron la estructura sistemática de la geometría como un cuerpo deductivo de conocimiento al centrarse en el análisis dialéctico de ciertas pruebas como cuestiones particulares y derivar hacia problemas conceptuales filosóficos y científicos de diverso orden; por otro lado, su intención de explicitar los supuestos determinantes de la fuerza demostrativa de tales pruebas les llevó a veces a nuevas construcciones sistemáticas, por ejemplo a la declaración de los postulados de la aritmética —no considerados siquiera por los libros VII-IX de los *Elementos*—. *Vid.* la versión de E. GRANT de las definiciones, postulados y axiomas propuestos por la *Arithmetica de* JORDANO DE NEMORE a mediados del siglo XIII, en el recién citado *A Source Book in Medieval Science*, § 21, págs. 102-103 en particular.

(cuya madurez se sitúa hacia 1140-1150) [55]. Comparte muchos arabismos con las versiones adelardianas, a los que añade algunos otros de cosecha propia (e. g., «aelman geme»/nociones comunes, «almukadina»/axioma, «chateti»/líneas —en I, 1—); también emplea ciertos términos latinos peculiares (e. g., «umbo» por «gnomon», «mutus» por «irrationalis» o «surdus»). A la luz de algunas coincidencias con *Adelardo II* cabe conjeturar que Hermann se apoyara en esta versión, pero no deja de incluir pruebas completas. Por lo demás, también comienza con una introducción en la que expone los «siete fundamentos de toda la disciplina», a saber: *preceptum* (el enunciado de la proposición); *exemplum* (el caso considerado en particular); *alteratio* (la apertura de una prueba indirecta o de una reducción al absurdo); *collatio* (confrontación); *divisio* (disyunción de las cuestiones); *argumentum* (prueba); *finis* (conclusión en la que todos convienen). La fundamentación en la que está pensando Hermann es, por consiguiente, una fundamentación dialéctica.

D. Por último, disponemos de otra versión del siglo XII que no sólo difiere de la tradición adelardiana, sino de la versión de Hermann y de la matriz habitual Al-Hajjāj; es posible que haya sido compuesta sobre la base de un texto de la tradición Ishāq-Thābit que integra elementos de Al-Hajjāj (una versión mixta como la representada por el ms. esculiarense árabe 907). Se trata de la traducción de Gerardo de Cremona (1114-1187), el más prolífico de los traductores de Toledo, asistido en esta ocasión por un

[55] *Vid.* H. L. BUSARD, ed., *The Translation of the Elements of Euclid from the Arabic into Latin by Hermann of Carinthia (?)*, Leiden, 1968 (libros I-VI); Amsterdam, 1977 (libros VII-XIII).

ayudante mozárabe («Gallipus mixtarabus») [56]. En general, es considerada la versión más acorde con la tradición griega de entre las versiones arábigo-latinas; pero no tuvo ni de cerca el influjo de las versiones adelardianas. Está libre de los arabismos que cunden por las demás versiones coetáneas y, entre otros rasgos distintivos, adopta *Et due recte non comprehendunt superficie* como última noción común, no como postulado, según venía siendo costumbre en las versiones adelardianas y en Hermann. Sin embargo, también se complace en una introducción donde presenta las partes de la prueba; son éstas: *propositum* (lo propuesto); *exemplum; contrarium* (ejemplo que corresponde a una reducción al absurdo); *dispositio; differentia* (distinción entre lo posible y lo imposible); *probatio; conclusio.*

3.3. *Versiones latinas renacentistas y académicas.* — La suerte de los *Elementos* en el latín de los humanistas y en el latín universitario o académico de los siglos XV y posteriores se refleja en cuatro momentos relativamente señalados: cabría considerar los dos primeros como fases sucesivas de un proceso de depuración en la versión latina del tratado; los otros dos obedecen a motivos menos relacionados con el texto euclídeo y más pendientes de los desarrollos y usos disciplinarios de la geometría euclidiana.

a) El primer momento corresponde a las últimas décadas del siglo XV y las primeras del siglo XVI. Viene marcado por dos acontecimientos: uno es la publicación de la reelaboración de la versión adelardiana realizada por Campano a mediados del siglo XIII: la publica E. Ratdolt (*Preclarissimum opus elementorum Euclidis megarensis una*

[56] *Vid.* H. L. BUSARD, ed., *The Latin Translation of the Arabic Version of Euclid's Elements Commonly Adscribed to Gerard of Cremona*, Leiden, 1984.

cum commentis Campani perspicacissimi in arte geometrica, Venecia, 1492) no sin esfuerzo —*non sine maximo labore*— debido a las complicaciones técnicas que entonces suponía la impresión ajustada de los diagramas geométricos. Constituye la primera edición impresa de Euclides —unos cincuenta años antes de que aparezca la *editio princeps* del texto griego—. Esta circunstancia y la reconocida competencia matemática de Campano contribuyeron a ampliar el ámbito de influencia de esta versión de los *Elementos*, a pesar de que no se ignoraban sus raíces árabes. El segundo acontecimiento fue la aparición de una nueva traducción hecha directamente a partir del griego por B. Zamberti (*Euclidis megarensis philosophi platonici mathematicaum disciplinarum Janitoris... elementorum libros xiii cum expositione Theonis insignis mathematici...*, Venecia, 1505). Zamberti emplea algún ms. teonino no identificado (que, según Heiberg, no parece ser uno de los mejores) y descalifica la edición de Campano, *interpres barbarissimus;* de hecho, Campano no había trabajado sobre una fuente griega. Zamberti, por su parte, consigue una traducción propiamente dicha de los *Elementos*, libre además de los arabismos de la tradición adelardiana; pero insiste en el error de atribuir las pruebas, la *expositio* o *commentum* de las proposiciones, al bueno de Teón. Poco después, Luca Pacioli reedita a Campano (Venecia, 1509) y, corrigiendo ciertos errores que atribuye a los copistas, trata de vindicarlo como *interpres fidelissimus*. Las secuelas de esta querella fueron algunas ediciones de los *Elementos* que recogían tanto la versión de Campano como las pruebas de «Teón-Zamberti» (e. g., la edición de J. Lefèvre, París, 1516) [57].

[57] *Vid.* H. WEISENBORN, *Die Uebersetzungen des Euklid durch Campano und Zamberti*, Halle, 1882; T. L. HEATH, *The Thirteen Books...*, ed. cit., 1926², 1, págs. 97-100.

b) La traducción de F. Commandino (*Euclidis elementorum libri XV, una cum scholiis antiquis,* Pesaro, 1572) señala el paso a una fase superior de este proceso de depuración y rigor creciente en la versión latina de los *Elementos.* Commandino utiliza sabiamente el material a su disposición: mss. teoninos y escolios extraídos de un ms. del tipo de Vaticano Gr. 192, la *editio princeps* de Grynaeus, comentarios de diversas fuentes; por lo demás, no se priva de añadir notas y observaciones propias. Su traducción es la base de la mayoría de las versiones posteriores hasta que aparece la edición de Peyrard (1814-1818); en particular, es la adoptada como versión latina de los *Elementos* en la edición de Gregory (Oxford, 1703) y es la asumida por las influyentes versiones latina e inglesa de Simson (Glasgow, 1756). Por añadidura, la edición de Commandino incluye en sus «Prolegomena» una especie de manifiesto metodológico sobre la índole de la geometría y el rigor de la prueba geométrica, aparte de consideraciones históricas y filosóficas que hoy pueden parecernos curiosas, aunque por entonces no dejaban de ser relativamente tópicas (e. g., la referencia a una matemática anterior al Diluvio universal, recogida por la tradición bíblica de los caldeos, hebreos y egipcios, que habiendo sido trasplantada por Tales y Pitágoras alcanza en el medio heleno un nivel superior de desarrollo gracias a la contribución de generaciones sucesivas de matemáticos griegos [58]); sin embargo, también tiene el mérito de despejar la confusión tradicional entre

[58] Una memoria histórica similar es la que muestra Tycho Brahe en su *De disciplinis mathematicis,* compuesto en 1574. Sobre el significado de estas historias, *vid.* N. JARDINE, *The Birth of History and Philosophy of Science,* Cambridge-Nueva York, 1984, cap. 8, págs. 261 ss. en particular.

el autor de los *Elementos* y su homónimo filósofico de Mégara, así como la atribución de las pruebas a Teón.

c) Al margen de esos procesos de edición y versión, otro momento del Euclides latino académico está representado por las exposiciones ampliamente comentadas de los *Elementos,* en su calidad de tratado básico de la geometría tradicional. La edición de C. Clavio: *Euclidis Elementorum libri XV* (Roma, 1574, 1589[2]) constituye una muestra harto elocuente: a las 468 proposiciones euclídeas agrega otras 671 de cuño propio, amén de numerosos corolarios, escolios y excursos, hasta concluir «in universum ergo 1234 propositiones in nostro Euclide demonstrantur». Este derroche de erudición le valió a Clavio el renombre de «Euclides de nuestro siglo» entre sus contemporáneos; por lo demás, su sombra se alargará hasta el lejano Oriente, donde propicia una versión china de los libros I-VI (1603-1607) de su discípulo y consocio M. Ricci, con la colaboración de Xu Guangki [59]. A la misma Compañía que Clavio y Ricci pertenece G. Saccheri, cuyo *Euclides ab omni naevo vindicatus* (Milán, 1733) brilla con luz propia dentro del

[59] En torno a la significación y el marco histórico de la contribución de Clavio, puede verse F. A. HOMANN, «Christopher Clavius and the Renaissance of Euclidean Geometry», *Archivum Historicum Societatis Jesu* 52 (1983), 233-246. Por lo demás, esa versión china inicial fue continuada en 1857 por A. Wylie y Li Shanlan, pero ya no sobre la base de la edición de Clavio, sino a partir de versiones inglesas, en particular la de Simson. Las versiones y la recepción chinas de la geometría euclídea son un caso sumamente ilustrativo de las dificultades lingüísticas y conceptuales que puede envolver su trasplante a otra cultura muy dispar y no «formada» o «educada» en unos usos metodológicos semejantes a los griegos —aunque los chinos dispusieron desde antiguo de sus propias pruebas matemáticas, e. g., de la prueba intuitiva del «teorema de Pitágoras»—; *vid.* al respecto J. C. MARTZLOFF, *Histoire des mathématiques chinoises,* París, 1988, págs. 104 ss., en especial.

género de exposiciones-comentarios, sobre todo por su discusión del postulado de las paralelas [60]. En fin, a las labores de este tipo se suman ciertos motivos de orden filosófico para que ésta sea también la hora en que Euclides, figura paradigmática del difundido «esprit de la géométrie», parezca enseñar lógica además de geometría.

d) Los *Elementos* conocen una suerte bien distinta cuando y donde las aplicaciones prácticas o los motivos pedagógicos prevalecen sobre las pretensiones de rigor en la traducción o sobre las pretensiones de rigor en el cultivo de una disciplina demostrativa. Pues el latín académico deviene asimismo una vía de conversión y reducción de los *Elementos* al formato de un manual claro, intuitivo, básico. Aunque estos usos tienen una vieja raigambre —e. g., en los *curricula* tanto escolásticos de la Edad Media como humanistas del Renacimiento—, un ejemplar digno de mención por su índole y su descendencia es Andrè Tacquet: *Elementa geometriae planae et solidae* (Amberes, 1654). Comprende los libros I-VI, XI-XII; trae pruebas reducidas a su expresión más asequible; se recrea en corolarios y escolios útiles desde un punto de vista didáctico y práctico; alcanza considerable difusión en el siglo XVIII y todavía cuenta 20 reediciones en las primeras décadas del siglo XIX.

[60] El interés de los historiadores de la lógica y de las matemáticas por Saccheri no sólo se hizo patente en el primer tercio de este siglo, sino que ha renacido en las dos últimas décadas. Basten como síntomas los trabajos de A. DOU, «Logical and historical remarks on Saccheri's Geometry», *Notre Dame Journal of Formal Logic* XI - 4 (1970), 385-415; I. ANGELELLI, «On Sacheri's use of the «Consequentia mirabilis», en *Akten II. Intern. Leibniz-Kongresses*, Wiesbaden, 1975, págs. 19-26; G. C. GIACOBBE, «La *Lògica demonstrativa* di Gerolamo Saccheri», en *Atti del Convegno Intern. di Storia della Lògica*, Bolonia, 1983, págs. 265-270; y, en fin, la reciente reimpresión de G. B. HALSTED, ed., *Gerolamo Saccheri's Euclides Vindicatus (I)*, 1920, Nueva York, 1986.

De otras muchas versiones escolares que dan en expurgar pruebas cuando no en expoliar proposiciones de la «Geometría de Euclides», más vale no hablar.

4. *El Euclides de las lenguas vernáculas*

Sólo mencionaré unas pocas versiones y traducciones en algunas lenguas próximas.

La versión de N. Tartaglia da lugar a la primera edición italiana (Venecia, 1543); Tartaglia procede sobre la base de Campano y Zamberti, «secondo le due tradittioni». El propio Commandino trasladó al italiano (Urbino, 1575) su versión latina anterior. En italiano también se encuentran dos ediciones importantes en el presente siglo: las dos son traducciones del texto griego establecido en la edición de Heiberg. Una es la realizada por F. Enriques y colaboradores: *Gli Elementi d'Euclide e la critica antica e moderna* (Roma-Bolonia, 1925-1935); sus consideraciones hermenéuticas son sugerentes, pero acusan las limitaciones de la época y del horizonte explorado hasta entonces por los pioneros de la historiografía de la ciencia griega (entre los que puede contarse el propio Enriques). Otra traducción más reciente y ajustada es la de A. Frajese y L. Maccioni, *Euclide. Gli Elementi* (Turín, 1970), cuyas indicaciones y notas también tienen notable interés.

La primera edición francesa de los *Elementos* aparece con la edición de P. Forcadel (París, 1564): comprende los libros I-IX y su autor se mueve en la órbita del afamado dialéctico humanista Petrus Ramus (que en su *Scholae mathematicae*, Francfort, 1559, había declarado su amor por Euclides al tiempo que lo sometía a los chispazos de una crítica metodológica impertinente —hay amores imposibles—). Una versión muy influyente fue la adapta-

ción de ocho libros' por parte de C. F. Milliet de Chales (Lyon, 1672). En nuestros días, tiene relieve la edición de J. Itard de los libros VII-IX: *Les livres arithmétiques d'Euclide,* París, 1961, así como su reedición de la versión francesa de Peyrard (1819): *Les oeuvres d'Euclide. Les Éléments,* París, 1966. Itard añade los *Données* en las págs. 517-603, *i. e.* la traducción de los *Datos* de Euclides, a partir del texto de Heiberg-Menge y de un estudio crítico de M. Michaux *(Le commentaire de Marinus aux Data d'Euclide,* Lovaina, 1947). La introducción de Itard es interesante por glosar la labor crítica de Peyrard. También cabe mencionar, en atención sobre todo a su introducción y a un breve anexo sobre las geometrías no euclidianas, la edición bilingüe (texto griego de Heiberg-Stamatis y versión francesa libre) de G. J. Kayas, *Euclide. Les éléments,* París, 1978 [60bis].

Las primeras versiones alemanas conocidas son la de los libros VII-IX a cargo de J. Scheubel (Tubinga, 1558) y la de los libros I-VI a cargo de W. Holtzmann (Basilea, 1562) —esta segunda, en particular, parece obedecer a propósitos prácticos hasta el punto de ser una versión bastante libre y prescindir de las pruebas—. (Hay, sin embargo, noticias de una *Geometría Teutsch* anónima, aparecida en Nuremberg, 1484). La primera traducción literal se hará esperar hasta 1773: es la de los seis primeros libros realizada por J. F. Lorenz. Dos ediciones comentadas de especial importancia son las de E. S. Unger, *Die Geometrie des Euklid* (Basilea, 1833), y M. Simon, *Euklid und die sechs planimetrischen Bücher* (Berlín, 1901; para entonces la geometría euclidiana ya había alcanzado su mayoría de edad formal con D. Hilbert, *Grundlagen der Geometrie,* Leipzig, 1899). La traducción más empleada hoy sigue siendo la de C. Thaer, *Euklid. Die Elemente* (Leipzig, 1933-1937), Darmstadt, 1980[8]; está realizada y anotada sobre la edición de Heiberg.

[60bis] Durante la corrección de pruebas, ha llegado a mis manos la versión: EUCLIDE, *Les Éléments,* vol. I *(Livres I à IV),* París, 1990, con introducción de M. CAVEING y traducción y comentarios de B. VITRAC. Aquí sólo puedo dejar constancia de su aparición.

La primera y, en opinión de Heath, más importante traducción inglesa es la de H. Billingsley, *The Elements of Geometrie of the most ancient Philosopher Euclide of Megara,* Londres, 1570. (La hipótesis de un Euclides en inglés en la primera mitad del siglo x, sugerida por unos antiguos versos [61], está completamente descartada). La versión de Billingsley acusa el influjo de las «tradiciones» Campano-Zamberti. Pero, además, sus notas se benefician de Proclo y de otros comentadores griegos y, por añadidura, la edición cuenta con un interesante prefacio de J. Dee que merece atención por sí mismo [62]. Otra contribución de importancia fue la realizada por R. Simson al hacer dos versiones, una latina y otra inglesa, publicadas en el mismo año (Glasgow, 1756), sobre la base de la edición latina de Commandino; las dos versiones, no muy literales, vienen acompañadas de varias notas críticas y matemáticas. Simson aprovechó las frecuentes reediciones de su versión inglesa para revisar el texto con la ayuda de

[61] *Vid.* T. L. HEATH, ed. cit., 1926[2], pág. 95; cf. F. A. YELDHAM, «The alleged early English version of Euclid», *Isis* 9 (1927), 234-238.

[62] *Vid.* J. DEE, *The Mathematical Preface to the Elements of Euclid of Megara (1570),* edic. de A. G. DEBUS, Nueva York, 1975. Commandino destaca la idea de demostración y el rigor deductivo del *ars disserendi,* mientras que J. Dee se interesa por la utilidad práctica del cálculo y la calidad heurística del *ars inveniendi.* Sobre el significado histórico de esta divergencia, *vid.* E. I. RAMBALDI, «John Dee and Federico Commandino: an English and an Italian interpretation of Euclid during the Renaissance», *Rivista di Storia della Filosofía* XLIV-2 (1989), 211-247.

un *learned gentleman*. Gracias a su difusión tanto dentro del Reino Unido (alcanzaría la edic. 26.ª antes de cumplir su centenario), como fuera de él (aquí en España, sin ir más lejos), se convirtió en la versión más conocida del siglo XVIII. Podemos terminar recordando una vez más la edición monumental de T. S. Heath, *The Thirteen Books of Euclid's Elements* (Cambridge, 1909, 1926[2] edic. revisada), Nueva York, 1956. Su traducción, hecha sobre el texto establecido por Heiberg, todavía representa hoy la traducción estándar de los *Elementos*, la versión común de referencia entre los historiadores de las matemáticas y de la ciencia en general. También ha sido Heath la autoridad principal a la hora de fijar las acepciones técnicas de los términos matemáticos griegos —según se reconoce en el prefacio de 1925 de H. S. Jones y en el *postscript* editorial de 1940 del Liddel & Scott: *Greek-English Lexicon*—. Por si esto fuera poco, con su detallada introducción y sus copiosas notas y referencias a lo largo del texto, esta edición constituye aún la fuente más completa de información histórico-matemática sobre el tratado de que disponemos. Naturalmente, el notable renacimiento del interés por la matemática griega en esta segunda mitad del siglo, las discusiones en torno a ciertas cuestiones y relaciones oscuras de los *Elementos* con la matemática anterior, el mejor conocimiento de algunos momentos de la transmisión del texto, han hecho sentir a esta edición el peso de la edad. Heath, por poner algún ejemplo, no podía caer entonces en la cuenta de la importancia y la complejidad de ciertos aspectos de las tradiciones pre-euclídeas, como los desvelados por Szabó, Knorr o Fowler; no podía contar con una visión de la estructura deductiva de los *Elementos* como la que hoy proporcionan las reconstrucciones parciales de algunos libros o la reconstrucción general de Mueller; tampoco

podía llegar al conocimiento del Euclides latino medieval que nos han suministrado los estudios de Clagett y Murdoch, el trabajo editorial de Busard. Pero ninguno de estos achaques afecta a la traducción de Heath; ni ha vuelto ocioso, por cierto, el fruto de su erudición, su acopio de la mejor información disponible en el primer tercio del siglo.

VI. Los «Elementos» en España

A diferencia de lo que ocurre en otras lenguas civilizadas, en español no ha aparecido hasta el momento ninguna traducción completa de los *Elementos*. Y por contraste con lo que sucede en otras partes, tampoco hay al parecer un catálogo cabal o un análisis comprensivo de las versiones y paráfrasis de Euclides en este país nuestro de las maravillas, aunque la suerte de los *Elementos* sea un aspecto significativo de la suerte que corren los estudios clásicos y matemáticos en el mundo hispánico durante los siglos XVI y XVII especialmente [63].

No se conoce ninguna versión en español de los *Elementos* anterior a los años 1570; la primera traducción de que hay noticia data de 1574 y está impresa en 1576. Esta versión y las posteriores son obra de matemáticos; ni el cultivo del griego atravesaba por un momento floreciente, ni nuestros filólogos o nuestros helenistas suelen estar para ocuparse de estas cosas (de autores como Euclides, Arquí-

[63] Puede que el reciente M. E. Piñero y M.ª I. Vicente, «Primeras versiones castellanas (1540-1640) de las obras de Euclides: su finalidad y sus autores», *Asclepio* XLI-1 (1989), 203-231, sea un anuncio de mejores tiempos. Da noticias bio-bibliográficas de las versiones aquí numeradas 1-3, 5-7 (vid. *infra*) sin entrar en su contenido, pero con referencias a su marco histórico.

medes, Tolemeo...) [64]. Por otro lado, a medida que avanza el siglo XVIII, la traducción de los *Elementos* deviene una empresa carente de sentido en vista de la modernización —asentada por fin— de las matemáticas en España [65]. Pues bien, dentro de este marco de los siglos XVI-XVIII, intentaré una recopilación provisional de las versiones existentes de los *Elementos* y haré unas breves observaciones para destacar algún rasgo sintomático o característico. Por

[64] *Vid.* C. HERNANDO, *Helenismo e ilustración. (El griego en el s. XVIII español)*, Madrid, 1975; L. GIL, *Panorama social del humanismo español (1500-1800)*, Madrid, 1981. Son estudios sumamente ilustrativos por la copiosa información que proporcionan acerca de la postración del griego en la cultura humanista española de estos siglos. Pero también son sintomáticos por su propia selección del material digno de consideración: versiones y ediciones de las «letras» griegas, sin hacer referencia a la otra dimensión, los textos «científicos», del legado clásico. Esta omisión sesga a veces los datos aducidos, e. g.: según Gil, durante todo el siglo XVII, sólo se darían 10 traducciones directas del griego y 15 indirectas, a través del latín y del italiano *(op. cit.*, pág. 217, sobre la base de un cómputo de BEARDSLEY). De ser así, no entran en estas cuentas las versiones impresas de Euclides (sólo de los *Elementos* hay 5) y de otros clásicos «científicos» (como Hipócrates, Tolemeo, Galeno). Por lo demás, tampoco se recuerdan las versiones aparecidas en el siglo anterior. Cf. J. M.ª LÓPEZ PIÑERO y T. F. GLICK, *Bibliometría e inventario (Hispaniae Scientia* I). Valencia, 1979, donde se recogen las impresiones de textos científicos españoles hasta 1600.

[65] Cabe hacerse una idea de la cultura matemática española de los siglos XVI-XVIII y de las instituciones y factores que intervienen en la situación, a través de J. M.ª LÓPEZ PIÑERO, *Ciencia y técnica en la sociedad española de los siglos XVI y XVII*, Barcelona, 1979. Entre los estudios con un enfoque más monográfico, *vid.* J. REY PASTOR, *Los matemáticos españoles del siglo XVI*, Madrid, 1926; P. PEÑALVER, *Bosquejo de la matemática española en los siglos de la decadencia*, Sevilla, 1930; J. M.ª SÁNCHEZ PÉREZ, «La matemática», en el colectivo *Estudios sobre la ciencia española del siglo XVII*, Madrid, 1935, págs. 597-633; S. GARMA, «Cultura matemática en la España de los ss. XVIII y XIX», en el colectivo *Ciencia y sociedad en España*, Madrid, 1988, págs. 93-127.

lo demás, no entraré en la consideración de otra literatura euclidiana frecuente por entonces, como la compuesta por extractos, comentarios, adaptaciones y usos varios [66].

Las versiones en español que he podido conocer [67] han sido las siguientes:

[66] Me refiero a obras del tipo de P. J. MONZÓ, *Elementa Arithmeticae et Geometriae ad disciplinas omnes, Aristotelis praesertim Dialecticam ac Philosophiam... ex Euclide decerpta*, Valencia, 1559; J. SEGURA, *Mathematicae quaedam selectae propositiones ex Euclidis, Boëtij, & antiquorum aliorum libris decerptae...*, Alcalá, 1566. O comentarios como los de J. MUÑOZ, *Interpretatio in sex libros Euclidis* (último tercio del siglo XVI); F. SÁNCHEZ «el escéptico», *Objectiones et erotemata super geometricas Euclidis demonstrationes* (1577(?)); R. DOSMA, *Annotationes in Euclidem, Archimedem et alios* (principios del siglo XVII). Pero también cabe incluir extractos y usos de Euclides en castellano como los aprovechados por JULIO CÉSAR FIRRUFINO, *Platica manual y breve compendio de artillería*, Madrid, 1626; o, en fin, reelaboraciones dentro de la tradición de la geometría especulativa-práctica, como las de J. ZARAGOZA, *Euclides Nuevo-Antiguo. Geometría especulativa y práctica de los planos y sólidos*, Valencia, 1671; A. PUIG, *Aritmética especulativa y práctica y arte de Álgebra con la explicación de los libros V y VII-X de Euclides*, Valencia, 1672. La literatura euclidiana de este tipo provinente de los medios universitarios tradicionales solía publicarse en latín, a pesar de que nuestros universitarios de entonces tampoco destacaban por sus latines.

[67] Una fuente común de información bibliográfica es F. PICATOSTE, *Apuntes para una biblioteca científica española del siglo XVI*, Madrid, 1891; también es accesible el «inventario» de M. MENÉNDEZ PELAYO, *La ciencia española* (en la edic. de M. ARTIGAS, Madrid, 1933, tomo II, págs. 342-369). Así mismo, cabe acudir a los trabajos ya citados de REY PASTOR (1926) y SÁNCHEZ PÉREZ (1935). REY PASTOR, precisamente, menciona como traductores de los libros I-VI en el siglo XVI a Julio César Firrufino y a Jerónimo Muñoz —junto con Rodrigo de Porras, Juan Cedillo y Nicolás Vibario—, op. cit., pág. 159. De la existencia de una «traducción» de J. C. Firrufino, PICATOSTE sólo tiene constancia indirecta (1891, op. cit., pág. 108, ref. 289); desde luego, no pueden llegar a tanto las siete definiciones del libro I y los resultados sobre perpendicula-

1 *Los seis libros primeros de la Geometria de Evclides, Traduzidos en lengua Española por Rodrigo çamorano Astrologo y Mathematico, y Cathedratico de Cosmographia por su Magestad en la casa de Contratacio de Seuilla*, Sevilla, 1576.

Es un volumen en 4.°, con 121 folios y fe de erratas. Tiene licencia real de Felipe II, fechado en marzo de 1574, y está dedicado a Luciano de Negron, canónigo de Sevilla. En su prólogo «Al curioso lector», Zamorano recoge la tradición que remonta el origen de la geometría a las necesidades prácticas de medición provocadas por las crecidas del Nilo, e insiste no sólo en sus aplicaciones y servicios a muy diversas artes y disciplinas, sino en su calidad de ciencia básica y demostrada. «Y finalmente a nadie podemos juzgar por docto, a nadie por perito en su ciencia o arte alguna: si carece del conocimiento de la Geometría, basis y fundamento de todas ellas» (fol. 7r). Así mismo dice que se limita a los seis primeros libros por ser los más necesarios y que renuncia a añadirles comentarios o escolios —«que pudiera»— por bastarse el propio texto. Las definiciones, el «primer género de principios» que abre el libro I, son 35 y de este tenor: «1. Punto es, cuya parte es ninguna», «4. Línea recta es la que igualmente está entre sus puntos», «35. Líneas rectas paralelas son las que estando en un mismo llano, y extendidas de ambas partes en infinito, en ninguna parte concurren». Su alto número obedece a la introducción de una Def. 19 (del segmento de círculo) y a la distribución de algunas definiciones euclídeas en los casos comprendidos (clases de ángulos, de triángulos, de rectángulos). El segundo género de principios, «las peticiones», consiste en los cinco postulados normales. El tercer género, «las comunes sentencias», suma diez nociones: incluye no sólo las tres primeras,

res, paralelas e inscripciones, que este autor emplea en su ya citado *Platica manual y breve compendio de artilleria*, 1626 (*vid.* fols. 1v-3r y 3r-6r, respectivamente); hay noticia, sin embargo, de una versión manuscrita de su padre, Julián Firrufino. Por último, me temo que la referencia de Rey Pastor a Jerónimo Muñoz da en tomar sus comentarios a Euclides (libros I-VI) por una traducción.

sino otras cuatro habituales en la tradición arábigo-latina [68]; más «8. Las ⟨cosas⟩ que entre sí convienen son iguales entre sí», «9. El todo es mayor que su parte», «10. Dos líneas rectas no cierran superficie». Tales principios están ilustrados además con figuras geométricas. El número de las proposiciones de los libros I-VI concuerda con el hoy establecido —las proposiciones presentan una doble numeración: la correspondiente a su orden entre los problemas y teoremas y la correspondiente a su orden dentro de la secuencia general del libro—. Las definiciones de V y VI presentan ciertas alteraciones típicas: por ejemplo, en V se introducen los conceptos de proporcionalidad «ordenada», «desordenada» y «apartada» (Deff. 18-20); en VI se añade la definición 5 de «razón compuesta». La estructura de las proposiciones sigue una pauta reconocible: enunciado, referencia a un caso dado («Sea...») y construcción subsiguiente (prueba iniciada con un «Porque...»), remate de la conclusión («Lo cual convino —convenía— hacerse» o «Lo cual convino —había de— demostrarse», según se trate de problemas o teoremas). En cada caso figura la ilustración gráfica pertinente y se mencionan los principios o las proposiciones previas que justifican los principales pasos de la prueba.

Esta versión de Zamorano es, en conjunto, excelente. Constituye la única traducción propiamente dicha que he encontrado en este período de los siglos XVI-XVIII. Puede que la edic. latina de Zamberti (1505) sea su principal texto fuente [69]. Se mueve den-

[68] Son las sentencias comunes 4: «Y si a desiguales se juntan cosas iguales los todos serán desiguales». 5: «Y si de desiguales se quitan cosas iguales las restas serán desiguales». 6: «Las cosas que son dobladas a una misma son iguales entre sí». 7: «Las cosas que son de una misma mitad son iguales entre sí».

[69] Vid. B. ZAMBERTI, *Euclidis megarensis...*, Venecia, 1505 (Biblioteca Nacional, 2, 24543). En las nociones clave como las definiciones, postulados y nociones comunes del libro I, y las definiciones del libro V, sólo cabe apreciar contadas diferencias terminológicas entre el español de Zamorano y el latín de Zamberti (a saber: «punto»/«signum», «lla-

tro de la tradición que considera a Euclides filósofo megarense y los *Elementos* compuestos por 15 libros. Por otra parte, varios de los rasgos antes señalados serán compartidos por las versiones españolas subsiguientes. Un detalle final, revelador de los usos académicos de la época, es que Zamorano tenga que defender su versión castellana frente a la opinión entonces vigente: «que el andar las scientias en lengua vulgar es hazerlas Mechanicas»; pocos años después, una orden de Felipe II contribuirá a que las vindicaciones de este tipo resulten innecesarias [70].

---

no»/«planus», «quadrangulo»/«altera pars longius», «cantidad»/«magnitudo», «semejanza»/«identitas»), amén de un cambio de orden de las definiciones 12 y 13 de V (razón permutada, razón conversa). Zamorano no parece seguir, en cambio, la versión latina de Campano editada por Ratdolt en Venecia, 1482, aunque podía haber entonces ejemplares disponibles (al menos, hoy se conservan en Madrid, Barcelona, Sevilla, Valladolid). Tampoco hay motivos para suponer que utilizara una edición griega (e.g. la de Grynaeus, 1533). Por otro lado, puede que la versión latina de Commandino (Pesaro, 1572) resultara demasiado reciente —en todo caso, Zamorano no se muestra en deuda con ella—; esta edición o la italiana revisada en 1575 será la seguida más tarde por otro matemático de la Casa de Contratación, Miguel Jerónimo de Sta. Cruz, en el estudio de las proporciones de la segunda parte (libro II) de su *Dorado Contador, Arithmetica especulativa y practica*, Madrid, 1594.

[70] En la licencia y privilegio del rey Felipe que autoriza —con fecha de 15 de septiembre de 1584— la publicación de *La Perspectiva y la Especvlaria de Evclides traduzidas en vulgar Castellano por Pedro Ambrosio Onderiz* (Madrid, 1585), se menciona la existencia de esta orden real: «que en nuestra corte se leyesen las matemáticas en lengua castellana». El propio Ondériz, al dirigirse al lector, se hace eco de la orden y recuerda que se había llamado a Juan Bautista Labaña para presidir esta tarea en la recien creada Academia de Matemáticas (1583). La Casa de Contratacion —en la que profesaba Zamorano— y esta misma Academia —en la que profesará Juan Cedillo Díaz y se formará Luis Carduchi— son instituciones con aires humanistas y unos intereses técnicos alejados de la escolastica latina de la universidad tradicional.

2. *Teorica y practica de fortificacion, conforme las medidas y defensas destos tiempos, repartida en tres partes, por el Capitan Cristoval de Rojas, Ingeniero del Rey nuestro Señor*, Madrid, 1598.

Es un volumen de 106 folios numerados, que incluye una dedicatoria —fechada en 1596— «al Príncipe nuestro Señor Don Felipe III», un prólogo y un grabado; al final agrega una tabla de contenidos. En el prólogo, Cristóbal de Rojas reconoce el estímulo que habían supuesto las lecciones sobre la geometría de Euclides dadas en la Academia Real por Julián Firrufino «que con la profesión de leyes acompaña la de las ciencias Mathematicas»; pero nuestro ingeniero atribuye un papel decisivo en la composición de su libro a don Francisco Arias de Bobadilla, Conde de Puñoenrostro, quien le había encomendado la enseñanza de la fortificación al procurar que otros también «leyeran» en la Academia Real otras materias relacionadas con la geometría. Este libro de Cristóbal de Rojas viene a ser una buena muestra de los usos y servicios que por entonces parecían asegurar la supervivencia de Euclides —de un cierto Euclides—. En el cap. II de la primera parte de la obra, el autor afirma que «la primera de las tres cosas que han de concurrir en el Ingeniero, es la Geometría» (las otras dos serán la aritmética y el reconocimiento de los sitios); allí mismo alude a las 35 definiciones, 5 peticiones y 10 comunes sentencias del libro I de los *Elementos*, y se refiere a algunas proposiciones principales de los libros I-VI, XI y XII. En el cap. 5, aparecen 15 definiciones con figuras ilustrativas al margen (fols. 4v-5v); algunas son versiones escolares de definiciones del libro I (e. g.: «Punto, cuya parte es ninguna», «Línea es la que tiene longitud y no latitud ni anchura»); otras son peculiares (e. g.: «Ángulos de Advertice son los de las dos CC y lo mismo son los de las dos DD», por referencia a una figura adjunta en la que CC y DD designan ángulos opuestos por el vértice; «Ángulos Deinceps son los destas dos letras EE», por referencia a una figura adjunta en la que EE designan ángulos adyacentes). En el mismo capítulo se recogen a continuación (fols. 5v-14r) 12 proposiciones del libro I (1, 3, 9-10, 12-13, 21, 23, 31-32,

46 47); 2 del II (3, 12); 2 del III (31, 36); 3 del IV (5, 10-11) más un «corolario de la penúltima»; 1 del V (16); 6 del VI (4, 12-13, 17, 25, 30); 1 del XI (14). Su presentación sigue esta pauta: el autor parafrasea el enunciado de la proposición en cuestión, encarece su importancia o su utilidad, da una prueba sucinta —por lo regular más sugerida que desarrollada—; en ocasiones (e. g., en el corolario o porisma de IV, 15; en VI, 4; en XI, 14), falta incluso esa especie de prueba. En todo caso, está claro que Cristóbal de Rojas pretende una iniciación geométrica elemental, no una versión (de fragmentos) de los *Elementos* de Euclides. Por lo demás, el resto del libro ya no guarda ni siquiera esa lejana relación con los *Elementos* euclídeos.

3. *Testos de los 15 libros de los Elementos de Euclides romanceados por don R.° de porras.*

Es un manuscrito en 4.° de 148 folios (algunos de ellos en blanco); los libros I-XIII cubren los fols. 1-65r (XIV-XV, fols. 65v-68v); siguen unas proposiciones geométricas demostradas por el propio Rodrigo de Porras (fols. 73r ss.), y aritméticas (fols. 121r ss.), hasta concluir con cuestiones de binomios resueltas algebraicamente (fols. 137r ss.). Se encuentra en la Biblioteca Nacional, Ms. 9285. Data probablemente de finales del siglo XVI o principios del siglo XVII.

El libro I, donde Porras alude a la condición de Euclides como «philosopho megarense», empieza con las definiciones; son 35, como en Zamorano, pero de distinto tenor: «1. Punto es una cosa indivisible porque carece de partes», «4. Línea recta es la brevísima extensión de punto a punto», «35. Líneas equidistantes o paralelas son las que están en una misma superficie y no concurren aunque sean alargadas en infinito por una y otra parte» [71].

[71] Esa noción de línea recta, procedente de Arquímedes, y su proyección sobre la noción de superficie plana habían sido difundidas por el comentario de Al-Nayrīzī. La asociación de «equidistantes» a «paralelas» también era común en la tradición arábigo-latina y subsistió durante mucho tiempo (como ya veremos, hasta bien entrado el siglo XVIII, e. g., en Simson, 1756, 1774), a pesar de la crítica a que Saccheri sometió esta definición «complessa», que colaba de rondón, en la definición de «líneas paralelas», una hipótesis de equidistancia necesitada de prueba.

Las «peticiones» o postulados se asemejan a los de Zamorano, pero al final aparece tachado «dos líneas rectas no cierran superficie» que, luego, constituye la «común sentencia» décima y última —estas comunes sentencias son también parejas a las de Zamorano—. Las proposiciones tienen la doble numeración que ya habíamos visto en la traducción anterior. Los libros V y VI tienen así mismo una disposición similar, aunque la def. VI, 5 de Porras menciona la proporcionalidad continua y se aproxima más a la tradición de Adelardo y Campano. En VII —donde Euclides sólo ofrece definiciones—, Porras introduce 4 peticiones y 2 comunes sentencias; las tres primeras peticiones son postulados procedentes de Jordano y la cuarta es una adición de Campano; las comunes sentencias son: «1. Toda parte es menor que un todo», «2. Todos aquellos números que fueren igualmente multíplices a un mismo número o a números iguales, serán iguales entre sí». También el libro X cuenta con una petición original que no he conseguido descifrar por completo: «Una cantidad puede ser multiplicada tantas veces que ⟨¿... ...?⟩ se da otra cualquiera cantidad del mismo género». Los libros VII-IX tienen todos ellos proposiciones añadidas; en cambio, XII se ha quedado sin tres.

No se trata de una traducción, sino más bien de una versión simplificada e incompleta: las proposiciones sólo tienen el enunciado y una referencia a la figura correlativa situada al margen, sin ninguna otra indicación o desarrollo de la prueba. E incluso a partir del final de IX, las ilustraciones desaparecen aunque se les reserva un espacio al margen en blanco. Quizás se trate de un material de trabajo puesto en limpio, pero no ultimado. El texto fuente es más difícil de adivinar aún que el de Zamorano; según todos los visos, Porras parte de una versión ecléctica que acusa ante todo el peso de la tradición arábigo-latina y la mediación de Campano.

4. Manuscrito anónimo, sin título ni fecha (conjeturo que ronda el final del siglo XVI o principios del siglo XVII). Contiene los libros I-VI, B. N., Ms. 9128; tamaño folio, 121 fols.

El libro «Primo» tiene 24 definiciones con una disposición semejante a la euclídea, pero no sin peculiaridades de contenido y de explicación gráfica, e. g., «3.ª Línea recta es la más breve distancia entre dos puntos como la línea A...», «24.ª Líneas paralelas son las que igualmente distan por todas sus partes, como las líneas ...». No hay postulados. Las llamadas «concepciones» son 9: las 7 primeras son las ya familiares; la octava reza «Si una cantidad se pone sobre otra cantidad y no excede la una de la otra, serán iguales entre sí», como si fuera una versión de *Adelardo I;* la novena explica «el todo es mayor que la parte, como el todo ...». Entre los libros I y II (fol. 29r) se intercalan unas curiosas «proposiciones de los números» que recogen nociones de raigambre medieval sobre múltiplos («multíplices»). Por contra, el libro V se ve libre de las interpolaciones habituales y el VI sólo introduce la Def. 5 de «composición de proporción». La estructura de la proposición distingue tres pasos principales: enunciado, «Sea ...» o consideración del caso dado, «Porque ...» o prueba propiamente dicha. Cabe pensar que se trata de una versión algo simplificada, hecha a partir de una fuente relativamente sesgada por la tradición arábigo-latina, aunque se vea libre de varias de sus contaminaciones características (especialmente en el libro V).

5. *Los Primeros seys libros de Euclides, agora nuebamente traduzidos de lengua latina en castellana, explicados, demostrados y en muchas partes ilustrados. Con muchos y agradables corolarios y añadimientos. Por Nicolas Vibario, Ingeniero de su Magestad en los estados de Flandes,* Amberes, 1616. Manuscrito en tamaño folio, 266 fols. B. N., Ms. 9214.

Tras la dedicatoria a D. Íñigo de Borja, maestre de campo de un tercio de infantería en Flandes, sigue un prólogo: «Que cosa sea Geometria y las Artes mathematicas por que assi llamadas y quienes fueron sus primeros mentores». Nada tiene de particular: la etimología que declara su condición de doctrina o disciplina [*mathéma*]; el canto a su excelencia, pues las matemáticas no conceden ninguna cosa falsa y no consienten sino principios claros y conclusiones demostradas; su origen egipcio y su aumento griego gracias a las agudísimas pruebas de Euclides megarense, entre otros. «El primer elemento de Euclides» comienza en el fol. 5r. Abunda en definiciones (39) y en explicaciones: e. g., a propósito de línea recta se insiste en «el más breve camino que hay de un punto a otro, como por ejemplo...»; a continuación, se introduce la noción de línea como «una medida», superficie como «dos medidas», cuerpos y cubos como «tres medidas». Siguen 3 «preguntas», *i. e.* postulados, que reproducen los tres primeros de Euclides. Luego vienen las «comunes noticias y sentencias del entendimiento», que suman 12 al incluir el postulado cuarto (sentencia décima), el quinto (undécima) y, en fin, una versión harto peculiar de la común sentencia 12 consabida: «Dos líneas derechas no tienen espacio». La inserción de algunos postulados entre las comunes sentencias ya se encuentra en Grynaeus, en Clavio, en Candalla. Todos estos principios se ilustran con figuras; llama la atención el gasto de plumilla en la representación de sólidos para dar una impresión vívida de volumen. Se intercalan a continuación (fols. 16r-17r) unas aclaraciones acerca de los tipos de principios y de proposiciones que componen la geometría. Las definiciones del libro V también cuentan con la glosa correspondiente; la definición de proporcionalidad en términos de equimúltiplos (Def. 6) tiene una ilustración numérica, al igual que ciertas operaciones entre proporciones. La estructura de las proposiciones sigue unas pautas declaradas previamente (fol. 17r); explicación *(i. e.* «exposición de lo que se da y lo que se requiere»), operación *(i. e.* «composición con ayuda del compás y la regla») y demostración, en el caso de los problemas; pero explicación y demostración únicamente en el de los teoremas. La versión de Vibario está simplificada y acusa un claro sesgo didáctico. No faltan huellas de la formación un tanto escolástica del autor. Es difícil determinar su fuente latina. En todo caso, evidencia una confección esmerada; es una especie de

manuscrito puesto en limpio con aires de cuaderno de maestro de escuela.

6. *Los seis libros primeros de la geometría de Euclides traducidos del latín en castellano por el doctor Juan Cedillo Diaz, Cosmographo Mayor de su Magestad y Catedratico de la Real Academia de Mathematicas que se lee en su Palacio,* 1620. Forma parte de un manuscrito en tamaño folio de 219 fols. (la referencia anterior se lee en el fol. 64v), B. N., Ms. 9091.

El ms. es en conjunto un verdadero revoltijo (quizás el favor de coserle las hojas se lo hiciera a Cedillo uno de sus peores enemigos). Contiene la traducción de dos libros del *Arte de navegar* de P. Núñez de Saa (fols. 2r-20v); un ensayo, *Idea astronomica de la fabrica del mundo,* distribuido así: dos libros (fols. 108r-177v), su revisión (fols. 182r ss.), el tercer libro (fols. 84r-107v); otro ensayo, *De la Idea y Cosmologia* (fols. 180r-181v); y, en fin, fragmentos de los libros I-VI de los *Elementos* diseminados casi al azar entre el fol. 1 y el fol. 83 —el único fragmento seguido es el de las definiciones y las proposiciones 1-15 del libro VI (fols. 60-63)—. Las definiciones de I son las tradicionales 35: «1. Punto es cuya parte es ninguna. O punto es una extensión sin cantidad», «4. Línea recta es la que igualmente está entre sus puntos. O así: línea recta es la más breve extensión o camino de un punto a otro punto», «35. Líneas rectas paralelas son aquellas que estando en un mismo plano y siendo producidas o extendidas en infinito por una y otra ⟨parte⟩ no se juntan jamás ni pueden». Las peticiones se reducen a los tres postulados primeros. Las comunes sentencias (repartidas entre los fols 1v y 77) son parejas a las 12 que ya hemos visto en Vibario. Las definiciones del libro V están libres en cambio de las contaminaciones habituales (*vid.* fol. 80v). Por lo demás, las proposiciones vienen glosadas, pero sus pruebas se han simplificado. Seguramente Cedillo se ha servido de diversas fuentes —una de ellas podría ser la empleada por Zamorano si no su traducción misma—, y no falta la obligada referencia de Campano. Según todos los visos, se trata de un borrador todavía incompleto de lo que podría ser una versión ecléctica e instructiva de los *Elementos* (I-VI).

7. *Elementos geometricos de Evclides philosopho megarense. Sus seys primeros libros. Traducido el texto y comentado por Lvis Cardvchi Mathematico de su Magestad,* Alcalá, 1637.

Es la segunda versión castellana de los *Elementos* que ve la imprenta. Se trata de un volumen en 4.° al que adornan unos poemillas de amigos del autor, una Advertencia de éste (que incluye fe de erratas), un prólogo en el que ensalza la geometría por sus múltiples aplicaciones prácticas sin olvidar de paso un elogio de la pintura —o, más bien, de los *Diálogos de la pintura,* de Vicencio Carduchi—, y una dedicatoria al Excmo. Sr. D. Gaspar de Guzmán, conde de Olivares. Es ilustrativa del espíritu de los tiempos la aprobación del P. Ricardo (Richard), a la sazón catedrático de matemáticas en el Colegio Imperial: «He leydo este comentario sobre los seis primeros libros de Euclides ... y me parecen dignos de ser impresos por la breuedad y claridad del comentario necessaria a todos los que quisieran sauer la Geometria». La versión de Carduchi (fols. 1-144) se inicia con unos breves preliminares sobre la geometría («ciencia contemplativa y descripción de formas y figuras de la cantidad continua, que muestra las propiedades de las figuras») y sobre sus especies (a saber: línea, superficie y cuerpo). Las definiciones de I son 29 y tienen una glosa explicativa o ilustrativa: «1. Punto es, cuya parte es ninguna», «Línea recta es la más breve extensión de un punto a otro —o fluxo de un punto a otro—», «28. Líneas paralelas son las que igualmente distan por todos sus puntos y extendidas en infinito jamás concurren», «29. Líneas paralelas hacen a una misma parte los ángulos iguales a dos rectos; y las que no lo son, hacen los ángulos a una parte menores que dos rectos». Tras una nota sobre las nociones de «petición» y de «sentencia común», siguen las peticiones. Son siete, Carduchi intercala como petición 2: «Concédase que de un punto a otro no puede haber más que una línea recta» y añade como petición 7: «Dos líneas rectas no cierran superficie». Las comunes sentencias quedan re-

ducidas entonces a nueve y no son otras que las ya familiares. La glosa de las peticiones y sentencias no es explicativa, sino puramente ilustrativa. En V y VI reproduce las definiciones usuales interpoladas, incluida la de proporcionalidad continua. La estructura de la proposición sigue la pauta presente de Zamorano, aunque las pruebas están por lo general resumidas. Es posible que Carduchi haya tenido en cuenta la traducción de Zamorano. Pero el peso de la tradición arábigo-latina y de Campano parece más acusado y su versión es muy inferior: no sólo por asemejarse a una paráfrasis de Euclides, de raíz ecléctica y con propósitos didácticos, sino porque su castellano incorpora —además de los grecismos habituales en aquel tiempo (triángulo «orthogonio», «ambligonio», «oxigonio»)— algún que otro latinismo (e. g., «adinvicem») y algún que otro arabismo («helmuayn», «helmuaerife»), inexistentes en las demás traducciones y versiones.

8. Lorenzo de San Nicolás, *Arte y uso de la arquitectura*, vol. I, Madrid, 1633; vol. II, Madrid, 1665.

Esta obra, modelo dentro de su género —fue el manual de arquitectura de mayor influjo en España y en Hispanoamérica durante el barroco—, contiene en su vol. II una versión del libro I a cargo de Antonio de Náxera: definiciones, postulados y axiomas en el cap. XVI, págs. 28-40, y proposiciones en un apéndice, págs. 220-287; en el vol. I, la versión del libro V del mismo traductor y una versión del libro VII a cargo de Juan de la Rocha, ambas en el cap. LXV, págs. 213-355. Precisamente en este último contexto, tras aludir a la versión de Antonio de Náxera de los libros I-IV y VI, el P. Lorenzo de San Nicolás declara refiriéndose a los libros V y VII: «De quien estos dos se han traducido y los cinco dichos es del Padre Cristóbal Clavio Bambergensi, de la Compañía de Jesús» (pág. 282). Suponemos que habría una versión de Antonio de Náxera de la que hoy sólo conocemos los libros aquí recogidos, a la que se añade la versión del libro VII por cuenta de Juan de la Rocha: las dos se mueven en el ámbito de influencia de la edición comentada de Clavio. El ejemplar que he manejado corresponde a la 4.ª imp., 1796.

En el caso de I, nos encontramos con las 35 definiciones de rigor («1. Punto es aquel que su parte es nada o que no tiene ninguna grandeza», «4. Línea recta es aquella que igualmente se interpone entre sus puntos» —también se alude a su condición de línea más corta—, «35. Líneas paralelas son aquellas que estando en un mismo plano y produciéndose en infinito, para una y otra parte, jamás se encontrarán una con otra»). Las peticiones o «postulata» son los tres postulados primeros a los que se agrega: «4. A cualquier grandeza dada se puede tomar otra grandeza, o mayor o menor»; Náxera reconoce además la existencia de otras que se mencionarán más adelante, cuando las requieran las pruebas de las proposiciones. Los axiomas o sentencias comunes son diez, el último es el acostumbrado: «Dos líneas rectas no comprehenden espacio». Si en I abundan los escolios, en V se presta especial atención a un catálogo de las proporciones y al hallazgo y uso de las proporciones racionales. Pero, en general, las proposiciones discurren sin hacer referencia a la justificación teórica de las pruebas (a los principios o a las proposiciones previamente sentadas), sino a su demostración práctica por medio de figuras y de construcciones diagramáticas. Dado el contexto en el que aparecen, es natural que en estas versiones prevalezca el interés por una exposición breve, clara e ilustrativa. Aun así, su valor —en particular, el de la versión del libro I—, resulta muy superior al de los extractos y las adaptaciones que solían incluir por entonces los manuales de matemática aplicada; baste recordar el precedente —ya referido— de la *Teorica y practica de fortificación*, de Cristóbal de Rojas (1598).

9. *Los seis primeros libros, onze y doze de los elementos geometricos de Euclides megarense, aumentados de muchas Proposiciones curiosas que facilitan su uso. Por don Sebastian Fernandez de Medrano, Capitán y Maestro de Metemáticas del ejército de los Paises Bajos*, Bruselas, 1688.

Se trata de una impresión en 8.° de 340 págs., y es la primera muestra hispana de la adición de XI-XII a los libros «fundamentales» I-VI, conforme al uso impuesto a mediados del siglo XVII,

sobre todo en Francia (e. g., por las ediciones de A. Tacquet, 1654, o Milliet de Chales, 1660). En I, a las 35 definiciones «de Euclides», Medrano agrega otras dos que, según él, incluyen los geómetras para más inteligencia; son las de *paralelogramo* y *cumplimiento* o *suplimiento, i.e.* el paralelogramo «por el que no pasa el diámetro de todo un paralelogramo». Las peticiones o «demandas» son las tres de rigor, pero también se menciona la cuarta de Clavio: «siendo dado un grandor o cantidad: que se puede tomar otra mayor o menor». Los axiomas, «máximas» o comunes sentencias suman, a su vez, trece: los 12 habituales en la línea de Clavio y «13. Dos líneas rectas que se encuentran indirectamente no tienen un mismo segmento». En el caso de 11, que recoge el postulado de las paralelas, Medrano alude a la necesidad de demostración que algunos intérpretes alegan y a su sustitución por este criterio: «las líneas que se terminan en dos paralelas, siendo perpendiculares a ellas, serán iguales» —formulación que tal vez se difundió a partir de la edición de G. Vitale: *Euclide restituto* (Roma, 1680)—. En el libro V, Medrano hace referencia a la generalidad y dificultad de la teoría de las proporciones, que presenta en versión simplificada y accesible «siguiendo a autores modernos». Tras las definiciones —cuya mayor peculiaridad es la introducción de una prueba de la definición euclídea de proporción en términos de equimúltiplos, prueba obediente también a ciertos gustos modernos—, explicita como postulado: «pídese que propuestas tres cantidades se conceda una cuarta < proporcional > ». Tanto el libro VI como los libros XI y XII son objeto de tratamiento sumario; se omiten no sólo ciertas pruebas que se consideran evidentes, sino algunas proposiciones. Por lo demás, al principio de XI explica la omisión de VII-X en razón de que versan sobre inconmensurables, y su estudio es más sencillo por medio de la llamada «álgebra especiosa o Analítica».

La versión de Medrano conoce una nueva edición, «amplificada», en Amberes, 1708, cuando el autor ha alcanzado el grado de Sargento General de Batalla. No he tenido acceso a ella. Años después aparece una reimpresión en 8.°, 327 págs. (Amberes, 1708), que Medrano —ahora Director de la Academia Real y Militar de los Países Bajos— justifica por haberse agotado las ediciones anteriores o haberse quemado los ejemplares que quedaban bajo el fuego graneado de un ataque a Bruselas. Un añadido que cabe reseñar es el de una proposición alternativa a I, 19, consistente en realidad en la proposición conversa de I, 27 *(vid.* Apéndice, págs. 324-327). Pero esta postrera edición mantiene, en general, los términos, los principios y las pruebas simplificadas de la edición primera (e incluso la trasnochada referencia a Euclides como filósofo megarense).

10. *Elementos geometricos de Evclides. Los seis primeros libros de los planos y los onzeno y dozeno de los solidos; con algunos selectos theoremas de Archimedes, traduzidos y explicados por el P. Jacobo Kresa de la Compañía de Jesvs,* Bruselas, 1689.

Es un volumen en 4.° con 459 págs. El P. Kresa, de origen austríaco, profesor durante algún tiempo del Colegio Imperial en Madrid, aduce para justificar esta versión el hecho de que la antigüedad o la excesiva brevedad de otras anteriores hacían los *Elementos* «arduos a los principiantes». La verdad es que no aporta grandes novedades, salvo la de elevar el número de los axiomas hasta 20 de manera que quepan entre ellos no sólo los tradicionales de Clavio, sino los explicitados por otros jesuitas editores de Euclides, en particular el P. Andrè Tacquet, quien también había sido llamado anteriormente para desempeñar la cátedra de matemáticas del Colegio Imperial. En el libro V se dejan oír los ecos de Campano, pero el más escuchado sigue siendo Tacquet, del que Kresa toma un apéndice sobre las operaciones y la composición de razones racionales e irracionales. La autoridad del «método» de Tacquet campea en el libro XII, donde es seguido al pie de la letra —según confesión del propio Kresa—, y en los teoremas extraídos de Arquímedes. En suma, representa no una versión, sino más bien una paráfrasis en la línea marcada por algunos comentadores euclidianos notables de la Compañía de Jesús.

11. *Los seys primeros libros de los elementos de Euclides. Explicados y demostrados por don Francisco Larrando de Mauleón, Capitan de Infanteria Española,* Barcelona, 1698.

Es un volumen en 32.° con 368 págs. Su dedicatoria a don Manuel A. de Velasco y Tobar puede ilustrar una de las tres vías principales de introducción del viejo Euclides en España. Me refiero a la de los profesionales de la milicia que alternan esta ocupación con la docencia más o menos regular de las matemáticas. Dice Larrando con su peculiar retórica militar: «Es la Geometría (especial blanco de mi pluma) una de las partes de la Matemática que es más conveniente a un Monarca y a un Capitán General y a un Príncipe. Toda la razón de estado es ciega sin el conocimiento de esta nobilísima ciencia. Todas las líneas del gobernar se han de tirar al centro de la utilidad común: Geómetra es el Príncipe o el Gobernador que ha de saber medir la distancia que hay del mandar al obedecer. Para todas las funciones del Arte Militar, es la que da más luz la Geometría, verdad tan innegable que la confiesan todas las Escuelas de Marte». Tras estos loores nos encontramos con una versión simplificada e ilustrada, que no tiene mucho de particular. Los postulados se reducen a tres. Los axiomas ascienden a trece: hasta el noveno, mantienen su venerable aspecto. El décimo es: «Dos líneas no pueden tener un segmento común» (similar al axioma o máxima 13 de Medrano). El undécimo consiste en el postulado de las paralelas y el duodécimo recuerda que dos líneas no pueden encerrar un espacio. En fin, el décimotercero es: «El todo es igual que todas sus partes juntas», una precisión debida a Clavio que Medrano ya había recogido como corolario de su axioma o máxima 9 (el todo es mayor que su parte). Las pruebas de las proposiciones se ofrecen abreviadas. Con esta versión de Larrando iniciamos la pendiente que nos lleva a una paráfrasis de los *Elementos* cuyo contenido geométrico o euclidiano es más bien trivial, está decididamente escolarizado; su mayor aliciente en esta perspectiva reside seguramente en el colorido de sus consideraciones preliminares y en el tono costumbrista de su contextualización y su glosa de los resultados tradicionales.

12. Manuscrito anónimo sin título ni fecha, denominado *Tratado de Geometría* y situado en el siglo XVIII por el *Inventario general de manuscritos de la Biblioteca Nacional* XI (Madrid, 1987); ref. 6104, pág. 105.

Comprende los folios numerados 7-73. Hasta el fol. 60r incluye los libros I-III de los *Elementos*. A partir del fol. 60v se introduce la «Materia de cuerdas senos rectos tangentes y secantes», con definiciones y proposiciones originales, y tiene lugar un cambio de mano o de escritura. Los fols. 65v-71r tratan de geometría esférica y vuelven a una escritura similar a la inicial. En 71v está escrito en posición invertida «Ad Euclidem». El fol. 72 está en blanco. El fol. 73v trae enmarcada una especie de plegaria y más abajo se remata con un «Finis». B. N., Ms. 6104.

No aparecen las definiciones, postulados y nociones comunes de I; el ms. comienza con la proposición 1: «*Super data recta linea terminata triangulum aequilaterum constituere.* Hacer un triángulo equilátero que tenga lo que quisieren por lado». La prueba está esquematizada; hay al margen referencias a postulados y nociones comunes bajo una curiosa clave, e. g., «C 10» (¿postulado iii?), «C 1» (¿noción común i*?). El resto de las proposiciones de I-III sigue en general una pauta parecida: tras el enunciado en latín y su versión libre en castellano viene una indicación muy somera de la prueba con algunas referencias a postulados y nociones comunes; en la carilla opuesta figura la ilustración correspondiente. En los libros II y III sí se recogen y explican las definiciones (fols. 31v y 40v, respectivamente) antes de pasar a las proposiciones. El ms. evidencia una fuente latina académica, una versión sumamente libre y una confección escolar esmerada.

13. *Euclides. Geometria especulativa y practica de los planos y solidos. Dispuesta por don Antonio Joseph Deu y Abella, Capitán agregado al estado mayor de Zaragoza,* Zaragoza, 1723.

Se trata en realidad de un manual de geometría tradicional compuesto en 4.° por 175 págs. Comprende los libros I-V, XI

y XII, aunque dentro de un marco que les es ajeno: el de la distinción entre la «Geometría especulativa», fundada en las definiciones y los axiomas (págs. 5-127), y la «Geometría práctica», abierta por los postulados (págs. 128-173) y consistente en problemas —es una distinción medieval que puede remontarse a Hugo de S. Victor (siglo XII)—.

Deu y Abella dedica su obra «a la Madre de Dios del Pilar» con un estilo que al parecer considera a la altura de tan alta instancia: «En esta escuela, Señora, donde sus *principios primeros* son las Armas, *definiciones* los echos, y *conclusiones* las Plazas, desconocía con la experiencia el pensamiento, y practicando sus avisos, principios prácticos del Militar estudio, estudié, para que con la pluma más se ennobleciesse de la espada mejorado su exercicio». (No sé si será eximente o agravante la pretensión del autor de abrir una Academia de Matemáticas en Zaragoza.) Tras la dedicatoria, Deu y Abella declara seguir «el methodo del P. Joseph Zaragoza» y no respetar por motivos didácticos el orden de las proposiciones de Euclides —aunque sí el de los libros—. Un proemio (págs. 2-4) recoge las ideas comunes sobre la importancia de las definiciones para la ciencia, la índole de la matemática (ciencia de la cantidad inteligible, esto es: mensurable o numerable) y, en fin, sobre su división (pura/no pura o física) y sus subdivisiones. La concepción de las definiciones, postulados y axiomas, expuesta al principio de la «Geometría especulativa», concuerda con la difundida por los manifiestos del «esprit géométrique»: las definiciones son nominales, los axiomas son verdades «tan claras y evidentes que sólo suponerlas es obligar al entendimiento a confesar su verdad». A continuación, presenta los axiomas (págs. 5-6) y las definiciones (págs. 6-11) con una formulación que sintetiza las enseñanzas euclídea y moderna; e. g.: «1. *Magnitud* o grandeza: cantidad continua y mensurable. Si es finita y terminada, sus términos son los extremos de la magnitud (del continuo)». «2. *Punto Mathematico,* en la magnitud es un signo o señal indivisible, sólo por imposición y consideración del entendimiento». Los postulados de la «Geometría práctica» tienen una presentación similar. Las pruebas discurren de

acuerdo con los propósitos sinópticos y pedagógicos de la obra. Por lo demás, no falta una breve recapitulación de los principios y las reglas generales de la Geometría. Son 14 y hacen referencia —por este orden— a las proporciones y los resultados de operar con ellas, a las fórmulas para hallar áreas y volúmenes, a las nociones comunes clásicas de igualdad con la adición final de «el todo es igual a sus partes juntas».

14. *Elementos geometricos de Euclides, dispuestos en methodo breve y facil para mayor comodidad de los aficionados, y uso del Real Seminario de Nobles de Madrid. Su autor el Padre Gaspar Álvarez de la Compañía de Jesús,* Madrid, 1739.

Es un volumen en 4.°, consta de 184 págs. y comprende los libros I-VI, XI y XII. Según se declara en la aprobación del P. Casani, el autor también cambia el orden y la secuencia de las proposiciones en sintonía con la intención moderna de facilitar la enseñanza; a fin de cuentas, los *Elementos* sólo son un magnífico manual, «una llave que abre la puerta a todas las facultades Mathematicas». Hablando de matemáticas, el P. Casani no deja de insistir en sus servicios de toda índole, algunos son los tradicionales (la fortificación y la estrategia militar, la navegación, la geografía...), pero otros tienen el discreto encanto de la Ilustración —en el Seminario de Nobles en particular—, e. g.: las matemáticas deparan «las curiosidades de la esfera para la conversación y familiar trato en donde se lucen las prendas; la discreción y el útil uso del tiempo; el manejo científico en la destreza; la diversión de la música. Son prendas todas de tanto adorno a la Nobleza que se echan mucho de menos cuando faltan» [72]. En las palabras «Al lector» del propio Gaspar Álvarez también abundan las referencias sintomáticas del espíritu de los tiempos. Tras aludir al escaso interés por las matemáticas en nuestro país

[72] La moda de la geometría en el mundo ilustrado había alcanzado no sólo a los filósofos, sino —desde finales del siglo XVII— a la buena sociedad. *Vid.* por ejemplo, P. HAZARD, *La crisis de la conciencia europea,* 1946², Madrid, 1952, págs. 283-284 en particular.

y a la paciencia que exige su constitución abstracta, procura justificar su nueva versión por la conveniencia de salvar la prolijidad de las demostraciones euclídeas e incluso la que padecen otras versiones disponibles como la del P. Kresa. Porque —alega— si bien el orden de Euclides es admirable, «el no poder variar en algún accidente y cansar a los principiantes probando con demostraciones largas verdades por sí patentes, lo tengo por un cierto género de superstición». A mayor abundamiento: «quien todo lo quisiere por extenso» podrá acudir a las ediciones del P. Bettino —quien, por cierto, había cambiado hasta el orden de los libros— o del P. Clavio. Gaspar Álvarez, por su parte, se apoyará en las innovaciones del P. Andrè Tacquet. La formulación de las definiciones, postulados y axiomas sigue la presentación estándar en las versiones de los jesuitas, especialmente a partir de Clavio y de Tacquet. La obra tiene el aderezo de algunos recursos didácticos (e. g., un mapa de la doctrina de los lados, ángulos y triángulos al final del libro I), pero no pasa de ser un compendio compuesto por proposiciones seleccionadas con sus demostraciones aliñadas en aras de la brevedad y la simplificación.

15. *Los seis libros primeros de los planos y los onze y doze de los solidos. Por D. Blas Martinez de Velasco.*

Es una publicación póstuma: Puerto de Santa María, 1747; 4.°, 192 págs. Siento no haber hallado un ejemplar de este libro, escrito por un discípulo de Antonio Hugo de Omerique (cuya reivindicación del «método analítico de los antiguos» en geometría mereció una elogiosa mención por parte de Newton) [73].

[73] El reconocimiento de Omerique es uno de los tópicos más repetidos a la hora de reanimar la matemática española del siglo XVII. El pasaje original de Newton puede verse en J. A. SÁNCHEZ PÉREZ, art. cit., 1935, pág. 630; hay otras referencias elogiosas de Montucla y de Chasles en las págs. 626 y 629. Por otra parte, KRESA, en su edición de los *Elementos* (1689), cita expresamente y utiliza resultados de Omerique (vid. *supra*, núm. 10, págs. 250, 264). De la versión de Martínez de Velasco sólo tengo noticia por el *Manual del librero hispanoamericano* (Barcelona, 1951²; tomo V, pág 199, ref. 84729); al parecer, todavía se encontraba algún ejemplar en venta en 1894.

16. *Los seis primeros libros y el undécimo y duodécimo de los Elementos de Euclides. Traducidos de nuevo sobre la versión latina de Federico Commandino conforme a la fiel, y correctísima edición de ella publicada modernamente por Roberto Simson Profesor de Matemática en la Universidad de Glasgow: e ilustrada con notas críticas y geométricas del mismo autor,* Madrid, 1774.

Es un volumen en 4.° de 360 págs. Por desgracia, no aparece el nombre del traductor. Se trata de la única versión que hace frente a la vulgarización escolar de los *Elementos* en el siglo XVIII español. En las Advertencias que encabezan la obra (págs. 1-16) se dice que ha dispuesto esta traducción el Conde de O'Reilly, Inspector General de Infantería del Rey, como parte de su plan para cimentar el estudio de la ciencia en las obras de los autores originales. También se deja constancia de la superioridad de esta nueva versión sobre las ya existentes, bien porque éstas parten de texto incorrecto o alterado, bien porque son compendios o rudimentos de geometría. El traductor declara su propósito de acomodar, llegado el caso, las expresiones de Simson a la terminología establecida en nuestro país y adelanta, en fin, algunas nociones y términos de uso corriente entre los geómetras españoles. No faltan indicios de que era una persona ilustrada y competente en matemáticas. El prólogo de Simson (págs. 17-20) denuncia varias alteraciones del texto introducidas por Teón, aparte de las añadidas luego por la tradición arábigo-latina y por Campano, y expresa su intención de ofrecer un Euclides depurado. Al mismo propósito obedecen las notas que Simson agrega al final (págs. 295-360), fundadas en el cotejo personal de fuentes griegas y en «la restitución de su fuerza a las pruebas», bases que le permitirán incluso corregir en algún punto al propio Commandino. Lo cierto es, sin embargo, que las intenciones de depuración y preservación del «texto original» cuajan con desigual fortuna: el interés por la reconstrucción racional de las pruebas euclídeas suele prevalecer sobre la sabiduría hermenéutica. Por

otra parte, Simson tampoco renuncia a cierto eclecticismo en las definiciones y en los axiomas del libro I, así como en su reelaboración del libro V. En el caso de I, por ejemplo: «1. Punto o signo es lo que no tiene partes, o lo que no tiene magnitud»; «7. Superficie plana es aquella en la cual tomados dos puntos cualesquiera, la recta terminada por ellos se halla toda en la misma superficie» (def. procedente de Herón) [74]; «35. Paralelas o equidistantes son las rectas que estando en un mismo plano, prolongadas por ambas partes al infinito, jamás se encontrarán». Los axiomas, por otro lado, se dicen referidos a cantidades y los dos últimos, 11 y 12, son, respectivamente, los postulados de la igualdad de los ángulos rectos y de las paralelas; a propósito de éste último, reconoce su denominación como «postulado» y asegura que ni debe contarse entre los axiomas ni admite demostración: lo único que precisa es explicación *(vid.* nota a I, 9, pág. 299). En el libro V, al que dedica especial atención, Simson introduce cuatro axiomas de propia cosecha sobre las relaciones de igualdad y desigualdad en términos de «multíplices» [75] y varias proposiciones adicionales. Por lo que concierne al cuerpo de las proposiciones, los cambios debidos a Simson son de tres clases principales: a) Enunciaciones reformuladas o aclaradas (e. g., I, 2, 7, 11, 42, 44, 47). b) Pruebas modificadas o introducidas en orden a la consideración de casos (e. g., III, 13, 21, 35; proposiciones A-D añadidas a V, 6); pruebas alteradas para evitar la consideración de casos (e. g., I, 24, 35); opción por pruebas alternativas (e. g., III, 23, en línea con Clavio; V, 18, para evitar

[74] Esta sustitución de la definición euclídea por otra «más clara», según alega SIMSON en la nota correspondiente *(op. cit.*, pág. 295), no es muy afortunada; cf. T. L. HEATH, ed. cit., 1926², 1, pág. 172.

[75] Los axiomas son éstos: «I. Las cantidades equimultíplices de una misma cantidad o de cantidades iguales son entre sí iguales», «II. Las cantidades, de quienes una misma cantidad es equimultíplice o cuyas equimultíplices son iguales, son también iguales entre sí», «III. La multíplice de una cantidad mayor es mayor que la equimultíplice de una menor», «IV. La cantidad, cuya multíplice es mayor que la equimultíplice de otra, es mayor que ésta».

la suposición ilegítima del cuarto término proporcional). c) Adición de corolarios (e. g., a I, 11, 15, 32, 45).

En suma, la versión de Simson no es tanto una traducción como una edición desarrollada sobre la base del texto latino de Commandino. Su traslado al español sitúa el conocimiento hispano de los *Elementos* al nivel de la crítica matemática e histórica europea del momento. Además tiene el interés añadido de revelarnos la existencia de una terminología geométrica relativamente asentada, libre de los arcaísmos tradicionales («ambligonio», «oxigonio», etc.), e incluso provista de cierta capacidad creadora a veces más afortunada (e. g., al formar la correlación «tántupla de... cuántupla de ...», a partir de «tan multíplice... cuan multíplice»), a veces menos (e. g., en el uso de «quindecágono» pese a conocer «pentadecágono»).

Las conclusiones que podamos entresacar de todo este material serán, desde luego, más significativas para la historia de la matemática española que para la historia de los *Elementos* euclídeos. Está clara, sin ir más lejos, la existencia de tres vías al menos de «ilustración» geométrica: la institucional o de inspiración real, representada por la Casa de Contratación (Zamorano) y la Academia de Matemáticas (Rojas, Cedillo, Carduchi); la militar (Vibario, Fernández Medrano; más tarde, Larrando, Deu); la desarrollada al hilo de la implantación académica de los jesuitas (en el Colegio Imperial que absorbe la Academia: Carduchi, Kresa; en el Seminario de Nobles: Álvarez). No son líneas completamente inconexas o aisladas, pero es una lástima que no tejan la trama de una verdadera tradición y cada traductor se crea autorizado —cuando no obligado— a comenzar de nuevo para cortar el traje, a fin de cuentas, por el mismo patrón: los libros I-VI (XI-XII), una geometría básica de planos y sólidos. Sería exagerado pintar la situación con el tenebrismo dolido y provocador de un Eche-

garay o un Rey Pastor, es decir: como si la matemática española de entonces sólo hubiera producido, en esta ámbito, las populares «geometrías de sastres» [76]. Pero, desde luego, bien podemos lamentar que unos principios tan prometedores como los sentados por la traducción de Zamorano no hallaran continuación en su momento bien por su parte, bien por la de otros.

La última fase de la historia de las traducciones y versiones en español de los *Elementos* está representada por las ediciones hechas en nuestro siglo antes de la edición presente. Son las tres que siguen.

F. ENRIQUES y colaboradores, *Los Elementos de Euclides y la crítica antigua y moderna.* Libros I-IV. Traducidos de la edición italiana por José MINGOT-SHELLEY, Madrid, C.S.I.C. (Colección Textos Clásicos), 1954.

Es una versión española del vol. I de esta edición italiana (Roma, 1925) dirigida por Federigo Enriques. El propio Enriques, en el prefacio (págs. 9-11), manifiesta un doble propósito: el de contribuir a la formación en la crítica histórica y comparativa de textos mediante el examen de mo-

[76] La descalificación de la producción matemática del siglo XVII en el esos términos, *i.e.* como si se hubiera reducido a «libros de cuentas y geometrías de sastres», procede del discurso de recepción de José Echegaray en la Academia de Ciencias (11 de marzo de 1866); *vid.* la compilación de Ernesto y Enrique GARCÍA CAMARERO, *La polémica de la ciencia española*, Madrid, 1970, pág. 176. Es reiterada al pie de la letra por REY PASTOR, op. cit., 1926, pág. 133. No estaría de más, siquiera para no precipitarse en caricaturas y en juicios de valor, que una revisión general e interdisciplinaria de la historia de la matemática española de estos siglos (XVI-XVIII) acompañara a las recientes revisiones de la modernización española en otros ámbitos.

delos clásicos; el de hacer una divulgación de los *Elementos* suficientemente fiel, para apreciar el sabor de la obra griega, y suficientemente libre, al adoptar expresiones geométricas más directas y familiares para nuestros oídos. El texto griego de base es la edición crítica de Heiberg. Una breve introducción (págs. 15-21) da noticia de la tradición cultural griega y de los *Elementos* de Euclides, y se cierra con una selección bibliográfica. Los tres tipos de principios (términos, postulados, nociones comunes), en general, y casi cada uno de estos principios en particular se presentan con referencias históricas y acotaciones críticas a la luz del desarrollo moderno de la geometría. Referencias similares acompañan así mismo a muchas proposiciones de los libros I-IV. Algunas peculiaridades de esta versión pueden ser: En la última definición o «término», 23: *paralelas*, se omite la determinación «indefinidamente (o al infinito)» de la prolongación de las rectas. Entre las nociones comunes, se incluyen sin reticencias las que Heiberg juzga interpoladas; la noción común 7 es singular: «Y las unidades (?) de una misma cosa son iguales entre sí» y la 8 tiene una lectura demasiado sabia: «Y las cosas que se superponen una a la otra son iguales entre sí»; en fin, a propósito de 9: «Y dos rectas no comprenden un espacio», Enriques considera que está posiblemente trastocada y debería ponerse entre los postulados como complemento del postulado 1 (por lo demás, la asume sin reparos en la prueba de I, 4). Esta edición, en general, guarda un lejano aire de familia con la edición de Heath, autor al que alguna vez también se cita. Pero las deudas de Enriques con la historiografía tradicional son más acusadas: mantiene la atribución del punto-mónada al pitagorismo antiguo (en la onda de Tannery); la relación de los postulados con las construcciones y demostraciones de existencia (en la onda

de Zeuthen); la afinidad del genio griego con el ideal de finitud-perfección (en la onda de los que venían popularizando la imagen de una Weltanschauung clásica apolínea). La traducción española, por último, no siempre es muy afortunada.

Euclides, *Obras completas. Elementos de Geometría.* Vol. I: Libros I y II. Versión directa, introducción y notas de J. David García Bacca, México, UNAM [Bibliotheca Scriptorum Graecorum et Romanorum Mexicana], 1944. Vol. II: Lib. III-V. Versión, prólogo y notas de José Álvarez Lasso, México, UNAM, 1956.

El vol. I comienza con una «Introducción filosófica» de García Bacca (págs. ix-xcii). Sigue una versión parcial y selectiva de D. Hilbert (1899): *Fundamentos de Geometría* (págs. 31-32). Luego, tras unas notas y advertencias sobre la traducción (págs. 37-40), aparece la edición bilingüe (en griego y español) de los dos primeros libros de los *Elementos* (pág. 4 ss.): el texto griego y el aparato crítico incluido a pie de página están tomados de Heiberg; pueden confrontarse con la traducción impresa en la página opuesta, donde salta a la vista el interés de García Bacca por explicitar cada uno de los pasos que conforman el proceso deductivo de las proposiciones. Las notas se remiten al final (págs. 167-178): en las noticias y comentarios se deja sentir el peso de Enriques.

El vol. II se inicia con un prólogo (págs. ix-xvii) que asegura la continuidad de la labor de Álvarez Lasso con el patrón establecido en el vol. I por García Bacca. Y efectivamente, se mantienen la edición bilingüe, la reproducción del texto y aparato de Heiberg, su traducción correlativa y el despliegue metódico de las proposiciones. Las notas (págs. 231-238), aunque suelen ser más circunspectas, añaden a la información procedente de Enriques algunas referencias a Simson.

Esta edición tiene dos rasgos llamativos: uno es el calado especulativo de la introducción filosófica del propio García Bacca —cuyas contribuciones a la filosofía actual de habla hispánica no es preciso resaltar aquí—. García Bacca ensaya una contextualización de tres formas o fases principales de la geometría («intuitiva» o preeuclídea, «formal» o euclídea, «transcendental» o no euclidiana) por referencia a ciertas categorías conceptuales: no es nada fácil reconocer la historia real en este teatro de ideas. A continuación dedica otro apartado al «Mundo euclídeo» con notables intuiciones y sugerencias, aunque el autor recaiga en el tópico de «la finitud y la delimitación» —como una presunta característica esencial de la mentalidad helénica—, y acuse el lastre de la hermenéutica tradicional heredada del pasado siglo. El segundo rasgo notable de esta edición es la pretensión —compartida por ambos traductores— de sacar a la luz y definir cabalmente la lógica que suponen subyacente en las pruebas euclídeas. Ahora los supuestos interpretativos que guían el análisis se tornan por su propio carácter técnico algo más precisos, pero no siempre resultan más verosímiles: por ejemplo, no tienen mucho fundamento ni el nexo alegado entre la deducción euclídea y la lógica estoica, ni la atribución a Euclides del uso de una regla de la moderna lógica cuantificacional como la regla de sustitución.

Por lo que se refiere a la versión del texto euclídeo, se plantea como una empresa rigurosa y de notable esfuerzo. Cabe anotar la inclusión formal de la noción común IX: «Y dos rectas no circundan región», y su empleo en I, 4, por más que estén desautorizados en el texto griego correlativo por el aparato crítico de Heiberg; lo mismo ocu-

rre en algún otro caso, como el de la proposición I, 40, que Heiberg (1903) había descartado a la luz del papiro Fayûm. Pero los autores no se paran a explicar los posibles desajustes de este tipo entre la fuente crítica griega y su versión española. La traducción se mueve, en general, entre la presión de la literalidad y la presión de la explicitación moderna. La primera se trasluce en algunos giros y expresiones de un español un tanto pintoresco. La segunda se advierte en ciertas versiones sesgadas o sabias, e. g., Def. 1: «punto es lo que ya no tiene partes» (cf. nota 1, pág. 167); noción común VII: «Y las cosas congruentes entre sí son iguales entre sí». En cualquier caso, el título general de la edición anuncia una obra todavía en ciernes pero de singular empeño dentro de la (escasa) literatura euclídea disponible en nuestra lengua.

Científicos griegos. Recopilación, estudio preliminar, preámbulos y notas de Francisco VERA, Madrid, Aguilar (Col. Grandes Obras), 1970; 2 vols. Euclides: *Elementos de Geometría,* vol. I, págs. 702-980.

Francisco Vera se ha dado a conocer especialmente por sus contribuciones al desarrollo de la historiografía de la matemática española. Esta versión viene precedida de un preámbulo con noticias sobre Euclides y sus obras, donde Vera, naturalmente, presta especial atención a la significación de los *Elementos* y a sus ediciones y versiones; hay al final una selección bibliográfica —no muy accesible ni actualizada—. Es una lástima que algunas erratas de imprenta den lugar a confusiones y a equivocaciones de bulto. Por fortuna, la versión de los *Elementos* ha tenido una impresión mucho más cuidada.

No se menciona el texto griego tomado como fuente de la traducción. Parece tratarse, en general, de la edición

de Heiberg; aunque Vera no reconoce la interpolación de las nociones comunes 4-6 y 9 —a propósito de esta última dice (nota 21, pág. 705): «Algunos editores, con excelente criterio, consideran esta noción común como un sexto postulado que completa el de paralelismo»—; tampoco hay ninguna observación acerca de otras proposiciones intercaladas, como I, 40. En las notas abundan las referencias a Enriques. Dentro ya del contexto de las proposiciones del libro I, el traductor declara seguir escrupulosamente los *Elementos* habiéndose limitado «a aligerar un poco el texto, suprimiendo palabras superfluas y, desde luego, la inútil repetición literal del enunciado antes de decir que era lo que se quería demostrar» (nota 26, pág. 708). El resultado es, sin embargo, una versión libre, sintética y modernizada por regla general. Esta versión se ajusta sustancialmente al texto estándar en los libros I-VI. Pero a partir de VII se acentúan las distancias motivadas por la simplificación; en la nota 1 (pág. 829), Vera habla de la «inelegancia» de algunas demostraciones, de la «dificultad de seguir el camino del razonamiento a través de la manigua de palabras»: es un aviso de que nos vamos a encontrar proposiciones con pruebas abreviadas y enunciados sin prueba. Al parecer, entre los motivos que obran en estas omisiones se cuentan: a) la sustitución de la prueba euclídea por una formulación moderna o por la indicación del procedimiento a seguir (e. g., VII, 3, 6-9, 11, 15, 19-22, 27-29, 36); b) la analogía de la prueba omitida con alguna otra anterior (e. g., VII, 2, 12, 33); c) el ser un teorema recíproco o converso de la proposición anterior (e. g., VII, 30); d) ningún motivo en especial (e. g., VII, 10, 17-18, 32, 35, 37-39). La situación en los libros VIII-IX es muy similar. Parece agravarse en el libro X (págs. 860 ss.), donde el editor alude a su importancia como documento histórí-

co, al tiempo que recuerda la posibilidad de obtener sus resultados por medio de unas pocas reglas algebraicas: en suma, la tercera parte de las proposiciones del libro se queda en los puros enunciados; ni una mención siquiera llega a merecer el sustancial corolario de X, 111 sobre las clases de rectas irracionales. La situación mejora en XI-XIII, aunque no falten omisiones ocasionales de pruebas, bien porque se juzgan innecesarias (e. g., XI, 7; XII, 6, 10, 17-18), bien por su analogía con pruebas anteriores (e. g., XI, 8, 34; XII, 8, 11; XIII, 2). Por lo demás, del tratado ha desaparecido la tradicional alusión a los principios o a proposiciones previas en determinados pasos de las pruebas y no se respeta mucho la lógica interna de las proposiciones euclídeas. En realidad, esta edición de los *Elementos de Geometría* de Euclides no constituye una traducción de los *Elementos*, sino una versión clara, suelta y aproximada al lector de hoy, con las simplificaciones que quizás se consideran convenientes en el marco de una edición antológica que procura facilitar el acceso del público hispano a los textos científicos griegos.

UNA SELECCIÓN BIBLIOGRÁFICA

En las notas ya he dado referencias bibliográficas sobre aspectos generales y sobre puntos concretos relacionados con Euclides y, en particular, con los *Elementos*. Por eso en esta breve selección final me limitaré a recordar algunos títulos de especial interés; serán, además, estudios que en mayor o menor medida difieren entre sí —aunque puedan complementarse—, con el fin de evitar duplicaciones innecesarias. Distribuiré estas obras en tres grupos:

A. Para hacerse una idea del contexto general en que aparecen los *Elementos*, cabe acudir en principio a una selección de textos matemáticos griegos, como la edición bilingüe (en griego e inglés) de I. THOMAS, ed., *Selections of Greek Mathematics*, Londres-Cambridge (Mass.), 1939, 1967 3.ª reimp.; también puede verse J. FAUVEL y F. GRAY, eds., *The History of Mathematics: A Reader*, Basingstoke-Londres, 1987, 2-6, págs. 46-239. El estudio de P. TANNERY (1887), *La géométrie grecque. 1: Histoire de la géométrie élémentaire*, Hildesheim-Zurich-Nueva York, 1988, reimp., tiene la intuición y los encantos del pionero. Pero hoy debería cotejarse con T. L. HEATH (1921): *A history of Greek Mathematics*, Nueva York, 1981, reimp., 2 vols.; CH. MUGLER, *Dictionnaire historique de la terminologie géométrique des Grecs*, París, 1958-1959 (disponible en microfichas: Klincksieck EC 28-9). También habrá que tener en cuenta los diversos aspectos de las tradiciones previas y del marco de los *Elementos* que han sugerido o desarrollado A. SZABÓ (1969): *The Beginnings of Greek Mathematics*, Dordrecht-Budapest, 1978; W. R. KNORR, *The Evolution of the Euclidean Elements*, Dordrecht-Boston, 1975, y *The Ancient Tradition of Geometric Problems*, Boston, 1986; D. H. FOWLER, *The Mathematics of Plato's Academy. A New Reconstruction*, Oxford, 1987. Acerca de las teorías, las tradiciones y las prácticas griegas de la prueba, puede verse L. VEGA, *La trama de la demostración. (Los griegos y la razón tejedora de pruebas)*, Madrid, 1990.

B. Sobre Euclides y su obra, siempre se encontrará información en cualquier Historia de las matemáticas. En español se puede utilizar, por ejemplo. C. B. BOYER (1968): *Historia de la matemática*, Madrid, 1986, cap. VII, págs. 141-163. Pero son singularmente instructivas las síntesis monográficas de F. O. HULTSCH, «Eukleides» en la PAULY-WISSOWA, *Real Enzyclopädie der klassischen Altertumswissenschaft*, Stuttgart, 1909, reimp. 1970, VI 1, cols. 1003-1052; I. BULMER-THOMAS, «Euclid, Life and Works», en CH. C. GILLISPIE, dir., *Dictionary of Scientific Biography*, Nueva York, 1970-1980, reimp. 1981, vol. 4, págs. 414-437; D. GJERS-

TEN, *The Classics of Science*, Nueva York, 1984, cap. 3, págs. 38-70; A. DOU, «Euclides», en AA.VV., *Historia de la matemática hasta el siglo XVIII*, Madrid, 1986, págs. 61-78.

C. En fin, por lo que concierne al texto y a la constitución interna de los *Elementos*, conviene recordar los trabajos clásicos de J. L. HEIBERG, «Paralipomena crítica», en la edición J. L. HEIBERG, H. MENGE, *Euclidis opera omnia*, tomo V (Leipzig, 1888), págs. xxiii-cxiii [recogido en la edición rev. de E. STAMATIS, *Euclidis Elementa*, tomo V (Leipzig, 1977), págs. xvi-lxxxix]; y «Paralipomena zu Euklid», *Hermes* 38 (1903), 46-74, 161-201, 321-356. De las observaciones y los procedimientos de Heiberg, también da noticia la fuente de información más completa que hoy existe sobre los *Elementos*, la edición inglesa de T. L. HEATH (1908, 1926): *The Thirteen Books of Euclid's Elements*, Nueva York, 1956, 3 vols. Algunas referencias han de actualizarse; por ejemplo, a la luz de J. E. MURDOCH, «Euclid. Transmission of the *Elements*», en CH. C. GILLISPIE, dir. (1970-1980, reimp. 1981), *Dictionary of Scientific Biography*, op. cit., vol. 4, págs. 437-459. La reconstrucción más comprensiva de la estructura deductiva interna de los *Elementos* es la ofrecida por I. MUELLER, *Philosophy of Mathematics and Deductive Structure in Euclid's Elements*, Cambridge (Mass.)-Londres, 1981.

LUIS VEGA

ELEMENTOS

NOTA SOBRE LA PRESENTE TRADUCCIÓN

La presente traducción sigue el texto griego fijado por J. L Heiberg en la edición de J. L. Heiberg y H. Menge, *Euclidis Opera omnia,* vols. I-IV, Leipzig, 1883-1886. También he utilizado la revisión hecha por E. S. Stamatis, *Euclidis Elementa,* vols. I-IV, Leipzig, 1969-73. El texto de Stamatis es ligeramente más conservador. En los muy contados casos de discrepancia entre Heiberg y Stamatis asumo las atétesis justificadas por Heiberg.

Dada la importancia sustancialmente histórica que hoy reviste para nosotros el lenguaje de los *Elementos,* he procurado mantener su expresividad original, en la medida de lo posible, y me he guardado de dar la falsa imagen de una terminología técnica que corresponderá más bien a momentos posteriores de desarrollo de la matemática, o de producir la falsa impresión de que el texto compartía ya la terminología estándar que hoy nos resulta familiar.

LIBRO PRIMERO

DEFINICIONES

1. Un punto es lo que no tiene partes [1].
2. Una línea es una longitud sin anchura.
3. Los extremos de una línea son puntos.

---

[1] La definición recoge la idea tradicional de punto como aquello que es indivisible en partes. Pero no incurre en el vicio que Aristóteles atribuye a las definiciones habituales en su tiempo, el de definir lo anterior por referencia a lo posterior: el punto como límite de la línea, la línea como límite de la superficie, la superficie como límite del cuerpo sólido (*Tóp.* VI 4, 14b15-27). También es significativa la opción de Euclides por el término *sēmeîon* para designar el punto. Los pitagóricos habían legado el término *mónas* para indicar la unidad tanto aritmética como geométrica: cuando el contexto lo exigía, la unidad aritmética podía precisarse como *mónas áthetos* (unidad sin posición) y la unidad geométrica como *mónas thésin ékhousa* (unidad con posición) (*vid.* ARISTÓTELES, *Metaf.* Δ 6, 1016b14-26; PROCLO, *Com.* 95, 12). Pero en la primera mitad del siglo IV a. C. habían empezado a imponerse otras dos denominaciones del punto tomadas al parecer del lenguaje corriente: *stigmḗ*, que significaba originariamente «marca», «punción», y *sēmeîon*, que tenía el significado de «signo», «señal» y otros asociados (e. g., «presagio», «síntoma»). Al parecer, en la tradición filosófica prevalece más bien el uso del primero, y en la tradición matemática el uso del segundo. Aristóteles suele emplear *stigmḗ* en el contexto de discusiones filosóficas (e. g., *Me-*

4. Una línea recta es aquella que yace por igual respecto de los puntos que están en ella [2].

taf. A 9, 992a20-24; *Fís.* 2, 1, 231a24-26) y recurrir a *sēmeîon* cuando cita proposiciones geométricas (e. g., *Meteor.* Γ 3; 373a3-5; 5, 376a8-9). Por otra parte, los grandes textos geométricos del siglo III, empezando por los *Elementos*, emplearán *sēmeîon* de forma casi sistemática —un uso parecido se encuentra ya en los tratados astronómicos de Autólico—. Más tarde, en los escritos de los comentadores, se olvida esta especialización técnica y reaparece el uso de *stigmḗ* en los contextos geométricos. Con todo, Alejandro de Afrodisia todavía puede recordar las vicisitudes de la tradición: «Los elementos primeros del plano son las líneas, y de las líneas los puntos [*stigmaí*], que los matemáticos llaman *puntos* [*sēmeîa*] y ellos (los pitagóricos) mónadas [*mónades*]» (*In Metaf.* 55, 20; Ross, *Arist. fragm.*, fr. 2). Para más detalles, *vid.* V. Vita, «Il punto nella terminologia matematica greca», *Archive for History of Exact Sciences* 27-2 (1982), 101-114; sobre otras definiciones antiguas y modernas del punto geométrico, *vid.* T. L. Heath, *The Thirteen Books...*, ed. cit., 1926², 1, págs. 155-158. A propósito de éste y de otros muchos términos, también puede cotejarse Ch. Mugler, *Dictionnaire historique de la terminologie géométrique des Grecs*, París, 1959.

[2] Los griegos se formaron tres representaciones básicas de la línea recta: la de un hilo tenso, la de un rayo de luz, la de un eje o lugar de los puntos que se mantienen inmóviles en un cuerpo fusiforme suspendido por ambos extremos, *vid.* Ch. Mugler, «Sur l'histoire de quelques définitions de la géométrie grecque», *L'antiquité classique* 26 (1957), 331-345. Las tres imágenes se combinan para describir la luz celeste y el huso de la necesidad en el mito de Er de la *República* de Platón (X 616b-c). De hecho, estas representaciones subsisten en algunas caracterizaciones de la línea recta que hoy conservamos de los griegos. La tensión [*tásis*] está asociada a la definición de la recta como una línea tendida o estirada [*tetaménḗ*] hacia los puntos (Proclo, *Com.* 110, 18) o hacia los extremos (Pappo, *Def.* 4). La imagen del rayo óptico está presente en la conocida definición platónica de la recta como la línea cuyo medio intercepta (eclipsa [*epíprosthen*]) ambos extremos, vid. *Parménides* 137e. La tercera imagen se trasluce en la definición de Herón recogida por Proclo: «una línea que permanece fija cuando sus extremos permanecen fijos» (*Com.* 110, 21-22). Suele considerarse que la definición de Euclides es una elaboración de la platónica, pues ésta contendría implícitamente

5. Una superficie es lo que sólo tiene longitud y anchura [3].

una alusión al sentido de la vista y supondría, asimismo, una asimilación del rayo visual al rayo óptico, connotaciones que Euclides procura evitar. *Vid.*, por ejemplo, T. L. Heath, ed. cit., 1, 1926², pág. 166; en cambio, Mugler, art. cit., 1957, pág. 336, discute esta implicación visual de la definición de Platón. En cualquier caso, la definición euclídea parece original, pues, aparte de la definición citada del *Parménides* de Platón, sólo hay constancia de otra definición anterior, recordada y criticada por Aristóteles: la línea es el límite de la superficie (*Tóp.* VI 4, 141b21). También ha llamado la atención la construcción lingüística de la definición: *ex ísou toîs eph' heautês sēmeíois keîtai*, en particular el empleo de la expresión adverbial *ex ísou*. Es una expresión frecuente en Platón y en Aristóteles, y suele tener el sentido de «por igual», «en pie de igualdad». Cabe considerar que el dativo *toîs... sēmeíois* está relacionado con *ex ísou* o con *keîtai:* en el primer caso, resulta la versión adoptada aquí; en el segundo caso se diría más bien que la línea recta está simétricamente determinada por sus puntos, noción que envuelve un aspecto de dirección. Los geómetras griegos, efectivamente, se refieren a la línea AB o a la línea BA como a una misma línea. Pero es dudoso que esta identificación intuitiva descanse en alguna suposición de simetría o de reversibilidad —la conciencia de éstos y otros presupuestos lógicos sólo irá aflorando con el desarrollo moderno de la geometría euclidiana—. Por lo demás, esta definición euclídea de línea recta es un intento de explicar una noción tan simple que se resiste a una formulación cabal y precisa, y la misma suerte espera a otros ensayos en el mismo sentido. No obstante, si alguna definición griega ha tenido especial fortuna, ha sido la aportada por Arquímedes como una de las asunciones de *Sobre la esfera y el cilindro:* «la recta es la más corta [*elakhístēn*] de todas las líneas que tienen los mismos extremos».

[3] La palabra *epipháneia* fue usada por Euclides y escritores posteriores para denotar «superficie» en general, mientras que *epípedon* se utilizó para significar «superficie plana» (vid. *infra*, Def. 7). *Epipháneia* parece haber designado originariamente la apariencia visible de un cuerpo sólido; los pitagóricos llaman a la superficie *(chro(i)á, i. e.* «piel», como testifica Aristóteles (*Acerca de la sensación* 3, 439a31), aunque olvida el significado anterior de *chro(i)á* y le da el de *color*, que había prevalecido posteriormente. Con todo, el primer testimonio del uso de *epipháneia*

6. Los extremos de una superficie son líneas.
7. Una superficie plana es aquella que yace por igual respecto de las líneas que éstan en ella [4].
8. Un ángulo plano es la inclinacion mutua de dos líneas que se encuentran una a otra en un plano y no están en linea recta [5].

---

en un contexto geométrico bien podría ser el del frag. 68 B 155 de Demócrito. Platón y Aristóteles emplean indistintamente este término y *epipedon* para designar la superficie. El uso de *epiphaneia* se fija con Autólico y Euclides. La definición sienta su carácter bidimensional y la Def. 2 del libro XI establecerá su condición de límite del cuerpo sólido: ambos aspectos de la superficie eran familiares para Aristóteles *(vid.*, e. g., *Top.* VI 4, 141b22; *Metaf.* M 9, 1085a11).

[4] Esta definición, como puede apreciarse fácilmente, está calcada sobre la Def. 4 *mutatis mutandis:* «línea recta» y «puntos» son reemplazados por «superficie plana» y «rectas», respectivamente. Por lo demás, no es extraño que las nociones asociadas a la línea recta se quieran extender a la superficie plana, e. g.: la superficie plana es «la tendida o estirada», «la menor superficie de todas las que tienen los mismos límites» *(vid.* PROCLO, *Com.* 117, 2-13). Sobre los problemas derivados y otras cuestiones conexas, *vid.* T. L. HEATH, ed. cit., 1926², 1, págs. 172-176.

[5] La expresión *mè ep' eutheias keiménōn*... («y no están en linea recta») resulta sorprendente, teniendo en cuenta que el ángulo puede estar formado por rectas o curvas. Parece como si Euclides hubiera pensado al principio definir el ángulo rectilíneo y hubiera tomado después en consideración los ángulos formados por una recta y una curva o por dos curvas, con lo que se ve llevado a sustituir «líneas rectas» por «líneas».

Al parecer, el uso de *klísis* «inclinación» es una innovación de Euclides, frente a la tradición anterior presente en Aristóteles que relacionaba el ángulo con la idea de «fractura» *klásis.* Para otras definiciones, *vid.* HEATH, ed. cit., 1926², 1, págs. 176 ss. Por otra parte, Euclides considera solamente los ángulos convexos (menores que dos rectos), lo mismo que en la definición 19 sólo se refiere a polígonos convexos. Aunque los griegos sólo tomaron en consideración los ángulos y polígonos convexos, se encuentra la distinción entre las dos clases en la definición 34 de Herón y en el comentario de Proclo (PROCLO, *Com.* 165, 21) se trata de polígonos cóncavos con ángulos entrantes.

9. Cuando las líneas que comprenden el ángulo son rectas el ángulo se llama rectilíneo.
10. Cuando una recta levantada sobre otra recta forma ángulos adyacentes [6] iguales entre sí, cada uno de los ángulos iguales es recto y la recta levantada se llama perpendicular a aquella sobre la que está.
11. Ángulo obtuso es el (ángulo) mayor que un recto.
12. Ángulo agudo es el (ángulo) menor que un recto.
13. Un límite es aquello que es extremo de algo [7].
14. Una figura es lo contenido por uno o varios límites.
15. Un círculo es una figura plana comprendida por una línea [que se llama circunferencia] tal que todas las rectas que caen sobre ella desde un punto de los que están dentro de la figura son iguales entre sí [8].

---

[6] *Ephexês, i. e.* «en serie», «uno tras otro» es la palabra utilizada generalmente para ángulos adyacentes.

[7] Límite: *hóros;* extremo: *péras.*

[8] *Hè kaleîtai periphéreia* («que se llama circunferencia») se considera una interpolación, al igual que la frase no traducida: *pros tén toû kyklou periphéreian* («hacia la circunferencia del círculo»), que aparece en algunos mss. hacia el final de la definición. Heiberg las suprime porque, a pesar de que esos mss. las tienen, las fuentes más antiguas (Proclo, Tauro, Sexto Empírico, Boecio) las omiten. El descubrimiento del papiro Herculanense núm. 1061 que presenta también la definición sin las palabras en cuestión confirma la postura de Heiberg *(vid.* J. L. HEIBERG, «Paralipomena zu Euklid», *Hermes* XXXVIII [1903], 47). Las palabras fueron añadidas seguramente debido a la aparición de *periphéreia* «circunferencia» en las definiciones 17 y 18 sin ninguna explicación. Pero en realidad no es necesaria. Pues, aunque la palabra no aparece en Platón, Aristóteles la utiliza varias veces:

1. En el sentido general de «contorno», sin un significado específicamente matemático.
2. En sentido matemático, para referirse tanto a la circunferencia como a un arco o a un círculo.

Por eso Euclides podía usar perfectamente la palabra en las definiciones 17 y 18 sin definirla, porque se trataba de una palabra universalmente

16. Y el punto se llama centro del círculo.
17. Un diámetro del círculo es una recta cualquiera trazada a través del centro y limitada en ambos sentidos por la circunferencia del círculo, recta que también divide el círculo en dos partes iguales [9].
18. Un semicírculo es la figura comprendida entre el diámetro y la circunferencia por él cortada. Y el centro del semicírculo es el mismo que el del círculo.

---

conocida y que no tenía de suyo referencia matemática. Debe añadirse que Al-Nayrīzī no incluye en su texto las palabras suprimidas por Heiberg e intenta explicar la omisión de la palabra circunferencia.

Por otra parte, la definición de círculo no contiene nada nuevo en sustancia. PLATÓN en *Parménides* 137e dice: *stróngylón ge poú esti toûto, hoû àn tà éschata pantachêi apò toû mésou ison apéchei*: «Redondo es aquello cuyos extremos están en todas las direcciones a igual distancia del medio». En Aristóteles encontramos las siguientes expresiones: «La figura plana rodeada por una línea circular» *peripherógramon (De caelo* II 4, 286b13-16); «Plano regular desde un centro [*epípedon tò ek toû mésou ison*]» *(Retórica* III 6, 1407b27) refiriéndose al círculo. Compara también con el círculo «cualquier otra figura que no tenga iguales las líneas desde el centro, por ejemplo una figura ovalada» *(Sobre el cielo* II 4, 287a19).

La palabra *kéntron*, «centro» se usaba también regularmente: *vid.* la referencia de Proclo a los *lógia* «oráculos» que declaran: «el centro desde el que todas las (líneas tendidas) hasta el borde son iguales...» *(Com.* 155, 4-5).

[9] *Diámetros* es la palabra empleada regularmente por Euclides también para cuadrados y paralelogramos. *Diagṓnios*, «diagonal» es un término tardío definido por HERÓN *(Def.* 67) como la línea recta dibujada de un ángulo a otro. Las últimas palabras «recta que también divide al círculo en dos partes iguales» son omitidas por Simson y los editores que le siguen. Pero son necesarias aunque no pertenecen a la definición sino que sólo expresan una propiedad del diámetro tal como ha sido definido. Pues, sin esta explicación, Euclides no habría tenido justificación al definir un semicírculo como una porción de círculo limitada por el diámetro y la circunferencia cortada por él.

19. Figuras rectilíneas son las comprendidas por rectas, triláteras las comprendidas por tres, cuadriláteras las comprendidas por cuatro, multiláteras las comprendidas por más de cuatro rectas [10].
20. De entre las figuras triláteras, triángulo equilátero es la que tiene los tres lados iguales, isósceles la que tiene sólo dos lados iguales, y escaleno la que tiene los tres lados desiguales [11].
21. Además, de entre las figuras triláteras, triángulo rectángulo es la que tiene un ángulo recto, obtusángulo la que tiene un ángulo obtuso, acutángulo la que tiene los tres ángulos agudos.
22. De entre las figuras cuadriláteras, cuadrado es la que es equilátera y rectangular, rectángulo la que es rectangular pero no equilátera, rombo la que es equilátera pero no rectangular, romboide la que tiene los ángulos y lados opuestos iguales entre sí, pero no es equilátera ni rectangular; y llámense trapecios las demás figuras cuadriláteras [12].

---

[10] La distinción entre figuras triláteras, cuadriláteras y multiláteras es probablemente una aportación del propio Euclides. *Tripleuron, tetrápleuron, polýpleuron* no aparecen en Platón ni en Aristóteles. Con el uso de *tetrápleuron*, Euclides acaba con la ambigüedad del término *tetrágōnon* que empieza a utilizarse restrictivamente para el *cuadrado*.

[11] *Isoskelés* («de piernas iguales») es usado por Platón y Aristóteles. *Skalenós* con su variante *skalenés* lo usa Aristóteles para referirse a un triángulo que no tiene dos lados iguales. Platón utiliza el término *scalenós* para un número impar por oposición a *isoskelés* para un número par. Proclo pone en relación *scalenós* con el verbo *skádsō* («cojear»). Otros lo relacionan con *skoliós, i. e.* «torcido», «defectuoso».

[12] Cuadrado: *tetrágōnon*. *Tetrágōnon* designaba ya el cuadrado entre los pitagóricos *(vid.* Aristóteles, *Metafísica* 986a26). Y este mismo significado es el más frecuente en Aristóteles; pero en *Sobre el alma* II 3, 414b31, parece significar «cuadrilátero»; por otra parte, la alusión de *Metafísica* 1054b2, a *tetrágōna* «iguales» y «equiangulares» ha de referirse a cuadri-

23. Son rectas paralelas las que estando en el mismo plano y siendo prolongadas indefinidamente en ambos sentidos, no se encuentran una a otra en ninguno de ellos [13].

láteros para que «equiangular» alcance a tener algún sentido. Aunque al introducir *tetrápleuron* para designar el cuadrilátero Euclides permitió evitar la ambigüedad, hay huellas del antiguo uso de *tetrágōnon* en autores mucho más tardíos.

La palabra *rómbos* («rombo») parece derivar de *rémbō* («dar vueltas») y se utilizaba para designar, entre otras cosas, la peonza. Arquímedes usa la expresión «rombo sólido» para referirse a la figura sólida formada por dos conos con una base circular común. Si los conos son iguales, la sección a través del eje común podía ser un rombo plano. Aunque en Arquímedes los conos no tienen por qué ser iguales, es posible que el sólido del que fue tomado el nombre originariamente estuviera formado por dos conos iguales y Arquímedes lo habría extendido a otros casos.

[13] *Parállēloí eisin eutheîai, haítines en tôi autôi epipédōi oûsai kai ekballómenai eis ápeiron eph' ekátera tà mérē epi mēdétera sympíptousin allēlais. Parállēlos*, como tal término compuesto, no aparece en Platón, pero ya tiene un uso habitual en tiempos de Aristóteles. En «*ekballómenai eis ápeiron* (siendo prolongadas *indefinidamente*)», la expresión *eis ápeiron* se debe tomar en el sentido adverbial indicado, no como si designara una región o apuntara hacia un lugar («al infinito»); lo mismo cabe decir de la expresión *ep' ápeiron* que a veces sustituye a la anterior. «En ambos sentidos» es la versión de *eph' ekátera tà mérē*, cuya traducción menos libre sería: «hacia ambas partes (por ambas partes)».

Euclides, como ya se señalaba en la introducción general, opta por la característica de no intersección o no encuentro como criterio determinante de las rectas coplanarias paralelas. No era la única opción conocida por los griegos ni, al parecer, fue la más frecuente. A primera vista resultaba más natural el criterio de equidistancia, preferido por Posidonio o Gémino, entre otros (PROCLO, *Com.* 176, 5 ss.). Conforme a este criterio, son paralelas las rectas coplanarias que no convergen ni divergen y, así, todas las perpendiculares trazadas de una a otra son iguales. También se ha especulado con un posible criterio de igual dirección, mencionado por Filopón (*vid.* T. L. HEATH, ed. cit., 1926², 1, págs. 191-192). En cualquier caso, la opción de Euclides por este criterio de no intersección y la asunción complementaria del postulado 5, con ser menos intuitivas, son más adecuadas y evitan la falacia de probar la existencia de rectas paralelas sobre la base de algún supuesto derivado justamente de esa misma existencia —Aristóteles había denunciado una petición de principio de este tipo en planteamientos coetáneos de las paralelas, *A. Pr.* II 16, 65a1-9—. El criterio de no intersección y el criterio de equidistancia convenientemente formulados, resultan lógicamente equivalentes. *Vid.* R. J. TRUDEAU, *The Non-Euclidean Revolution*, Boston-Basilea-Stuttgart, 1987, cap. 4, págs. 118 ss.

POSTULADOS

1. Postúlese el trazar una línea recta desde un punto cualquiera hasta un punto cualquiera [14].
2. Y el prolongar continuamente una recta finita en línea recta [15].
3. Y el describir un círculo con cualquier centro y distancia [16].
4. Y el ser todos los ángulos rectos iguales entre sí.
5. Y que si una recta al incidir sobre dos rectas hace los ángulos internos del mismo lado menores que

[14] *Apò pantòs sēmeíou epì pân semeîon*, lit. dice literalmente: «de todo punto a todo punto». Los griegos, en afirmaciones generales de este tipo, no decían, como solemos hacer nosotros, «cualquier punto», «cualquier triángulo», etc., sino «todo punto», «todo triángulo», etc.

[15] Traduzco *peperasméné* por «finita» porque los términos «delimitada» o «limitada», aplicados a una recta, se refieren más bien a lo que nosotros entendemos por un segmento de recta.

[16] *Diástēma*, «distancia», es la palabra usada en geometría para el radio determinado con que se describe un círculo. Su uso se debe a que todos los puntos de la circunferencia están a la misma distancia del centro. Los griegos no tienen un término equivalente a *radio*, normalmente utilizan la expresión *hai ek toû kéntrou (agómenai grammaí), i. e.* «las (líneas trazadas) desde el centro». El término «radio», *radius*, aparece por primera vez en Pierre de la Ramée (P. Ramus, siglo XVI) y se usará ya comúnmente desde Vieta. Por lo demás, en geometría esférica, *diástēma* hace referencia no sólo al círculo, sino también a la esfera. (AUTÓLICO, 6).

dos rectos, las dos rectas prolongadas indefinidamente se encontrarán en el lado en el que están los (ángulos) menores que dos rectos [17].

[17] Ninguna proposición de los *Elementos* ha tenido una vida tan agitada como la de este célebre postulado, según se apuntaba ya en la introducción general. Por lo regular, ha corrido el albur de un estatuto incierto. Fue asumido como postulado por la tradición adelardiana y por Campano, así como por varios editores y comentadores renacentistas, e. g.: ZAMBERTI (1505), LUCA PACCIOLI (1509), TARTAGLIA (1543), COMMANDINO (1572). Otros muchos, quizás a partir de la *editio princeps* de GRYNAEUS (1533), prefieren incluirlo entre las nociones comunes, e. g.: F. CANDALLA (1556), CLAVIO (1574), VITALE (1682), GREGORY (1703). Pero el trance más radical, el de pasar por una proposición necesitada de prueba, fue una amenaza que se cernió sobre él desde un principio. «Debe ser borrado por completo de los postulados porque se trata de un teorema henchido de dificultades, que Tolemeo se propuso resolver en un libro, y su demostración requiere varias definiciones y teoremas. Más aún: La proposición conversa es efectivamente demostrada por el propio Euclides como un teorema» (PROCLO, *Com.* 191, 21 ss.). Proclo alude al parecer al teorema I 17: la suma de dos ángulos cualesquiera de un triángulo es menor que dos rectos, pues el postulado 5 equivale a decir que las rectas, al llegar a encontrarse por el lado correspondiente a los ángulos cuya suma es menos que dos ángulos rectos, forman un triángulo. «En el caso presente —continúa Proclo un poco más adelante—, el hecho de que las rectas convergen cuando los ángulos rectos son minorados, es cierto y necesario; por contra, la afirmación de que como convergen más y más a medida que se prolongan, llegarán alguna vez a encontrarse, es una afirmación verosímil pero no es necesaria a falta de un argumento que pruebe que esto es verdad acerca de las líneas rectas. Pues el hecho de que haya algunas líneas que se aproximan indefinidamente pero permanecen sin tocarse [*asýmptotoi*], por más improbable y paradójico que parezca, también es cierto y está completamente comprobado en relación con líneas de otro tipo. ¿Por qué en el caso de las rectas no es posible lo mismo que ocurre con las líneas mentadas?» (*Com.* 192, 13-22). Sobre las fases principales del intento fallido de reducir el postulado, hasta desembocar en el alumbramiento de las geometrías no euclidianas, puede recordarse la nota (19) de la introducción general. Aquí parece más indicado mencionar algunas de las proposiciones, lógicamente equivalentes al postulado euclideo, que se fueron explicitando a lo largo del proceso. Se suelen destacar las siguientes (junto con algún portavoz caracterizado):

(i) La suma de los ángulos de cualquier triángulo es igual a dos ángulos rectos (resultado cuya relación con la teoría de las paralelas ya era conocida en tiempos de Aristóteles, cf. *Elementos* I 32; SACCHERI, 1733).

(ii) Las rectas paralelas son equidistantes (atribuido a Posidonio); todos los puntos equidistantes de una línea recta, situados a un lado determinado de ella, constituyen una línea recta (CLAVIO, 1574).

(ii') Dos rectas paralelas guardan entre sí una distancia finita (Proclo).

(ii'') Las rectas no equidistantes convergen en una dirección y divergen en la opuesta (THĀBIT IBN QURRA, h. 826-901; CATALDI, 1603).

(iii) Por un punto exterior a una recta sólo cabe trazar una paralela (TOLEMEO; ALHAZEN, h. 965-1041; popularizado por J. Playfair a finales del siglo XVIII).

(iv) Sobre una recta finita siempre se puede construir un triángulo semejante a un triángulo dado (J. WALLIS, 1663; A. M. LEGENDRE, 1824); existe un par de triángulos no congruentes, pero semejantes —con sus respectivos ángulos iguales— (SACCHERI, 1733).

(v) En todo cuadrilátero que tenga tres ángulos rectos, el cuarto ángulo también es recto (A. C. CLAIRAUT, 1741; J. H. LAMBERT, 1766).

(vi) Se puede construir un triángulo cuya área sea mayor que cualquier área dada (K. F. GAUSS, 1799).

(vii) Dados tres puntos no alineados, siempre será posible construir un círculo que pase por todos ellos (A. M. LEGENDRE, 1824; F. BOLYAI, 1832).

(viii) No hay patrón métrico absoluto de longitud (K. F. GAUSS, 1816).

Para más detalles sobre algunas de estas proposiciones y sobre otras de análoga condición, puede verse T. L. HEATH, ed. cit., 1926², 1, págs. 204-220.

NOCIONES COMUNES

1. Las cosas iguales a una misma cosa son también iguales entre sí [18].

[18] Esta proposición es seguramente la más conocida de las nociones comunes o axiomas —según Proclo, ambas expresiones son sinónimas para Aristóteles y para los geómetras, aunque los *Elementos* nunca ha-

2. Y si se añaden cosas iguales a cosas iguales, los totales son iguales.
3. Y si de cosas iguales se quitan cosas iguales, los restos son iguales [19].

blan de «axiomas [*axiómata*]» y Aristóteles prefiere hablar de unos principios [*archai*] o premisas comunes [*tà koinà*] o primordiales [*tà prôta*]—. Sin embargo, la citada con mayor frecuencia por parte de Aristóteles, en particular, es la noción común tercera. Las nociones comunes se distinguen, como es bien sabido, por su calidad de principios no sólo verdaderos y palmarios, sino indemostrables. Con todo y con esto, Apolonio al parecer trató de demostrarlos —confusión que lamenta PROCLO (*Com.* 194, 10-12)—. Su prueba de esta noción común discurriría como sigue: «Sea A igual a B y esta última igual a Γ. Digo que A también es igual a Γ. Dado que A, siendo igual a B, ocupa el mismo espacio que ella, y dado que B, siendo igual a Γ, ocupa el mismo espacio que ella, A también ocupa el mismo espacio que Γ. Por consiguiente, son iguales entre sí» (*Com.* 195, 1-5). Proclo señala que este argumento supone dos asunciones previas: a) las cosas que ocupan el mismo espacio, son iguales; b) dos cosas que ocupan el mismo espacio que una tercera, ocupan a su vez el mismo espacio la una que la otra; por último, hace notar que ni a) ni b) son proposiciones más claras e inmediatas que las proposiciones que tratan de establecer. Ante la desmesura de venir a demostrar lo indemostrable, K. von Fritz ha sugerido que la intención de Apolonio era más bien la de probar la transitividad de la congruencia en el caso de las líneas precisamente a partir de la noción común 1 (*vid.* su art. ya mencionado: «Die ARXAI in der griechischen Mathematik», *Archive für Begriffsgeschichte* 1 (1955), págs. 65, 100). En los modernos sistemas lógicos de identidad, la ley correspondiente a esta noción común 1 es derivable de otras leyes de la relación de identidad como la simetría y la transitividad (*vid.*, por ejemplo, el clásico A. TARSKI (1940), *Introducción a la lógica y a la metodología de las ciencias deductivas*, Madrid, 1968²; § 187, pág. 84).

[19] Heiberg atetiza las nociones comunes 4, 5, 6:

4. «Y si se añaden cosas iguales a cosas desiguales los totales son desiguales».
5. «Y los dobles de una misma cosa son iguales entre sí».
6. «Y las mitades de una misma cosa son iguales entre sí».

Proclo, observando que no se deben multiplicar los axiomas sin necesidad, indica que deben ser rechazados todos los que se siguen de los cinco admitidos por él (las cinco nociones comunes recogidas en el texto). Por ejemplo, menciona la noción común 5 como una de las que deben ser excluidas por seguirse de la primera. Según una observación de Simplicio recogida por Al-Nayrīzī, sólo había tres axiomas en los mss. más antiguos, pero se fueron añadiendo otros con posterioridad. El furor por la explicitación de supuestos y la falta de elegancia deductiva llevarán a algún comentador moderno a formular no menos de treinta axiomas euclidianos.

[20] Heiberg suprime también la noción común 9: «Dos rectas no encierran un espacio». Proclo la cita como uno de los ejemplo de multiplicación innecesaria de axiomas y señala además como objeción a éste que pertenece al ámbito específico de la geometría y no es de carácter general. El axioma es innecesario por estar incluido en el significado del postulado 1.

Seguramente se incluyó a partir del pasaje interpolado de I 4: «Y si la base BΓ no coincide con la base EH, dos rectas encerrarán un espacio, lo cual es imposible» (*vid.* nota 30). Algunos mss. incluyen esta noción común: V, b, p. E incluso otros, P y F, lo introducen después del postulado 5, proceder que sigue Campano. Stamatis lo mantiene en este lugar.

7. Y las cosas que coinciden entre sí son iguales entre sí.
8. Y el todo es mayor que la parte [20].

PROPOSICIÓN 1

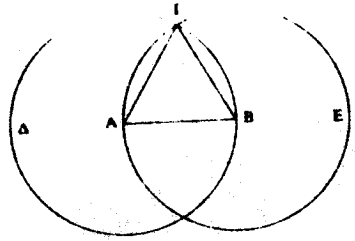
Construir un triángulo equilátero sobre una recta finita dada.

Sea AB la recta finita dada.

Así pues, hay que construir sobre la recta AB un triángulo equilátero.

Descríbase con el centro A y la distancia AB el círculo BΓΔ [Post. 3], y con el centro B y la distancia BA descríba-

se a su vez el círculo AFE [Post. 3], y a partir del punto Γ donde los círculos se cortan entre sí, trácense las rectas ΓA, ΓB hasta los puntos A, B [Post. 1].



Y puesto que el punto A es el centro del círculo ΓΔB, AΓ es igual a AB [Def. 15]; puesto que el punto B es a su vez el centro del círculo ΓAE, BΓ es igual a BA [Def. 15]; pero se ha demostrado que ΓA es igual a AB; por tanto, cada una de las (rectas) [21] ΓA, ΓB es igual a AB. Ahora bien, las cosas iguales a una misma cosa son también iguales entre sí [N. C. 1]; por tanto, ΓA es también igual a ΓB; luego las tres ΓA, AB, BΓ son iguales entre sí.

Por consiguiente, el triángulo ABΓ es equilátero y ha sido construido sobre la recta finita dada AB. (Que es) lo que había que hacer [22].

[21] Euclides omite con frecuencia las palabras «punto» «línea» «recta» «figura», etc. En griego, por lo general, basta con el artículo para saber a qué se refiere en cada caso por ser dichas palabras de diferente género; no ocurre lo mismo en español. Por ello, en los casos en los que me parece necesaria la palabra omitida para la comprensión de texto, la pongo entre paréntesis.

[22] La prueba de esta proposición suele oficiar como paradigma de la demostración euclídea. No sólo tiene una claridad meridiana, sino que discurre por los pasos canónicos que Proclo identifica (proposición, exposición, especificación, preparación, demostración propiamente dicha, conclusión; *Com.* 203, 1-14; *vid. supra*, introducción general, II, § 2), y además procede con la informalidad característica de la geometría clásica. Esta informalidad, el uso de supuestos no declarados, le ha valido objeciones y reparos. Las objeciones se remontan a críticos antiguos, como el epicúreo Zenón de Sidón (hacia finales del siglo II y principios del siglo I a. C.), quien observó que el problema no quedaba resuelto a menos que se diera por descontado que dos líneas rectas no pueden tener un segmento común (PROCLO, *Com.* 214, 18 ss.). Los reparos son de origen más moderno y se centran en la necesidad de un postulado de continuidad que asegure la existencia de un punto de intersección entre los círculos; el punto de intersección, que viene a ser uno de los vértices del triángulo —con los otros dos situados en los centros de los círculos—, aparece en esta prueba por obra y gracia de la representación diagramática del problema, pero no se halla previsto ni cubierto por las bases teóricas de su construcción. Pues bien, el postulado de continuidad —en una versión similar a la propuesta por DEDEKIND (1872)— prevé que si todos los puntos de una recta se distribuyen en dos clases tales que cada punto de la primera clase está a la izquierda de todo punto de la segunda clase, existe un solo punto que produce esta división de todos los puntos en dos clases, la división de la recta en dos partes. De este postulado se desprende que si, en un plano dado, un círculo C tiene un punto P dentro, y otro punto Q fuera, de otro círculo C', ambos círculos tendrán dos puntos de intersección. Para más referencias, *vid.* T. L. HEATH, ed. cit., 1926², 1, págs. 234-240, 242.

PROPOSICIÓN 2

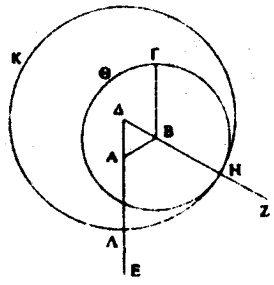
Poner en un punto dado (como extremo) [23] una recta *igual a una recta dada.*

Sea A el punto dado y BΓ la recta dada.

Así pues, hay que poner en el punto A una recta igual a la recta dada BΓ.

[23] Añado «(como extremo)» porque el enunciado «poner en un punto» no queda suficientemente claro en español. En griego la preposición *pros (tō dothénti sēmeioi)* da una idea más clara de lo que se pretende.

Trácese [24], pues, desde el punto A hasta el punto B la recta AB [Post. 1] y constrúyase sobre ella el triángulo equilátero ΔAB [I, 1], y sean AE, BZ el resultado de prolongar en línea recta las rectas ΔA, ΔB [25] [Post. 2]; y con el centro B y la distancia BΓ descríbase el círculo ΓHΘ, y a su vez con el centro Δ y la distancia ΔH, descríbase el círculo HKΛ [Post. 3].



Así pues, como el punto B es el centro del círculo ΓHΘ, BΓ es igual a BH. Como a su vez el punto Δ es el centro del círculo HKΛ, ΔΛ es igual a ΔH, cuyas partes respectivas ΔA y ΔB son iguales. Luego la parte restante AΛ es igual a la parte restante BH [N. C. 3]. Pero se ha demostrado que también BΓ es igual a BH; por tanto, cada una de las (rectas) AΛ, BΓ es igual a BH. Y las cosas iguales a una misma cosa son también iguales entre sí [N. C. 1]; luego AΛ es también igual a BΓ.

Por consiguiente, en el punto dado A se ha puesto la recta AΛ igual a la recta dada BΓ Q. E. F. [26].

[24] El verbo utilizado aquí es *epedseúchthō*, que tiene el significado de «unir». Euclides lo utiliza cuando se trata de unir dos puntos trazando una recta. Para trazar líneas rectas en general utiliza el verbo *agō*.

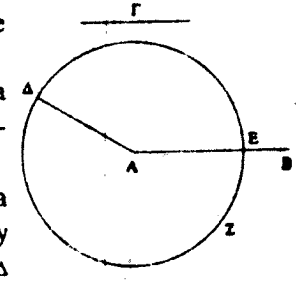
[25] El primer lugar donde Euclides emplea una expresión semejante a la nuestra: «prolónguese uno de sus lados BΓ hasta Δ», es en I, 16: *prosekbeblēsthō autoû mía pleurà hē* BΓ *epì tò* Δ.

[26] A partir de aquí utilizaré la abreviatura «Q. E. F.» *(Quod erat faciendum* = Que es lo que había que hacer) para señalar la conclusión de problemas.

Proposición 3

Dadas dos rectas desiguales, quitar de la mayor una recta igual a la menor.

Sean AB, Γ las dos rectas desiguales dadas, sea AB la mayor de ellas.



Así pues, hay que quitar de la mayor, AB una recta igual a la menor, Γ.

Colóquese sobre el punto A la recta AΔ igual a la recta Γ [I, 2] y con el centro A y la distancia AΔ descríbase el círculo ΔEZ [Post. 3], y como el punto A es el centro del círculo ΔEZ, AE es igual a AΔ; pero también Γ es igual a AΔ; luego cada una de las (rectas) AE, Γ es igual a AΔ; de modo que también AE es igual a Γ [N. C. 1].

Por consiguiente, dadas dos rectas desiguales: AB, Γ, se ha quitado de la mayor, AB, la recta AE igual a la menor, Γ. Q. E. F.

Proposición 4

Si dos triángulos tienen dos lados del uno iguales a dos lados del otro y tienen iguales los ángulos comprendidos por las rectas iguales [27]*, tendrán también las respectivas*

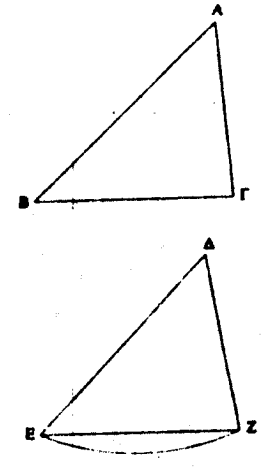
[27] Euclides habla de «rectas» y no de «lados» por ceñirse a la definición de ángulo como la inclinación mutua de dos líneas.

bases iguales [28], *y un triángulo será igual al otro, y los ángulos restantes, a saber: los subtendidos por lados iguales, serán también iguales respectivamente.*

Sean ABΓ, ΔEZ, dos triángulos que tienen los dos lados AB, AΓ iguales a ΔE, ΔZ, respectivamente, es decir AB a ΔE y AΓ a ΔZ, y el ángulo BAΓ igual al ángulo EΔZ.

Digo que también la base BΓ es igual a la base EZ y que el triángulo ABΓ será igual al triángulo ΔEZ y los ángulos restantes, subtendidos por los lados iguales, serán también iguales respectivamente, es decir, el (ángulo) [29] ABΓ (igual) al (ángulo) ΔEZ, y el (ángulo) AΓB (igual) al (ángulo) ΔZE.

Pues si se aplica el triángulo ABΓ al triángulo ΔEZ y el punto A se coloca sobre el punto Δ y la recta AB sobre la recta ΔE, coincidirá también el punto B sobre el punto E por ser igual AB a ΔE; al coincidir también AB con ΔE, la recta AΓ coincidirá también con ΔZ por ser igual el ángulo BAΓ al EΔZ; de modo que también el punto Γ coincidirá con el punto Z por ser igual a su vez AΓ a ΔZ. Pero también el punto B había coincidido con el punto E; de modo que la base BΓ coincidirá con la base EZ [pues si habiendo coincidido el punto B con el punto E y el punto Γ con el punto Z, no coindice la base BΓ con la base EZ, dos rectas encerrarán un espacio, lo cual es imposible. Por tanto, coincidirá la base BΓ con la base EZ] [30] y será igual a ella [N. C. 4]; de modo que también el triángulo entero ABΓ coincidirá con el triángulo entero ΔEZ y será igual a él, y los ángulos restantes coincidirán con los ángulos restantes y serán iguales a ellos, el (ángulo) ABΓ al (ángulo) ΔEZ y el (ángulo) AΓB al (ángulo) ΔZE.



Por consiguiente, si dos triángulos tienen dos lados del uno iguales a dos lados del otro y tienen iguales los ángulos comprendidos por las rectas iguales, tendrán también las respectivas bases iguales y un triángulo será igual al otro, y los ángulos restantes, a saber: los subtendidos por lados iguales, serán también iguales, respectivamente. (Que es lo que había que demostrar.)

[28] Es la primera vez que utiliza la palabra *básis* (base) en los *Elementos*. Proclo explica que se refiere al lado que está al mismo nivel que la vista, o al tercer lado cuando se han mencionado ya los otros dos. No es una explicación satisfactoria, porque el término aparece también en contextos donde dicho significado no tendría sentido y se aplica también a la base de los paralelogramos. El uso del término debe provenir más bien de la práctica de trazar el lado en cuestión horizontalmente y el resto de la figura sobre él. Se hablaba de la base de una figura en el mismo sentido que de la base de cualquier otra cosa: por ej., el pedestal de una columna.

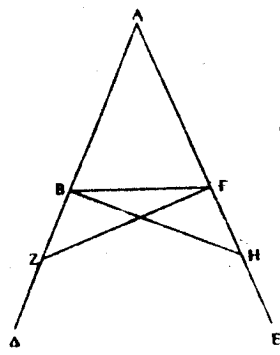
[29] Euclides emplea con frecuencia braquilogías para referirse a los ángulos. La expresión más frecuente es *hē hypò* seguido de las tres letras que definen las rectas que comprenden el ángulo: por ej., *he hypò* BAΓ. La expresión más completa sería *hē hypò tôn* BA, AΓ *periekhoménē gōnía* («el ángulo comprendido por las rectas BA, AΓ»). Otra expresión utilizada con frecuencia es *hē prὸs tôi* seguido de la letra del vértice, por ej., *he prὸs tô* A «el ángulo correspondiente a A».

[30] Heiberg atetiza el texto traducido entre corchetes considerándolo una interpolación temprana. El texto de Al-Nayrīzī no da las palabras en este lugar sino después de la cláusula que marca la conclusión «Q. E. D.», señal de que se trata de un escolio.

Proposición 5

En los triángulos isósceles los ángulos de la base son iguales entre sí, y prolongadas las dos rectas iguales, los ángulos situados bajo la base serán iguales entre sí.

Sea ABΓ el triángulo isósceles que tiene el lado AB igual al lado AΓ, y sean BΔ, ΓE el resultado de prolongar en línea recta las rectas AB, AΓ [Post. 2].



Digo que el ángulo ABΓ es igual al (ángulo) AΓB y el (ángulo) ΓBΔ es igual al (ángulo) BΓE.

Pues tómese al azar un punto Z en la recta BΔ y quítese de la mayor AE, la recta AH igual a la menor AZ [I, 3], y trácense las rectas ZΓ, HB. [Post. 1].

Ahora bien, como AZ es igual a AH y AB a AΓ, las dos rectas ZA, AΓ son respectivamente iguales a las dos rectas HA, AB y comprenden el ángulo común ZAH; por tanto, la base ZΓ es igual a la base HB, y el triángulo AZΓ será igual al triángulo AHB, y los ángulos restantes subtendidos por lados iguales serán también iguales respectivamente, el (ángulo) AΓZ al (ángulo) ABH y el (ángulo) AZΓ al (ángulo) AHB [I, 4]. Como la recta entera AZ es igual a la recta entera AH, cuyas respectivas partes AB y AΓ son iguales, entonces la parte restante BZ es igual a la parte restante ΓH. Pero se ha demostrado también que ZΓ es igual a HB; entonces las dos rectas BZ, ZΓ son iguales a las dos rectas ΓH, HB, respectivamente: y el ángulo BZΓ es igual al (ángulo) ΓHB y su base común es BΓ; y el triángulo BZΓ será, por tanto, igual al triángulo ΓHB, y los ángulos restantes subtendidos por lados iguales serán también iguales respectivamente; así que es igual el ángulo ZBΓ al ángulo HΓB y el ángulo BΓZ al ángulo ΓBH. Así pues, como se ha demostrado que el ángulo entero ABH es igual al ángulo entero AΓZ cuyas partes respectivas ΓBH, BΓZ son iguales, entonces el ángulo restante ABΓ es igual al ángulo restante AΓB y están en la base del triángulo ABΓ. Pero se ha demostrado que también el ángulo ZBΓ es igual al ángulo HΓB; y son los situados debajo de la base.

Por consiguiente, en los triángulos isósceles, los ángulos que están en la base son iguales entre sí y, prolongadas las rectas iguales, los ángulos situados debajo de la base serán iguales entre sí. Q. E. D. [31].

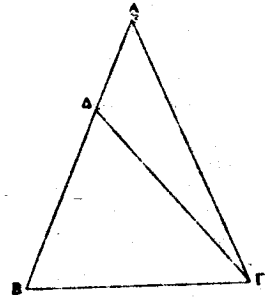
Proposición 6

Si dos ángulos de un triángulo son iguales entre sí, también los lados que subtienden a los ángulos iguales serán iguales entre sí.

[31] Proclo *(Com.* 250, 20) se hace eco de la tradición que atribuye a Tales el descubrimiento de este teorema (otros resultados conocidos desde antiguo serían I 12, 15, 23, 26, 47). Como muestra de una prueba preuclídea de I 5, *vid.* Aristóteles, *A. Pr.* 1, 24, 41b13-22, donde se hace uso del ángulo curvilíneo formado por un arco y una cuerda.

A partir de aquí emplearé la abreviatura usual «Q. E. D.» *(quod erat demonstrandum* = que es lo que había que demostrar) para señalar la conclusión de teoremas. Los griegos conocían una abreviatura similar: *oi* para la cláusula *hóper édei deîxai (deikhthênai), vid.* por ejemplo, la edic. de Arquímedes: *El Método* (Madrid, 1986), proposición 3, pág. 53, nota 12.

Sea el triángulo ABΓ que tiene el ángulo ABΓ igual al ángulo AΓB.



Digo que también el lado AB es igual al lado AΓ. Pues si AB no es igual a AΓ, uno de ellos es mayor. Sea AB el (lado) mayor. Y del (lado) mayor AB quítese la (recta) AB igual al (lado) menor AΓ, y trácese ΔΓ.

Ahora bien, como AB es igual a AΓ y BΓ es común, también los dos lados ΔB, BΓ son iguales a los dos lados AΓ, ΓB, respectivamente, y el ángulo ΔBΓ es igual al ángulo AΓB; por tanto, la base ΔΓ es igual a la base AB, y el triángulo ΔBΓ será igual al triángulo AΓB, el menor al mayor; lo cual es absurdo; entonces los lados AB y AΓ no son desiguales; luego son iguales.

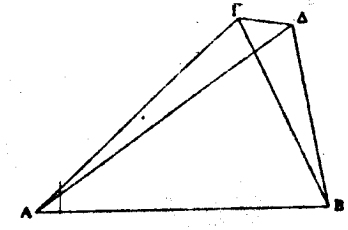
Por consiguiente, si dos ángulos de un triángulo son iguales entre sí, también los lados que subtienden a los ángulos iguales serán iguales entre sí Q. E. D.

Proposición 7

No se podrán levantar sobre la misma recta otras dos rectas iguales respectivamente a dos rectas dadas, de modo que se encuentren en dos puntos distintos por el mismo lado y con los mismos extremos que las rectas dadas [32].

[32] *Synístasthai* es el verbo empleado comúnmente por los geómetras griegos para referirse a la construcción de un triángulo en particular. Por tanto, los griegos podían entender fácilmente la expresión *systathé-*

Pues, si es posible (hacerlo), constrúyanse sobre una misma recta AB, otras dos rectas AΔ, ΔB iguales respectivamente a dos rectas dadas AΓ, ΓB, que se encuentren en dos puntos diferentes Γ y Δ, por el mismo lado y que tengan los mismos extremos, de modo que ΓA sea igual a ΔA y tenga el mismo extremo que ella: A, y ΓB sea igual a ΔB y tenga el mismo extremo que ella: B, y trácese ΓΔ.



Ahora bien, puesto que AΓ es igual a AΔ, también es igual el ángulo AΓΔ al ángulo AΔΓ [I, 5], por tanto, el ángulo AΔΓ es mayor que el ángulo ΔΓB; luego el ángulo ΓΔB es mucho mayor que el ángulo ΔΓB. Puesto que ΓB es a su vez igual a ΔB, también es igual el ángulo ΓΔB al ángulo ΔΓB. Pero se ha demostrado que también es mucho mayor que él; lo cual es imposible.

Por consiguiente, no se podrán levantar sobre una misma recta otras dos rectas iguales respectivamente a dos rectas dadas de modo que se encuentren en dos puntos distintos por el mismo lado y con los mismos extremos que las rectas dadas. Q.E.D.

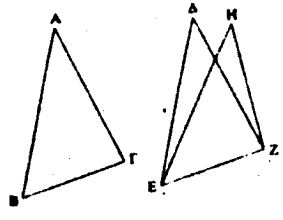
sontai epí con el sentido de construir o levantar dos rectas formando un triángulo sobre otra recta. En español habría que precisar el sentido del término, de ahí mi traducción de *epì tês autês eutheías... ou systathésontai... pròs állōi kaí állōi sēmeíōi* como: «no se podrán levantar sobre la misma recta ... de modo que se encuentren en puntos distintos».

Proposición 8

Si dos triángulos tienen dos lados del uno iguales respectivamente a dos lados del otro y tienen también iguales sus bases respectivas, también tendrán iguales los ángulos comprendidos por las rectas iguales.

Sean ΑΒΓ, ΔΕΖ dos triángulos que tienen los dos lados ΑΒ, ΑΓ iguales, respectivamente, a ΔΕ, ΔΖ, (es decir) el lado ΑΒ al lado ΔΕ y el lado ΑΓ al lado ΔΖ; y tengan también la base ΒΓ igual a la base ΕΖ.

Digo que el ángulo ΒΑΓ es también igual al ángulo ΕΔΖ.



Pues si se aplica el triángulo ΑΒΓ al triángulo ΔΕΖ y se pone el punto Β sobre el punto Ε y la recta ΒΓ sobre la recta ΕΖ, coincidirá también el punto Γ con el punto Ζ por ser igual ΒΓ a ΕΖ; y al coincidir ΒΓ con ΕΖ coincidirán también ΒΑ, ΓΑ con ΕΔ, ΔΖ. Pues si coinciden la base ΒΓ con la base ΕΖ y los lados ΒΑ, ΑΓ no coinciden con los lados ΕΔ, ΔΖ sino que se desvían como ΕΗ, ΗΖ, podrán ser construidas sobre una misma recta otras dos rectas iguales respectivamente a dos rectas dadas que se encuentren en puntos distintos por el mismo lado y con los mismos extremos. Pero no pueden construirse [I, 7]; por tanto, no es posible que, aplicada la base ΒΓ a la base ΕΖ, no coincidan los lados ΒΑ, ΑΓ con los lados ΕΔ, ΔΖ. Luego coincidirán; de modo que también el ángulo ΒΑΓ coincidirá con el ángulo ΕΔΖ y será igual a él.

Por consiguiente, si dos triángulos tienen dos lados del uno iguales respectivamente a dos lados del otro y tienen las bases respectivas iguales, también tendrán iguales los ángulos comprendidos por las rectas iguales. Q. E. D.

Proposición 9

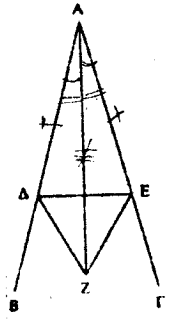
Dividir en dos partes iguales un ángulo rectilíneo dado.

Sea ΒΑΓ el ángulo rectilíneo dado.

Así pues, hay que dividirlo en dos partes iguales.

Tómese al azar un punto Δ en la recta ΑΒ y quítese de ΑΓ la recta ΑΕ igual a ΑΔ [I, 3] y trácese ΔΕ, y constrúyase sobre ΔΕ el triángulo equilátero ΔΕΖ, y trácese ΑΖ.

Digo que el ángulo ΒΑΓ ha sido dividido en dos partes iguales por la recta ΑΖ.



Pues como la (recta) ΑΔ es igual a la (recta) ΑΕ y ΑΖ es común, las dos (rectas) ΔΑ, ΑΖ son iguales respectivamente a las dos (rectas) ΕΑ, ΑΖ. Y la base ΔΖ es igual a la base ΕΖ; por tanto, el ángulo ΔΑΖ es igual al ángulo ΕΑΖ [I, 8].

Por consiguiente, el ángulo rectilíneo dado ΒΑΓ ha sido dividido en dos partes iguales por la recta ΑΖ. Q. E. F.

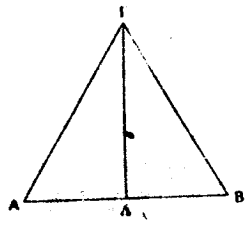
Proposición 10

Dividir en dos partes iguales una recta finita dada.

Sea ΑΒ la recta finita dada.

Así pues, hay que dividir en dos partes iguales la recta finita ΑΒ.

Constrúyase sobre ella el triángulo equilátero ABΓ [I, 1] y divídase en dos partes iguales el ángulo AΓB mediante la recta ΓΔ [I, 9].



Digo que la recta AB ha sido dividida en dos partes iguales en el punto Δ.

Pues como la recta AΓ es igual a la recta ΓB y la recta ΓΔ es común, entonces las dos rectas AΓ, ΓΔ son iguales respectivamente a las dos rectas BΓ, ΓΔ; y el ángulo AΓΔ es igual al ángulo BΓΔ; por tanto, la base AΔ es igual a la base BΔ [I, 4].

Por consiguiente, la recta finita dada AB ha sido dividida en dos partes iguales en el punto Δ. Q. E. F.

Proposición 11

Trazar una línea recta que forme ángulos rectos [33] *con una recta dada, desde un punto dado en ella.*

Sea AB la recta dada y sea Γ el punto dado en ella.

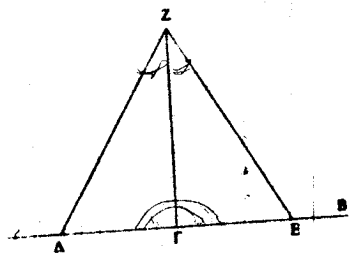
Así pues, hay que trazar una línea recta que forme ángulos rectos con la recta AB desde el punto Γ.

Tómese un punto Δ al azar sobre la recta AΓ, y hágase [34] ΓE igual a ΓΔ [I, 3], y constrúyase sobre ΔE el triángulo equilátero ZΔE [I, 1], y trácese ZΓ.

[33] Traduzco *pròs orthàs gonías* como «que forme ángulos rectos» para distinguirlo de *káthetos*, «perpendicular».

[34] La forma verbal *keístho* de *keímai*, verbo que se utiliza para la voz pasiva de *títhemi*, «poner», «colocar», se emplea en este contexto en el sentido de «hacer una línea igual a otra».

Digo que ha sido trazada la línea recta ZΓ que forma ángulos rectos con la recta dada AB, desde el punto Γ dado en ella.



Pues como ΔΓ es igual a ΓE, y ΓZ es común, los dos (lados) ΔΓ, ΓZ son iguales respectivamente a los dos (lados) EΓ, ΓZ; y la base ΔZ es igual a la base ZE; por tanto, el ángulo ΔΓZ es igual al ángulo EΓZ [I, 8]; y son ángulos adyacentes. Y cuando una recta levantada sobre otra recta hace los ángulos adyacentes iguales entre sí, cada uno de los ángulos iguales es recto [Def. 10], por tanto cada uno de los ángulos ΔΓZ y ZΓE es recto.

Por consiguiente, ha sido trazada la línea recta ΓZ que forma ángulos rectos con la recta dada AB, desde el punto Γ dado en ella. Q. E. F.

Proposición 12

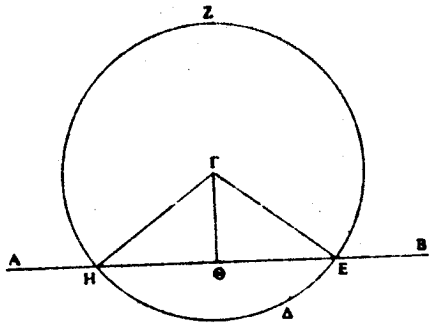
Trazar una línea recta perpendicular a una recta infinita [35] *dada desde un punto dado que no esté en ella.*

[35] *Ápeiron*. La precisión «indefinida» que presentan otras versiones españolas (e. g., J. D. García Bacca, ed. cit., 1944, pág. 37; F. Vera, ed. cit., 1970, pág. 712) no parece necesaria en este contexto.

Sea AB la recta infinita dada y Γ el punto dado que no está en ella.

Así pues, hay que trazar una línea recta perpendicular a la recta infinita dada AB desde el punto dado Γ que no está en ella.

Tómese, pues, al azar un punto Δ al otro lado de la recta AB, y con el centro Γ y la distancia ΓΔ descríbase el círculo EZH [Post. 3], y divídase en dos partes iguales la recta EH en Θ, y trácense las rectas ΓH, ΓΘ, ΓE [Post. 1].



Digo que ha sido trazada la recta ΓΘ perpendicular a la recta infinita dada AB desde el punto dado Γ que no está en ella.

Pues como HΘ es igual a ΘE y ΘΓ es común, los dos (lados) HΘ, ΘΓ son iguales respectivamente a los dos (lados) EΘ, ΘΓ; y la base ΓH es igual a la base ΓE; por tanto, el ángulo ΓΘH es igual al ángulo EΘΓ [I, 8]. Y son adyacentes. Ahora bien, cuando una recta levantada sobre otra recta hace los ángulos adyacentes iguales entre sí, cada uno de los ángulos iguales es recto, y la recta que se ha levantado se llama perpendicular a aquella sobre la que está (Def. 10).

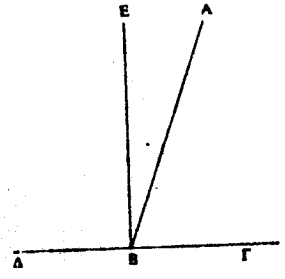
Por consiguiente, se ha trazado la recta ΓΘ perpendicular a la recta infinita dada AB, desde el punto dado Γ que no está en ella. Q. E. F.

Proposición 13

Si una recta levantada sobre otra recta forma ángulos, o bien formará dos rectos o bien (ángulos) iguales a dos rectos.

Así pues, forme una recta cualquiera AB levantada sobre la recta ΓΔ los ángulos ΓBA, ABΔ.

Digo que los ángulos ΓBA, ABΔ son o bien dos rectos o iguales a dos rectos.



En efecto, si ΓBA es igual a ABΔ, son dos rectos [Def. 10]. Pero si no, trácese desde el punto B la recta BE que forme ángulos rectos con ΓΔ [I, 11]; entonces los ángulos ΓBE, EBΔ son dos rectos, y dado que el (ángulo) ΓBE es igual a los dos (ángulos) ΓBA, ABE, añádase al uno y a los otros [36] el ángulo EBΔ; entonces los (ángulos) ΓBE, EBΔ

[36] *Koinè proskeísthō hē hypò* EBΔ, literalmente: «añádase el ángulo EBΔ común». Del mismo modo se utiliza *koinè apherésthō* para quitar

son iguales a los tres (ángulos) ΓBA, ABE, EBA [N. C. 2]. Como el (ángulo) ABA es igual a su vez a los dos (ángulos) ABΓ, EBA añádase al uno y a los otros el (ángulo) ABΓ; entonces los (ángulos) ABA, ABΓ son iguales a los tres (ángulos) ABE, EBA, ABΓ [N. C. 2]. Pero se ha demostrado que también los (ángulos) ΓBE, EBA son iguales a esos mismos tres; ahora bien, las cosas iguales a una misma cosa son también iguales entre sí [N. C. 1]; por tanto, los (ángulos) ΓBE, EBA son también iguales a los (ángulos) ABA, ABΓ. Pero los (ángulos) ΓBE, EBA son dos rectos; por tanto, los (ángulos) ABA, ABΓ son también iguales a dos rectos.

Por consiguiente, si una recta levantada sobre otra recta forma ángulos, formará o bien dos rectos o bien (ángulos) iguales a dos rectos. Q. E. D.

Proposición 14

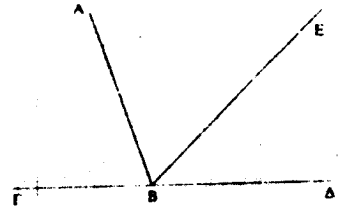
Si dos rectas forman con una recta cualquiera y en un punto de ella ángulos adyacentes iguales a dos rectos y no están en el mismo lado (de ella), ambas rectas estarán en línea recta.

Así pues, sean dos rectas BΓ, BΔ que, con una recta cualquiera AB y en un punto de ella B y sin estar colocadas en el mismo lado (de la recta AB), formen dos ángulos adyacentes ABΓ, ABΔ iguales a dos rectos.

Digo que BΔ está en línea recta con ΓB.

un ángulo o una línea recta de otros. En español, la expresión «añadir el ángulo común» resultaría ambigua, por ello traduzco *koiné* por «a uno y a los otros», o «a ambos», cuando se trata de dos.

Pues si BΔ no está en línea recta con BΓ, esté BE en línea recta con ΓB.



Así pues, dado que la recta AB ha sido levantada sobre la recta ΓBE, entonces los ángulos ABΓ, ABE son iguales a dos rectos [I, 13]; pero también los (ángulos) ABΓ, ABΔ son iguales a dos rectos; por tanto los (ángulos) ΓBA, ABE son iguales a los (ángulos) ΓBA, ABΔ [Post. 4 y N. C. 1]. Quítese de ambos el (ángulo) común ΓBA; luego el ángulo restante ABE es igual al ángulo restante ABΔ [N. C. 3], el menor al mayor; lo cual es imposible. Por tanto, BE no está en línea recta con ΓB. Y de modo semejante demostraríamos esto de cualquier otra que no sea la recta BΔ. Por tanto, ΓB está en línea recta con BΔ.

Por consiguiente, si dos rectas forman con una recta cualquiera y en un punto de ella ángulos adyacentes iguales a dos rectos y no están en el mismo lado (de ella), ambas rectas están en línea recta. Q. E. D.

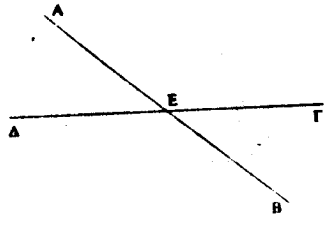
Proposición 15

Si dos rectas se cortan, hacen los ángulos del vértice [37] *iguales entre sí.*

[37] *Hai katà koryphèn gōniai.* Se refiere a los ángulos opuestos por el vértice.

Así pues, córtense las dos rectas AB, ΓΔ en el punto E.

Digo que el ángulo AEΓ es igual al (ángulo) ΔEB y el (ángulo) ΓEB al (ángulo) AEΔ.



Pues, dado que la recta AE ha sido levantada sobre la recta ΓΔ formando los ángulos ΓEA, AEΔ, entonces los ángulos ΓEA, AEΔ son iguales a dos rectos [I, 13]. Dado que la recta ΔE ha sido levantada a su vez sobre la recta AB formando los ángulos AEΔ, ΔEB, entonces los ángulos AEΔ, ΔEB son iguales a dos rectos. Pero se ha demostrado que los (ángulos) ΓEA, AEΔ son iguales a dos rectos; luego los (ángulos) ΓEA, AEΔ son iguales a los (ángulos) AEΔ, ΔEB [Post. 4 y N. C. 1]. Quítese de ambos AEΔ; entonces, el (ángulo) restante ΓEA es igual al (ángulo) restante BEΔ [N. C. 3]; de manera semejante demostraríamos [38] que también los (ángulos) ΓEB, ΔEA son iguales.

Por consiguiente, si dos rectas se cortan, hacen los ángulos del vértice iguales entre sí. Q. E. D. [39].

[38] En griego aparece el futuro *deixomen*, «demostraremos». Considero más adecuado el condicional en español porque, en este contexto, esa expresión no representa un compromiso real con una demostración futura, sino la indicación de otras posibles aplicaciones del mismo procedimiento.

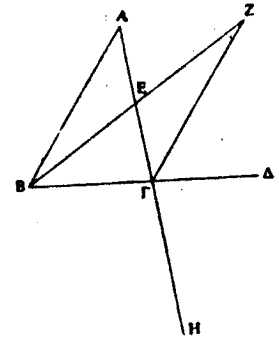
[39] Aunque no está en los mss. más acreditados, Proclo y Pselo reconocen el siguiente porisma:

Proposición 16

En todo triángulo, si se prolonga uno de sus lados, el ángulo externo es mayor que cada uno de los ángulos internos y opuestos.

Sea ABΓ el triángulo y prolónguese uno de sus lados, BΓ, hasta Δ.

Digo que el ángulo externo AΓΔ es mayor que cada uno de los ángulos internos y opuestos, ΓBA y BAΓ.



Córtese en dos partes iguales AΓ por el punto E y, trazada BE prolónguese en línea recta hasta Z y hágase EZ igual a BE [I, 3], y trácese ZΓ [Post. 1], y prolónguese por el otro lado [40] AΓ hasta H [Post. 2].

Así pues, como AE es igual a EΓ, y BE a EZ, los dos (lados) AE, EB son iguales respectivamente a los dos (lados) ΓE, EZ; y el ángulo AEB es igual al ángulo ZEΓ; pues son opuestos por el vértice; entonces, la base AB es igual a la base ZΓ y el triángulo ABE es igual

[Porisma: A partir de esto queda claro que si dos rectas se cortan hacen los ángulos de la intersección iguales a cuatro rectos.]

Las proposiciones de este género no tienen en los *Elementos* el carácter que al parecer presentaban en un tratado perdido de Euclides titulado precisamente *Porismata*. Si allí eran una especie de proposiciones intermedias entre teoremas y problemas, aquí representan lo que hoy conocemos como corolarios.

[40] En esta ocasión, Euclides utiliza el verbo *didgomai (diékhthō)* en lugar de *ekbeblésthō*, que es el empleado habitualmente para «prolongar». La traducción literal de *didgein* sería «trazar a través de», «trazar pasando por un punto».

al triángulo ZEΓ, y los ángulos restantes, a saber: los subtendidos por lados iguales son respectivamente iguales [I, 4], entonces es igual el (ángulo) BAE al (ángulo) EΓZ. Pero el (ángulo) EΓΔ es mayor que el ángulo EΓZ [N. C. 5]; luego el ángulo AΓΔ es mayor que el ángulo BAE. Así pues, de manera semejante, si se divide en dos la recta BΓ, se demostrará que también el ángulo BΓH, es decir el ángulo AΓΔ [I, 15] es mayor que él ángulo ABΓ.

Por consiguiente, en todo triángulo, si se prolonga uno de sus lados, el ángulo externo es mayor que cada uno de los ángulos internos y opuestos. Q. E. D. [41].

Proposición 17

En todo triángulo dos ángulos tomados juntos de cualquier manera [42] *son menores que dos rectos.*

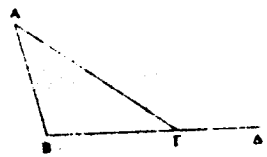
Sea ABΓ el triángulo.

[41] Heath observa, ed. cit., 1, pág. 280, que esta proposición no tiene vigencia universal a la luz de la hipótesis de Riemann de un espacio que se extienda sin límite pero no sea infinito. Puede verse ahí el anuncio de una relación entre I 16 y la teoría euclídea de las paralelas —aparte de la relación con otros supuestos característicos, e. g., el de que dos rectas no encierran un espacio—. Esa relación se confirmará más adelante cuando la prueba de I 27, que sienta una de las condiciones de paralelismo, recurra implícitamente a esta proposición I 16. Pero si esta proposición se toma por una exclusión temprana de la posibilidad de una geometría riemanniana, I 16 vendrá a incurrir en una suerte de paralogismo, *vid.* A. Dou: «Los paralogismos de Euclides y Saccheri en la teoría de las paralelas», *Revista de la Real Academia de Ciencias* 61 (1967), 155-174.

[42] *Pántei metalambanómenai*

Digo que dos ángulos del triángulo ABΓ tomados juntos de cualquier manera son menores que dos rectos.

Prolónguese BΓ hasta Δ [Post. 2]. Y puesto que el ángulo AΓΔ es un ángulo externo del triángulo ABΓ, es mayor que el interno y opuesto ABΓ [I, 16]. Añádase a ambos AΓB; entonces los (ángulos) AΓΔ, AΓB son mayores que los (ángulos) ABΓ, BΓA. Pero los (ángulos) AΓΔ, AΓB son iguales a dos rectos [I, 13]; por tanto, los (ángulos) ABΓ, BΓA son menores que dos rectos. Y de manera semejante demostraríamos que también los (ángulos) BAΓ, AΓB son menores que dos rectos así como los ángulos ΓAB, ABΓ.



Por consiguiente, en todo triángulo dos ángulos tomados juntos de cualquier manera son menores que dos rectos. Q. E. D.

Proposición 18

En todo triángulo el lado mayor subtiende al ángulo mayor [43].

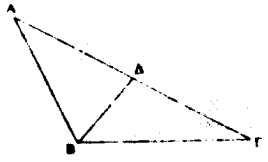
[43] Los enunciados de las proposiciones 18 y 19 son:

18. *Pantòs trigónou hē meídsōn pleurà tèn meídsona gōnían hypoteínei.*

19. *Pantòs trigónou hypò tèn meídsona gōnían he meídsōn pleurà hypoteínei.*

Sólo se diferencian en el orden de palabras y en la repetición de la preposición *hypó* con el régimen de *hypoteínei* en la segunda. La diferencia en el orden de formulación se debería a que en la proposición 18, el dato de partida es el lado del triángulo, mientras que en la 19 es el ángulo.

Sea, pues, ΑΒΓ el triángulo que tiene el lado ΑΓ mayor que ΑΒ.



Digo que el ángulo ΑΒΓ es también mayor que el ángulo ΒΓΑ.

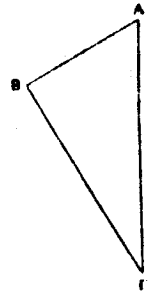
Pues como ΑΓ es mayor que ΑΒ, hágase ΑΔ igual a ΑΒ [I, 3], y trácese ΒΔ.

Y puesto que ΑΔΒ es un ángulo externo del triángulo ΒΓΔ, es mayor que el interno y opuesto ΔΓΒ [I, 16]; pero el (ángulo) ΑΔΒ es igual al (ángulo) ΑΒΔ, puesto que el lado ΑΒ es también igual a ΑΔ; por tanto, el (ángulo) ΑΒΔ también es mayor que el (ángulo) ΑΓΒ; luego el (ángulo) ΑΒΓ es mucho mayor que ΑΓΒ.

Por consiguiente, en todo triángulo el lado mayor subtiende al ángulo mayor. Q. E. D.

Proposición 19

En todo triángulo al ángulo mayor lo subtiende el lado mayor [43bis].



Sea ΑΒΓ el triángulo que tiene el ángulo ΑΒΓ mayor que el ángulo ΒΓΑ.

Digo que al lado ΑΓ es también mayor que el lado ΑΒ.

Porque si no, o bien ΑΓ es igual a ΑΒ o menor; ahora bien, ΑΓ no es igual a ΑΒ; pues entonces sería también igual el ángulo ΑΒΓ al (ángulo) ΑΓΒ [I, 5], pero no lo es; por tanto, ΑΓ no es igual a ΑΒ. Ni tampoco ΑΓ es menor que ΑΒ; pues entonces sería menor el ángulo ΑΒΓ que el (ángulo) ΑΓΒ [I, 18]; pero no lo es; por tanto, ΑΓ no es menor que ΑΒ. Pero se ha demostrado que tampoco es igual. Por tanto, ΑΓ es mayor que ΑΒ.

Por consiguiente, en todo triángulo al ángulo mayor lo subtiende el lado mayor. Q. E. D.

[43bis] Cf. nota 43.

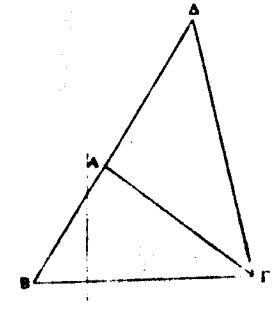
Proposición 20

En todo triángulo dos lados tomados juntos de cualquier manera son mayores que el restante

Sea, pues, ΑΒΓ el triángulo.

Digo que dos lados del triángulo ΑΒΓ tomados juntos de cualquier manera son mayores que el restante, los lados ΒΑ, ΑΓ (mayores) que ΒΓ, los lados ΑΒ, ΒΓ (mayores) que ΑΓ, y los lados ΒΓ, ΓΑ (mayores) que ΑΒ.

Prolónguese por el otro lado ΒΑ hasta el punto Δ, y hágase ΑΔ igual a ΓΑ y trácese ΔΓ.



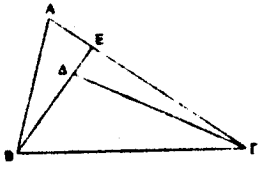
Entonces como ΔΑ es igual a ΑΓ, también el ángulo ΑΔΓ es igual al (ángulo) ΑΓΔ [I, 5]; por tanto, el (ángulo) ΒΓΔ es mayor que el (ángulo) ΑΔΓ [N. C. 5]; y puesto que ΔΓΒ es un triángulo que tiene el ángulo ΒΓΔ mayor que el ángulo ΒΔΓ, y al ángulo mayor lo subtiende el lado mayor [I, 19], entonces ΔΒ es mayor que ΒΓ. Pero ΔΑ es igual que ΑΓ; por tanto, los (lados) ΒΑ, ΑΓ son mayores que ΒΓ; de manera semejante demostraríamos que los (lados) ΑΒ, ΒΓ son también mayores que ΓΑ y los (lados) ΒΓ, ΓΑ (mayores) que ΑΒ.

Por consiguiente, en todo triángulo dos lados tomados juntos de cualquier manera son mayores que el restante. Q. E. D. [44].

Proposición 21

Si a partir de los extremos de uno de los lados de un triángulo se construyen dos rectas que se encuentren [45] *en el interior (de él), las (rectas) construidas serán menores que los dos lados restantes del triángulo, pero comprenderán un ángulo mayor.*

Sobre ΒΓ, uno de los lados del triángulo ΑΒΓ, a partir de los extremos Β, Γ constrúyanse dos rectas ΒΔ, ΔΓ que se encuentren en el interior de él.



Digo que ΒΔ, ΔΓ son menores que los dos lados restantes del triángulo ΒΑ, ΑΓ, pero comprenden el ángulo ΒΔΓ mayor que el (ángulo) ΒΑΓ.

Prolónguese ΒΔ hasta Ε. Y puesto que en todo triángulo dos lados son mayores que el restante [I, 20], entonces los dos lados,

[44] Proclo dice que los epicúreos acostumbraban a ridiculizar este teorema porque «era evidente incluso para un asno y no requería prueba». Su afirmación de que el teorema era comprensible *(gnórimon)* incluso para un asno, se basaba en el hecho de que, si se colocaba forraje en un vértice y el asno en el otro, el hambriento animal no iría en busca de su pitanza a través de dos lados del triángulo, sino sencillamente a través de aquel que le separaba de la comida (argumento que hizo exclamar a Savile que sus autores eran *digni ipsi, qui cum asino foenum essent*: «dignos de compartir el heno con el asno»). Proclo replica que la mera percepción de la verdad de un teorema es algo diferente de una prueba científica del mismo y del conocimiento de la razón por la cual es verdad. Por otra parte, recuerda Simson a este respecto, no debe multiplicarse el número de axiomas sin necesidad.

[45] El verbo empleado es *systathósin (synistamai)*. Ver nota 32.

ΑΒ, ΑΕ del triángulo ΑΒΕ son mayores que ΒΕ; añádase a uno y a los otros el (lado) ΕΓ, entonces ΒΑ, ΑΓ son mayores que ΒΕ, ΕΓ. Asimismo, puesto que los lados ΓΕ, ΕΔ del triángulo ΓΕΔ son mayores que ΓΔ, añádase a uno y a los otros el (lado) ΔΒ; entonces ΓΕ, ΕΒ son mayores que ΓΔ, ΔΒ. Pero se ha demostrado que ΒΑ, ΑΓ son mayores que ΒΕ, ΕΓ; entonces ΒΑ, ΑΓ son mucho mayores que ΒΔ, ΔΓ.

Asimismo, como en todo triángulo, el ángulo externo es mayor que el interno y opuesto [I, 16]; entonces en el triángulo ΓΔΕ el ángulo externo ΒΔΓ es mayor que el ángulo ΓΕΔ. Por la misma razón, también en el triángulo ΑΒΕ el ángulo externo ΓΕΒ es mayor que el (ángulo) ΒΑΓ. Pero se ha demostrado que el ángulo ΒΔΓ es mayor que el ángulo ΓΕΒ; luego el (ángulo) ΒΔΓ es mucho mayor que el ángulo ΒΑΓ.

Por consiguiente, si a partir de los extremos de uno de los lados de un triángulo se construyen dos rectas que se encuentren en el interior (de él), las (rectas) construidas son menores que los dos lados restantes del triángulo, pero comprenden un ángulo mayor. Q. E. D.

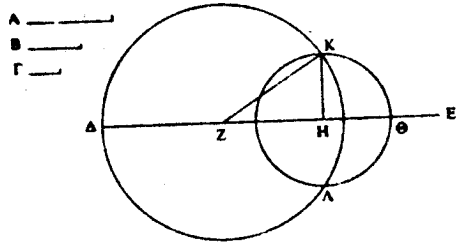
Proposición 22

Construir un triángulo con tres rectas que son iguales a tres rectas dadas. Pero es necesario que dos (de las) rectas tomadas juntas de cualquier manera sean mayores que la restante [46].

[46] Éste es el primer caso de *diorismós* en el contexto de un problema que aparece en los *Elementos*. En este contexto tiene el sentido de condi-

Sean A, B, Γ, las tres rectas dadas, y dos de éstas, tomadas de cualquier manera, sean mayores que la restante: A, B (mayores) que Γ; A, Γ (mayores) que B; y además B, Γ (mayores) que A.

Así pues, hay que construir un triángulo con (rectas) iguales a A, B, Γ.



Póngase una recta ΔE, limitada por Δ e ilimitada en la dirección de E, y hágase ΔZ igual a A, y ZH igual a B y HΘ igual a Γ [I, 3]; y con el centro Z y la distancia ZΔ descríbase el círculo ΔKΛ; asimismo, con el centro H y la distancia HΘ descríbase el círculo KΛΘ, y trácense KZ, KH.

Digo que se ha construido el triángulo KZH con tres rectas iguales a A, B, Γ.

Pues como el punto Z es el centro del círculo ΔKΛ, ZΔ es igual a ZK; pero ZΔ es igual a A, entonces KZ es igual a A. Y como el punto H es el centro del círculo ΛKΘ, HΘ es igual a HK: pero HΘ es igual a Γ; entonces KH es también igual a Γ. Pero también ZH es igual a B; luego las tres rectas KZ, ZH, HK, son iguales a las tres rectas A, B, Γ.

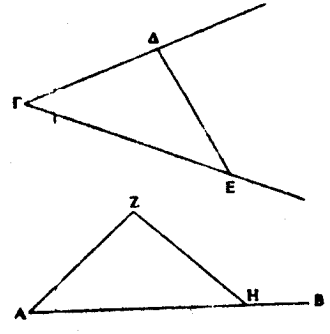
Por consiguiente, se ha construido el triángulo KZH con las tres rectas KZ, ZH, HK que son iguales a las tres rectas dadas A, B, Γ. Q. E. F.

ción o restricción que determina la posibilidad de solución del problema. En los mss. y en algunos comentarios sigue al *diorismás* la explicación: «Porque, en todo triángulo, dos lados tomados juntos de cualquier manera son mayores que el restante». Pero esta explicación tiene todos los visos de una glosa. De hecho, Proclo, y Campano la omiten, y Heiberg y Stamatis la consideran una interpolación.

Proposición 23

Construir un ángulo rectilíneo igual a un ángulo rectilíneo dado, sobre una recta dada y en uno de sus puntos.

Sea AB la recta dada, A uno de sus puntos y ΔΓE el ángulo rectilíneo dado.



Así pues, hay que construir un ángulo rectilíneo igual al ángulo dado ΔΓE sobre la recta dada AB y en su punto A.

Tómense al azar los puntos Δ, E en las rectas ΓΔ, ΓE, respectivamente, y trácese ΔE. Y con tres rectas que son iguales a las tres rectas ΓΔ, ΔE, ΓE, constrúyase el triángulo AZH, de modo que ΓΔ sea igual a AZ, ΓE a AH, y además ΔE a ZH [I, 22].

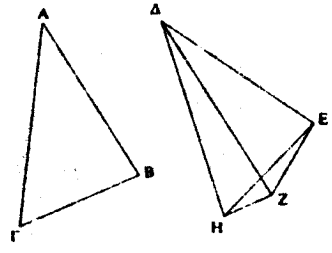
Pues bien, dado que las dos (rectas) ΑΓ, ΓΕ son iguales a las dos (rectas) ΖΑ, ΑΗ, respectivamente, y la base ΔΕ es igual a la base ΖΗ, entonces el ángulo ΑΓΕ es igual al ángulo ΖΑΗ [I, 8].

Por consiguiente, ha sido construido el ángulo rectilíneo ΖΑΗ igual al ángulo rectilíneo dado ΔΓΕ en la recta dada ΑΒ y en su punto Α. Q. E. F.

Proposición 24

Si dos triángulos tienen dos lados del uno iguales respectivamente a dos lados del otro, pero uno tiene el ángulo comprendido por las rectas iguales mayor que el otro, también tendrá la base mayor que la otra.

Sean ΑΒΓ, ΔΕΖ dos triángulos que tienen los dos lados ΑΒ, ΑΓ, iguales a los dos lados ΔΕ, ΔΖ, respectivamente: ΑΒ a ΔΕ y ΑΓ a ΔΖ, pero el ángulo correspondiente a Α sea mayor que el ángulo correspondiente a Δ.



Digo que la base ΒΓ también es mayor que la base ΕΖ.

Pues como el ángulo ΒΑΓ es mayor que el ángulo ΕΔΖ, constrúyase en la recta ΔΕ y en su punto Δ el (ángulo) ΕΔΗ igual al ángulo ΒΑΓ [I, 23], y hágase ΔΗ igual a una de las dos rectas ΑΓ, ΔΖ, y trácense ΕΗ, ΖΗ.

Pues bien, como ΑΒ es igual a ΔΕ, y ΑΓ a ΔΗ, las dos (rectas) ΒΑ, ΑΓ son iguales respectivamente a las dos (rectas) ΕΔ, ΔΗ; y el ángulo ΒΑΓ es igual al ángulo ΕΔΗ; por tanto, la base ΒΓ es igual a la base ΕΗ [I, 4]. Asimismo, como ΔΖ es igual a ΔΗ, el ángulo ΔΗΖ es también igual al ángulo ΔΖΗ [I, 5]; por tanto, el (ángulo) ΔΖΗ es mayor que el (ángulo) ΕΗΖ; entonces el (ángulo) ΕΖΗ es mucho mayor que el ángulo ΕΗΖ. Y dado que ΕΖΗ es un triángulo que tiene el ángulo ΔΖΗ mayor que el (ángulo) ΕΗΖ y al ángulo mayor lo subtiende el lado mayor [I, 19], entonces el lado ΕΗ es también mayor que el lado ΕΖ. Pero ΕΗ es igual a ΒΓ; por tanto, ΒΓ es mayor que ΕΖ.

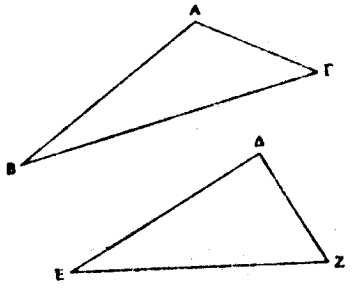
Por consiguiente, si dos triángulos tienen dos lados del uno iguales respectivamente a dos lados del otro, pero tienen uno de los ángulos comprendidos por las rectas iguales mayor que el otro, también tendrán la base mayor que la otra. Q. E. D.

Proposición 25

Si dos triángulos tienen dos lados del uno iguales respectivamente a dos lados del otro, pero tienen la base (del uno) mayor que la base (del otro), también tendrán el ángulo comprendido por las rectas iguales (del uno) mayor que el del otro.

Sean ΑΒΓ, ΔΕΖ dos triángulos que tienen los dos lados ΑΒ, ΑΓ iguales a los dos lados ΔΕ, ΔΖ, respectivamente: ΑΒ a ΔΕ y ΑΓ a ΔΖ; pero sea la base ΒΓ mayor que la base ΕΖ.

Digo que el ángulo BAΓ también es mayor que el ángulo EΔZ.



Pues si no, o bien es igual a él o bien menor; ahora bien, el (ángulo) BAΓ no es igual al (ángulo) EΔZ; pues la base BΓ también sería igual a la base EZ [I, 4]; pero no lo es. Por tanto, el ángulo BAΓ no es igual al (ángulo) EΔZ; pero el (ángulo) BAΓ tampoco es menor que el (ángulo) EΔZ; pues la base BΓ también sería menor que la base EZ [I, 24]; pero no lo es; por tanto, el ángulo BAΓ tampoco es menor que el (ángulo) EΔZ. Pero se ha demostrado que tampoco es igual; luego el (ángulo) BAΓ es mayor que el (ángulo) EΔZ.

Por consiguiente, si dos triángulos tienen dos lados del uno iguales, respectivamente, a dos lados del otro, pero tienen la base (del uno) mayor que la base (del otro), también tendrán el ángulo comprendido por las rectas iguales (del uno) mayor que el del otro. Q. E. D.

Proposición 26

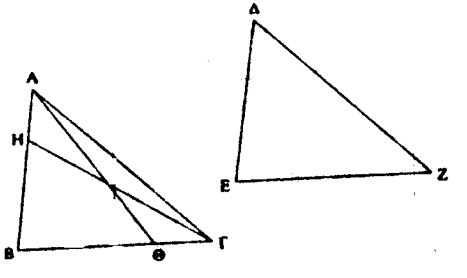
Si dos triángulos tienen dos ángulos del uno iguales respectivamente a dos ángulos del otro y un lado del uno igual a un lado del otro: ya sea el correspondiente a los

ángulos iguales o el que subtiende uno de los ángulos iguales, tendrán también los lados restantes iguales a los lados restantes y el ángulo restante (igual) al ángulo restante.

Sean ABΓ, ΔEZ dos triángulos que tienen los dos ángulos ABΓ, ΓBA iguales respectivamente a los dos ángulos ΔEZ, EZΔ: ABΓ igual a ΔEZ y BΓA igual a EZΔ; y tengan también un lado igual a un lado, en primer lugar el correspondiente a los ángulos iguales: BΓ igual a EZ.

Digo que también tendrán los lados restantes iguales respectivamente a los lados restantes: AB a ΔE y AΓ a ΔZ y el ángulo restante al ángulo restante: BAΓ a EΔZ.

Pues si AB no es igual a ΔE, uno de ellos es mayor. Sea el mayor AB, y hágase BH igual a ΔE, y trácese HΓ.



Así pues, como BH es igual a ΔE, y BΓ a EZ, los dos lados BH, BΓ son iguales respectivamente a los dos lados ΔE, EZ; y el ángulo HBΓ es igual al ángulo ΔEZ; por tanto, la base HΓ es igual a la base ΔZ, y el triángulo HBΓ es igual al triángulo ΔEZ, y los ángulos restantes, subtendidos por los lados iguales, serán también iguales [I, 4]; por tanto, el ángulo HΓB es igual al ángulo ΔZE. Pero se ha supuesto que el ángulo ΔZE es igual al ángulo BΓA; por tanto, el ángulo BΓH es también igual al ángulo BΓA, el menor al mayor; lo cual es imposible. Por tanto, AB no

es desigual a ΔE. Luego es igual. Pero también BΓ es igual a EZ; entonces, los dos (lados) AB, BΓ son iguales respectivamente a los dos (lados) ΔE, EZ; y el ángulo ABΓ es igual al ángulo ΔEZ; por tanto, la base AΓ es igual a la base ΔZ, y el ángulo restante BAΓ es igual al ángulo restante EΔZ [I, 4].

Pero sean iguales a su vez los lados que subtienden a los ángulos iguales, como AB a ΔE.

Digo, asimismo, que los lados restantes también serán iguales a los lados restantes: AΓ a ΔZ y BΓ a EZ y además el ángulo restante BAΓ es igual al ángulo restante EΔZ.

Pues si BΓ no es igual a EZ, uno de ellos es mayor. Sea el mayor, si es posible, BΓ y hágase BΘ igual a EZ, y trácese AΘ. Y puesto que BΘ es igual a EZ y AB a ΔE, los dos lados AB, BΘ son iguales respectivamente a los dos (lados) ΔE, EZ; y comprenden ángulos iguales; por tanto, la base AΘ es igual a la base ΔZ y el triángulo ABΘ es igual al triángulo ΔEZ, y los ángulos restantes, subtendidos por los lados iguales serán también iguales [I, 4]; por tanto, el ángulo BΘA es igual al (ángulo) EZΔ. Pero el (ángulo) EZΔ es igual al (ángulo) BΓA. Entonces el ángulo externo BΘA del triángulo AΘΓ es igual al (ángulo) interno y opuesto BΓA; lo cual es imposible; por tanto, BΓ no es desigual a EZ; luego es igual. Pero también AB es igual a ΔE. Entonces los dos lados AB, BΓ son iguales respectivamente a los dos lados ΔE, EZ; y comprenden ángulos iguales; por tanto, la base AΓ es igual a la base ΔZ, y el triángulo ABΓ es igual al triángulo ΔEZ, y el ángulo restante BAΓ es igual al ángulo restante EΔZ [I, 4].

Por consiguiente, si dos triángulos tienen dos ángulos del uno iguales respectivamente a dos ángulos del otro y un lado del uno igual a un lado del otro: ya sea el correspondiente a los ángulos iguales o el que subtiende uno de

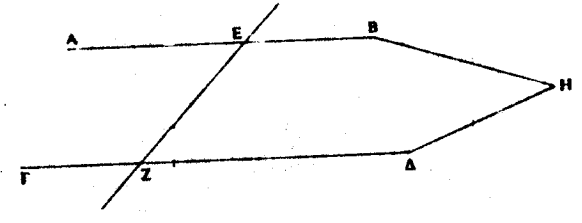
los ángulos iguales, tendrán también los lados restantes iguales a los lados restantes y el ángulo restante (igual) al ángulo restante. Q. E. D.

Proposición 27

Si una recta al incidir sobre dos rectas [47] *hace los ángulos alternos* [48] *iguales entre sí, las dos rectas serán paralelas entre sí.*

Así pues, al incidir sobre las dos rectas AB, ΓΔ, haga la recta EZ los ángulos alternos AEZ, EZΔ iguales entre sí.

Digo que AB es paralela a ΓΔ.



[47] *Eis dúo eutheías eutheîa empiptousa.* Se trata de la misma expresión utilizada en el postulado 5 para referirse a una transversal.

[48] La expresión utilizada para los ángulos alternos es *hai enallàx gōniai.* Según Proclo, Euclides emplea *enallάx* («alternativamente») en dos contextos:

1. En el de cierta transformación de una proporción en el libro V y en los libros sobre aritmética.

2. Como aquí, en el caso de una recta que corta paralelas formando ángulos con ellas. *Hai enallàx gōniai* son los ángulos internos que, quedando en lados distintos de la transversal, no son adyacentes. Seguramente, el uso de *enallάx* responde a que si se imaginan los ángulos internos contados en círculo, los alternos serían cualesquiera de los cuatro no sucesivos sino separados por otro de ellos.

Pues si no, prolongadas AB, ΓΔ, se encontrarán o bien en el sentido de B, Δ, o bien en el sentido de A, Γ. Prolónguense y encuéntrense en el sentido de B, Δ en H. Entonces el ángulo externo AEZ del triángulo HEZ es igual al ángulo interno y opuesto EZH; lo cual es imposible [I, 16]; por tanto, AB, ΓΔ prolongadas no se encuentran en el sentido de B, Δ. De manera semejante demostraríamos que tampoco en el sentido de A, Γ; pero las rectas que no se encuentran en ninguno de los dos sentidos son paralelas [Def. 23]; por tanto, AB es paralela a ΓΔ.

Por consiguiente, si una recta al incidir sobre dos rectas hace los ángulos alternos iguales entre sí, las dos rectas serán paralelas. Q. E. D. [49].

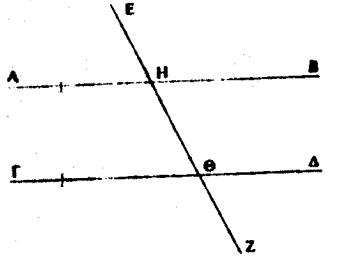
Proposición 28

Si una recta al incidir sobre dos rectas hace el ángulo externo igual al interno y opuesto del mismo lado, o los dos internos del mismo lado iguales a dos rectos, las rectas serán paralelas entre sí.

[49] Con esta proposición comienza la segunda sección del libro I. Hasta aquí Euclides ha tratado principalmente de los triángulos, su construcción y propiedades en el sentido de la relación entre sus partes, lados y ángulos, y la comparación entre triángulos diferentes en lo que se refiere a sus partes y a sus áreas en los casos particulares en los que son congruentes. La segunda sección lleva a la tercera, a partir de la proposición 35, en la que se pasa a la relación entre áreas de triángulos, paralelogramos y cuadrados con una nueva concepción de igualdad de áreas en la que la igualdad no depende de la congruencia. Todo esto requiere el uso de las paralelas. Consecuentemente la segunda sección que empieza en I 27 establece la teoría de las paralelas, introduce el concepto correlativo de la igualdad de la suma de los ángulos de un triángulo a dos rectos (I 32), y finaliza con dos proposiciones que forman la transición a la tercera sección (I 33, y I 34). Éstas introducen el paralelogramo por primera vez.

Así pues, al incidir sobre las dos rectas AB, ΓΔ, haga la recta EZ el ángulo externo EHB igual al interno y opuesto HΘΔ, o los ángulos internos del mismo lado: BHΘ, HΘΔ iguales a dos rectos.

Digo que AB es paralela a ΓΔ [I, 27].



Pues como el (ángulo) EHB es igual al (ángulo) HΘΔ, mientras que el (ángulo) EHB es igual al (ángulo) AHΘ [I, 15]; entonces el (ángulo) AHΘ también es igual al (ángulo) HΘΔ; y son alternos; por tanto, AB es paralela a ΓΔ [I, 27].

Como los (ángulos) BHΘ, HΘΔ son iguales a su vez a dos rectos, también los ángulos AHΘ, BHΘ son iguales a dos rectos, por tanto, los (ángulos) AHΘ, BHΘ son iguales a los (ángulos) BHΘ, HΘΔ; quítese de ambos el ángulo BHΘ; entonces el (ángulo) restante AHΘ es igual al (ángulo) restante HΘΔ; y son alternos. Por tanto, AB es paralela a ΓΔ [I, 27].

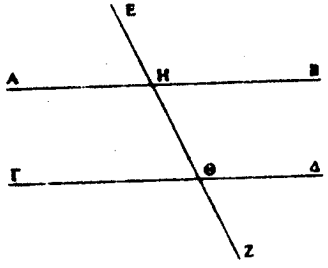
Por consiguiente, si una recta al incidir sobre dos rectas hace el ángulo externo igual al interno y opuesto del mismo lado o los dos internos del mismo lado iguales a dos rectos, las rectas serán paralelas entre sí. Q. E. D.

Proposición 29

La recta que incide sobre rectas paralelas hace los ángulos alternos iguales entre sí, y el (ángulo) externo igual al interno y opuesto, y los (ángulos) internos del mismo lado iguales a dos rectos.

Incida la recta EZ sobre las rectas paralelas AB, ΓΔ.

Digo que hace iguales los ángulos alternos AHΘ, HΘΔ, y el ángulo externo EHB igual al interno y opuesto HΘΔ, y los internos del mismo lado: BHΘ, HΘΔ, iguales a dos rectos.



Pues si AHΘ no es igual a HΘΔ, uno de ellos es mayor. Sea AHΘ el mayor; añádase a ambos el (ángulo) BHΘ; entonces los (ángulos) AHΘ, BHΘ son mayores que los (ángulos) BHΘ, HΘΔ. Pero los (ángulos) AHΘ, BHΘ son iguales a dos rectos [I, 13]. Por tanto, los (ángulos) BHΘ, HΘΔ son menores que dos rectos. Pero las rectas prolongadas indefinidamente a partir de (ángulos) menores que dos rectos se encuentran [Post. 5]; luego las rectas AB, ΓΔ, prolongadas indefinidamente se encontrarán; pero no se encuentran, porque se las ha supuesto paralelas; por tanto,

el (ángulo) AHΘ no es desigual al (ángulo) HΘΔ; luego es igual. Pero el (ángulo) AHΘ es igual al (ángulo) EHB [I, 15]; por tanto, el (ángulo) EHB es también igual al (ángulo) HΘΔ [N. C. 1]. Añádase a ambos BHΘ; entonces los (ángulos) EHB, BHΘ son iguales a los (ángulos) BHΘ, HΘΔ [N. C. 2]. Pero los ángulos EHB, BHΘ son iguales a dos rectos [I, 13]; por tanto, los (ángulos) BHΘ, HΘΔ son también iguales a dos rectos.

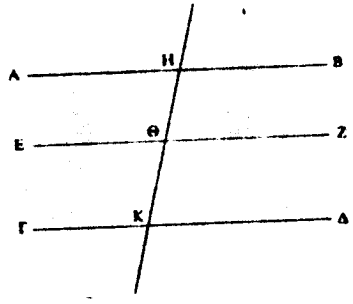
Por consiguiente, la recta que incide sobre rectas paralelas hace los ángulos alternos iguales entre sí, y el (ángulo) externo igual al interno y opuesto, y los (ángulos) internos del mismo lado iguales a dos rectos. Q. E. D. [50].

[50] Las cuatro proposiciones I 17 —que en principio se supone independiente del postulado quinto—; I 28; I 29 y el postulado mismo de las paralelas constituyen un núcleo de la teoría euclídea comúnmente denominado «cuadro de las paralelas» que ha atraído la atención de los comentadores italianos de Euclides. Es un cuadro de relaciones de implicación de modo que supone la formulación de todas esas proposiciones en términos condicionales y en el contexto de una recta que cae sobre otras dos rectas. Según esto, I 17 se lee: «si la recta al caer forma un triángulo, los ángulos internos del mismo lado suman menos que dos rectos» y reviste —digamos— la forma 'si α, β'. De donde, por contraposición de la implicación, se puede deducir I 28 con la forma 'si no β, no α'. Por otra parte, el postulado quinto se deja leer entonces como el condicional converso de I 17: «si al caer la recta los lados internos del mismo lado suman menos que dos rectos, forma un triángulo», *i. e.* bajo la forma 'si β, α'. Pues bien, del postulado quinto así formulado, por contraposición, se deduce I 29, *i. e.* 'si no α, no β'. Podría ser una muestra de la conversión geométrica que alaba Proclo cuando enumera las excelencias de los *Elementos* (*Com.* 69, 21-22).

Proposición 30

Las paralelas a una misma recta son también paralelas entre sí.

Sean cada una de las (rectas) AB, ΓΔ paralelas a EZ. Digo que también AB es paralela a ΓΔ.



Pues incida sobre ellas la recta HK. Y dado que la recta HK ha incidido sobre las rectas paralelas AB, EZ, entonces el (ángulo) AHK es igual al (ángulo) HΘZ [I, 29]. Como a su vez la recta HK ha incidido sobre las rectas paralelas EZ, ΓΔ, el (ángulo) HΘZ es igual al ángulo HKΔ [I, 29]. Pero se ha demostrado también que el (ángulo) AHK es igual al ángulo HΘZ. Por tanto, el ángulo AHK también es igual al ángulo HKΔ [N. C. 1]; y son alternos. Por tanto, AB es paralela a ΓΔ.

[Por consiguiente, las paralelas a una misma recta son también paralelas entre sí] Q. E. D. [51].

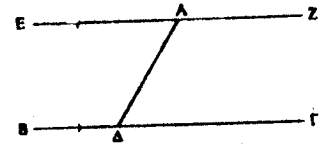
[51] El enunciado de la conclusión no aparece en los principales mss.

Proposición 31

Por un punto dado trazar una línea recta paralela a una recta dada.

Sea A el punto dado y BΓ la recta dada.

Hay que trazar por el punto A una línea recta paralela a BΓ.



Tómese al azar un punto Δ en BΓ y trácese AΔ; y constrúyase en la recta ΔA y en el punto A de ella el ángulo ΔAE igual al (ángulo) AΔΓ [I, 23]; y sea AZ el resultado de prolongar en línea recta EA.

Y dado que la recta AΔ al incidir sobre las dos rectas BΓ, EZ ha hecho iguales los ángulos alternos EAΔ, AΔΓ, entonces EAZ es paralela a BΓ [I, 27].

Por consiguiente, se ha trazado la línea recta EAZ paralela a la recta dada BΓ por el punto dado A. Q. E. F.

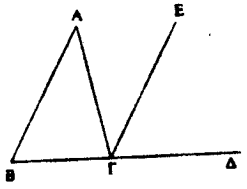
Proposición 32

En todo triángulo, si se prolonga uno de los lados, el ángulo externo es igual a los dos ángulos internos y opuestos, y los tres ángulos internos del triángulo son iguales a dos rectos.

Sea ABΓ el triángulo, y prolónguese uno de sus lados, BΓ, hasta Δ.

Digo que el ángulo externo ΑΓΔ es igual a los dos internos y opuestos, ΓΑΒ, ΑΒΓ, y los tres ángulos internos del triángulo, ΑΒΓ, ΒΓΑ, ΓΑΒ son iguales a dos rectos.

Pues trácese por el punto Γ (la recta) ΓΕ paralela a la recta ΑΒ [I, 31].



Y puesto que ΑΒ es paralela a ΓΕ y ΑΓ ha incidido sobre ellas, los ángulos alternos ΒΑΓ, ΑΓΕ son iguales entre sí [I, 29]. Puesto que, a su vez, ΑΒ es paralela a ΓΕ y la recta ΒΔ ha incidido sobre ellas, el (ángulo) externo ΕΓΔ es igual al interno y opuesto ΑΒΓ [I, 29]. Pero se ha demostrado que el (ángulo) ΑΓΕ es también igual al (ángulo) ΒΑΓ; por tanto, el ángulo entero ΑΓΔ es igual a los dos internos y opuestos ΒΑΓ, ΑΒΓ.

Añádase al uno y a los otros el ángulo ΑΓΒ; entonces los (ángulos) ΑΓΔ, ΑΓΒ son iguales a los tres (ángulos) ΑΒΓ, ΒΓΑ, ΓΑΒ. Pero los (ángulos) ΑΓΔ, ΑΓΒ son iguales a dos rectos [I, 13]; por tanto, los (ángulos) ΑΓΒ, ΓΒΑ, ΓΑΒ son también iguales a dos rectos.

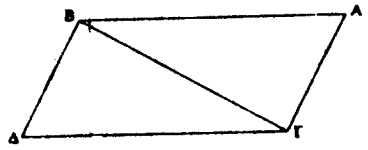
Por consiguiente, en todo triángulo, si se prolonga uno de los lados, el ángulo externo es igual a los dos ángulos internos y opuestos, y los tres ángulos internos del triángulo son iguales a dos rectos. Q. E. D.

Proposición 33

Las rectas que unen por (los extremos que están en) el mismo lado a (rectas) iguales y paralelas son también ellas mismas iguales y paralelas [52].

[52] *Hai tàs ísas te kai parallélous epì tà autà mére epidseugnýusai eutheîai...* dice literalmente: «las rectas que unen por el mismo lado a iguales y paralelas». He añadido las palabras entre paréntesis con vistas a una mayor claridad, pues la versión literal del enunciado resulta bastante confusa.

Sean ΑΒ, ΓΔ las (rectas) iguales y paralelas y trácense uniéndolas por (los extremos de) el mismo lado las rectas ΑΓ, ΒΔ.



Digo que ΑΓ, ΒΔ son también iguales y paralelas.

Trácese ΒΓ. Y puesto que ΑΒ es paralela a ΓΔ, y ΒΓ ha incidido sobre ellas, los ángulos alternos ΑΒΓ, ΒΓΔ son iguales entre sí [I, 29]. Y puesto que ΑΒ es igual que ΓΔ y ΒΓ es común, las dos (rectas) ΑΒ, ΒΓ son iguales a las dos (rectas) ΒΓ, ΓΔ; y el ángulo ΑΒΓ es igual al ángulo ΒΓΔ; por tanto, la base ΑΓ es igual a la base ΒΔ, y el triángulo ΑΒΓ es igual al triángulo ΒΓΔ, y los ángulos restantes, subtendidos por los lados iguales, serán también iguales respectivamente [I, 4]; por tanto, el ángulo ΑΓΒ es igual al ángulo ΓΒΔ. Y dado que la recta ΒΓ que incide sobre las dos rectas ΑΓ, ΒΔ ha hecho iguales entre sí los ángulos alternos, entonces ΑΓ es paralela a ΒΔ [I, 27]. Pero se ha demostrado que también es igual a ella.

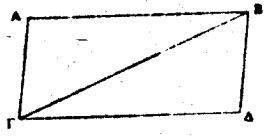
Por consiguiente, las rectas que unen por (los extremos que están en) el mismo lado a (rectas) iguales y paralelas, son también ellas mismas iguales y paralelas. Q. E. D.

Proposición 34

En las áreas de paralelogramos [53] *los lados y los ángulos opuestos son iguales entre sí, y la diagonal* [54] *las divide en dos partes iguales.*

Sea ΑΓΔΒ el área de paralelogramo, y su diagonal ΒΓ.

Digo que los lados y los ángulos opuestos del paralelogramo ΑΓΔΒ son iguales entre sí, y (que) la diagonal ΒΓ lo divide en dos partes (iguales).



Pues como ΑΒ es paralela a ΓΔ, y la recta ΒΓ ha incidido sobre ellas, los ángulos alternos ΑΒΓ, ΒΓΔ son iguales entre sí [I, 29]. Como a su vez ΑΓ es paralela a ΒΔ, y ΒΓ ha incidido sobre ellas, los ángulos alternos ΑΓΒ, ΓΒΔ son iguales entre sí [I, 29]. Entonces ΑΒΓ, ΒΓΔ son dos triángulos que tienen los dos ángulos ΑΒΓ, ΒΓΑ iguales respectivamente a los dos ángulos ΒΓΔ, ΓΒΔ y un lado igual a un lado, el correspondiente a los ángulos iguales, común a ellos: ΒΓ; así pues, también tendrá los lados restantes iguales respectivamente a los lados restantes, y el ángulo restante (igual) al ángulo restante [I, 26]; por tanto, el lado ΑΒ es igual al (lado) ΓΔ, el (lado) ΑΓ al ΒΔ, y además el ángulo ΒΑΓ es igual al (ángulo) ΓΔΒ. Y dado que el ángulo ΑΒΓ es igual al (ángulo) ΒΓΔ, y el (ángulo) ΓΒΔ al (ángulo) ΑΓΒ, entonces el (ángulo) entero ΑΒΔ es igual al (ángulo) entero ΑΓΔ [N. C. 2]. Pero se ha demostrado que el ángulo ΒΑΓ también es igual al ángulo ΓΔΒ.

[53] La primera vez que aparecen mencionados los paralelogramos Euclides utiliza *parallelográmmōn chorión* («áreas paralelogramas»). Se refiere a áreas limitadas por rectas paralelas. El uso euclídeo impone además la restricción de que esta denominación sólo se aplique a figuras de cuatro lados. Proclo dice que la palabra «paralelogramo» fue introducida por primera vez por Euclides, que su uso estaba sugerido en I, 33, y que la formación de *parallelógrammon* (de líneas paralelas) fue análoga a la de *euthygrammos* (de líneas rectas).

[54] La palabra empleada por Euclides es *diámetros*, que en este contexto significa obviamente «diagonal».

Por consiguiente, en las áreas de paralelogramos los lados y ángulos opuestos son iguales entre sí.

Digo entonces que también la diagonal las divide en dos partes iguales.

Pues como ΑΒ es igual a ΓΔ y ΒΓ es común, las dos ΑΒ, ΒΓ son iguales respectivamente a las dos ΓΔ, ΒΓ; y el ángulo ΑΒΓ es igual al ángulo ΒΓΔ. Por tanto la base ΑΓ es también igual a ΔΒ. Y el triángulo ΑΒΓ es también igual al triángulo ΒΓΔ [I, 4].

Por consiguiente, la diagonal ΒΓ divide en dos partes (iguales) el paralelogramo ΑΒΓΔ. Q. E. D.

Proposición 35

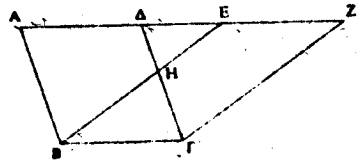
Los paralelogramos que están sobre la misma base y entre las mismas paralelas son iguales entre sí.

Sean ΑΒΓΔ, ΕΒΓΖ los paralelogramos (que están) sobre la misma base ΒΓ y entre las mismas paralelas ΑΖ, ΒΓ.

Digo que (el paralelogramo) ΑΒΓΔ es igual al paralelogramo ΕΒΓΖ.

Pues como ΑΒΓΔ es un paralelogramo, ΑΔ es igual a ΒΓ [I, 34]. Por lo mismo, ΕΖ también es igual a ΒΓ; de

modo que también ΑΔ es igual a ΕΖ [N. C. 1]; y ΔΕ es común; por tanto, la (recta) entera ΑΕ es igual a la recta entera ΔΖ [N. C. 2]. Y ΑΒ es también igual a ΔΓ [I, 34]; entonces los dos (lados) ΕΑ, ΑΒ son iguales respectivamente a los dos (lados) ΖΔ, ΔΓ; y el ángulo ΖΔΓ es igual al ángulo ΕΑΒ, el externo al interno [I, 29]; por tanto, la base ΕΒ es igual a la base ΖΓ, y el triángulo ΕΑΒ será igual al triángulo ΔΖΓ [I, 4]; quítese de ambos el triángulo ΑΗΕ; entonces el trapecio restante ΑΒΗΔ es igual al trapecio restante ΕΗΓΖ [N. C. 3]; añádase a ambos el triángulo ΗΒΓ; entonces el paralelogramo entero ΑΒΓΔ es igual al paralelogramo entero ΕΒΓΖ [N. C. 2].



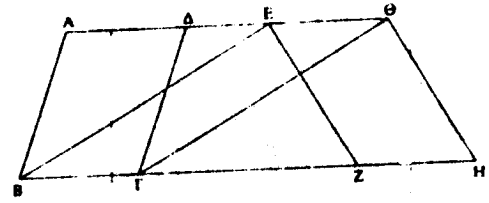
Por consiguiente, los paralelogramos que están sobre la misma base y entre las mismas paralelas son iguales entre sí. Q. E. D.

Proposición 36

Los paralelogramos que están sobre bases iguales y entre las mismas paralelas son iguales entre sí.

Sean ΑΒΓΔ, ΕΖΗΘ los paralelogramos que están sobre bases iguales ΒΓ, ΖΗ y entre las mismas paralelas ΑΘ, ΒΗ.

Digo que el paralelogramo ΑΒΓΔ es igual al paralelogramo ΕΖΗΘ.



Trácense, pues, ΒΕ y ΓΘ. Y dado que ΒΓ es igual a ΖΗ, mientras que ΖΗ es igual a ΕΘ, entonces ΒΓ es también igual a ΕΘ [N. C. 1]. Pero son también paralelas. Y ΕΒ, ΘΓ las unen; pero las (rectas) que unen por (los extremos de) el mismo lado (rectas) iguales y paralelas son iguales y paralelas [I, 33]; por tanto, ΕΒΓΘ es un paralelogramo [I, 34]. Y es igual a ΑΒΓΔ: pues tiene la misma base que él, ΒΓ y está entre las mismas paralelas que él: ΒΓ, ΑΘ [I, 35]. Por lo mismo, ΕΖΗΘ es también igual al mismo ΕΒΓΘ [I, 35]. De modo que el paralelogramo ΑΒΓΔ es también igual a ΕΖΗΘ [N. C. 1].

Por consiguiente, los paralelogramos que están sobre bases iguales y entre las mismas paralelas son iguales entre sí. Q. E. D.

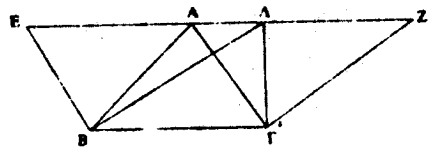
Proposición 37

Los triángulos que están sobre la misma base y entre las mismas paralelas son iguales entre sí.

Sean ΑΒΓ, ΔΒΓ los triángulos que están sobre la misma base ΒΓ y entre las mismas paralelas ΑΔ, ΒΓ.

Digo que el triángulo ΑΒΓ es igual al triángulo ΔΒΓ.

Prolónguese AΔ en ambos sentidos hasta E, Z y por el (punto) B trácese BE paralela a ΓΔ [I, 31], y por el punto Γ trácese ΓZ paralela a BΔ [I, 31]. Entonces cada una de las (figuras) EBΓΔ, ΔBΓZ es un paralelogramo; y son iguales: porque están sobre la misma base BΓ y entre las mismas paralelas BΓ, EZ [I, 35]; y el triángulo ABΓ es la mitad del paralelogramo EBΓΔ: porque la diagonal AB lo divide en dos partes (iguales) [I, 34]; y el triángulo ΔBΓ es la mitad del paralelogramo ΔBΓZ: porque la diagonal ΔΓ lo divide en dos partes (iguales) [I, 34]. [Pero las mitades de cosas iguales son iguales entre sí] [55]. Por tanto, el triángulo ABΓ es igual al triángulo ΔBΓ.



Por consiguiente, los triángulos que están sobre la misma base y entre las mismas paralelas son iguales entre sí. Q. E. D. [55bis].

[55] Aquí y en la proposición siguiente, Heiberg suprime las palabras *tà tôn isōn hemisē ísa allélois estín*. La noción común a que hace referencia es una interpolación muy antigua (anterior a Teón); por lo tanto, es probable que estas palabras fueran interpoladas en la misma época.

Los teoremas I, 35, y I, 37 muestran que dos figuras planas pueden tener la misma superficie sin contar con el mismo perímetro. Este resultado muestra el error de los navegantes que creían calcular el área de una isla por el tiempo empleado en bordear su perímetro (TUCÍDIDES, *Historia de la Guerra del Peloponeso* VI 1).

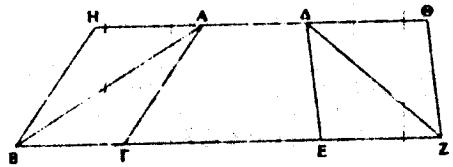
[55bis] Heiberg considera esta frase como una interpolación sobre la base de *Fayûm*, pág. 96, núm. 9 (ver HEATH, ed. cit., 1, 1926², pág. 337). Stamatis la mantiene sin la menor reserva.

PROPOSICIÓN 38

Los triángulos que están sobre bases iguales y entre las mismas paralelas son iguales entre sí.

Sean ABΓ, ΔEZ los triángulos (que están) sobre las bases iguales BΓ, EZ y entre las mismas paralelas BZ, AΔ.

Digo que el triángulo ABΓ es igual al triángulo ΔEZ.



Prolónguese, pues, AΔ en ambos sentidos hasta H, Θ, y por el (punto) B trácese BH paralela a ΓΔ [I, 31], y por el (punto) Z trácese ZΘ paralela a ΔE. Entonces cada una de las (figuras) HBΓΔ, ΔEZΘ es un paralelogramo; y HBΓΔ es igual a ΔEZΘ: porque está sobre las bases iguales BΓ, EZ y entre las mismas paralelas BZ, HΘ [I, 36]; y el triángulo ABΓ es la mitad del paralelogramo HBΓΔ: porque la diagonal AB lo divide en dos partes iguales [I, 34]; y el triángulo ZEΔ es la mitad del paralelogramo ΔEZΘ: porque la diagonal ΔZ lo divide en dos partes iguales [I, 34]; [y las mitades de las mismas cosas son iguales entre sí] [56]. Por tanto, el triángulo ABΓ es igual al triángulo ΔEZ.

Por consiguiente, los triángulos que están sobre bases iguales y entre las mismas paralelas son iguales entre sí. Q. E. D.

[56] Cf. nota 55 bis.

Proposición 39

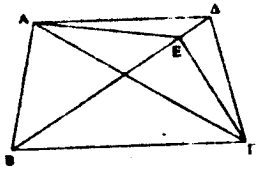
Los triángulos iguales que están sobre la misma base y en el mismo lado, están también entre las mismas paralelas.

Sean ΑΒΓ, ΔΒΓ triángulos iguales que están sobre la misma base ΒΓ y en el mismo lado de ella.

[Digo que también están entre las mismas paralelas] [56bis].

[Pues] trácese ΑΔ.

Digo que ΑΔ es paralela a ΒΓ.



Pues si no trácese por el (punto) Α la recta ΑΕ paralela a ΒΓ [I, 31], y trácese ΕΓ. Entonces el triángulo ΑΒΓ es igual al triángulo ΕΒΓ: porque está sobre la misma base que él, ΒΓ, y entre las mismas paralelas [I, 37]. Pero ΑΒΓ es igual a ΔΒΓ; por tanto, ΔΒΓ es también igual a ΕΒΓ [N. C. 1], el mayor al menor; lo cual es imposible; por tanto, ΑΕ no es paralela a ΒΓ. De manera semejante demostraríamos que ninguna otra (lo es) excepto ΑΔ; por tanto, ΑΔ es paralela a ΒΓ.

Por consiguiente, los triángulos iguales que están sobre la misma base y en el mismo lado, están también entre las mismas paralelas. Q. E. D.

[56bis] Cf. nota 55bis.

Proposición 40

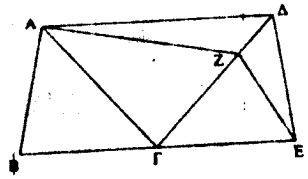
Los triángulos iguales que están sobre bases iguales y en el mismo lado, están también entre las mismas paralelas.

Sean ΑΒΓ, ΓΔΕ triángulos iguales que están sobre las bases iguales ΒΓ, ΓΕ y en el mismo lado.

Digo que también están entre las mismas paralelas.

Pues trácese ΑΔ.

Digo que ΑΔ es paralela a ΒΕ.



Pues si no, por el punto Α trácese ΑΖ paralela a ΒΕ [I, 31], trácese además ΖΕ. Entonces el triángulo ΑΒΓ es igual al triángulo ΖΓΕ: porque están sobre las bases iguales ΒΓ, ΓΕ y entre las mismas paralelas ΒΕ, ΑΖ [I, 38]. Pero el triángulo ΑΒΓ es igual al [triángulo] ΔΓΕ; por tanto, el [triángulo] ΔΓΕ es también igual al triángulo ΖΓΕ [N. C. 1], el mayor al menor: lo cual es imposible; por tanto, ΑΖ no es paralela a ΒΕ. De manera semejante demostraríamos que ninguna otra (lo es) excepto ΑΔ; por tanto, ΑΔ es paralela a ΒΕ.

Por consiguiente, los triángulos iguales que están sobre bases iguales y en el mismo lado, están también entre las mismas paralelas [Q. E. D.] [57].

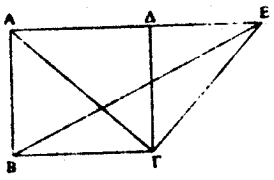
[57] Heiberg considera que esta proposición es una interpolación introducida por alguien que pensó que debía haber una proposición que si-

Proposición 41

Si un paralelogramo tiene la misma base que un triángulo y está entre las mismas paralelas, el paralelogramo es el doble del triángulo.

Tenga el paralelogramo ABΓΔ la misma base, BΓ, que el triángulo EBΓ y esté entre las mismas paralelas BΓ, AE.

Digo que el paralelogramo ABΓΔ es el doble del triángulo BEΓ.



Pues trácese AΓ. Entonces el triángulo ABΓ es igual al triángulo EBΓ: porque está sobre la misma base que él, BΓ, y entre las mismas paralelas BΓ, AE [I, 37]. Pero el paralelogramo ABΓΔ es el doble del triángulo ABΓ: porque la diagonal AΓ lo divide en dos partes (iguales) [I, 34]; de modo que el paralelogramo ABΓΔ es también el doble del triángulo EBΓ.

Por consiguiente, si un paralelogramo tiene la misma base que un triángulo y está entre las mismas paralelas, el paralelogramo es el doble del triángulo. Q. E. D.

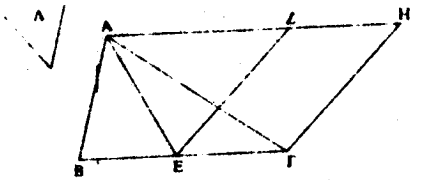
Proposición 42

Construir en un ángulo rectilíneo dado un paralelogramo igual a un triángulo dado.

Sea ABΓ el triángulo dado y Δ el ángulo rectilíneo dado.

guiera a I 39 y estuviera relacionada con ella de la misma manera que I 38 está relacionada con I 37, y I 36 con I 35. Sin embargo, aparece en los principales mss. y en la edición de Stamatis.

Así pues, hay que construir un paralelogramo igual al triángulo ABΓ en el ángulo rectilíneo Δ.



Divídase BΓ en dos por el punto E, trácese AE, y constrúyase en la recta EΓ y en su punto E el (ángulo) ΓEZ igual al ángulo Δ [I, 23], y por el (punto) A trácese AH paralela a EΓ [I, 31], y por el (punto) Γ trácese ΓH paralela a EZ; entonces ZEΓH es un paralelogramo. Y como BE es igual a EΓ, el triángulo ABE es también igual al triángulo AEΓ: porque están sobre las bases iguales BE, EΓ y entre las mismas paralelas BΓ, AH [I, 38]; por tanto, el triángulo ABΓ es el doble del triángulo AEΓ. Pero también el paralelogramo ZEΓH es el doble del triángulo AEΓ: porque tiene la misma base que él y está entre las mismas paralelas [I, 41]; por tanto, el paralelogramo ZEΓH es igual al triángulo ABΓ. Y tiene el ángulo ΓEZ igual al (ángulo) dado Δ.

Por consiguiente, se ha construido el paralelogramo ZEΓH igual al triángulo dado ABΓ, en el ángulo ΓEZ, que es igual al Δ. Q. E. F.

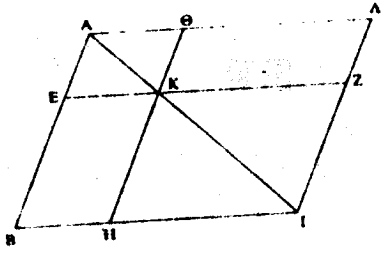
Proposición 43

En todo paralelogramo los complementos [58] *de los paralelogramos situados en torno a la diagonal son iguales entre sí.*

[58] *Paraplerômata*, «complementos», son las figuras que quedan entre los paralelogramos situados en torno a la diagonal y completan el parale-

Sea ΑΒΓΔ el paralelogramo y ΑΓ su diagonal, y sean ΕΘ, ΖΗ los paralelogramos (situados) en torno a ΑΓ, y sean ΒΚ, ΚΔ los llamados complementos.

Digo que el complemento ΒΚ es igual al complemento ΚΔ.



Pues, como ΑΒΓΔ es un paralelogramo y ΑΓ su diagonal, el triángulo ΑΒΓ es igual al triángulo ΑΓΔ [I, 34]; como ΕΘ es a su vez un paralelogramo y ΑΚ es su diagonal, el triángulo ΑΕΚ es igual al triángulo ΑΘΚ. Por la misma razón, el triángulo ΚΖΓ es también igual al triángulo ΚΗΓ. Así pues, como el triángulo ΑΕΚ es igual al triángulo ΑΘΚ, y el triángulo ΚΖΓ al ΚΗΓ, el triángulo ΑΕΚ junto con ΚΗΓ es igual al triángulo ΑΘΚ junto con ΚΖΓ [N. C. 2]; pero también el triángulo entero ΑΒΓ es igual al (triángulo) entero ΑΔΓ; por tanto, el complemento restante ΒΚ es igual al complemento restante ΚΔ [N. C. 3].

Por consiguiente, en todo paralelogramo los complementos de los paralelogramos (situados) en torno a la diagonal son iguales entre sí. Q. E. D.

---

logramo. La fra[illegible] de la *ékthesis* «los llamados complementos» indica que el uso técnico de *paraplerómata* no es nuevo para Euclides, aunque tal vez no fuera de dominio público común.

PROPOSICIÓN 44

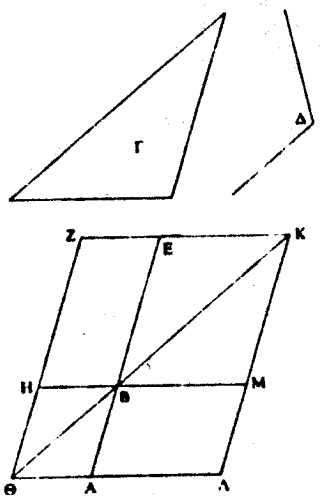
Aplicar a una recta dada en un ángulo rectilíneo dado, un paralelogramo igual a un triángulo dado [59].

---

[59] Merece destacarse este resultado que permite la transformación de un paralelogramo de cualquier tamaño en otro con el mismo ángulo y de área igual, pero con un lado de cualquier longitud dada. También es interesante el procedimiento empleado, el método de aplicación de áreas, cuyo uso se remonta a los pitagóricos. Proclo le dedica una larga nota (*Com.* 419, 15-420, 23) que comienza así: «Estas cosas, según Eudemo, son antiguas y fueron invención de la Musa de los pitagóricos, quiero decir la aplicación de áreas por yuxtaposición [*paraboló tôn chorîôn*], por exceso [*hyperboló*] y por defecto [*élleipsis*]. Los geómetras posteriores tomaron estas denominaciones de los pitagóricos y las trasladaron a las líneas llamadas *cónicas*, de modo que una de ellas nombra la parábola, otra la hipérbola y otra la elipse, mientras que los hombres de la Antiguedad, semejantes a dioses, veían que estos términos significaban la construcción de áreas, en el plano, sobre una línea recta finita. Pues cuando se tiene una línea recta y se extiende el área a todo lo largo de la línea, dicen que se aplica o yuxtapone el área; pero cuando se hace la longitud del área mayor que la propia línea recta, dicen que la excede, y cuando se hace menor, en cuyo caso hay alguna parte de la línea recta que sobrepasa el área trazada, entonces dicen que es deficiente». En otras palabras, el procedimiento por defecto consiste en aplicar un rectángulo (o más en general un paralelogramo) a un segmento de modo que resulte deficiente en un cuadrado (o en una figura rectilínea dada), y por lo regular requiere el uso de un *diorismós*; el procedimiento por exceso corresponde al caso de que el área aplicada exceda del segmento dado de modo similar a como se producía en el caso anterior; el procedimiento de aplicación o yuxtaposición utiliza un paralelogramo de área igual a la figura rectilínea dada. Éste es el empleo en las proposiciones I, 44-45. El uso de *parábola*, *hipérbola* y *elipse* en el estudio de las secciones cónicas no será simplemente un desplazamiento terminológico a otro dominio de referencia, sino que, desde Apolonio al menos, representará así mismo un cambio de marco conceptual.

Sea AB la recta dada, Γ el triángulo dado y Δ el ángulo rectilíneo dado.

Así pues, hay que aplicar a la recta dada AB, en un ángulo igual a Δ, un paralelogramo igual al triángulo dado Γ.



Constrúyase el paralelogramo BEZH igual al triángulo Γ en el ángulo EBH, que es igual al Δ [I, 42]; y hágase de manera que BE esté en línea recta con AB y prolónguese hacia el otro lado ZH hasta Θ, y por el (punto) A trácese AΘ paralela a una de las dos rectas BH, EZ [I, 31], y trácese ΘB. Y dado que la recta ΘZ incide sobre las paralelas AΘ, EZ, entonces los ángulos AΘZ, ΘZE son iguales a dos rectos [I, 29]. Por tanto, los ángulos BΘH, HZE son menores que dos rectos y las rectas prolongadas indefinidamente a partir de ángulos menores que dos rectos, se encuentran [Post. 5]; luego ΘB, ZE prolongadas se encontrarán. Prolónguense y encuéntrense en K, y por el punto

K trácese KΛ paralela a las dos rectas EΛ, ZΘ [I, 31] y prolónguense ΘΛ, HB hasta los puntos KΛ, M. Entonces ZΛKΘ es un paralelogramo y ΘK su diagonal, y AH, ME los paralelogramos (situados) en torno a ΘK, y ΛB, BZ los llamados complementos; por tanto, ΛB es igual a BZ [I, 43]. Pero BZ es igual al triángulo Γ; luego ΛB es también igual a Γ [N. C. 1]. Y como el ángulo HBE es igual al (ángulo) ABM [I, 15], mientras que el ángulo HBE es igual a Δ, entonces el ángulo ABM es también igual al ángulo Δ.

Por consiguiente, se ha aplicado a la recta dada AB el paralelogramo ΛB igual al triángulo dado Γ, en el ángulo ABM que es igual a Δ. Q. E. F.

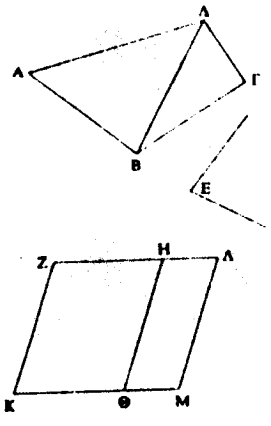
PROPOSICIÓN 45

Construir en un ángulo rectilíneo dado, un paralelogramo igual a una (figura) rectilínea dada [60]

[60] Si las proposiciones I 43-45 nos introducen en el ámbito metódico de la transformación de áreas, también abren la perspectiva de lo que se ha dado en llamar «el álgebra geométrica» de los griegos (vid. *supra*, Introducción, III, § 2). Si un rectángulo corresponde en álgebra al producto de dos magnitudes, se puede considerar la aplicación a una recta dada de un rectángulo igual a una área determinada como el correlato geométrico de una división algebraica del producto de dos cantidades por una tercera. El procedimiento permite así mismo añadir o sustraer áreas rectilíneas y representar la suma o la diferencia por un rectángulo de cualquier longitud dada, de modo parejo a lo que sería la obtención de un factor común. Proclo, a propósito de I 45, sugiere que en su opinión fue la solución de este problema la que llevó a los antiguos geómetras a investigar la cuadratura del círculo. «Pues si cabe hallar un paralelogramo igual a otra figura rectilínea cualquiera, vale la pena plantearse si no será posible también probar la existencia de una figura rectilí-

Sea ΑΒΓΔ la (figura) rectilínea dada y Ε el ángulo rectilíneo dado.

Así pues, hay que construir en el ángulo dado Ε, un paralelogramo igual a la figura rectilínea dada ΑΒΓΔ.



Trácese ΔΒ, y constrúyase en el ángulo ΘΚΖ, que es igual a Ε, el paralelogramo ΖΘ igual al triángulo ΑΒΔ [I, 42]; y aplíquese a la recta ΗΘ, el paralelogramo ΗΜ igual al triángulo ΑΒΓ en el ángulo ΗΘΜ, que es igual a Ε [I, 44]. Y puesto que el ángulo Ε es igual a cada uno de los dos (ángulos) ΘΚΖ, ΗΘΜ, también el (ángulo) ΘΚΖ es, por tanto, igual al (ángulo) ΗΘΜ [N. C. 1]. Añádase a ambos el ángulo ΚΘΗ; entonces los (ángulos) ΖΚΘ, ΚΘΗ son iguales a los (ángulos) ΚΘΗ, ΗΘΜ. Pero los (ángulos) ΖΚΘ, ΚΘΗ son iguales a dos rectos [I, 29]; por tanto, los (ángulos) ΚΘΗ, ΗΘΜ son también iguales a dos rectos. Entonces en una recta cualquiera ΗΘ y en un punto de ella Θ, las dos rectas ΚΘ, ΘΜ, no colocadas en el mismo lado, hacen los ángulos adyacentes iguales a dos rectos; luego ΚΘ está en línea recta con ΘΜ [I, 14]; y como la recta ΘΗ incide sobre las paralelas ΚΜ, ΖΗ, los ángulos alternos ΜΘΗ, ΘΗΖ son iguales entre sí [I, 29]; añádase a ambos el (ángulo) ΘΗΛ; entonces los (ángulos) ΜΘΗ, ΘΗΛ son iguales a los ángulos ΘΗΖ, ΘΗΛ [N. C. 2]. Pero los (ángulos) ΜΘΗ, ΘΗΛ son iguales a dos rectos [I, 29]; luego los ángulos ΘΗΖ, ΘΗΛ son también iguales a dos rectos [N. C. 1]; por tanto, ΖΗ está en línea recta con ΗΛ [I, 14]. Y dado que ΖΚ es igual y paralela a ΘΗ [I, 34], pero también ΘΗ a ΜΛ, entonces también ΚΖ es igual y paralela a ΜΛ [N. C. 1; I, 30]; y las rectas ΚΜ, ΖΛ las unen (por sus extremos); luego también ΚΜ, ΖΛ son iguales y paralelas [I, 33]; por tanto, ΚΖΛΜ es un paralelogramo. Y dado que el triángulo ΑΒΔ es igual al paralelogramo ΖΘ, y el (triángulo) ΔΒΓ al (paralelogramo) ΗΜ, entonces la (figura) rectilínea entera ΑΒΓΔ es igual al paralelogramo entero ΚΖΛΜ.

Por consiguiente, se ha construido en el ángulo ΖΚΜ, que es igual al (ángulo) dado Ε, el paralelogramo ΚΖΛΜ igual a la (figura) rectilínea dada ΑΒΓΔ. Q. E. F.

nea igual a una área circular» (*Com.* 422, 24-423, 2). Puede que este problema de la cuadratura del círculo se planteara inicialmente como una extensión de los métodos elementales de la geometría plana, como si se tratara de un problema *plano*, antes de caer en la cuenta de que la empresa estaba condenada al fracaso. *Vid.* T. L. HEATH (1981²): *A History of Greek Mathematics*..., op. cit., I, vii, págs. 218 ss.

PROPOSICIÓN 46

Trazar un cuadrado a partir de una recta dada [61].

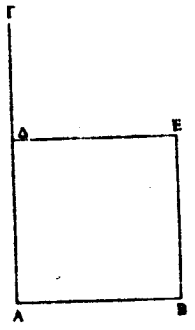
Sea ΑΒ la recta dada.

Así pues, hay que trazar un cuadrado a partir de ΑΒ.

Trácese la recta ΑΓ que forme ángulos rectos con la recta ΑΒ desde su punto Α, y hágase ΑΔ igual a ΑΒ; y por

[61] Euclides utiliza distintos verbos para referirse a la construcción de distintas figuras geométricas. Para dibujar un cuadrado a partir de un lado la expresión usada es *anagrápsai apó*, mientras que en el caso del triángulo utiliza *systésasthai*, o para describir un círculo, *gráphesthai*. *Anagrápho apó* indica la acción de dibujar repetidamente a partir de una recta dada (un lado) las demás rectas (lados) que cierran un cuadrado.

el punto Δ trácese ΔE paralela a AB, y por el punto B trácese BE paralela a AΔ [I, 31]. Entonces AΔEB es un paralelogramo; por tanto, AB es igual a ΔE, y AΔ a BE [I, 34]. Pero AB es igual a AΔ; luego las cuatro rectas BA, AΔ, ΔE, EB son iguales entre sí; entonces el paralelogramo AΔEB es equilátero.



Además digo que también es rectangular. Pues, dado que la recta AΔ incide sobre las paralelas AB, ΔE, entonces los ángulos BAΔ, AΔE son iguales a dos rectos [I, 29]. Pero el ángulo BAΔ es recto; por tanto, el (ángulo) AΔE también es recto. Ahora bien, en las áreas de paralelogramos los lados y ángulos opuestos son iguales entre sí [I, 34]; por tanto, cada uno de los ángulos opuestos ABE, BEΔ también es recto; luego AΔEB es rectangular. Pero se ha demostrado que también es equilátero.

Por consiguiente, es un cuadrado; y está trazado a partir de la recta AB. Q. E. F.

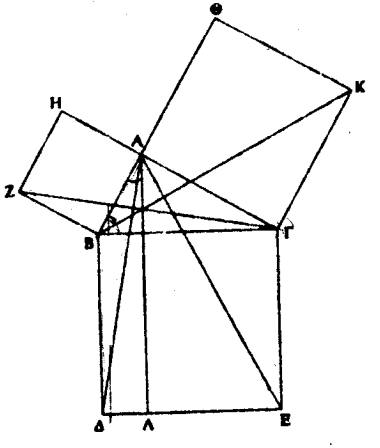
Proposición 47

En los triángulos rectángulos el cuadrado del lado que subtiende el ángulo recto es igual a los cuadrados de los lados que comprenden el ángulo recto.

Sea ABΓ el triángulo rectángulo que tiene el ángulo recto BAΓ.

Digo que el cuadrado de BΓ es igual a los cuadrados de BA, AΓ.

Trácese pues a partir de BΓ el cuadrado BΔEΓ, y a partir de BA, AΓ los cuadrados HB, ΘΓ [I, 46], y por el (punto) A trácese AΔ paralela a una de las dos (rectas) BΔ, ΓE; y trácense AΔ, ZΓ. Y dado que cada uno de los ángulos BAΓ, BAH es recto, entonces en una recta cualquiera BA y por un punto de ella, A, las dos rectas AΓ, AH, no colocadas



en el mismo lado, hacen los ángulos adyacentes iguales a dos rectos; por tanto, IA está en línea recta con AH [I, 14]. Por la misma razón BA también está en línea recta con AΘ. Y como el ángulo ΔBΓ es igual al (ángulo) ZBA —porque cada uno (de ellos) es recto— añádase a ambos el (ángulo) ABΓ; entonces el (ángulo) entero ΔBA es igual al (ángulo) entero ZBΓ [N. C. 2]; y como ΔB es igual a BΓ, y ZB a BA, los dos (lados) ΔB, BA son iguales respectivamente a los dos (lados) ZB, BΓ; y el ángulo ΔBA es igual al ángulo ZBΓ; entonces la base AΔ es igual a la base ZΓ, y el triángulo ABΔ es igual al triángulo ZBΓ [I, 4]; y el paralelogramo BΛ es el doble del triángulo ABΔ: porque

tienen la misma base BΔ y están entre las mismas paralelas BΔ, AΛ [I, 41]; pero el cuadrado HB es el doble del triángulo ZBΓ: porque tienen a su vez la misma base ZB y están entre las mismas paralelas ZB, HΓ [I, 41]; [pero los dobles de cosas iguales son iguales entre sí] [62]; por tanto, el paralelogramo BΔ es también igual al cuadrado HB. De manera semejante, trazando las (rectas) AE, BK se demostraría que también el paralelogramo ΓΛ es igual al cuadrado ΘΓ; por tanto, el cuadrado entero BΔEΓ es igual a los cuadrados HB, ΘΓ [N. C. 2]. Asimismo, el cuadrado BΔEΓ ha sido trazado a partir de BΓ, y los (cuadrados) HB, ΘΓ a partir de BΔ, AΓ. Por tanto, el cuadrado del lado BΓ es igual a los cuadrados de los lados BΔ, AΓ.

Por consiguiente, en los triángulos rectángulos el cuadrado del lado que subtiende el ángulo recto es igual a los cuadrados de los lados que comprenden el ángulo recto. Q. E. D. [63].

---

[62] Heiberg suprime las palabras *tà dè tôn isôn diplásia isa allêlois estín*. Aluden a una noción común interpolada. Ver nota 19.

[63] Según PROCLO (*Com.* 426, 6-15): «Si escuchamos a quienes gustan de narrar cosas antiguas, hallaremos que atribuyen este teorema a Pitágoras y dicen que sacrificó un buey por su descubrimiento. Por mi parte, aunque admiro a los que conocieron primero la verdad de este teorema, más me maravilla el autor de los *Elementos*, no sólo por establecerlo mediante una clara demostración, sino por haber sentado en el libro sexto (VI, 31) una proposición aún más general con las pruebas incontestables de la ciencia». Esta apreciación de Proclo se ha hecho incluso más sabia con el paso del tiempo. Hoy podemos reconocer que las nociones tendenciosamente llamadas «triángulo pitagórico» *ti. e.* el triángulo rectángulo en el que los tres lados son proporcionales a 3 enteros x, y, z que satisfagan la condición del «teorema de Pitágoras»: $x^2 + y^2 = z^2$), o «triplo pitagórico» *i. e.* cualquier triplo de enteros $< x, y, z >$ que satisfaga la misma condición—, eran nociones familiares para las matemáticas prehelenica y no estaban ausentes de otras culturas muy

PROPOSICIÓN 48

Si en un triángulo el cuadrado de uno de los lados es igual a los cuadrados de los dos lados restantes del triángulo, el ángulo comprendido por esos lados restantes del triángulo es recto.

Sea, pues, el cuadrado del lado BΓ del triángulo ABΓ, igual a los cuadrados de los lados BA, AΓ.

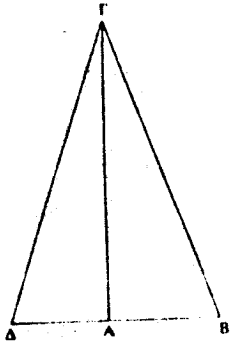
Digo que el ángulo BAΓ es recto.

Pues trácese la (recta) AΔ que forme ángulos rectos con la recta AΓ desde el punto A y hágase AΔ igual a BA, y trácese ΔΓ. Puesto que ΔA es igual a AB, el cuadrado de

---

dispares. Hay usos de los triángulos y de los triplos «pitagóricos» entre los *harpedonáptai* egipcios (p. Berlín 6619), en el texto cuneiforme Plimpton 322 (que corresponde a la dinastía de Hammurabi); en los *Sulvasutras*, manuales indios de construcción de altares (compuestos hacia 500-200 a. C., aunque se hagan eco de conocimientos anteriores); en el capítulo noveno de la colección china de problemas «Nueve capítulos sobre el arte de la matemática» (escrito durante el período Han, entre el 200 a. C. y el 220 d. C.). Van der Waerden, a la luz de estos datos, ha llegado a conjeturar la existencia de una tradición ramificada (egipcia, babilonia, india, griega, china), derivada de un origen común que se remontaría al neolítico europeo, *vid.* B. L. VAN DER WAERDEN, *Geometry and Algebra in Ancient Civilizations*, Berlín-Heidelberg-Nueva York, 1983, caps. 1-2, págs. 1-69. Su conjetura también se remite a la hipótesis de Seidenberg del origen ritual de la geometría, en particular en el caso indio, *vid.* A. SEIDENBERG, «The ritual origin of Geometry», *Archive for History of Exact Sciences*, 1 (1962), 69-82; «The origin of Mathematics», *ibid.* 18 (1978), 301-342. Pero no hay constancia de que las comprobaciones y reglas prácticas que acompañan a veces a los triplos babilonios e indios fueran demostraciones propiamente dichas; tampoco la hay de que el resultado o «teorema» de Pitágoras se sentara deductivamente como un teorema efectivo.

ΔΑ es también igual al cuadrado de ΑΒ. Anádase a ambos el cuadrado de ΑΓ; entonces los cuadrados de ΔΑ, ΑΓ son iguales a los cuadrados de ΒΑ, ΑΓ. Pero el cuadrado de ΑΓ es igual a los cuadrados de ΔΑ, ΑΓ: porque el ángulo ΔΑΓ es recto [I, 47]; pero el cuadrado de ΒΓ es igual a los cuadrados de ΒΑ, ΑΓ: porque esto es lo que se ha supuesto; por tanto, el cuadrado de ΑΓ es igual al cuadrado de ΒΓ; de modo que también el lado ΑΓ es igual al (lado) ΒΓ; y como ΔΑ es igual a ΑΒ, y ΑΓ es común, los dos (lados) ΔΑ, ΑΓ son iguales a los dos (lados) ΒΑ, ΑΓ; y la base ΔΓ es igual a la base ΒΓ; por tanto, el (ángulo) ΔΑΓ es igual al (ángulo) ΒΑΓ [I, 8]. Pero ΔΑΓ es recto; luego también es recto el (ángulo) ΒΑΓ.



Por consiguiente, si en un triángulo el cuadrado de uno de los lados es igual a los cuadrados de los dos restantes del triángulo, el ángulo comprendido por esos lados restantes del triángulo es recto. Q. E. D.

LIBRO SEGUNDO

DEFINICIONES

1) De todo paralelogramo rectangular se dice que está comprendido por las dos rectas que comprenden el ángulo recto.
2) En toda área de paralelogramo llámese gnomon a uno cualquiera de los paralelogramos situados en torno a su diagonal junto con los dos complementos [64]

[64] La referencia al *gnômōn* más antigua que conocemos se encuentra en HERÓDOTO (II 109): «... los griegos aprendieron de los babilonios el *pólos*, el *gnômōn* y la división del día en doce partes».

Según Suidas, fue Anaximandro quien introdujo el *gnómion* en Grecia. Se trataba de un instrumento consistente en una vara perpendicular al horizonte, cuya sombra proyectada sobre un plano o círculo servía para medir el tiempo: la sombra más corta del día corresponde al mediodía, la más corta del año al solsticio de verano y la más larga al de invierno. Fue muy utilizado hasta la época alejandrina. El *pólos* era un instrumento parecido cuyo uso no se basaba en la longitud de la sombra sino en la dirección marcada por la translación (aparente) del sol.

Enópides de Quíos llamaba *katá gnômona* a la perpendicular trazada desde un punto a una recta. Más tarde encontramos el término utilizado para designar un intrumento en forma de L que servía para dibujar ángu-

Proposición 1

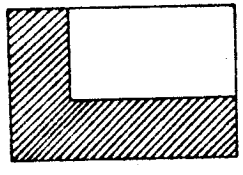
Si hay dos rectas y una de ellas se corta en un número cualquiera de segmentos, el rectángulo comprendido por las dos rectas es igual a los rectángulos comprendidos por la (recta) no cortada y cada uno de los segmentos [65].

Sean A, BΓ dos rectas y córtese al azar BΓ en los puntos Δ, E.

---

los rectos. Es natural que a partir de la forma de dicho instrumento se llamara *gnómōn* a la figura que queda después de quitar, de la esquina de un cuadrado, otro cuadrado más pequeño (o, según Aristóteles, a la figura que, añadida a un cuadrado, aumenta sus lados pero no altera su forma). Los pitagóricos usan el término en este sentido y, en aritmética, para referirse al número que hay que añadir para pasar de un número figurado al siguiente de la misma naturaleza.

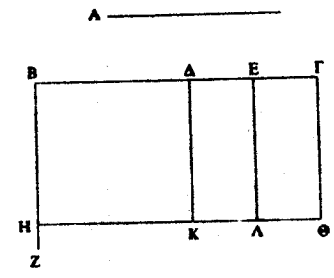
Euclides amplía el significado de *gnómōn* aplicándolo a paralelogramos en general. Más tarde encontramos otra definición en Herón de Alejandría (Def. 58) como cualquier figura que, añadida a otra, hace la figura entera semejante a aquella a la que ha sido añadida. Según esto, si el paralelogramo es un rectángulo, la parte rayada de la figura adjunta sería un *gnómōn*.



[65] Los pitagóricos y matemáticos griegos posteriores hablan de distintas clases de números según las distintas figuras geométricas que formen (números triangulares, cuadrados, planos, sólidos etc.). Un número plano sería el número resultante de la multiplicación de otros dos números a los que se refieren como «lados», a veces como «longitud» y «anchura» respectivamente, del número que es su producto.

Por otra parte, este libro II ha sido uno de los pretextos más socorridos para hablar del «álgebra geométrica» de los griegos. En particular las diez primeras proposiciones suelen representarse en esta perspectiva con la forma de ecuaciones. La fórmula correspondiente a la presente proposición sería: $a(b+c+d\ldots) = ab+ac+ad\ldots$

Digo que el rectángulo comprendido por A, BΓ es igual al rectángulo comprendido por A, BΔ y el (comprendido) por A, ΔE y además el (comprendido) por A, EΓ [66].



Pues trácese a partir del (punto) B la (recta) BZ que forme ángulos rectos con BΓ [I, 11], y hágase BH igual a Λ [I, 3], y por el punto H trácese HΘ paralela a BΓ [I, 31], y por los (puntos) Δ, E, Γ trácense ΔK, EΛ, ΓΘ paralelas a BH.

Entonces el (rectángulo) [66bis] BΘ es igual a los (rectángulos) BK, ΔΛ, EΘ. Ahora bien, BΘ es el (rectángulo comprendido) por A, BΓ: porque está comprendido por HB, BΓ, y BH es igual a A; pero BK es el (rectángulo comprendido) por A, BΔ: porque está comprendido por HB, BΔ, y BH es

---

[66] La expresión más común para referirse a un rectángulo es *tò hypò tôn* seguido de las letras que definen los dos lados que forman el ángulo recto, por ej., *tò hypò tôn* AB, BΓ, expresión abreviada a partir de: *tò hypò tôn* AB, BΓ *periechómenon orthogónion*. En griego el artículo neutro es suficiente para saber que se trata de un rectángulo y no de un ángulo, que llevaría artículo femenino. En español, sin embargo, es necesario incluir las palabras «rectángulo comprendido» para entender a qué se refiere el texto. Cuando dichas palabras no aparecen en griego, las pongo entre paréntesis. Otra manera de referirse tanto a rectángulos como a cuadrados es citando su diagonal.

[66bis] Cf. nota 66.

igual a A. Y ΛΛ es el (rectángulo comprendido) por A, ΔE: porque ΔK, es decir BH [I, 34] es igual a A. Y además, del mismo modo EΘ es el (rectángulo comprendido) por A, EΓ; así pues, el (rectángulo comprendido) por A, BΓ es igual al (rectángulo comprendido) por A, BΔ y el (comprendido) por A, ΔE, y además el (comprendido) por A, EΓ.

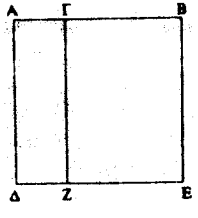
Por consiguiente, si hay dos rectas y una de ellas se corta en un número cualquiera de segmentos, el rectángulo comprendido por las dos rectas es igual a los rectángulos comprendidos por la (recta) no cortada y cada uno de los segmentos. Q. E. D.

Proposición 2

Si se corta al azar una línea recta, el rectángulo comprendido por la (recta) entera y cada uno de los segmentos es igual al cuadrado de la (recta) entera [67].

Córtese, pues, la recta AB al azar en el punto Γ.

Digo que el rectángulo comprendido por AB, BΓ junto con el rectángulo comprendido por BA, AΓ es igual al cuadrado de AB [68].



Pues constrúyase a partir de AB el cuadrado AΔEB [I, 46], y trácese por el punto Γ la (recta) ΓZ paralela a una de las dos (rectas) AΔ, BE [I, 31].

Entonces el (rectángulo) AE es igual a los (rectángulos) AZ, ΓE. Ahora bien, AE es el cuadrado de AB, pero AZ es el rectángulo comprendido por BA, AΓ: porque está comprendido por ΔA, AΓ, y AΔ es igual a AB; y ΓE es el (rectángulo comprendido) por AB, BΓ: porque BE es igual a AB. Por tanto, el (rectángulo comprendido) por BA, AΓ junto con el (rectángulo comprendido) por AB, BΓ es igual al cuadrado de AB.

Por consiguiente, si se corta al azar una línea recta, el rectángulo comprendido por la (recta) entera y cada uno de los segmentos es igual al cuadrado de la (recta) entera. Q. E. D.

[67] La expresión algebraica de este teorema es $(a+b)a + (a+b)b = (a+b)^2$.

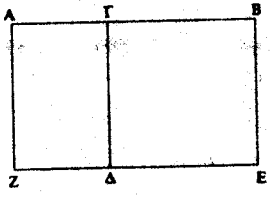
[68] *Tò apò tês* BI *tetrágōnon*. Euclides utiliza aquí la expresión completa para referirse a un cuadrado. Es muy frecuente que la palabra *tetrágōnon* se omita. En español es necesario incluirla. La pongo entre paréntesis cuando no aparece en el texto griego.

Proposición 3

Si se corta al azar una línea recta, el rectángulo comprendido por la recta entera y uno de los segmentos es igual al rectángulo comprendido por los segmentos y el cuadrado del segmento primeramente dicho [69].

Córtese, pues, al azar la recta AB en Γ.

Digo que el rectángulo comprendido por AB, BΓ es igual al rectángulo comprendido por AΓ, ΓB junto con el cuadrado de BΓ.



Pues constrúyase a partir de ΓB el cuadrado ΓΔEB [I, 46], y prolónguese EΔ hasta Z, y por el (punto) A trácese AZ paralela a una de las dos (rectas) ΓΔ, BE [I, 31]. Entonces AE es igual a AΔ, ΓE; ahora bien, AE es el rectángulo comprendido por AB, BΓ: porque está comprendido por AB, BE, y

[69] Expresión algebraica: $(a+b)a = ab + a^2$.

ΒΓ es igual a ΒΙ; pero ΆΛ es el (rectángulo comprendido) por ΑΙ, ΓΒ: porque ΑΓ es igual a ΓΒ; y ΔΒ es el cuadrado de ΓΒ; así pues, el rectángulo comprendido por ΑΒ, ΒΓ es igual al rectángulo comprendido por ΑΓ, ΓΒ junto con el cuadrado de ΒΓ.

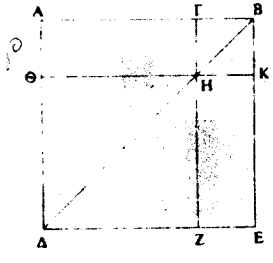
Por consiguiente, si se corta al azar una línea recta, el rectángulo comprendido por la recta entera y uno de los segmentos es igual al rectángulo comprendido por los segmentos y el cuadrado del segmento primeramente dicho. Q. E. D.

Proposición 4

Si se corta al azar una línea recta, el cuadrado de la (recta) entera es igual a los cuadrados de los segmentos y dos veces el rectángulo comprendido por los segmentos [70].

Córtese, pues, al azar la línea ΑΒ en el (punto) Γ.

Digo que el cuadrado de ΑΒ es igual a los cuadrados de ΑΓ, ΓΒ y dos veces el rectángulo comprendido por ΑΓ, ΓΒ.



Pues constrúyase a partir de ΑΒ el cuadrado ΑΔΕΒ [I, 46], y trácese ΒΔ, y por el (punto) Γ trácese ΓΖ paralela a una de las dos (rectas) ΑΔ, ΕΒ, y por el (punto) Η trácese ΘΚ paralela a las dos (rectas) ΑΒ, ΔΕ [I, 31]. Ahora bien, como ΓΖ es paralela a ΑΔ, y ΒΔ ha incidido sobre ellas, el ángulo externo ΓΗΒ es igual al interno y opuesto ΑΔΒ [I, 29]. Pero el (ángulo) ΑΔΒ es igual al (ángulo)

[70] La expresión algebraica de este teorema es: $(a+b)^2 = a^2 + b^2 + 2ab$.

ΑΒΔ, puesto que el lado ΒΑ es también igual al (lado) ΑΔ [I, 5]; por tanto, el ángulo ΓΗΒ es también igual al (ángulo) ΗΒΓ; de modo que el lado ΒΓ es también igual al lado ΓΗ [I, 6]; pero ΓΒ es igual a ΗΚ y ΓΗ a ΚΒ [I, 34]; por tanto, ΗΚ es también igual a ΚΒ; luego ΓΗΚΒ es equilátero.

Digo, en fin, que también es rectangular. Pues como ΓΗ es paralela a ΒΚ [y ΓΒ ha incidido sobre ellas], entonces los ángulos ΚΒΓ, ΗΓΒ son iguales a dos rectos [I, 29]. Pero el (ángulo) ΚΒΓ es recto; por tanto, el (ángulo) ΒΓΗ es también recto; de modo que los (ángulos) opuestos ΓΗΚ, ΗΚΒ son, asimismo, rectos [I, 34]. Luego ΓΗΚΒ es rectangular; pero se ha demostrado que también es equilátero; por tanto, es un cuadrado; y es el (cuadrado) de ΓΒ. Por lo mismo, efectivamente ΘΖ es también un cuadrado; y es el (cuadrado) de ΘΗ, es decir de ΑΓ [I, 34]; por tanto, ΘΖ, ΚΓ son los (cuadrados) de ΑΓ, ΓΒ. Y como ΑΗ es igual a ΗΕ y ΑΗ es el (rectángulo comprendido) por ΑΓ, ΓΒ: porque ΗΓ es igual a ΓΒ; entonces ΗΕ es también igual al (rectángulo comprendido) por ΑΓ, ΓΒ; por tanto, ΑΗ, ΗΕ son iguales a dos veces el (rectángulo comprendido) por ΑΓ, ΓΒ. Pero ΘΖ, ΓΚ son también los cuadrados de ΑΓ, ΓΒ; por tanto, los cuatro ΘΖ, ΓΚ, ΑΗ, ΗΕ son iguales a los cuadrados de ΑΓ, ΓΒ y dos veces el rectángulo comprendido por ΑΓ, ΓΒ. Ahora bien, ΘΖ, ΓΚ, ΑΗ, ΗΕ son el (cuadrado) entero ΑΔΕΒ, que es el cuadrado de ΑΒ; por tanto, el cuadrado de ΑΒ es igual a los cuadrados de ΑΓ, ΓΒ y dos veces el rectángulo comprendido por ΑΓ, ΓΒ.

Por consiguiente, si se corta al azar una línea recta, el cuadrado de la (recta) entera es igual a los cuadrados de los segmentos y dos veces el rectángulo comprendido por los segmentos. Q. E. D.

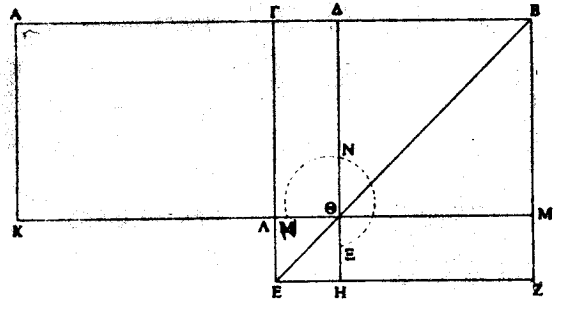
[Porisma:

A partir de esto, queda claro que en las áreas de cuadrados, los paralelogramos situados en torno a la diagonal son cuadrados] [71].

Proposición 5

Si se corta una línea recta en (segmentos) iguales y desiguales, el rectángulo comprendido por los segmentos desiguales de la (recta) entera junto con el cuadrado de la (recta que está) entre los puntos de sección [72], *es igual al cuadrado de la mitad* [73].

Córtese, pues, una recta cualquiera AB en (segmentos) iguales en el punto Γ, y en (segmentos) desiguales en el (punto) Δ.



Digo que el rectángulo comprendido por AΔ, ΔB junto con el cuadrado de ΓΔ es igual al cuadrado de ΓB.

[71] Este porisma es, sin duda, una interpolación, como ha probado Heiberg en «Paralipomena zu Euklid», *Hermes* 38 (1905), 48.

[72] *Toû apò tês metaxù tôn tómon tetragónou.* No es posible una traducción literal al español. Se refiere, sin duda, a la recta que está entre los puntos de sección.

[73] Expresión algebraica: $ab + \left(\frac{a+b}{2} - b\right)^2 = \left(\frac{a+b}{2}\right)^2$.

Pues constrúyase a partir de ΓB el cuadrado ΓEZB [I, 46], y trácese BE, y por el (punto) Δ trácese ΔH paralela a una de las dos (rectas) ΓE, BZ, y por el (punto) Θ trácese a su vez KM paralela a una de las dos (rectas) AB, EZ, y por el (punto) A trácese, asimismo, AK paralela a una de las dos (rectas) ΓΛ, BM [I, 31]. Y como el complemento ΓΘ es igual al complemento ΘZ [I, 43], añádase a ambos ΔM; por tanto, el (rectángulo) entero ΓM es igual al rectángulo entero ΔZ. Pero el (rectángulo) ΓM es igual al (rectángulo) ΛΛ, puesto que la (recta) AΓ es igual a la (recta) ΓB [I, 36]; por tanto, el (rectángulo) AΛ es también igual al (rectángulo) ΔZ. Añádase a ambos ΓΘ; entonces el (rectángulo) entero AΘ es igual al gnomon MNΞ [74]. Pero AΘ es el (rectángulo comprendido) por AΔ, ΔB: porque ΔΘ es igual a ΔB; entonces el gnomon MNΞ es igual al (rectángulo comprendido) por AΔ, ΔB. Añádase a ambos ΛH, que es igual al cuadrado de ΓΔ; entonces el gnomon MNΞ y ΛH es igual al rectángulo comprendido por AΔ, ΔB y el cuadrado de ΓΔ. Ahora bien, el gnomon MNΞ y ΛH es el cuadrado entero ΓEZB, que es el cuadrado de ΓB; por tanto, el rectángulo comprendido por AΔ, ΔB junto con el cuadrado de ΓΔ es igual al cuadrado de ΓB.

Por consiguiente, si se corta una línea recta en (segmentos) iguales y desiguales, el rectángulo comprendido por los segmentos desiguales de la (recta) entera junto con el cuadrado de la (recta que está) entre los puntos de sección, es igual al cuadrado de la mitad. Q. E. D.

[74] El *gnómon* se señala en el dibujo mediante tres letras y una línea punteada. En el dibujo que acompaña a este teorema, la letra M se utiliza dos veces, una para el extremo de la recta KM y la otra para el interior del gnomon MNΞ.

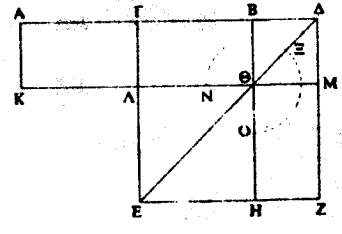
Proposición 6

Si se divide en dos partes iguales una línea recta y se le añade, en línea recta, otra recta, el rectángulo comprendido por la (recta) entera con la (recta) añadida y la (recta) añadida junto con el cuadrado de la mitad es igual al cuadrado de la (recta) compuesta por la mitad y la (recta) añadida [75].

Divídase, pues, la recta AB en dos por el (punto) Γ, y añádasele, en línea recta, otra recta, BΔ. ...

Digo que el rectángulo comprendido por AΔ, ΔB junto con el cuadrado de ΓB es igual al cuadrado de ΓΔ.

Pues constrúyase a partir de ΓΔ el cuadrado ΓEZΔ [I, 46], y trácese ΔE, y por el punto B trácese BH paralela a una de las dos (rectas) EΓ, ΔZ, y por el punto Θ trácese KM paralela a una de las dos (rectas) AB, EZ, y además por el (punto) A trácese AK paralela a una de las dos (rectas) ΓΛ, ΔM [I, 31].



Entonces, como AΓ es igual a ΓB, también el (rectángulo) AΛ es igual al ΓΘ [I, 36]. Pero el (rectángulo) ΓΘ es igual al ΘZ [I, 43]. Por tanto, el (rectángulo) AΛ es también igual al ΘZ. Añádase a ambos el (rectángulo) ΓM; entonces el (rectángulo) entero AM es igual al gnomon NΞO. Pero AM es el (rectángulo comprendido) por AΔ, ΔB: porque ΔM es igual a ΔB; por tanto, el gnomon NΞO es también igual al [rectángulo comprendido] por AΔ, ΔB. Añádase a ambos ΛH, que es igual al cuadrado de BΓ; entonces el rectángulo comprendido por AΔ, ΔB junto con el cuadrado de ΓB es igual al gnomon NΞO y ΛH. Pero el gnomon NΞO y ΛH son el cuadrado entero ΓEZΔ, que es el cuadrado de ΓΔ; por tanto, el rectángulo comprendido por AΔ, ΔB junto con el cuadrado de ΓB es igual al cuadrado de ΓΔ.

Por consiguiente, si se divide en dos una línea recta y se le añade, en línea recta, otra recta, el rectángulo comprendido por la (recta) entera con la (recta) añadida y la (recta) añadida junto con el cuadrado de la mitad es igual al cuadrado de la (recta) compuesta a partir de la mitad y la (recta) añadida. Q. E. D.

[75] Expresión algebraica: $(2a + b)b + a^2 = (a + b)^2$.

Proposición 7

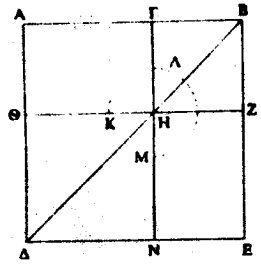
Si se corta al azar una línea recta, el cuadrado de la (recta) entera y el de uno de los segmentos tomados conjuntamente son iguales a dos veces el rectángulo comprendido por la (recta) entera y el segmento antedicho más el cuadrado del segmento restante [76].

Córtese, pues, al azar una recta AB por el punto Γ.

Digo que los cuadrados de AB, BΓ son iguales a dos veces el rectángulo comprendido por AB, BΓ más el cuadrado de ΓA.

[76] Expresión algebraica: $(a + b)^2 + a^2 = 2(a + b)a + b^2$.

Pues constrúyase a partir de AB el cuadrado AΔEB [I, 46]; e inscríbase la figura [77].



Así pues, como el (rectángulo) AH es igual al HE [I, 43], añádase a ambos ΓZ; entonces el (rectángulo) entero AZ es igual al (rectángulo) entero ΓE; por tanto, los (rectángulos) AZ, ΓE son el doble de AZ. Pero los (rectángulos) AZ, ΓE son el gnomon KΛM más el cuadrado ΓZ; por tanto, el gnomon KΛM y ΓZ son el doble de AZ. Ahora bien, el doble de AZ es también dos veces el (rectángulo comprendido) por AB, BΓ, porque BZ es igual a BΓ; por tanto, el gnomon KΛM y el cuadrado ΓZ son iguales a dos veces el (rectángulo comprendido) por AB, BΓ. Añádase a ambos ΔH, que es el cuadrado de AΓ; entonces el gnomon KΛM y los cuadrados BH, HΔ son iguales a dos veces el rectángulo comprendido por AB, BΓ y el cuadrado de AΓ. Pero el gnomon KΛM y los cuadrados BH, HΔ son el cuadrado entero AΔEB y el (cuadrado) ΓZ, que son los cuadrados de AB, BΓ; por tanto, los cuadrados de AB, BΓ son .guales a dos veces el rectángulo comprendido por AB, BΓ junto con el cuadrado de AΓ.

Por consiguiente, si se corta al azar una línea recta, el cuadrado de la (recta) entera y el de uno de los segmentos tomados conjuntamente son iguales a dos veces el rectángulo comprendido por la (recta) entera y el segmento antedicho más el cuadrado del segmento restante. Q. E. D.

[77] Conforme al método empleado en proposiciones anteriores.

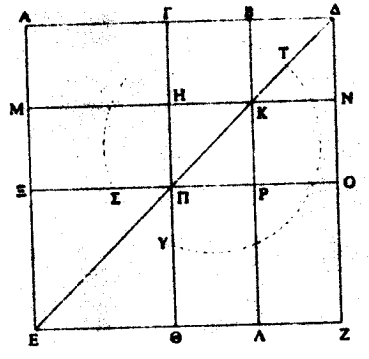
Proposición 8

Si se corta al azar una línea recta, cuatro veces el rectángulo comprendido por la (recta) entera y uno de los segmentos junto con el cuadrado del segmento restante es igual al cuadrado construido a partir de la (recta) entera y del segmento primeramente dicho, tomados como una sola recta [78].

Córtese, pues, al azar una recta AB por el punto Γ.

Digo que cuatro veces el rectángulo comprendido por AB, BΓ junto con el cuadrado de AΓ es igual al cuadrado construido a partir de AB, BΓ tomados como una sola recta.

Pues prolónguese BΔ en línea recta [con la recta AB], y hágase BΔ igual a ΓB, y constrúyase a partir de AΔ el cuadrado AEZΔ [I, 46], e inscríbase la figura duplicada



Así pues, como ΓB es igual a BΔ, mientras que ΓB es igual a HK, y BΔ a KN, entonces HK es también igual a

[78] Expresión algebraica: $4(a+b)a+b^2 = [(a+b) + a]^2$.

KN. Por lo mismo, ΠP también es igual a PO. Y como BΓ es igual a BΔ y HK a KN, entonces el (cuadrado) ΓK es también igual al (cuadrado) KΔ, y el (cuadrado) HP al PN [I, 36]. Pero ΓK es igual a PN: porque son complementos del paralelogramo ΓO [I, 43]; por tanto, el (cuadrado) KΔ es también igual al (cuadrado) HP; luego los cuatro ΔK, ΓK, HP, PN son iguales entre sí. Por tanto, los cuatro son el cuádruple de ΓK. Como ΓB es a su vez igual a BΔ, mientras que BΔ es igual a BK, es decir a ΓH, pero ΓB es igual a HK, es decir a HΠ, entonces ΓH es también igual a HΠ. Y como ΓH es igual a HΠ, y ΠP a PO, el (rectángulo) AH es también igual al (rectángulo) MΠ, y el (rectángulo) ΠΛ al (rectángulo) PZ [I, 36]. Pero el (rectángulo) MΠ es igual al (rectángulo) ΠΛ: porque son complementos del paralelogramo MΛ [I, 43]; por tanto, el (rectángulo) AH es igual al (rectángulo) PZ; luego los cuatro AH, MΠ, ΠΛ, PZ son iguales entre sí; por tanto, los cuatro son el cuádruple de AΠ. Ahora bien, se ha demostrado que los cuatro ΓK, KΔ, HP, PN son también el cuádruple de ΓK; por tanto, los ocho que comprenden el gnomon ΣTY son el cuádruple de AK. Y como AK es el (rectángulo comprendido) por AB, BΔ: porque BK es igual a BΔ; entonces cuatro veces el (rectángulo comprendido) por AB, BΔ es el cuádruple de AK. Pero se ha demostrado que el gnomon ΣTY es también el cuádruple de AK; por tanto, cuatro veces el (rectángulo comprendido) por AB, BΔ es igual al gnomon ΣTY. Añádase a ambos ΞΘ, que es igual al cuadrado de AΓ; entonces cuatro veces el rectángulo comprendido por AB, BΔ junto con el cuadrado de AΓ es igual al gnomon ΣTY y el (cuadrado) ΞΘ. Pero el gnomon ΣTY y el (cuadrado) ΞΘ son el cuadrado entero AEZΔ, que es el (cuadrado) de AΔ; luego cuatro veces el (rectángulo comprendido) por AB, BΔ junto con el (cuadrado) de AΓ es igual al cuadrado de AΔ; ahora bien,

BΔ es igual a BΓ. Por tanto, cuatro veces el rectángulo comprendido por AB, BΓ junto con el cuadrado de AΓ es igual al (cuadrado) de AΔ, es decir al (cuadrado) construido a partir de AB y BΓ tomados como una sola (recta).

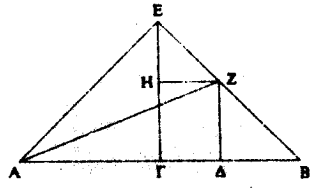
Por consiguiente, si se corta al azar una línea recta, cuatro veces el rectángulo comprendido por la (recta) entera y uno de los segmentos junto con el cuadrado del segmento restante es igual al cuadrado construido a partir de la (recta) entera y el segmento primeramente dicho tomados como una sola (recta). Q. E. D.

Proposición 9

Si se corta una línea recta en partes iguales y desiguales, los cuadrados de los segmentos desiguales de la (recta) entera son el doble del cuadrado de la mitad más el cuadrado de la (recta situada) entre los (puntos) de sección [79].

Córtese, pues, una (recta) AB en (segmentos) iguales por el (punto) Γ, y en desiguales por el (punto) Δ.

Digo que los cuadrados de AΔ, ΔB son el doble de los cuadrados de AΓ, ΓΔ.



Trácese, pues, por el (punto) Γ la (recta) ΓE que forme ángulos rectos con AB, y hágase ΓE igual a cada una de

[79] Expresión algebraica: $a^2 + b^2 = 2\left[\left(\frac{a+b}{2}\right)^2 + \left(\frac{a+b}{2} - b\right)\right]$

las (rectas) AΓ, ΓB, y trácense EA, EB, y por el (punto) Δ trácese ΔZ paralela a ΕΓ, y por el (punto) Z, ZH (paralela) a AB, y trácese AZ.

Y como AΓ es igual a ΓE, el (ángulo) EAΓ es igual al (ángulo) AEΓ. Y como es recto el ángulo correspondiente a Γ, entonces los (ángulos) restantes EAΓ, AEΓ son iguales a un recto [I, 32]; y son iguales; por tanto, cada uno de los (ángulos) ΓEA, ΓAE son la mitad de un recto. Por lo mismo, en efecto, cada uno de los ángulos ΓEB, EBΓ es también la mitad de un recto; por tanto, el (ángulo) entero AEB es recto. Y como el (ángulo) HEZ es la mitad de un recto y el (ángulo) EHZ es recto —porque es igual al interno y opuesto EΓB [I, 29]— entonces el (ángulo) restante EZH es la mitad de un recto [I, 32]; por tanto, el ángulo HEZ es igual al (ángulo) EZH; de modo que el lado EH también es igual al (lado) HZ [I, 6]. Como, asimismo, el ángulo correspondiente a B es la mitad de un recto, y el (ángulo) ZΔB es recto: porque es a su vez igual al interno y opuesto EΓB [I, 29]; entonces el (ángulo) restante BZΔ es la mitad de un recto [I, 32]; luego el (ángulo) correspondiente a B es igual al (ángulo) ΔZB; de modo que también el lado ZΔ es igual al lado ΔB [I, 6]. Y como AΓ es igual a ΓE, el (cuadrado) de AΓ también es igual al de ΓE; por tanto, los cuadrados de AΓ, ΓE son el doble del de AΓ. Pero el cuadrado de EA es igual a los de AΓ, ΓE —porque el ángulo AΓE es recto [I, 47]— por tanto, el (cuadrado) de EA es el doble del de AΓ. Como EH es a su vez igual a HZ, el cuadrado de EH es también igual al (cuadrado) de HZ; entonces los cuadrados de EH, HZ son el doble del cuadrado de HZ. Pero el cuadrado de EZ es igual a los cuadrados de EH, HZ; por tanto, el cuadrado de EZ es el doble del de HZ. Pero HZ es igual a ΓΔ [I, 34]; por tanto, el (cuadrado) de EZ es el doble del (cuadrado) de ΓΔ. Pero

el (cuadrado) de EA es también el doble del de AΓ; por tanto, los cuadrados de AE, EZ son el doble de los cuadrados de AΓ, ΓΔ. Pero el cuadrado de AZ es igual a los cuadrados de AE, EZ: porque el ángulo AEZ es recto [I, 47]; por tanto, el cuadrado de AZ es el doble de los (cuadrados) de AΓ, ΓΔ. Ahora bien, los (cuadrados) de AΔ, ΔZ son iguales al (cuadrado) de AZ: porque el ángulo correspondiente a Δ es recto [I, 47]; luego los (cuadrados) de AΔ, ΔZ son el doble de los cuadrados de AΓ, ΓΔ. Pero ΔZ es igual a ΔB; por tanto, los cuadrados de AΔ, ΔB son el doble de los cuadrados de AΓ, ΓΔ.

Por consiguiente, si se corta una línea recta en partes iguales y desiguales, los cuadrados de los segmentos desiguales de la (recta) entera son el doble del cuadrado de la mitad y el cuadrado de la (recta situada) entre los (puntos) de sección. Q. E. D.

Proposición 10

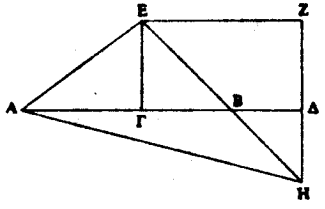
Si se divide en dos partes iguales una línea recta y se le añade, en línea recta, otra recta, el cuadrado de la (recta) entera con la (recta) añadida y el (cuadrado) de la añadida, tomados conjuntamente, son el doble del (cuadrado) de la mitad y el cuadrado construido a partir de la (recta) compuesta por la mitad y la (recta) añadida, tomadas como una sola recta [80].

Divídase, pues, en dos partes iguales una recta AB por el (punto) Γ, y añádase a ella, en línea recta, otra recta BΔ.

---

[80] Expresión algebraica: $(2a+b)^2 + b^2 = 2[a^2 + (a+b)^2]$.

Digo que los cuadrados de ΑΔ, ΔΒ son el doble de los cuadrados de ΑΓ, ΓΔ.



Pues trácese a partir del punto Γ la (recta) ΓΕ que forme ángulos rectos con ΑΒ [I, 11], y hágase igual a cada una de las (rectas) ΑΓ, ΓΒ [I, 3], y trácense ΕΑ, ΕΒ; y por el (punto) Ε trácese ΕΖ paralela a ΑΔ, y por el (punto) Δ trácese ΖΔ paralela a ΓΕ [I, 31]. Y dado que una recta ΕΖ ha incidido sobre las rectas paralelas ΕΓ, ΖΔ, entonces los (ángulos) ΓΕΖ, ΕΖΔ son iguales a dos rectos [I, 29]; por tanto, los (ángulos) ΖΕΒ, ΕΖΔ son menores que dos rectos; pero las rectas prolongadas a partir de (ángulos) menores que dos rectos se encuentran [I, Post. 5]; por tanto, las (rectas) ΕΒ, ΖΔ prolongadas en el sentido de Β, Δ se encontrarán. Prolónguense y encuéntrense en Η, y trácese ΑΗ. Y como ΑΓ es igual a ΓΕ, el ángulo ΕΑΓ es también igual al (ángulo) ΑΕΓ [I, 5], y el ángulo correspondiente a Γ es recto; por tanto, cada uno de los (ángulos) ΕΑΓ, ΑΕΓ es la mitad de un recto [I, 32]. Por lo mismo, cada uno de los (ángulos) ΓΕΒ, ΕΒΓ es la mitad de un recto; por tanto, el (ángulo) ΑΕΒ es recto. Y como el (ángulo) ΕΒΓ es la mitad de un recto, entonces ΔΒΗ es también la mitad de un recto [I, 15]. Pero el (ángulo) ΒΔΗ es también recto: pues es igual al (ángulo) ΔΓΕ, porque son alternos [I, 29]; por tanto, el ángulo restante ΔΗΒ es la mitad de un recto [I, 32]: luego el ángulo ΔΗΒ es igual al (ángulo) ΔΒΗ; de modo que el lado ΒΔ es también igual al lado ΗΔ [I, 6]. Como a su vez el ángulo ΕΗΖ es la mitad de un recto, y el (ángulo) correspondiente a Ζ es recto —porque es igual al (ángulo) opuesto, el correspondiente a Γ [I, 34]— entonces el ángulo restante ΖΕΗ es la mitad de un recto [I, 32]; por tanto, el ángulo ΕΗΖ es igual al (ángulo) ΖΕΗ; de modo que también el lado ΗΖ es igual al lado ΕΖ [I, 6]. Ahora bien, dado que el cuadrado de ΕΓ es igual al cuadrado de ΓΑ, entonces los cuadrados de ΕΓ, ΓΑ son el doble del cuadrado de ΓΑ. Pero el (cuadrado) de ΕΑ es igual a los (cuadrados) de ΕΓ, ΓΑ [I, 47]; por tanto, el cuadrado de ΕΑ es el doble del cuadrado de ΑΓ [N. C. 1]. Como ΖΗ es a su vez igual a ΕΖ, el (cuadrado) de ΖΗ es también igual al de ΖΕ, entonces los (cuadrados) de ΗΖ, ΖΕ son el doble del (cuadrado) de ΕΖ. Pero el (cuadrado) de ΕΗ es igual a los (cuadrados) de ΗΖ, ΖΕ [I, 47]; por tanto, el (cuadrado) de ΕΗ es el doble del (cuadrado) de ΕΖ. Y ΕΖ es igual a ΓΔ [I, 34]; por tanto, el (cuadrado) de ΕΗ es el doble del de ΓΔ. Ahora bien, se ha demostrado que el (cuadrado) de ΕΑ es también el doble del de ΑΓ; por tanto, los cuadrados de ΑΕ, ΕΗ son el doble de los cuadrados de ΑΓ, ΓΔ. Pero el (cuadrado) de ΑΗ es igual a los (cuadrados) de ΑΕ, ΕΗ [I, 47]; por tanto, el (cuadrado) de ΑΗ es el doble de los (cuadrados) de ΑΓ, ΓΔ. Pero los (cuadrados) de ΑΔ, ΔΗ son iguales al cuadrado de ΑΗ [I, 47]; por tanto, los (cuadrados) de ΑΔ, ΔΗ son el doble de los (cuadrados) de ΑΓ, ΓΔ. Pero ΔΗ es igual a ΔΒ; luego los (cuadrados) de ΑΔ, ΔΒ son el doble de los cuadrados de ΑΓ, ΓΔ.

Por consiguiente, si se divide en dos partes iguales una línea recta, y se le añade, en línea recta, otra recta, el cuadrado de la (recta) entera con la añadida y el cuadrado de la añadida, tomados conjuntamente, son el doble del

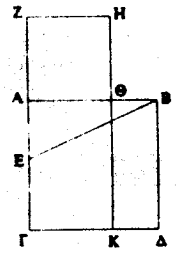
cuadrado de la mitad y el cuadrado construido a partir de la recta compuesta por la mitad y la añadida, tomadas como una sola recta. Q. E. D.

Proposición 11

Dividir una recta dada de manera que el rectángulo comprendido por la (recta) entera y uno de los segmentos sea igual al cuadrado del segmento restante.

Sea AB la recta dada.

Así pues, hay que dividir AB de modo que el rectángulo comprendido por la (recta) entera y uno de los segmentos sea igual al cuadrado del segmento restante.



Pues constrúyase a partir de AB el cuadrado ABΔΓ [I, 46] y divídase en dos AΓ por el punto E y trácese BE y prolónguese ΓA hasta Z, y hagáse EZ igual a BE, y constrúyase a partir de AZ el cuadrado ZΘ, y prolónguese HΘ hasta K.

Digo que AB ha sido cortada en Θ, de modo que hace el rectángulo comprendido por AB, BΘ igual al cuadrado AΘ.

Pues como la recta AΓ ha sido dividida en dos por el (punto) E y se le ha añadido ZA, entonces el rectángulo comprendido por ΓZ, ZA junto con el cuadrado de AE es igual al cuadrado de EZ [II, 6]. Pero EZ es igual a EB; por tanto, el (rectángulo comprendido) por ΓZ, ZA junto con el (cuadrado) de AE es igual al cuadrado de EB. Pero los (cuadrados) de BA, AE son iguales al (cuadrado) de EB, porque el ángulo correspondiente a A es recto [I, 47]; por tanto, el (rectángulo comprendido) por ΓZ, ZA junto con el (cuadrado) de AE es igual a los cuadrados de BA, AE. Quítese de ambos el (cuadrado) de AE; entonces el rectángulo restante comprendido por ΓZ, ZA es igual al (cuadrado) de AB. Ahora bien, el (rectángulo comprendido) por ΓZ, ZA es ZK: porque AZ es igual a ZH; pero el cuadrado de AB es AΔ; por tanto, ZK es igual a AΔ. Quítese de ambos AK; entonces el (cuadrado) restante ZΘ es igual a ΘΔ. Y ΘΔ es el (rectángulo comprendido) por AB, BΘ: porque AB es igual a BΔ; pero ZΘ es el cuadrado de AΘ; por tanto, el rectángulo comprendido por AB, BΘ es igual al cuadrado de ΘA.

Por consiguiente, la recta dada AB ha sido dividida en Θ de modo que hace el rectángulo comprendido por AB, BΘ igual al cuadrado de ΘA. Q. E. F.

Proposición 12

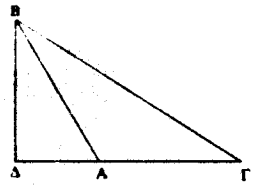
En los triángulos obtusángulos el cuadrado del lado que subtiende al ángulo obtuso es mayor que los cuadrados de los lados que comprenden el ángulo obtuso en dos veces el rectángulo comprendido por un (lado) de los del ángulo obtuso sobre el que cae la perpendicular y la (recta) exterior cortada por la perpendicular, hasta el ángulo obtuso [81].

Sea ABΓ el triángulo obtusángulo que tiene el (ángulo) obtuso BAΓ, y trácese a partir del punto B, BΔ perpendicular a ΓA prolongada.

---

[81] Una versión más libre y llana sería: «y su segmento exterior desde el punto de corte con la perpendicular hasta el ángulo obtuso». Esta referencia al segmento (exterior o interior) desde el punto de corte puede extenderse a los demás casos en que aparece la misma expresión.

Digo que el cuadrado de ΒΓ es mayor que los cuadrados de ΒΑ, ΑΓ en dos veces el rectángulo comprendido por ΓΑ, ΑΔ.



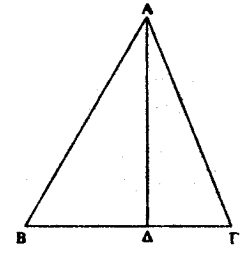
Pues dado que la recta ΓΑ ha sido cortada al azar en el punto Α, entonces el (cuadrado) de ΔΓ es igual a los cuadrados de ΓΑ, ΑΔ y dos veces el (rectángulo comprendido) por ΓΑ, ΑΔ [II, 4]. Añádase a ambos el cuadrado de ΔΒ; entonces los (cuadrados) de ΓΑ, ΔΒ son iguales a los cuadrados de ΓΑ, ΑΔ, ΔΒ y dos veces el [rectángulo comprendido] por ΓΑ, ΑΔ. Pero el (cuadrado) de ΓΒ es igual a los (cuadrados) de ΓΔ, ΔΒ: porque el ángulo correspondiente a Δ es recto [I, 47]. Pero el cuadrado de ΑΒ es igual a los (cuadrados) de ΑΔ, ΔΒ [I, 47]; por tanto, el cuadrado de ΓΒ es igual a los cuadrados de ΓΑ, ΑΒ y dos veces el rectángulo comprendido por ΓΑ, ΑΔ; de modo que el cuadrado de ΓΒ es mayor que los cuadrados de ΓΑ, ΑΒ es dos veces el rectángulo comprendido por ΓΑ, ΑΔ.

Por consiguiente, en los triángulos obtusángulos el cuadrado del lado que subtiende al ángulo obtuso es mayor que los cuadrados de los lados que comprenden el ángulo obtuso en dos veces el rectángulo comprendido por un (lado) de los del ángulo obtuso sobre el que cae la perpendicular y la (recta) exterior cortada por la perpendicular hasta el ángulo obtuso. Q. E. D.

Proposición 13

En los triángulos acutángulos, el cuadrado del lado que subtiende al ángulo agudo es menor que los cuadrados de los lados que comprenden el ángulo agudo en dos veces el rectángulo comprendido por uno de los lados del ángulo agudo sobre el que cae la perpendicular y la (recta) interior cortada por la perpendicular hasta el ángulo agudo.

Sea ΑΒΓ el triángulo acutángulo que tiene agudo el ángulo correspondiente a Β, y trácese a partir del punto Α, ΑΔ perpendicular a ΒΓ.



Digo que el cuadrado de ΑΓ es menor que los cuadrados de ΓΒ, ΒΑ en dos veces el rectángulo comprendido por ΓΒ, ΒΔ.

Pues dado que ΓΒ ha sido cortada al azar en Δ, entonces los cuadrados de ΓΒ, ΒΔ son iguales a dos veces el rectángulo comprendido por ΓΒ, ΒΔ y el cuadrado de ΔΓ [II, 7]. Añádase a ambos el (cuadrado) de ΔΑ; entonces los cuadrados de ΓΒ, ΒΔ, ΔΑ son iguales a dos veces el rectángulo comprendido por ΓΒ, ΒΔ y los cuadrados de ΑΔ, ΔΓ. Pero el cuadrado de ΑΒ es igual a los (cuadrados) de ΒΔ, ΔΑ: porque el ángulo correspondiente a Δ es recto [I, 47]; pero el (cuadrado) de ΑΓ es igual a los (cuadrados) de ΑΔ, ΔΓ; por tanto los cuadrados de ΓΒ, ΒΑ son iguales al cuadrado de ΑΓ y dos veces el (rectángulo comprendido) por ΓΒ, ΒΔ; de modo que el cuadrado de ΑΓ sólo es menor que los cuadrados de ΓΒ, ΒΑ en dos veces el rectángulo comprendido por ΓΒ, ΒΔ.

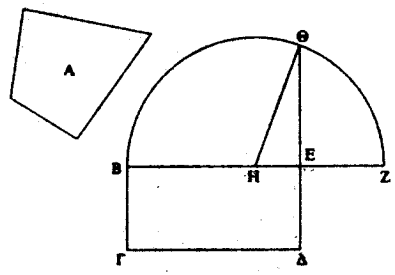
Por consiguiente, en los triángulos acutángulos, el lado que subtiende el ángulo agudo es menor que los lados que comprenden el ángulo agudo en dos veces el rectángulo comprendido por uno de los lados del ángulo agudo sobre el que cae la perpendicular y la (recta) interior cortada por la perpendicular hasta el ángulo agudo. Q. E. D.

Proposición 14

Construir un cuadrado igual a una figura rectilínea dada.

Sea A la figura rectilínea dada.

Así pues, hay que construir un cuadrado igual a la figura rectilínea A.



Constrúyase, pues, el paralelogramo rectángulo BΔ igual a la figura rectilínea A [I, 45]. Entonces, si BE es igual a EΔ, se habrá hecho lo propuesto: porque se ha construido el cuadrado BΔ igual a la figura rectilínea A; pero si no, una de las (rectas) BE, EΔ es mayor. Sea mayor BE y prolónguese hasta Z, y hágase EZ igual a EΔ, y divídase en dos partes iguales BZ en H, y con el centro H y una de

las (rectas) HB, HZ como distancia descríbase el semicírculo BΘZ, y prolónguese ΔE hasta Θ, y trácese HΘ.

Así pues, como la recta BZ ha sido cortada en partes iguales en H y en desiguales en E, entonces el rectángulo comprendido por BE, EZ junto con el cuadrado de EH es igual al cuadrado de HZ [II, 5]. Pero HZ es igual a HΘ; por tanto, el (rectángulo comprendido) por BE, EZ junto con el (cuadrado) de HE es igual al (cuadrado) de HΘ. Pero los cuadrados de ΘE, EH son iguales al (cuadrado) de HΘ [I, 47]; por tanto, el (rectángulo comprendido) por BE, EZ junto con el (cuadrado) de HE es igual a los (cuadrados) de ΘE, EH. Quítese de ambos el cuadrado de HE; entonces el rectángulo restante comprendido por BE, EZ es igual al cuadrado de EΘ. Ahora bien, el (rectángulo comprendido) por BE, EZ es BΔ: porque EZ es igual a EΔ; por tanto, el paralelogramo BΔ es igual al cuadrado de ΘE. Pero BΔ es igual a la figura rectilínea A. Luego A es igual al cuadrado que puede formarse a partir de EΘ.

Por consiguiente, se ha construido un cuadrado, el que puede formarse a partir de EΘ, igual a la figura rectilínea dada A. Q. E. F.

LIBRO TERCERO

DEFINICIONES

1. Círculos iguales son aquellos cuyos diámetros son iguales, o cuyos radios son iguales [82].
2. Se dice que es tangente [83] a un círculo la recta que, tocando el círculo y siendo prolongada, no corta el círculo.

---

[82] Los griegos no contaban con una palabra específica para «radio», utilizaban la expresión: *he ek toû kéntrou (eutheîa)* «la recta trazada a partir del centro». Dice Simson: «Ésta no es definición sino teorema de patente verdad; porque colocados los círculos, cuyos radios son iguales, uno sobre otro, de tal suerte que ajusten sus centros, ajustarán también los mismos círculos». *Los seis primeros libros y el undécimo y duodécimo de los Elementos...*, ed. cit., pág. 59.

[83] Los geómetras griegos establecen por lo general una distinción entre *háptesthai* y *epháptesthai*, utilizando el primero en el sentido de «encontrarse» y el último con el significado de «tocarse, ser tangente». Se trata de un uso regular, aunque no observado siempre por el propio Euclides en los *Elementos*, quien, por ejemplo, en IV, Def. 5 utiliza *háptesthai* en el sentido de «tocar». Así lo hace también Arquímedes en alguna ocasión. Por otra parte, Aristóteles usa *epháptesthai* en ciertos casos en los que se esperaría *háptesthai (Meteorológica* III 5, 376b9 y 376a6). En suma, al no tener el uso técnico establecido de nuestro «ser tangente» mantendré por lo regular la versión «tocar».

3. Se dice que son tangentes entre sí los círculos que, tocándose mutuamente, no se cortan.
4. En un círculo se dice que las rectas están a la misma distancia del centro, cuando las perpendiculares trazadas desde el centro hasta ellas son iguales.
5. Se dice que está a mayor distancia aquella (recta) sobre la que cae la perpendicular mayor.
6. Un segmento de un círculo es la figura comprendida por una recta y una circunferencia [84] de un círculo.
7. Un ángulo de un segmento es el comprendido por una recta y una circunferencia de un círculo.
8. Ángulo en un segmento es el ángulo que, cuando se toma un punto sobre la circunferencia del segmento y se trazan rectas desde él hasta los extremos de la recta que es la base del segmento, está comprendido por las rectas trazadas.
9. Cuando las rectas que comprenden el ángulo cortan una circunferencia se dice que el ángulo está sobre ella.
10. Un sector [85] de un círculo es la figura que, cuando se construye un ángulo en el centro del círculo, está comprendida por las rectas que comprenden el ángulo y la circunferencia cortada por ellas.

[84] *Periphéreia* (circunferencia) se refiere, en realidad, a un arco de circunferencia muy frecuentemente. Cuando se moderniza esta definición en los términos: «un segmento de un círculo es la figura comprendida por una cuerda y el arco de círculo subtendido por ella», se está traduciendo la versión arábigo-latina de la definición, no la versión euclídea. Siguiendo esta línea, en la definición 8 cabría entender que el ángulo en un segmento es el ángulo inscrito en un arco.

[85] Un escoliasta dice que fue el tranchete de zapatero *(skytotomikós tomeús)* lo que dio lugar al nombre de «sector» del círculo.

11. Son segmentos de círculo semejantes los que admiten ángulos iguales, o aquellos en que los ángulos son iguales entre sí.

Proposición 1

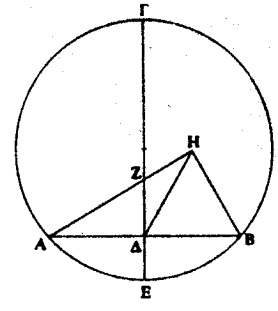
Hallar el centro de un círculo dado.

Sea el círculo dado ΑΒΓ.

Hay que hallar el centro del círculo ΑΒΓ.

Trácese en él al azar una recta ΑΒ, y divídase en dos por el punto Δ, y a partir de Δ trácese ΔΓ perpendicular a ΑΒ y prolónguese hasta Ε, y divídase en dos partes iguales ΓΕ en Ζ.

Digo que Ζ es el centro del [círculo] ΑΒΓ.



Pues supongamos que no [86], entonces si es posible sea Η (el centro), y trácense ΗΑ, ΗΔ, ΗΒ. Ahora bien, como ΑΔ es igual a ΔΒ y ΔΗ es común, los dos (lados) ΑΔ, ΔΗ son iguales respectivamente a los dos (lados) ΗΔ, ΔΒ; y la base ΗΑ es igual a la base ΗΒ, pues son radios; por tanto, el ángulo ΑΔΗ es igual al ángulo ΗΔΒ [I, 8]; pero cuando una recta levantada sobre otra recta hace los ángulos adyacentes iguales entre sí, cada uno de los ángulos iguales es recto [I, Def. 10]; por tanto, el ángulo ΗΔΒ es recto. Pero también es recto el ángulo ΖΔΒ; por tanto, el ángulo ΖΔΒ es igual al ángulo ΗΔΒ, el mayor al menor; lo cual es imposible. Luego Η no es el centro del círculo

[86] *Mè gar.*

ABΓ. De la misma manera demostraríamos que ningún otro (lo es) excepto Z.

Por consiguiente, el punto Z es el centro del (círculo) ABΓ.

Porisma:

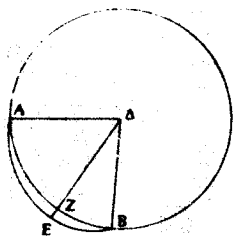
A partir de esto queda claro que si en un círculo una recta divide en dos partes iguales a otra recta formando ángulos rectos, el centro del círculo está sobre la recta que corta. Q. E. F. [87].

Proposición 2

Si se toman dos puntos al azar en la circunferencia de un círculo, la recta que une los dos puntos caerá dentro del círculo.

Sea ABΓ el círculo, y sobre su circunferencia tómense al azar dos puntos A, B.

Digo que la recta trazada desde A hasta B caerá dentro del círculo.



Pues supongamos que no, entonces, si es posible, caiga fuera como AEB, y tómese el centro del círculo ABΓ [III, 1], y sea Δ, y trácense ΔA, ΔB, y prolónguese ΔZE. Así pues, como ΔA es igual a ΔB entonces también el ángulo ΔAE es igual al ángulo ΔBE [I, 5]; y como un lado AEB del triángulo ΔAE ha sido prolongado, entonces el ángulo ΔEB es mayor que el ángulo ΔAE [I, 16]. Pero el (ángulo) ΔAE es igual al (ángulo) ΔBE. Por tanto, el (ángulo) ΔEB es mayor que el (ángulo) ΔBE. Ahora bien, al ángulo mayor lo subtiende el lado mayor [I, 19]; entonces ΔB es mayor que ΔE. Pero ΔB es igual a ΔZ. Por tanto, ΔZ es mayor que ΔE, el menor que el mayor; lo cual es imposible. Entonces la recta trazada de A a B no caerá fuera del círculo. De la misma manera demostraríamos que tampoco (caerá) sobre la misma circunferencia; por tanto (caerá) dentro.

Por consiguiente, si se toman dos puntos al azar en la circunferencia de un círculo, la recta que une los puntos caerá dentro del círculo. Q. E. D.

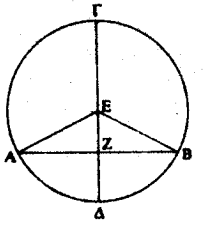
[87] Este porisma se ha introducido entre el problema y la cláusula *hóper édei poiêsai*, que, por supuesto, se refiere al problema.

Proposición 3

Si en un círculo una recta (trazada) a través del centro divide en dos partes iguales a otra recta no (trazada) a través del centro, la corta formando también ángulos rectos; y si la corta formando ángulos rectos, la divide también en dos partes iguales.

Sea ABΓ el círculo, y en él una recta ΓΔ (trazada) a través del centro divida en dos partes iguales a otra recta AB no (trazada) a través del centro, por el punto Z.

Digo que también la corta formando ángulos rectos.



Tómese, pues, el centro del círculo ABΓ y sea E, y trácense EA, EB.

Ahora bien, como AZ es igual a ZB y ZE es común, los dos (lados) son iguales a los dos (lados). Y la base EA es igual a la base EB; por tanto, el (ángulo) AZE es igual al (ángulo) BZE [I, 8]. Pero cuando una recta levantada sobre otra recta hace los

ángulos adyacentes iguales entre sí, cada uno de los ángulos iguales es recto [I, Def. 10]; por tanto, cada uno de los (ángulos) AZE, BZE es recto. Luego ΓΔ (trazada) a través del centro dividiendo en dos partes iguales a AB no (trazada) a través del centro, la corta también formando ángulos rectos.

Pero ahora corte ΓΔ a AB formando ángulos rectos.

Digo que también la divide en dos partes iguales, es decir, que AZ es igual a ZB.

Pues siguiendo la misma construcción, como EA es igual a EB, el ángulo EAZ es también igual al (ángulo) EBZ [I, 5]. Pero el (ángulo) recto AZE es igual al (ángulo) recto BZE; por tanto, EAZ, EZB son dos triángulos que tienen dos ángulos iguales a dos ángulos, y un lado igual a un lado, el común a ambos EZ, que subtiende a uno de los ángulos iguales; luego tendrá también los lados restantes iguales a los lados restantes [I, 26]; así pues, AZ es igual a ZB.

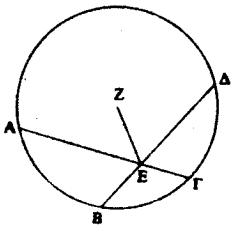
Por consiguiente, si en un círculo una recta (trazada) a través del centro divide en dos partes iguales a otra recta no (trazada) a través del centro, la corta también formando ángulos rectos; y si la corta formando ángulos rectos, la divide también en dos partes iguales. Q. E. D.

Proposición 4

Si en un círculo se cortan entre sí dos rectas que no pasan por el centro, no se dividen entre sí en dos partes iguales.

Sea ABΓΔ el círculo y en él las dos rectas AΓ, BΔ que no pasan por el centro córtense en el punto E.

Digo que no se dividen entre sí en dos partes iguales.



Pues, si fuera posible, divídanse entre sí en dos partes iguales de modo que AE sea igual a EΓ y BE a EΔ; y tómese el centro del círculo ABΓΔ [III, 1] y sea Z, y trácese ZE.

Así pues, como una recta que pasa por el centro ZE divide en dos a otra recta que no pasa por el centro AΓ, también la corta formando ángulos rectos [III, 3]; por tanto, el (ángulo) ZEA es recto; como, a su vez, una recta ZE divide en dos a otra recta BΔ, también la corta formando ángulos rectos [III, 3]; por tanto, el (ángulo) ZEB es recto. Pero se ha demostrado que también el (ángulo) ZEA es recto; luego el (ángulo) ZEA es igual al (ángulo) ZEB, el menor al mayor; lo cual es imposible. Por tanto, las (rectas) AΓ, BΔ no se dividen entre sí en dos partes iguales.

Por consiguiente, si en un círculo se cortan entre sí dos rectas que no pasan por el centro, no se dividen entre sí en dos partes iguales. Q. E. D.

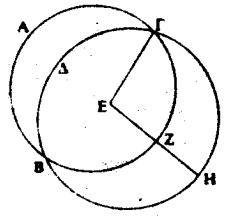
Proposición 5

Si dos círculos se cortan entre sí, su centro no será el mismo.

Córtense, pues, los círculos ABΓ, ΓΔH entre sí en los puntos B, Γ. Digo que su centro no será el mismo.

Pues, si es posible, sea E y trácese EΓ, y trácese al azar EZH. Ahora bien, como el punto E es el centro del círculo

ΑΒΓ, ΕΓ es igual a ΕΖ [I, Def. 15]. Como, a su vez, el punto Ε es el centro del círculo ΓΔΗ, ΕΓ es igual a ΕΗ; pero se ha demostrado que ΕΓ es igual a ΕΖ; por tanto, ΕΖ también es igual a ΕΗ, la menor a la mayor; lo cual es imposible. Por tanto, el punto Ε no es el centro de los círculos ΑΒΓ, ΓΔΗ.

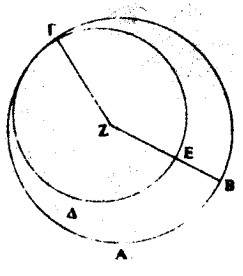


Por consiguiente, si dos círculos se cortan entre sí, su centro no es el mismo. Q. E. D.

Proposición 6

Si dos círculos se tocan uno a otro, su centro no será el mismo.

Tóquense, pues, los dos círculos ΑΒΓ, ΓΔΕ en el punto Γ. Digo que su centro no será el mismo.



Pues, si fuera posible, sea Ζ, y trácese ΖΓ, y trácese al azar ΖΕΒ.

Así pues, como el punto Ζ es el centro del círculo ΑΒΓ, ΖΓ es igual a ΖΒ. Como a su vez el punto Ζ es el centro del círculo ΓΔΕ, ΖΓ es igual a ΖΕ. Pero se ha demostrado que ΖΓ es igual a ΖΒ; por tanto, ΖΕ también es igual a ΖΒ, la menor a la mayor: lo cual es imposible. Por tanto, el punto Ζ no es el centro de los círculos ΑΒΓ, ΓΔΕ.

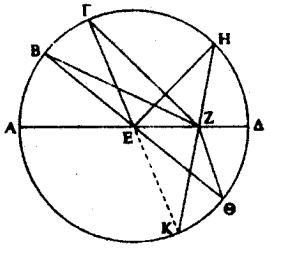
Por consiguiente, si dos círculos se tocan uno a otro, su centro no será el mismo. Q. E. D.

Proposición 7

Si se toma un punto en el diámetro de un círculo que no sea el centro del círculo y desde él hasta el círculo caen algunas rectas, será la mayor aquella en la que está el centro, y la menor la restante y de las demás la más cercana a la que pasa por el centro es siempre mayor que la más lejana, y sólo caerán dos iguales del punto al círculo a uno y otro lado de la más pequeña.

Sea ΑΒΓΔ el círculo, y sea su diámetro ΑΔ, y tómese sobre ΑΔ un punto Ζ, que no sea el centro del círculo, y sea el centro del círculo Ε, y desde Ζ hasta el círculo ΑΒΓΔ caigan algunas rectas: ΖΒ, ΖΓ, ΖΗ.

Digo que la mayor es ΖΑ, la menor ΖΔ, y de las demás ΖΒ mayor que ΖΓ, y ΖΓ que ΖΗ.



Trácense, pues, ΒΕ, ΓΕ, ΗΕ. Y como en todo triángulo dos lados son mayores que el restante [I, 20], entonces los (lados) ΕΒ, ΕΖ son mayores que ΒΖ. Pero ΑΕ es igual a ΒΕ; por tanto, ΑΖ es mayor que ΒΖ. Como ΒΕ es a su vez igual a ΓΕ, y ΖΕ es común, los dos (lados) ΒΕ, ΕΖ son iguales a los dos (lados) ΓΕ, ΕΖ. Ahora bien, el ángulo ΒΕΖ también es mayor que el (ángulo) ΓΕΖ. Luego la base ΒΖ es mayor que la base ΓΖ [I, 24]. Por lo mismo, ΓΖ es también mayor que ΖΗ.

Como, a su vez, los (lados) ΗΖ, ΖΕ son mayores que ΕΗ y ΕΗ es igual a ΕΔ, entonces los (lados) ΗΖ, ΖΕ son mayores que ΕΔ. Quítese de ambos ΕΖ; entonces el restante ΗΖ es mayor que el restante ΖΔ. Por tanto, ΖΑ es la mayor y ΖΔ la menor, y ΖΒ es mayor que ΖΓ, y ΖΓ que ΖΗ.

Digo también que sólo caerán dos rectas iguales del punto Z al círculo ABΓΔ a uno y otro lado de la más pequeña ZΔ.

Constrúyase, pues, sobre la recta EZ y en su punto E el ángulo ZEΘ igual al ángulo HEZ [I, 23] y trácese ZΘ. Así pues, como HE es igual a EΘ y EZ es común, los dos (lados) HE, EZ son iguales a los dos (lados) ΘE, EZ; y el ángulo HEZ es igual al ángulo ΘEZ; por tanto, la base ZH es igual a la base ZΘ [I, 4].

Digo, en fin, que no caerá del punto Z al círculo otra recta igual a ZH.

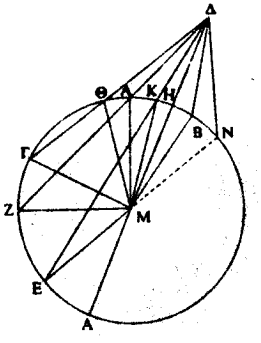
Pues, si fuera posible, caiga ZK. Ahora bien, como ZK es igual a ZH, mientras que ZΘ [es igual] a ZH, entonces ZK es también igual a ZΘ, la más cercana a la que pasa por el centro igual a la más lejana: lo cual es imposible. Luego no podrá caer del punto Z al círculo otra (recta) igual a HZ. Por tanto sólo una.

Por consiguiente, si se toma un punto en el diámetro de un círculo que no sea el centro del círculo y desde el punto hasta el círculo caen algunas rectas, será la mayor aquella en la que está el centro y la menor la restante, y de las demás la más cercana a la que pasa por el centro es siempre mayor que la más lejana y sólo caerán dos iguales del punto al círculo a uno y otro lado de la más pequeña. Q. E. D.

Proposición 8

Si se toma un punto exterior a un círculo y del punto al círculo se trazan algunas rectas, una de las cuales pasa por el centro y las demás al azar, de las rectas que caen en la parte cóncava de la circunferencia, la mayor es la que pasa por el centro, y de las demás siempre la más cercana a la que pasa por el centro es mayor a la más lejana; pero de las que caen en la parte convexa de la circunferencia la menor es la que está entre el punto y el diámetro, y de las demás la más cercana a la más pequeña es siempre menor que la más lejana, y sólo caerán dos iguales del punto al círculo a uno y otro lado de la más pequeña.

Sea ABΓ el círculo, y tómese un punto Δ exterior a ABΓ, y a partir de él trácense algunas rectas ΔA, ΔE, ΔZ, ΔΓ, y sea ΔA la que pasa por el centro.



Digo que las rectas que caen en la parte cóncava de la circunferencia AEZΓ, la mayor es ΔA, que pasa por el centro, y ΔE es mayor que ΔZ, y ΔZ que ΔΓ, pero de las rectas que caen en la parte convexa de la circunferencia ΘΛKH, la menor es ΔH que está entre el punto y el diámetro AH, y siempre la más cercana a la menor ΔH es menor que la más lejana, ΔK (menor que) ΔΛ, y ΔΛ (menor que) ΔΘ.

Tómese, pues, el centro del círculo ABΓ [III, 1] y sea M; y trácense ME, MZ, MΓ, MK, MΛ, MΘ.

Y como AM es igual a EM, añádase a ambas MΔ; entonces AΔ es igual a EM, MΔ. Pero EM, MΔ son mayores que EΔ [I, 20]; entonces AΔ es también mayor que EΔ. Como, a su vez, ME es igual a MZ, y MΔ es común, entonces EM, MΔ son iguales a ZM, MΔ; y el ángulo EMΔ es mayor que el ángulo ZMΔ. Por tanto, la base EΔ es mayor que la base ZΔ [I, 24]. De manera semejante demostraríamos que ZΔ es también mayor que ΓΔ; por tanto, la mayor es ΔA, mientras que ΔE es mayor que ΔZ, y ΔZ, que ΔΓ.

Y como MK, KΔ son mayores que MΔ [I, 20], mientras que MH es igual a MK, entonces la restante KΔ es mayor que la restante HΔ; de modo que HΔ es menor que KΔ; y como sobre uno de los lados MΔ del triángulo MΛΔ se han construido dos rectas MK, KΔ (que se encuentran) en su interior, entonces MK, KΔ son menores que MΛ, ΛΔ [I, 21]; pero MK es igual a MΛ; por tanto, la restante ΔK es menor que la restante ΔΛ. De manera semejante demostraríamos que ΔΛ es también menor que ΔΘ; por tanto, ΔH es la menor, mientras que ΔK es menor que ΔΛ, y ΔΛ que ΔΘ.

Digo también que sólo dos (rectas) iguales caerán del punto Δ al círculo a uno y otro lado de la menor ΔH.

Constrúyase en la recta MΔ y en su punto M el ángulo ΔMB igual al ángulo KMΔ y trácese ΔB. Y como MK es igual a MB, y MΔ común, los dos (lados) KM, MΔ son iguales a los dos (lados) BM, MΔ, respectivamente; y el ángulo KMΔ es igual al ángulo BMΔ; por tanto, la base ΔK es igual a la base ΔB [I, 4].

Digo que no caerá otra (recta) igual a la recta ΔK hasta el círculo desde el punto Δ.

Pues, si fuera posible, caiga y sea ΔN. Así pues, como ΛK es igual a ΔN, mientras que ΔK es igual a ΔB, entonces ΔB también es igual a ΔN, la más cercana a la menor ΔH igual a la más lejana; lo cual se ha demostrado (que es) imposible. Por tanto, no caerán más de dos (rectas) iguales desde el punto Δ hasta el círculo ABΓ a uno y otro lado de la menor ΔH.

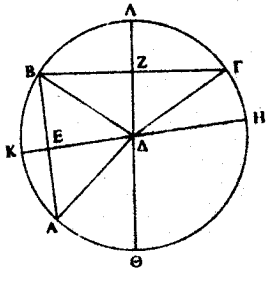
Por consiguiente, si se toma un punto exterior a un círculo y del punto al círculo se trazan algunas rectas, de las cuales una pasa por el centro y las demás al azar, de las rectas que caen en la parte cóncava de la circunferencia, la mayor es la que pasa por el centro, y de las demás siempre la más cercana a la que pasa por el centro es mayor que la más lejana; pero de las rectas que caen en la parte convexa de la circunferencia, la menor es la que está entre el punto y el diámetro, y de las demás, la más cercana a la más pequeña es siempre menor que la más lejana, y sólo caerán dos (rectas) iguales del punto al círculo a uno y otro lado de la más pequeña. Q. E. D.

Proposición 9

Si se toma un punto dentro de un círculo y del punto al círculo caen más de dos rectas iguales, el punto tomado es el centro del círculo.

Sea ABΓ el círculo y Δ un punto dentro de él, y desde Δ hasta el círculo ABΓ caigan más de dos rectas iguales ΔA, ΔB, ΔΓ.

Digo que el punto Δ es el centro del círculo ABΓ.



Trácense, pues, AB, BΓ y divídanse en dos partes iguales por los puntos E, Z, y, trazadas EΔ, ZΔ, prolónguense hasta los puntos H, K, Θ, Λ.

Entonces, como AE es igual a EB y EΔ es común, los dos (lados) AE, EΔ son iguales a los dos (lados) BE, EΔ; y la base ΔA es igual a la base ΔB; por tanto, el ángulo AEΔ es igual al ángulo BEΔ [I, 8]; luego cada uno de los ángulos AEΔ, BEΔ es recto [I, Def. 10]; así pues, HK divide AB en dos partes iguales formando ángulos rectos. Y como si en un círculo una recta divide en dos a otra recta formando ángulos rectos, el centro del círculo está en la (recta) que corta [III, 1, Por.],

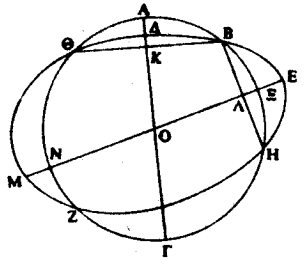
entonces el centro del círculo está en HK. Por lo mismo, el centro del círculo ABΓ está también en ΘΛ. Y HK y ΘΛ no tienen ningún otro punto común que el punto Δ; por tanto, el punto Δ es el centro del círculo ABΓ.

Por consiguiente, si se toma un punto dentro de un círculo, y del punto al círculo caen más de dos rectas iguales, el punto tomado es el centro del círculo. Q. E. D.

Proposición 10

Un círculo [88] *no corta a otro círculo en más de dos puntos.*

Pues, si fuera posible, corte el círculo ABΓ al círculo ΔEZ en más de dos puntos: B, H, Z, Θ, y, una vez trazadas BΘ, BH divídanse en dos partes iguales en los puntos K, Λ, y a partir de K, Λ trácense las rectas KΓ, ΛM formando ángulos rectos con BΘ, BH y prolónguense hasta los puntos A, E.



Así pues, como en el círculo ABΓ, una recta AΓ divide en dos partes iguales a otra recta BΘ formando ángulos rectos, entonces el centro del círculo ABΓ está en AΓ [III, 1, Por.]. Como, a su vez, en el mismo círculo ABΓ una

[88] La palabra *kýklos* (círculo) está aquí tomada en el sentido de «circunferencia».

recta NΞ divide en dos partes iguales a otra recta BH formando ángulos rectos, entonces el centro del círculo ABΓ está en NΞ. Pero se ha demostrado que está también en AΓ, y en ningún otro punto se encuentran las rectas AΓ, NΞ salvo en el (punto) O. Por tanto, el punto O es el centro del círculo ABΓ. De esta misma manera demostraríamos que también el centro del círculo ΔEZ es O; por tanto, el mismo punto O es el centro de los dos círculos que se cortan ABΓ, ΔEZ; lo cual es imposible [III, 5].

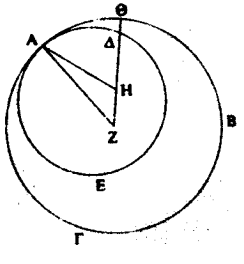
Por consiguiente, un círculo no corta a otro círculo en más de dos puntos. Q. E. D.

Proposición 11

Si dos círculos se tocan uno a otro por dentro, y se toman sus centros, la recta que une sus centros prolongada caerá sobre el punto de contacto de los círculos.

Tóquense, pues, los círculos ABΓ, AΔE uno a otro por dentro en el punto A, y tómese el centro Z del (círculo) ABΓ, y el (centro) H del (círculo) AΔE.

Digo que la recta trazada de H a Z prolongada caerá sobre A.



Pues supongamos que no, entonces, si es posible, caiga como la (recta) ZHΘ, y trácense AZ, AH.

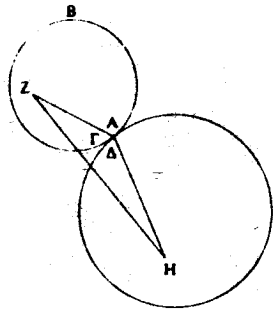
Así pues, como AH, HZ son mayores que ZA, es decir ZΘ, quítese de cada una ZH; entonces la restante AH es mayor que HΘ, pero AH es igual a HΔ; por tanto, HΔ es también mayor que HΘ, la menor que la mayor: lo cual es imposible; luego la recta trazada desde Z a H no caerá fuera; por tanto, caerá sobre el punto de contacto A.

Por consiguiente, si dos círculos se tocan uno a otro por dentro (y se toman sus centros), la recta que une los centros (prolongada) caerá sobre el punto de contacto de los círculos. Q. E. D.

Proposición 12 [89]

Si dos círculos se tocan uno a otro por fuera, la recta que une sus centros pasará a través del punto de contacto.

Tóquense, pues, los dos círculos ABΓ, AΔE por fuera en el punto A, y tómese el centro Z del círculo ABΓ, y el (centro) H del (círculo) AΔE.



Digo que la recta trazada de Z a H pasará por el punto de contacto A [90].

Pues supongamos que no, entonces, si es posible, pase como ZΓΔH, y trácense AZ, AH.

Así pues, como el punto Z es el centro del círculo ABΓ, ZA es igual a ZΓ. Como, a su vez, el punto H es el centro del círculo AΔE, HA es igual a HΔ. Pero se ha demostrado que también ZA es igual a ZΓ; por tanto, ZA, AH son iguales a ZΓ, HΔ; de modo que la (recta) entera ZH es mayor que las (rectas) ZA, AH; pero también menor [I, 20]; lo cual es imposible. Por tanto, no es posible que la recta trazada de Z a H no pase por el punto de contacto A; luego (pasará) a través de él.

Por consiguiente, si dos círculos se tocan por fuera, la [recta] que une sus centros pasará a través del punto de contacto. Q. E. D.

[89] Heath apunta la posibilidad de que esta proposición sea la interpolación de una prueba de Herón añadida a la proposición anterior (III, 11). Herón dice que Euclides se atuvo en la proposición 11 al caso en el que los círculos se tocan por dentro, y añade después que él se ocupará del caso en el que se tocan por fuera. La prueba y la figura que presenta son sustancialmente las de III, 12. También Itard la considera sospechosa, ver su edición de F. Peyrard, *Les oeuvres d'Euclides*, op. cit., pág. xii.

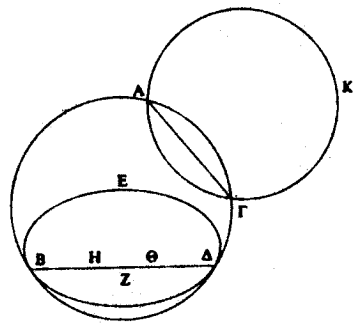
[90] La palabra usada para «punto de contacto» en esta proposición es *epaphé*, mientras que en la proposición anterior se utiliza *synaphé*.

Proposición 13

Un círculo no toca a otro círculo en más de un punto, ya sea por dentro o por fuera.

Pues, si es posible, toque el círculo ABΓΔ al círculo EBZΔ en primer lugar por dentro en más de un punto, a saber: en Δ, B.

Y tómese el centro H del círculo ABΓΔ, y el (centro) Θ del (círculo) EBZΔ.



Entonces la recta trazada de H a Θ caerá sobre B, Δ [III, 11]. Caiga como la recta BHΘΔ. Y como el punto H

es el centro del círculo ΑΒΓΔ, ΒΗ es igual a ΗΔ; entonces ΒΗ es mayor que ΘΔ; por tanto, ΒΘ es mucho mayor que ΘΔ. Como, a su vez, el punto Θ es el centro del círculo ΕΒΖΔ, ΒΘ es igual a ΘΔ; pero se ha demostrado que también es mucho mayor que ella; lo cual es imposible; por tanto, un círculo no toca a otro círculo por dentro en más de un punto.

Digo, además, que tampoco por fuera.

Pues, si es posible, toque el círculo ΑΓΚ al círculo ΑΒΓΔ por fuera en más de un punto, a saber: en Α, Γ, y trácese ΑΓ.

Así pues, como se han tomado dos puntos al azar Α, Γ, sobre las circunferencias de los círculos ΑΒΓΔ, ΑΓΚ, la recta que une los puntos caerá en el interior de cada uno [III, 2]. Ahora bien, ha caído dentro del (círculo) ΑΒΓΔ, pero fuera de ΑΓΚ [III, Def. 3], lo cual es absurdo; por tanto, un círculo no toca otro círculo por fuera en más de un punto. Pero se ha demostrado que tampoco por dentro.

Por consiguiente, un círculo no toca a otro círculo en más de un punto ya sea por dentro o por fuera. Q. E. D.

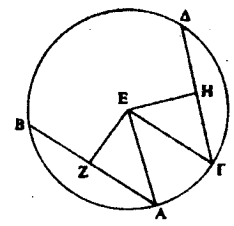
Proposición 14

En un círculo las rectas iguales están a la misma distancia del centro, y las que están a la misma distancia del centro son iguales entre sí.

Sea ΑΒΓΔ el círculo, y sean iguales en él las rectas ΑΒ, ΓΔ.

Digo que ΑΒ, ΓΔ están a la misma distancia del centro.

Tómese, pues, el centro del círculo ΑΒΓΔ [III, 1] y sea Ε, y trácense ΕΖ, ΕΗ perpendiculares a ΑΒ, ΓΔ, y trácense ΑΕ, ΕΓ.



Y como una recta que pasa por el centro, la ΕΖ, corta formando ángulos rectos a otra recta que no pasa por el centro, la ΑΒ, también la divide en dos partes iguales [III, 3]. Por tanto, ΑΖ es igual a ΖΒ; luego ΑΒ es el doble de ΑΖ. Por lo mismo, ΓΔ es también el doble de ΓΗ; ΑΒ es, asimismo, igual a ΓΔ; por tanto, ΑΖ es también igual a ΓΗ. Y como ΑΕ es igual a ΕΓ, el (cuadrado) de ΑΕ también es igual al (cuadrado) de ΕΓ. Pero los cuadrados de ΑΖ, ΕΖ son iguales al (cuadrado) de ΑΕ: porque el ángulo correspondiente a Ζ es recto; y los (cuadrados) de ΕΗ, ΗΓ son iguales al (cuadrado) de ΕΓ: porque el ángulo correspondiente a Η es recto [I, 47]; por tanto, los cuadrados de ΑΖ, ΖΕ son iguales a los cuadrados de ΓΗ, ΗΕ, de los cuales el cuadrado de ΑΖ es igual al (cuadrado) de ΓΗ: porque ΑΖ es igual a ΓΗ; luego el (cuadrado) restante de ΖΕ es igual al cuadrado de ΕΗ; así pues, ΕΖ es igual a ΕΗ. Pero en un círculo se dice que dos rectas están a la misma distancia del centro, cuando las perpendiculares trazadas desde el centro hasta ellas son iguales [III, Def. 4]; por tanto, ΑΒ, ΓΔ están a la misma distancia del centro.

Pero ahora estén a la misma distancia del centro las rectas ΑΒ, ΓΔ, es decir sea ΕΖ igual a ΕΗ.

Digo que ΑΒ también es igual a ΓΔ.

Pues, siguiendo la misma construcción, demostraríamos de modo similar que ΑΒ es el doble de ΑΖ, y ΓΔ (el doble) de ΓΗ; y como ΑΕ es igual a ΓΕ, el cuadrado de ΑΕ es igual al cuadrado de ΓΕ; pero los cuadrados de ΕΖ, ΖΑ son

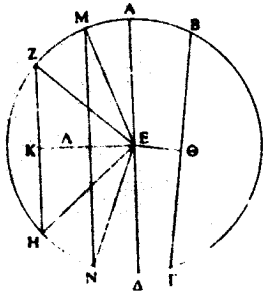
iguales al cuadrado de AE, y los cuadrados de EH, HΓ son iguales al cuadrado de ΓE [I, 47]. Por tanto, los cuadrados de EZ, ZA son iguales a los cuadrados de EH, HΓ; de los cuales el cuadrado de EZ es igual al cuadrado de EH; porque EZ es igual a EH; luego el cuadrado restante de AZ es igual al (cuadrado) de ΓH; así pues, AZ es igual a ΓH; y AB es el doble de AZ, y ΓΔ es el doble de ΓH; por tanto, AB es igual a ΓΔ.

Por consiguiente, en un círculo las rectas iguales están a la misma distancia del centro, y las que están a la misma distancia del centro son iguales entre sí. Q. E. D.

Proposición 15

En un círculo el diámetro es la recta mayor y de las demás, la más cercana al centro es siempre mayor que la más lejana.

Sea ABΓΔ el círculo y sea su diámetro AΔ, y el centro E, y la más cercana al diámetro AΔ sea BΓ, y la más lejana ZH.



Digo que AΔ es la mayor, y que BΓ es mayor que ZH.

Trácense, pues, a partir del centro E las (rectas) EΘ, EK perpendiculares a BΓ, ZH. Y como BΓ es más cercana al centro y ZH más lejana, entonces EK es mayor que EΘ [III, Def. 5]. Hágase EΛ igual a EΘ, y prolónguese a través de Λ la (recta) ΛM, trazada formando ángulos rectos con EK, hasta N, y trácense ME, EN, ZE, EH.

Y como EΘ es igual a EΛ, BΓ es también igual a MN [III, 14]. Como, a su vez, AE es igual a EM, y EΔ a EN, entonces AΔ es igual a ME, EN. Pero ME, EN son mayores que MN [I, 20] y MN es igual a BΓ; por tanto, AΔ es mayor que BΓ. Y como los dos (lados) ME, EN son iguales a los dos (lados) ZE, EH, y el ángulo MEN es mayor que el ángulo ZEH, entoces la base MN es mayor que la base ZH [I, 24]. Pero se ha demostrado que MN es igual a BΓ. Por tanto, la mayor es el diámetro AΔ, y BΓ es mayor que ZH.

Por consiguiente, en un círculo el diámetro es la (recta) mayor y, de las demás, la más cercana al centro es siempre mayor que la más lejana. Q. E. D.

Proposición 16

La (recta) trazada por el extremo del diámetro de un círculo formando ángulos rectos (con el mismo) caerá fuera del círculo, y no se interpondrá otra recta en el espacio entre la recta y la circunferencia; y el ángulo del semicírculo es mayor y el restante menor que cualquier ángulo rectilíneo agudo.

Sea el círculo ABΓ en torno al centro Δ y el diámetro AB.

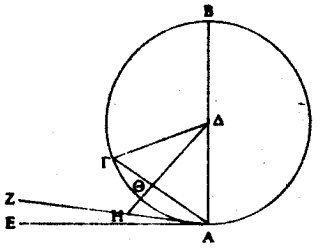
Digo que la (recta) trazada por el extremo A formando ángulos rectos con AB caerá fuera del círculo.

Pues, supongamos que no, entonces, si es posible, caiga dentro como ΓA, y trácese ΔΓ.

Como ΔA es igual a ΔΓ, el ángulo ΔAΓ es también igual al ángulo AΓΔ [I, 5]. Pero el ángulo ΔAΓ es recto; por tan-

to, el ángulo ΑΓΔ es también recto; entonces los dos ángulos ΑΔΓ, ΑΓΔ del triángulo ΑΓΔ son iguales a dos rectos; lo cual es imposible [I, 17]. Por tanto, la (recta) trazada a partir del punto Α formando ángulos rectos con ΒΑ no caerá dentro del círculo. De manera semejante demostraríamos que tampoco caerá sobre la circunferencia; por tanto (caerá) fuera.

Caiga como ΑΕ.



Digo, además, que en el espacio entre la recta ΑΕ y la circunferencia ΓΘΑ no se interpondrá otra recta.

Pues, si es posible, interpóngase como ΖΑ, y trácese a partir del punto Δ la (recta) ΔΗ perpendicular a ΖΑ. Y como el (ángulo) ΑΗΔ es recto y el ángulo ΔΑΗ es menor que un recto, entonces ΑΔ es mayor que ΔΗ [I, 19]. Pero ΔΑ es igual a ΔΘ; entonces ΔΘ es mayor que ΔΗ, la menor que la mayor; lo cual es imposible. Por tanto, en el espacio entre la recta y la circunferencia no se interpondrá otra recta.

Digo también que el ángulo del semicírculo, comprendido por la recta ΒΑ y la circunferencia ΓΘΑ es mayor que cualquier ángulo rectilíneo agudo, y el restante, comprendido por la circunferencia ΓΘΑ y la recta ΑΕ es menor que cualquier ángulo rectilíneo agudo.

Pues si un ángulo rectilíneo es mayor que el (ángulo) comprendido por la recta ΒΑ y la circunferencia ΓΘΑ, pero menor que el comprendido por la circunferencia ΓΘΑ y la recta ΑΕ, en el espacio entre la circunferencia ΓΘΑ y la recta ΑΕ se interpondrá una recta que hará mayor el ángulo comprendido por las rectas que el comprendido por la recta ΒΑ y la circunferencia ΓΘΑ, y menor que el comprendido por la circunferencia ΓΘΑ y la recta ΑΕ. Pero no se interpone; por tanto, el ángulo agudo comprendido por rectas no será mayor que el ángulo comprendido por la recta ΒΑ y la circunferencia ΓΘΑ, ni menor que el comprendido por la circunferencia ΓΘΑ y la recta ΑΕ.

Porisma:

A partir de esto queda claro que la recta trazada por el extremo del diámetro de un círculo formando ángulos rectos (con el mismo) toca el círculo. Q. E. D. [91].

Proposición 17

Desde un punto dado trazar una línea recta tangente a un círculo dado.

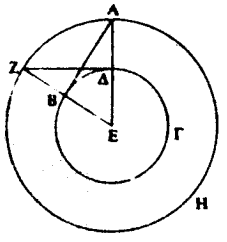
Sea el punto dado Α y el círculo dado ΒΓΔ.

Así pues, hay que trazar desde el punto Α una línea recta tangente al círculo ΒΓΔ.

---

[91] Jordano de Nemore (siglo XIII) dio el nombre de «ángulo de contingencia» al ángulo formado por el círculo y la tangente. La cuestión de si los ángulos de contingencia son nulos o tienen una magnitud que ha de tratarse por medios muy distintos a los empleados en la determinación de ángulos rectilíneos, fue un problema que ocupó a los comentadores de Euclides hasta el siglo XVIII (e. g., Proclo, Campano, Clavio, Wallis). *Vid.* Heath, ed. cit., 2, págs. 39-43.

Tómese, pues, el centro del círculo, E, y trácese AE y con el centro E y la distancia EA descríbase el círculo AZH, y a partir del (punto) Δ trácese ΔZ formando ángulos rectos con EA, y trácense EZ, AB.



Digo que ha sido trazada AB desde el punto A tangente al círculo BΓΔ.

Pues como E es el centro de los círculos BΓΔ, AZH, entonces EA es igual a EZ, y EΔ a EB; así pues, los dos (lados) AE, EB son iguales a los dos (lados) ZE, EΔ; y comprenden el ángulo común correspondiente a E; por tanto, la base ΔZ es igual a la base AB, y el triángulo ΔEZ es igual al triángulo EBA, y los ángulos restantes son iguales a los ángulos restantes [I, 4]; por tanto, el (ángulo) EΔZ es igual al ángulo EBA. Pero el (ángulo) EΔZ es recto; luego el (ángulo) EBA también es recto. Y EB es un radio; ahora bien, la (recta) trazada por el extremo del diámetro de un círculo formando ángulos rectos (con el mismo) toca el círculo [III, 16, Por.]; por tanto, AB es tangente al círculo BΓΔ.

Por consiguiente, a partir del punto dado A se ha trazado la línea recta AB tangente al círculo BΓΔ. Q. E. F.

Proposición 18

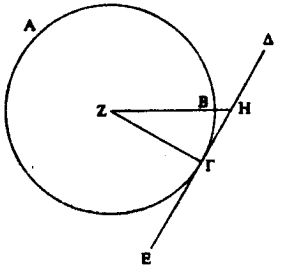
Si una recta toca un círculo, y se traza una recta desde el centro hasta el punto de contacto, la (recta) trazada será perpendicular a la tangente [92].

[92] La tangente: *hē ephaptoménē*. Es la primera vez que este término aparece sustantivado. En la proposición anterior, por ejemplo, se emplea como adjetivo.

Toque, pues, una recta ΔE al círculo ABΓ en el punto Γ, y tómese el centro Z del círculo ABΓ, y de Z a Γ trácese ZΓ.

Digo que ZΓ es perpendicular a ΔE.

Pues si no, trácese, desde Z, ZH perpendicular a ΔE.



Así pues, como el ángulo ZHΓ es recto, entonces el (ángulo) ZΓH es agudo [I, 17]; pero al lado mayor lo subtiende el ángulo mayor [I, 19]; por tanto, ZΓ es mayor que ZH; pero ZΓ es igual a ZB, por tanto, ZB es mayor que ZH, la menor que la mayor; lo cual es imposible; por tanto, ZH no es perpendicular a ΔE. Entonces de manera semejante demostraríamos que ninguna otra (es perpendicular) excepto ZΓ; por tanto, ZΓ es perpendicular a ΔE.

Por consiguiente, si una recta toca un círculo, y se traza una recta desde el centro hasta el punto de contacto, la recta trazada será perpendicular a la tangente. Q. E. D.

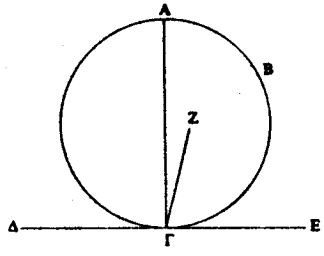
Proposición 19

Si una recta toca un círculo, y desde el punto de contacto se traza una línea recta formando ángulos rectos con la tangente, el centro del círculo estará en la recta trazada.

Toque, pues, una recta ΔE el círculo ABΓ en el punto Γ y desde Γ trácese ΓA formando ángulos rectos con ΔE.

Digo que el centro del círculo está en AΓ.

Pues, supongamos que no, entonces, si es posible, sea Z (el centro) y trácese ΓZ.



Como una recta ΔE toca el círculo ABΓ y ZΓ ha sido trazada desde el centro al punto de contacto, entonces ZΓ es perpendicular a ΔE [III, 18]; por tanto, el (ángulo) ZΓE es recto. Pero el (ángulo) AΓE también es recto; luego el ángulo ZΓE es igual al (ángulo) AΓE, el menor al mayor; lo cual es imposible. Por tanto, Z no es el centro del círculo ABΓ. De manera semejante demostraríamos que ningún otro (punto) salvo (un punto) en AΓ.

Por consiguiente, si una recta toca un círculo, y desde el punto de contacto se traza una línea recta formando ángulos rectos con la tangente, el centro del círculo estará en la recta trazada. Q. E. D.

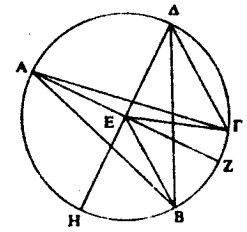
Proposición 20

En un círculo, el ángulo correspondiente al centro es el doble del correspondiente a la circunferencia cuando los ángulos tienen como base la misma circunferencia.

Sea el círculo ABΓ y sea el ángulo correspondiente a su centro BEΓ, y el (ángulo) BAΓ el correspondiente a la circunferencia, y tengan como base la misma circunferencia BΓ.

Digo que el ángulo BEΓ es el doble del (ángulo) BAΓ.

Una vez trazada la recta AE, prolónguese hasta Z.



Así pues, como EA es igual a EB, el ángulo EAB es también igual al (ángulo) EBA [I, 5]; por tanto, los ángulos EAB, EBA son el doble de EAB. Pero el (ángulo) BEZ es igual a los (ángulos) EAB, EBA [I, 32]; por tanto, el (ángulo) BEZ es también el doble de EAB. Por lo mismo, el (ángulo) ZEΓ es el doble del (ángulo) EAΓ. Luego el (ángulo) entero BEΓ, es el doble del (ángulo) entero BAΓ.

Trácese de nuevo una recta quebrada [93] y sea el otro ángulo BΔΓ, y, una vez trazada ΔE, prolónguese hasta H. De manera semejante demostraríamos que el ángulo HEΓ es el doble del (ángulo) EΔΓ, de cuyas partes el (ángulo) HEB es el doble del (ángulo) EΔB; por tanto, el (ángulo) restante BEΓ es el doble de BΔΓ.

Por consiguiente, en un círculo el ángulo correspondiente al centro es el doble del correspondiente a la circunferencia cuando los ángulos tienen como base la misma circunferencia. Q. E. D.

---

[93] *Keklásthō dè pálin.* Lit.: «quiébrese de nuevo (otra línea)». El verbo *kláō*, «romper, quebrar», era el término técnico para dibujar desde un punto una línea recta (quebrada) que se encuentra con otra línea recta o curva y después vuelve hacia atrás a otro punto, o para dibujar líneas

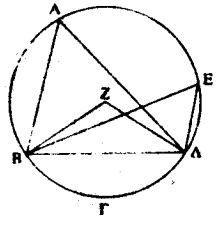
rectas desde dos puntos que se encuentran en otro punto, línea recta o curva. *Keklásthai* es uno de los términos geométricos cuya definición debe ser asumida según Aristóteles *(Apo.* I 10, 76b9).

Proposición 21

En un círculo los ángulos en el mismo segmento son iguales entre sí.

Sea el círculo ABΓΔ y en el mismo segmento BAEΔ estén los ángulos BAΔ, BEΔ.

Digo que los ángulos BAΔ, BEΔ son iguales entre sí.



Tómese, pues, el centro del círculo ABΓΔ y sea Z, y trácense BZ, ZΔ.

Y como el ángulo BZΔ es el correspondiente al centro y el (ángulo) BAΔ el correspondiente a la circunferencia, y tienen como base la misma circunferencia BΓΔ, entonces el ángulo BZΔ es el doble del (ángulo) BAΔ [III, 20]. Por lo mismo, el (ángulo) BZΔ es también el doble del (ángulo) BEΔ; luego el (ángulo) BAΔ es igual al ángulo BEΔ.

Por consiguiente, en un círculo los ángulos en el mismo segmento son iguales entre sí. Q. E. D.

Proposición 22

Los ángulos opuestos de los cuadriláteros en los círculos son iguales a dos rectos.

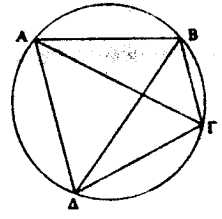
Sea el círculo ABΓΔ, y sea el cuadrilátero en el mismo ABΓΔ.

Digo que los ángulos opuestos son iguales a dos rectos.

Trácense AΓ, BΔ.

Así pues, como en todo triángulo los tres ángulos son iguales a dos rectos [I, 32], entonces los tres ángulos ΓAB, ABΓ, BΓA del triángulo ABΓ son iguales a dos rectos.

Pero el ángulo ΓAB es igual al (ángulo) BΔΓ: porque están en el mismo segmento BAΔΓ [III, 21]; y el (ángulo) AΓB (es igual) al (ángulo) AΔB: porque están en el mismo segmento AΔΓB; por tanto, el (ángulo) entero AΔΓ es igual a los (ángulos) BAΓ, AΓB. Añádase a ambos el (ángulo) ABΓ; entonces los (ángulos) ABΓ, BAΓ, AΓB son iguales a los (ángulos) ABΓ, AΔΓ; pero los ángulos ABΓ, BAΓ, AΓB son iguales a dos rectos. Por tanto, los (ángulos) ABΓ, AΔΓ son también iguales a dos rectos. De manera semejante demostraríamos que los ángulos BAΔ, ΔΓB son también iguales a dos rectos.

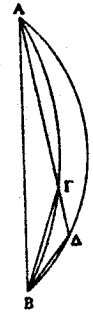


Por consiguiente, los ángulos opuestos de los cuadriláteros en los círculos son iguales a dos rectos. Q. E. D.

Proposición 23

Sobre la misma recta no se podrán construir dos segmentos circulares semejantes y desiguales en el mismo lado.

Pues, si fuera posible, sobre la misma recta AB constrúyanse dos segmentos circulares semejantes y desiguales en el mismo lado, los (segmentos) AΓB, AΔB, y trácese la (recta) AΓΔ, y trácense ΓB, ΔB.



Así pues, como el segmento AΓB es semejante al segmento AΔB, y segmentos semejantes de círculos son los que admiten ángulos iguales [III, Def. 11], entonces el ángulo AΓB es igual al (án-

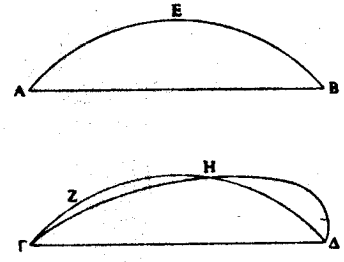
gulo) AΔB, el externo al interno, lo cual es imposible [I, 16].

Por consiguiente, sobre la misma recta no se podrán construir dos segmentos circulares semejantes y desiguales en el mismo lado. Q. E. D.

Proposición 24

Los segmentos circulares semejantes que están sobre rectas iguales son iguales entre sí.

Estén, pues, sobre las rectas iguales AB, ΓΔ los segmentos circulares semejantes AEB, ΓZΔ.



Digo que el segmento AEB es igual al segmento ΓZΔ.

Pues si se aplica el segmento AEB al (segmento) ΓZΔ y se pone el punto A sobre el punto Γ y la recta AB sobre la (recta) ΓΔ, el punto B coincidirá también con el punto Δ por ser igual AB a ΓΔ; y habiendo coincidido AB con ΓΔ coincidirá también el segmento AEB con el (segmento) ΓZΔ. Porque si la recta AB coincide con la (recta) ΓΔ pero el segmento AEB no coincide con el (segmento) ΓZΔ, o bien caerá dentro de él o fuera, o se desviará como el (segmento) ΓHΔ, y un círculo cortará a otro círculo en más de dos puntos; lo cual es imposible [III, 10]. Por tanto, no es

posible que coincidiendo la recta AB con la recta ΓΔ no coincida también el segmento AEB con el segmento ΓZΔ; luego coincidirá con él y será igual a él.

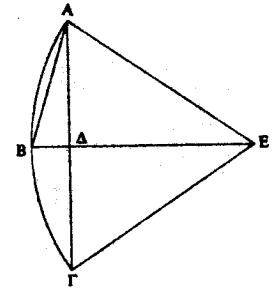
Por consiguiente, segmentos circulares semejantes que están sobre rectas iguales son iguales entre sí. Q. E. D.

Proposición 25

Dado un segmento de círculo completar el trazado del círculo [94] *del que es segmento.*

Sea ABΓ el segmento de círculo dado.

Así pues, hay que completar el trazado del círculo del que el segmento ABΓ es segmento.



Divídase, pues, en dos partes iguales AΓ en el punto Δ, y trácese a partir del punto Δ la (recta) ΔB formando ángulos rectos con AΓ, y trácese AB; entonces el ángulo ABΔ es o mayor o igual o menor que el (ángulo) BAΔ.

En primer lugar, sea mayor, y constrúyase en la recta BA y en su punto A el (ángulo) BAE igual al (ángulo) ABΔ, y prolónguese ΔB hasta E, y trácese EΓ.

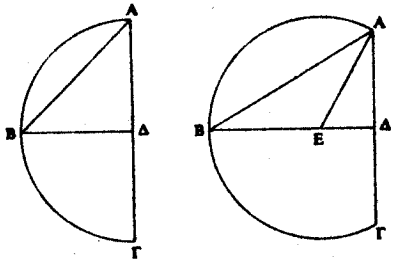
Así pues, como el ángulo ABE es igual al (ángulo) BAE, entonces la recta EB es también igual a EA [I, 6]. Y como AΔ es igual a ΔΓ, y ΔE es común, los dos lados AΔ, ΔE son iguales respectivamente a los dos (lados) ΓΔ, ΔE; y el ángulo AΔE es igual al ángulo ΓΔE: porque cada uno (de

[94] *Prosanagrápsai.*

ellos) es recto; por tanto, la base AE es igual a la base EΓ. Pero se ha demostrado que AE es igual a BE; por tanto, BE es igual a ΓE; luego las tres (rectas) AE, EB, EΓ son iguales entre sí; así pues, el círculo descrito con el centro E y como distancia una de las (rectas) AE, EB, EΓ pasará también por los restantes puntos y se habrá completado su trazado [III, 9].

Por tanto, se ha completado el trazado del círculo del segmento de círculo dado. Y es evidente que el segmento ABΓ es menor que el semicírculo por encontrarse fuera de él el centro E.

De forma semejante, si el ángulo ABΔ es igual al (ángulo) BAΔ, al ser igual AΔ a cada una de las (rectas) BΔ, ΔΓ, las tres ΔA, ΔB, ΔΓ serán iguales entre sí y Δ será el centro del círculo completo y, evidentemente, ABΓ será un semicírculo.



Pero si el (ángulo) ABΔ es menor que el (ángulo) BAΔ, y construimos en la recta BA y en su punto A (un ángulo) igual al ángulo ABΔ, el centro caerá dentro del segmento ABΓ sobre la (recta) ΔB, y el segmento ABΓ será claramente mayor que el semicírculo.

Por consiguiente, dado un segmento de círculo, se ha completado el trazado del círculo. Q. E. F.

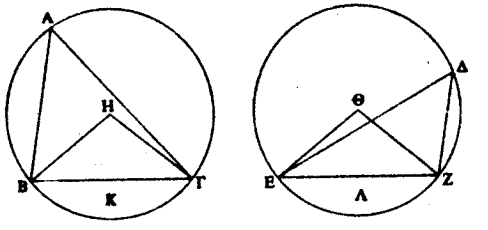
Proposición 26

En los círculos iguales los ángulos iguales están sobre circunferencias iguales, ya estén en los centros o en las circunferencias.

Sean ABΓ, ΔEZ círculos iguales y estén en ellos los ángulos iguales BHΓ, EΘZ en los centros, y en las circunferencias los (ángulos iguales) BAΓ, EΔZ.

Digo que la circunferencia BKΓ es igual a la circunferencia EΛZ.

Trácense, pues, BΓ, EZ.



Y como los círculos ABΓ, ΔEZ son iguales, los radios son iguales; entonces las dos (rectas) BH, HΓ son iguales a las dos (rectas) EΘ, ΘZ; y el ángulo correspondiente a H es igual al ángulo correspondiente a Θ; por tanto, la base BΓ es igual a la base EZ [I, 4]. Y como el ángulo correspondiente a A es también igual al ángulo correspondiente a Δ, entonces el segmento BAΓ es semejante al segmento EΔZ [III, Def. 11]; y están sobre rectas iguales [BΓ, EZ]; ahora bien, los segmentos circulares semejantes (que están) sobre rectas iguales son iguales entre sí [III, 24]; por tanto, el segmento BAΓ es igual al segmento EΔZ. Pero el círculo entero ABΓ es igual al círculo entero ΔEZ; por

tanto, la circunferencia restante BKΓ es igual a la circunferencia EΛZ.

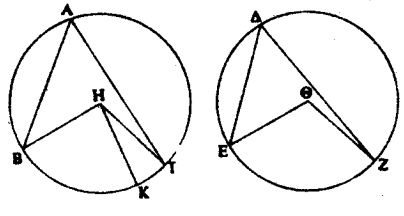
Por consiguiente, en los círculos iguales los ángulos iguales están sobre circunferencias iguales, ya estén en los centros o en las circunferencias. Q. E. D.

Proposición 27

En los círculos iguales, los ángulos que están sobre circunferencias iguales son iguales entre sí, ya estén en los centros o en las circunferencias.

Estén, pues, en los círculos iguales ABΓ, ΔEZ, sobre las circunferencias iguales BΓ, EZ, los (ángulos) BHΓ, EΘZ, en los centros H, Θ, y en las circunferencias los (ángulos) BAΓ, EΔZ.

Digo que el ángulo BHΓ es igual al (ángulo) EΘZ, y el (ángulo) BAΓ es igual al (ángulo) EΔZ.



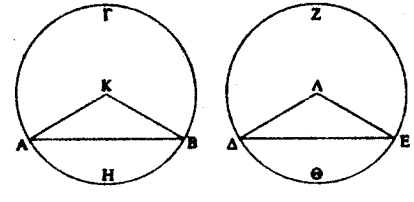
Pues si el (ángulo) BHΓ no es igual al (ángulo) EΘZ, uno de ellos es mayor. Sea mayor el (ángulo) BHΓ, y constrúyase en la recta BH y en su punto H el (ángulo) BHK igual al ángulo EΘZ [I, 23]; pero los ángulos iguales están sobre circunferencias iguales cuando están en los centros [III, 26]; por tanto, la circunferencia BK es igual a la circunferencia EZ. Pero EZ es igual a BΓ; entonces BK es también igual a BΓ, la menor a la mayor; lo cual es imposible. Por tanto, los ángulos BHΓ, EΘZ no son desiguales; luego son iguales y el (ángulo) correspondiente a A es la mitad del (ángulo) BHΓ, y el correspondiente a Δ la mitad del (ángulo) EΘZ [III, 20]; por tanto, el ángulo correspondiente a A es también igual al correspondiente a Δ.

Por consiguiente, en los círculos iguales los ángulos que están sobre circunferencias iguales son iguales entre sí, ya estén en los centros o en las circunferencias. Q. E. D.

Proposición 28

En los círculos iguales las rectas iguales cortan circunferencias iguales, la mayor (igual) a la mayor y la menor a la menor.

Sean ABΓ, ΔEZ círculos iguales, y en los círculos estén las rectas iguales AB, ΔE que cortan como circunferencias mayores AΓB, ΔZE y como menores AHB, ΔΘE.



Digo que la circunferencia mayor AΓB es igual a la circunferencia mayor ΔZE, y la circunferencia menor AHB a la (circunferencia) ΔΘE.

Tómense, pues, los centros de los círculos K, Λ y trácense AK, KB, ΔΛ, ΛE.

Y como son círculos iguales los radios también son iguales; entonces los dos (lados) AK, KB son iguales a los dos (lados) ΛΛ, ΛΕ; y la base AB igual a la base ΔΕ; por tanto, el ángulo AKB es igual al ángulo ΔΛΕ [I, 8].

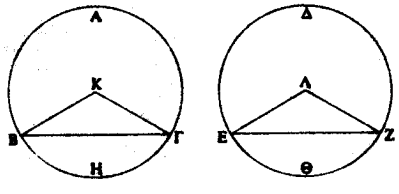
Pero los ángulos iguales están sobre circunferencias iguales, cuando están en los centros [III, 26]; por tanto, la circunferencia AHB es igual a la (circunferencia) ΔΘΕ. Pero el círculo entero ABΓ es también igual al círculo entero ΔΕΖ; por tanto, la circunferencia restante AΓB es también igual a la circunferencia restante ΔΖΕ.

Por consiguiente, en los círculos iguales, las rectas iguales cortan circunferencias iguales, la mayor (igual) a la mayor y la menor a la menor. Q. E. D.

Proposición 29

En los círculos iguales las rectas iguales subtienden circunferencias iguales.

Sean ABΓ, ΔΕΖ círculos iguales, y en ellos córtense las circunferencias iguales BHΓ, ΕΘΖ, y trácense las rectas BΓ, ΕΖ.



Digo que BΓ es igual a ΕΖ.

Tómense, pues, los centros de los círculos y sean K, Λ, y trácense BK, KΓ, ΕΛ, ΛΖ.

Ahora bien, como la circunferencia BHΓ es igual a la circunferencia ΕΘΖ, el ángulo BKΓ es también igual al (ángulo) ΕΛΖ [III, 27].

Y como los círculos ABΓ, ΔΕΖ son iguales, los radios son también iguales; entonces las dos (rectas) BK, KΓ son iguales a las dos (rectas) ΕΛ, ΛΖ; y comprenden ángulos iguales; por tanto, la base BΓ es igual a la base ΕΖ [I, 4].

Por consiguiente, en los círculos iguales las rectas iguales subtienden circunferencias iguales. Q. E. D.

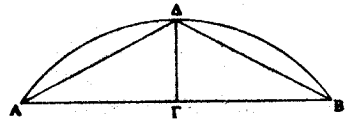
Proposición 30

Dividir en dos partes iguales una circunferencia dada.

Sea AΔB la circunferencia dada.

Así pues, hay que dividir en dos partes iguales la circunferencia AΔB.

Trácese, pues, AB y divídase en dos partes iguales por el (punto) Γ, y a partir del punto Γ trácese ΓΔ formando ángulos rectos con la recta AB, y trácense AΔ, ΔB.



Ahora bien, como AΓ es igual a ΓB, y ΓΔ es común, los dos (lados) AΓ, ΓΔ son iguales a los dos (lados) BΓ, ΓΔ; y el ángulo AΓΔ es igual al ángulo BΓΔ: porque cada uno (de ellos) es recto; entonces la base AΔ es igual a la base ΔB [I, 4]. Pero las rectas iguales cortan circunferencias iguales la mayor (igual) a la mayor y la menor a la menor

[III, 28]; y cada una de las circunferencias AΔ, ΔB es menor que un semicírculo; por tanto, la circunferencia AΔ es igual a la circunferencia ΔB.

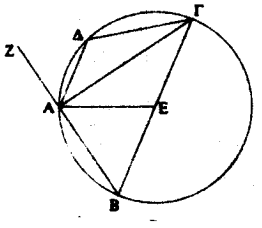
Por consiguiente, se ha dividido en dos partes iguales por el punto Δ la circunferencia dada. Q. E. F.

Proposición 31

En un círculo el ángulo en el semicírculo es recto, el (ángulo) en el segmento mayor es menor que un recto, el (ángulo) en el segmento menor es mayor que un recto; y además el ángulo del segmento mayor es mayor que un recto y el ángulo del segmento menor es menor que un recto.

Sea ABΓΔ el círculo y sea BΓ su diámetro y el centro E, y trácense BA, AΓ, AΔ, ΔΓ.

Digo que el ángulo BAΓ en el semicírculo BAΓ es recto, y el ángulo ABΓ en el segmento ABΓ mayor que el semicírculo es menor que un recto, y el ángulo AΔΓ en el segmento AΔΓ menor que el semicírculo es mayor que un recto.



Trácese AE, y prolónguese BA hasta Z.

Y como BE es igual a EA, el ángulo ABE es también igual al (ángulo) BAE [I, 5]. Como a su vez ΓE es igual a EA, el (ángulo) AΓE es también igual al (ángulo) ΓAE [I, 5]; por tanto, el (ángulo) entero BAΓ es igual a los dos (ángulos) ABΓ, AΓB. Pero el (ángulo) ZAΓ exterior al triángulo ABΓ es también igual a los dos ángulos ABΓ, AΓB [I, 32]; por tanto, el ángulo BAΓ es también igual al (ángulo) ZAΓ; luego (cada uno de ellos) es recto, [I, Def. 10]; así pues, el ángulo BAΓ en el semicírculo BAΓ es recto.

Además, como los dos ángulos ABΓ, BAΓ del triángulo ABΓ son menores que dos rectos [I, 17], y el (ángulo) BAΓ es recto, entonces el ángulo ABΓ es menor que un recto; y está en el segmento ABΓ mayor que el semicírculo.

Y como ABΓΔ es un cuadrilátero en un círculo, y los ángulos opuestos de los cuadriláteros en los círculos son iguales a dos rectos [III, 22], y el (ángulo) ABΓ es menor que un recto, entonces el ángulo restante AΔΓ es mayor que un recto; y está en el segmento AΔΓ menor que el semicírculo.

Digo también que el ángulo del segmento mayor, el comprendido por la circunferencia ABΓ y la recta AΓ es mayor que un recto, y que el ángulo del segmento menor, el comprendido por la circunferencia AΔ[Γ] y la recta AΓ es menor que un recto. Y esto es claro de suyo. Pues como el (ángulo) comprendido por las rectas BA, AΓ es recto, entonces el comprendido por la circunferencia ABΓ y la recta AΓ es mayor que un recto. Asimismo, como el (ángulo comprendido) por las rectas AΓ, AZ es recto, entonces el comprendido por la recta ΓA y la circunferencia AΔ[Γ] es menor que un recto.

Por consiguiente, en un círculo el ángulo en el semicírculo es recto, el (ángulo) en el segmento mayor es menor que un recto, y el ángulo en el [segmento] menor es mayor que un recto, y además el [ángulo] del segmento mayor, mayor que un recto, y el del segmento menor, menor que un recto. Q. E. D.

[Porisma:

A partir de esto queda claro que si un ángulo de un triángulo es igual a los (otros) dos, el ángulo es recto por

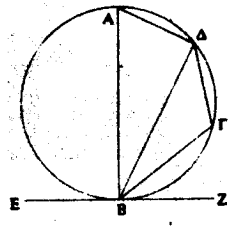
ser el ángulo exterior a aquel igual a los mismos; y si los ángulos adyacentes son iguales, son rectos] [95].

Proposición 32

Si una recta toca un círculo, y desde el punto de contacto hasta el círculo se traza una recta que corte el círculo, los ángulos que forma con la recta tangente serán iguales a los ángulos en los segmentos alternos del círculo.

Toque, pues, la recta EZ al círculo ABΓΔ en el punto B, y a partir del punto B trácese una recta BΔ en el círculo ABΓΔ que lo corte.

Digo que los ángulos que forma BΔ con la tangente EZ serán iguales a los ángulos en los segmentos alternos del círculo, es decir, que el ángulo ZBΔ es igual al ángulo construido en el segmento BAΔ, y el ángulo EBΔ es igual al ángulo construido en el segmento ΔΓB.



Trácese, pues, a partir de B la (recta) BA formando ángulos rectos con EZ, y tómese al azar un punto Γ en la circunferencia BΔ, y trácense AΔ, ΔΓ, ΓB.

Y como una recta EZ es tangente al círculo ABΓΔ en (el punto) B, y desde el punto de contacto se ha trazado BA formando ángulos rectos con la tangente, entonces el centro del círculo ABΓΔ está en BA [III, 19]. Por

[95] Este porisma es omitido por Heiberg como una interpolación de fecha posterior a Teón. No hay necesidad de este corolario y, por otra parte, la claúsula *hóper édei deîxai* aparece antes del porisma, y no después, como es habitual en estos casos en los *Elementos*.

tanto, BA es diámetro del círculo ABΓΔ; luego el ángulo AΔB, siendo un (ángulo) en un semicírculo, es recto [III, 31]. Por tanto, los ángulos restantes BAΔ, ABΔ son iguales a un recto [I, 32]. Pero el (ángulo) ABZ es recto; por tanto, el (ángulo) ABZ es igual a los (ángulos) BAΔ, ABΔ. Quítese de ambos el (ángulo) ABΔ; entonces el ángulo restante ΔBZ es igual al ángulo BAΔ en el segmento alterno del círculo. Asimismo, como ABΓΔ es un cuadrilátero en un círculo, sus ángulos opuestos son iguales a dos rectos [III, 22]. Pero los (ángulos) ΔBZ, ΔBE son también iguales a dos rectos; luego los (ángulos) ΔBZ, ΔBE son iguales a los (ángulos) BAΔ, BΓΔ, de los cuales el (ángulo) BAΔ se ha demostrado que es igual al (ángulo) ΔBZ; por tanto, el ángulo ΔBE es igual al (ángulo) ΔΓB en el segmento alterno del círculo ΔΓB.

Por consiguiente, si una recta toca un círculo y desde el punto de contacto hasta el círculo se traza una recta que corte el círculo, los ángulos que forma con la recta tangente serán iguales a los ángulos en los segmentos alternos del círculo. Q. E. D.

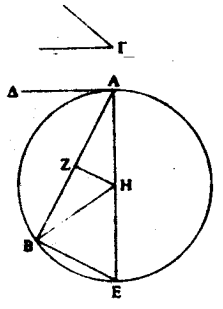
Proposición 33

Sobre una recta dada, describir un segmento de círculo que admita un ángulo igual a un ángulo rectilíneo dado.

Sea AB la recta dada, y el ángulo rectilíneo dado el correspondiente a Γ.

Así pues, hay que describir sobre la recta dada AB un segmento de círculo que admita un ángulo igual al correspondiente a Γ.

El ángulo correspondiente a Γ es entonces o agudo, o recto u obtuso; sea en primer lugar agudo, y como en la primera figura, constrúyase en (la recta) AB y en su punto A el (ángulo) BAΔ igual al ángulo correspondiente a Γ; entonces el (ángulo) BAΔ es también agudo. Trácese AE formando ángulos rectos con ΔA, y divídase en dos partes iguales AB en el (punto) Z, y trácese a partir del punto Z, ZH formando ángulos rectos con AB, y trácese HB.

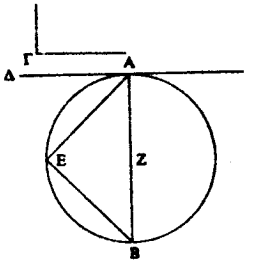


Y como AZ es igual a ZB, y ZH es común, los dos (lados) AZ, ZH son iguales a los dos (lados) BZ, ZH; y el (ángulo) AZH es igual al (ángulo) BZH; luego la base AH es igual a la base BH [I, 4]. Por tanto, el círculo descrito con el centro H y la distancia HA pasará también por B. Descríbase y sea ABE, y trácese EB. Así pues, como AΔ forma ángulos rectos con AE a partir del extremo A del diámetro AE, entonces AΔ toca el círculo ABE [III, 16, Por.]; así pues, como una recta AΔ toca el círculo ABE, y desde el punto de contacto A ha sido trazada una recta AB hasta el círculo ABE, entonces el ángulo ΔAB es igual al (ángulo) AEB en el segmento alterno del círculo [III, 32]. Pero el ángulo ΔAB es igual al correspondiente a Γ; luego el (ángulo) correspondiente a Γ es también igual al (ángulo) AEB.

Por consiguiente, sobre la recta dada AB ha sido descrito el segmento de círculo AEB que admite el ángulo AEB igual al ángulo dado correspondiente a Γ.

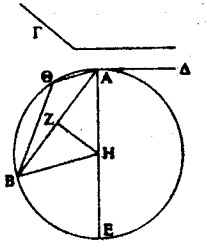
Ahora bien, sea recto el ángulo correspondiente a Γ; y haya que describir asimismo sobre AB un segmento de círculo que admita un ángulo igual al ángulo recto correspondiente a Γ.

Constrúyase el (ángulo) BAΔ igual al ángulo recto correspondiente a Γ como está en la segunda figura y divídase en dos partes iguales AB en el (punto) Z, y con el centro Z y como distancia una de las rectas ZA, ZB descríbase el círculo AEB.



Entonces la recta AΔ toca el círculo ABE por ser recto el ángulo correspondiente a A [III, 16, Por.]. Y el ángulo BAΔ es igual al (ángulo) en el segmento AEB: porque siendo un ángulo en un semicírculo también él es recto [III, 31]. Pero el (ángulo) BAΔ es también igual al correspondiente a Γ. Entonces el (ángulo) en (el segmento) AEB es igual al correspondiente a Γ.

Por consiguiente, también se ha descrito sobre AB el segmento de círculo AEB que admite un ángulo igual al correspondiente a Γ.



Ahora bien, sea obtuso el ángulo correspondiente a Γ; y constrúyase sobre la recta AB y en su punto A el ángulo BAΔ igual a él como está en la tercera figura; trácese AE formando ángulos rectos con AΔ, y divídase de nuevo en dos partes iguales AB en el (punto) Z; trácese ZH formando ángulos rectos con AB, y trácese HB.

Y como AZ es de nuevo igual a ZB, y ZH es común, los dos (lados) AZ, ZH son iguales a los dos (lados) BZ, ZH; y el ángulo AZH es igual al ángulo BZH; luego la base AH es igual a la base BH [I, 4]; por tanto, el círculo descrito con el centro H y la distancia HA pasará también por B. Pase como AEB. Y como AΔ forma ángulos rectos con

el diámetro AE a partir de un extremo, entonces AΔ es tangente al círculo AEB [III, 16, Por.]. Y AB se ha trazado a partir del punto de contacto A; por tanto, el ángulo BAΔ es igual al ángulo construido en el segmento alterno de círculo AΘB [III, 32]. Pero el ángulo BAΔ es igual al correspondiente a Γ. Así pues, el ángulo en el segmento AΘB también es igual al correspondiente a Γ.

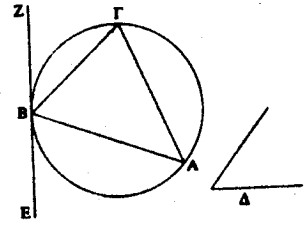
Por consiguiente, sobre la recta dada AB se ha descrito el segmento de círculo AΘB que admite un ángulo igual al correspondiente a Γ. Q. E. F.

Proposición 34

A partir de un círculo dado cortar un segmento que admita un ángulo igual a un ángulo rectilíneo dado.

Sea ABΓ el círculo dado y el ángulo rectilíneo dado el correspondiente a Δ.

Así pues, hay que cortar un segmento del círculo ABΓ que admita un ángulo igual al ángulo rectilíneo dado correspondiente a Δ.



Trácese EZ tangente al (círculo) ABΓ en el punto B, y constrúyase en la recta ZB y en su punto B el (ángulo) ZBΓ igual al ángulo correspondiente a Δ [I, 23].

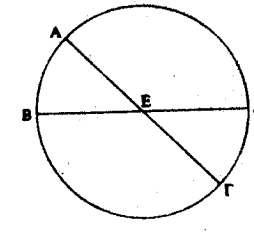
Así pues, como una recta EZ toca el círculo ABΓ, y a partir del punto de contacto B se ha trazado BΓ, entonces el (ángulo) ZBΓ es igual al ángulo construido en el segmento alterno BAΓ [III, 32]. Pero el ángulo ZBΓ es igual al correspondiente a Δ; por tanto, el (ángulo) en el segmento BAΓ es también igual al correspondiente a Δ.

Por consiguiente, a partir del círculo dado ABΓ se ha cortado el segmento BAΓ que admite un ángulo igual al ángulo rectilíneo dado correspondiente a Δ. Q. E. F.

Proposición 35

Si en un círculo se cortan dos rectas entre sí, el rectángulo comprendido por los segmentos de una es igual al rectángulo comprendido por los segmentos de la otra.

Córtense, pues, en el círculo ABΓΔ las dos rectas AΓ, BΔ en el punto E.

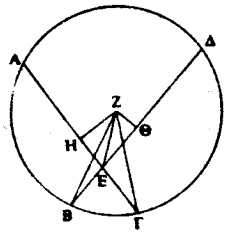


Digo que el rectángulo comprendido por AE, EΓ es igual al rectángulo comprendido por ΔE, EB.

Así pues, si AΓ, BΔ pasan por el centro de modo que E sea el centro del círculo ABΓΔ, queda claro que siendo iguales AE, EΓ, ΔE, EB el rectángulo comprendido por AE, EΓ es también igual al rectángulo comprendido por ΔE, EB.

No pasen por el centro AΓ, ΔB y tómese el centro del círculo ABΓΔ y sea Z, y trácense a partir de Z, ZH, ZΘ perpendiculares a las rectas AΓ, ΔB y trácense ZB, ZΓ, ZE.

Y puesto que una recta que pasa por el centro HZ corta formando ángulos rectos a otra recta que no pasa por el centro AΓ, la divide también en dos partes iguales [III, 3]; por tanto, AH es igual a HΓ. Así pues, como la recta AΓ se ha cortado en partes iguales en H y en desiguales en E, entonces el rectángulo comprendido por AE, EΓ junto con el cuadrado de EH es igual al (cuadrado) de HΓ [II, 5]; añádase el (cuadrado) de HZ; entonces el (rectángulo comprendido) por AE, EΓ junto con los (cuadrados) de HE, HZ es igual a los (cuadrados) de ΓH, HZ. Pero el (cuadrado) de ZE es igual a los (cuadrados) de EH, HZ, y el (cuadrado) de ZΓ es igual a los (cuadrados) de ΓH, HZ [I, 47]; entonces el (rectángulo comprendido) por AE, EΓ junto con ZE es igual al (cuadrado) de ZΓ. Pero ZΓ es igual a ZB; entonces el (rectángulo comprendido) por AE, EΓ junto con el (cuadrado) de EZ es igual al (cuadrado) de ZB. Por lo mismo, el (rectángulo comprendido) por ΔE, EB junto con el (cuadrado) de ZE es también igual al (cuadrado) de ZB. Pero se ha demostrado asimismo, que el (rectángulo comprendido) por AE, EΓ junto con el (cuadrado) de ZE es igual al (cuadrado) de ZB; entonces el (rectángulo comprendido) por AE, EΓ junto con el (cuadrado) de ZE es igual al (rectángulo comprendido) por ΔE, EB junto con el cuadrado de ZE. Quítese de ambos el (cuadrado) de ZE; entonces el rectángulo restante comprendido por AE, EΓ es igual al rectángulo comprendido por ΔE, EB.



Por consiguiente, si en un círculo se cortan dos rectas entre sí, el rectángulo comprendido por los segmentos de una es igual al rectángulo comprendido por los segmentos de la otra. Q. E. D.

Proposición 36

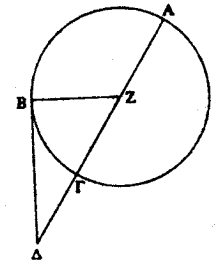
Si se toma un punto fuera de un círculo y de él al círculo caen dos rectas, y una de ellas corta el círculo y la otra lo toca, el (rectángulo comprendido) por la secante entera y la (parte) exterior tomada entre el punto y la circunferencia convexa es igual al cuadrado de la tangente.

Tómese, pues, un punto Δ fuera del círculo ABΓ, y de Δ al círculo caigan las dos rectas ΔΓ[A], ΔB; y corte una, ΔΓA, el círculo ABΓ, y tóquelo la otra BΔ.

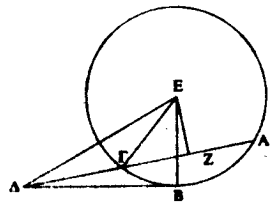
Digo que el rectángulo comprendido por AΔ, ΔΓ es igual al cuadrado de ΔB.

Entonces [Δ] ΓA o pasa por el centro o no.

En primer lugar, pase por el centro y sea Z el centro del círculo ABΓ, y trácese ZB; entonces el (ángulo) ZBΔ es recto [III, 18]. Y como la recta AΓ ha sido dividida en dos partes iguales en Z, y ΓΔ se ha añadido a ella, entonces el (rectángulo comprendido) por AΔ, ΔΓ junto con el (cuadrado) de ZΓ es igual al (cuadrado) de ZΔ [II, 6]. Pero ZΓ es igual a ZB; por tanto, el (rectángulo comprendido) por AΔ, ΔΓ junto con el (cuadrado) de ZB es igual al (cuadrado) de ZΔ. Pero los (cuadrados) de ZB, BΔ son iguales al (cuadrado) de ZΔ [I, 47]; luego el (rectángulo comprendido) por AΔ, ΔΓ junto con el (cuadrado) de ZB es igual a los (cuadrados) de ZB, BΔ. Quítese de ambos el (cuadrado) de ZB; entonces el (rectángulo) restante (comprendido) por AΔ, ΔΓ es igual al (cuadrado) de la tangente ΔB.



Ahora bien, no pase ΔΓA por el centro del círculo ABΓ, y tómese el centro E, y, desde E trácese EZ perpendicular a AΓ y trácense EB, EΓ, EΔ; entonces el ángulo EBΔ es recto [III, 18]. Y como una recta EZ que pasa por el centro corta formando ángulos rectos a una recta AΓ que no pasa por el centro, también la divide en dos partes iguales [III, 3]; entonces AZ es igual a ZΓ. Y como la recta AΓ se ha dividido en dos partes iguales en el punto Z, y se ha añadido a ella ΓΔ, el (rectángulo comprendido) por AΔ, ΔΓ junto con el (cuadrado) de ZΓ es igual al (cuadrado) de ZΔ [II, 6]. Añádase a ambos el (cuadrado) de ZE; entonces el (rectángulo comprendido) por AΔ, ΔΓ junto con los (cuadrados) de ΓZ, ZE es igual a los (cuadrados) de ZΔ, ZE. Pero el (cuadrado) de EΓ es igual a los (cuadrados) de ΓZ, ZE: porque el (ángulo) EZΓ es recto [I, 47]; pero el (cuadrado) de EΔ es igual a los (cuadrados) de ΔZ, ZE; por tanto, el (rectángulo comprendido) por AΔ, ΔΓ junto con el (cuadrado) de EΓ es igual al (cuadrado) de EΔ. Y EΓ es igual a EB; luego el (rectángulo comprendido) por AΔ, ΔΓ junto con el (cuadrado) de EB es igual al (cuadrado) de EΔ. Pero los (cuadrados) de EB, BΔ son iguales al (cuadrado) de EΔ: porque el ángulo EBΔ es recto [I, 47¹; por tanto, el (rectángulo comprendido) por AΔ, ΔΓ junto con el (cuadrado) de EB es igual a los (cuadrados) de EB, BΔ. Quítese de ambos el (cuadrado) de EB; entonces el (rectángulo) restante (comprendido) por AΔ, ΔΓ es igual al (cuadrado) de ΔB.



Por consiguiente, si se toma un punto fuera de un círculo, y de él al círculo caen dos rectas, y una de ellas corta el círculo y la otra lo toca, el (rectángulo comprendido) por la secante entera y la (parte) exterior tomada entre el punto y la circunferencia convexa es igual al cuadrado de la tangente. Q. E. D.

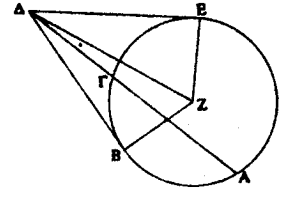
Proposición 37

Si se toma un punto fuera de un círculo y del punto al círculo caen dos rectas, y una de ellas corta el círculo, y la otra cae (sobre él), y además el (rectángulo comprendido) por la secante entera y la (parte) exterior tomada entre el punto y la circunferencia convexa es igual al cuadrado de la que cae, la (recta) que cae tocará el círculo.

Tómese, pues, un punto Δ fuera del círculo ABΓ, y de Δ al círculo ABΓ caigan las dos rectas ΔΓA, ΔB; corte ΔΓA el círculo, y caiga ΔB, y sea el (rectángulo comprendido) por AΔ, ΔΓ igual al (cuadrado) de ΔB.

Digo que ΔB toca el círculo ABΓ.

Pues trácese ΔE tangente a ABΓ, y tómese el centro del círculo ABΓ, y sea Z, y trácense ZE, ZB, ZΔ. Entonces el (ángulo) ZEΔ es recto [III, 18]. Y como ΔE toca el círculo ABΓ y ΔΓA lo corta, entonces el (rectángulo comprendido) por AΔ, ΔΓ es igual al cuadrado de ΔE. Pues el rectángulo comprendido por AΔ, ΔΓ era también igual al (cuadrado) de ΔB; entonces el (cuadrado) de ΔE es igual al (cuadrado) de ΔB; por tanto, ΔE es igual a ΔB. Pero ZE es también igual a ZB; entonces los dos (lados) ΔE, EZ son iguales a los dos (lados) ΔB, BZ; y su base ZΔ es común; por tanto, el ángulo ΔEZ es igual al ángulo ΔBZ [I, 8]. Pero ΔEZ es recto;



por tanto, el (ángulo) ΔBZ es recto. Y ZB prolongada es un diámetro; pero la (recta) trazada formando ángulos rectos con el diámetro de un círculo en un extremo toca el círculo [III, 16, Por.]; por tanto, ΔB toca el círculo ABΓ. De manera semejante se demostraría si el centro se encontrara en AΓ.

Por consiguiente, si se toma un punto fuera de un círculo y del punto al círculo caen dos rectas, y una de ellas corta el círculo, y la otra cae (sobre él), y además el (rectángulo comprendido) por la secante entera y la (parte) exterior tomada entre el punto y la circunferencia convexa es igual al cuadrado de la que cae, la que cae tocará el círculo. Q. E. D.

LIBRO CUARTO

DEFINICIONES

1. Se dice que una figura rectilínea está inscrita en otra figura rectilínea, cuando cada uno de los ángulos [96] de la figura inscrita toca los lados respectivos de aquella en la que se inscribe.
2. De manera semejante, se dice que una figura está circunscrita en torno a otra figura, cuando cada lado de la (figura) circunscrita toca los ángulos respectivos de aquella en torno a la cual se circunscribe.
3. Se dice que una figura rectilínea está inscrita en un círculo, cuando cada ángulo de la (figura) inscrita toca la circunferencia del círculo.
4. Se dice que una figura rectilínea está circunscrita en torno a un círculo, cuando cada lado de la (figura) circunscrita toca la circunferencia del círculo.
5. De manera semejante, se dice que un círculo está inscrito en una figura, cuando la circunferencia del círculo toca cada lado de la figura en la que está inscrito.

---

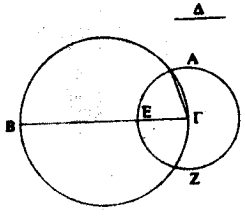
[96] Se refiere a los vértices.

6. Se dice que un círculo está circunscrito en torno a una figura, cuando la circunferencia del círculo toca cada ángulo de la (figura) en torno a la que está circunscrito.
7. Se dice que una recta está adaptada a un círculo, cuando sus extremos están en la circunferencia del círculo.

Proposición 1

Adaptar a un círculo dado una recta igual a una recta dada que no sea mayor que el diámetro del círculo.

Sea ABΓ el círculo dado y Δ la recta dada no mayor que el diámetro del círculo.



Así pues, hay que adaptar al círculo ABΓ una recta igual a la recta Δ.

Trácese el diámetro BΓ del círculo ABΓ. Entonces, si BΓ es igual a Δ, ya se habrá hecho lo propuesto: porque al círculo ABΓ se ha adaptado BΓ igual a la recta Δ. Pero si BΓ es mayor que Δ, hágase ΓE igual a Δ, y con el centro Γ y la distancia ΓE descríbase el círculo EAZ, y trácese ΓA.

Entonces, como el punto Γ es el centro del círculo EAZ, ΓA es igual a ΓE. Pero ΓE es igual a Δ; por tanto, Δ es igual a ΓA.

Por consiguiente, al círculo dado ABΓ se ha adaptado ΓA igual a la recta Δ. Q. E. F.

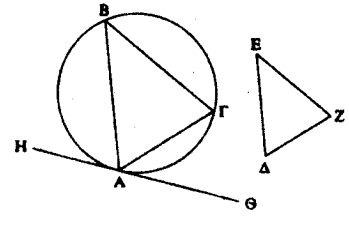
Proposición 2

Inscribir en un círculo dado un triángulo de ángulos iguales a los de un triángulo dado.

Sea ABΓ el círculo dado y ΔEZ el triángulo dado.

Así pues, hay que inscribir en el círculo ABΓ un triángulo de ángulos iguales a los del triángulo ΔEZ.

Trácese HΘ tangente al círculo ABΓ en el (punto) A [III, 16, Por.], y constrúyase en la recta AΘ y en su punto A el (ángulo) ΘAΓ igual al ángulo ΔEZ, y en la recta AH y en su punto A el (ángulo) HAB igual al (ángulo) ΔZE [I, 23], y trácese BΓ.



Pues bien, como una recta AΘ toca al círculo ABΓ, y desde el punto de contacto A hasta el círculo ha sido trazada la recta AΓ, entonces el (ángulo) ΘAΓ es igual al (ángulo) ABΓ en el segmento alterno del círculo [III, 32]. Pero el (ángulo) ΘAΓ es igual al ángulo ΔEZ; entonces el ángulo ABΓ es también igual al (ángulo) ΔEZ. Por lo mismo, el (ángulo) AΓB es también igual al (ángulo) ΔZE; por tanto, el (ángulo) restante BAΓ es igual al (ángulo) restante EΔZ [I, 32].

Por consiguiente, se ha inscrito en el círculo dado un triángulo de ángulos iguales a los del triángulo dado. Q. E. F.

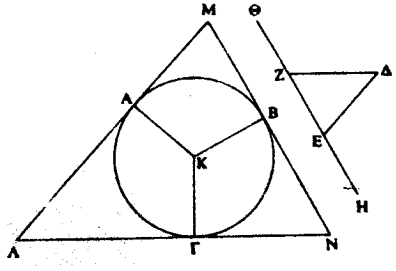
Proposición 3

Circunscribir en torno a un círculo dado un triángulo de ángulos iguales a los de un triángulo dado.

Sea ΑΒΓ el círculo dado, y ΔΕΖ el triángulo dado.

Así pues, hay que circunscribir en torno al círculo ΑΒΓ un triángulo de ángulos iguales a los del triángulo ΔΕΖ.

Prolónguese ΕΖ en ambos sentidos hasta los puntos Η, Θ, y tómese el centro Κ del círculo ΑΒΓ [III, 1], y trácese una recta cualquiera ΚΒ, y constrúyase en la recta ΚΒ y en su punto Κ el ángulo ΒΚΑ igual al ángulo ΔΕΗ, y el ángulo ΒΚΓ igual al ángulo ΔΖΘ [I, 23], y a través de los puntos Α, Β, Γ, trácense ΛΑΜ, ΜΒΝ, ΝΓΛ tangentes al círculo ΑΒΓ [III, 16, Por.].



Ahora bien, como las (rectas) ΛΜ, ΜΝ, ΝΛ tocan el círculo ΑΒΓ en los puntos Α, Β, Γ, y ΚΑ, ΚΒ, ΚΓ se han trazado desde el centro Κ hasta los puntos Α, Β, Γ, entonces los ángulos correspondientes a los puntos Α, Β, Γ son rectos [III, 18]. Y como los cuatro ángulos del cuadrilátero ΑΜΒΚ son iguales a cuatro rectos, puesto que ΑΜΒΚ se divide en dos triángulos y los ángulos ΚΑΜ, ΚΒΜ son rectos, entonces los (ángulos) restantes ΑΚΒ, ΑΜΒ son iguales a dos rectos. Pero los (ángulos) ΔΕΗ, ΔΕΖ son también iguales a dos rectos [I, 13]; entonces los (ángulos) ΑΚΒ, ΑΜΒ son iguales a los (ángulos) ΔΕΗ, ΔΕΖ, de los cuales ΑΚΒ es igual a ΔΕΗ; por tanto, el (ángulo) restante ΑΜΒ es igual al restante ΔΕΖ. De manera semejante, se demostraría que el (ángulo) ΑΝΒ es también igual al (ángulo) ΔΖΕ; por tanto, también el (ángulo) restante ΜΛΝ es igual al (ángulo) ΕΔΖ [I, 32]. Luego el triángulo ΑΜΝ es de ángulos iguales a los del triángulo ΔΕΖ; y está circunscrito en torno al círculo ΑΒΓ.

Por consiguiente, se ha circunscrito en torno al círculo dado un triángulo de ángulos iguales a los del triángulo dado. Q. E. F.

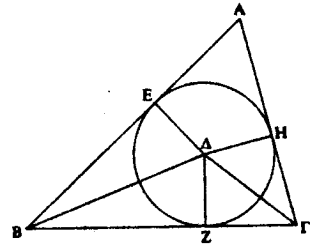
Proposición 4

Inscribir un círculo en un triángulo dado.

Sea ΑΒΓ el triángulo dado.

Así pues, hay que inscribir un círculo en el triángulo ΑΒΓ.

Divídanse en dos partes iguales los ángulos ΑΒΓ, ΑΓΒ con las rectas ΒΔ, ΓΔ [I, 9] y encuéntrense éstas en el punto Δ; trácense desde el punto Δ hasta las rectas ΑΒ, ΒΓ, ΓΑ las perpendiculares ΔΕ, ΔΖ, ΔΗ.



Ahora bien, como el ángulo ABΔ es igual al (ángulo) ΓBΔ, y el (ángulo) recto BEΔ es también igual al (ángulo) recto BZΔ, entonces EBΔ, ZBΔ son dos triángulos que tienen dos ángulos iguales a dos ángulos, y un lado igual a un lado, a saber: BΔ que subtiende a uno de los ángulos iguales y es común a ellos (a los dos triángulos), entonces tendrán también los lados restantes iguales a los lados restantes [I, 26]; por tanto, ΔE es igual a ΔZ. Por lo mismo, ΔH es igual a ΔZ. Luego las tres rectas ΔE, ΔZ, ΔH son iguales entre sí; por tanto, el círculo descrito con el centro Δ y como distancia una de las (rectas) (Δ)E, (Δ)Z, (Δ)H [97] pasará también por los puntos restantes y tocará las rectas AB, BΓ, ΓA por ser rectos los ángulos correspondientes a los puntos E, Z, H. Pues, si las cortara, la (recta) trazada perpendicular al diámetro de un círculo en un extremo caería dentro del círculo; lo cual se ha demostrado que es absurdo [III, 16]; por tanto, el círculo descrito con el centro Δ y como distancia una de las (rectas) (Δ)E, (Δ)Z, (Δ)H no cortará las rectas AB, BΓ, ΓA; luego las tocará y el círculo estará inscrito en el triángulo ABΓ [III, Def. 5]. Inscríbase como ZHE.

Por consiguiente, se ha inscrito el círculo EZH [97bis] en el triángulo dado ABΓ. Q. E. F.

[97] En el texto: ... *kaì diastémati henì tôn* E, Z, H, «... y como distancia una de las E, Z, H». No hay más remedio que incluir la letra Δ para saber a lo que se refiere Euclides. Se utiliza frecuentemente la misma expresión a lo largo del libro IV; como en otros casos, pongo entre paréntesis lo que no aparece en el texto griego.

[97bis] A pesar de la diferencia de orden de las letras en la denominación (ZHE, EZH), se trata del mismo círculo.

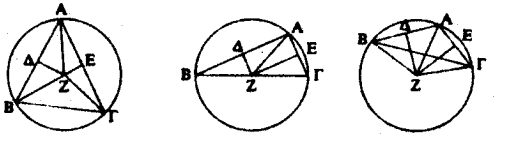
Proposición 5

Circunscribir un círculo en torno a un triángulo dado.

Sea ABΓ el triángulo dado.

Así pues, hay que circunscribir un círculo en torno al triángulo dado ABΓ.

Divídanse en dos partes iguales las rectas AB, AΓ por los puntos Δ, E [I, 10], y a partir de los puntos Δ, E trácense ΔZ, EZ formando ángulos rectos con AB, AΓ; entonces coincidirán o bien dentro del triángulo ABΓ o sobre la recta BΓ o fuera de BΓ.



Coincidan en primer lugar dentro en el punto Z, y trácense ZB, ZΓ, ZA. Y como AΔ es igual a ΔB, y ΔZ es común y (forma) ángulos rectos, entonces la base AZ es igual a la base ZB [I, 4]. De manera semejante demostraríamos que ΓZ es también igual a AZ, de modo que ZB es también igual a ZΓ; por tanto, las tres (rectas) ZA, ZB, ZΓ son iguales entre sí. Luego el círculo descrito con el centro Z y como distancia una de las (rectas) (Z)A, (Z)B, (Z)Γ pasará también por los puntos restantes, y el círculo estará circunscrito en torno al triángulo ABΓ. Circunscríbase como ABΓ.

Ahora bien, coincidan ΔZ, EZ sobre la recta BΓ en el punto Z como está en el segundo dibujo, y trácese AZ.

De manera semejante demostraríamos que el punto Z es el centro del círculo circunscrito en torno al triángulo ABΓ.

Ahora bien, coincidan ΔZ, EZ fuera del triángulo ABΓ en el punto Z de nuevo, como está en el tercer dibujo, y trácense AZ, BZ, ΓZ. Y como AΔ es de nuevo igual a ΔB y ΔZ es común y (forma) ángulos rectos, entonces la base AZ es igual a la base BZ [I, 4]. De manera semejante demostraríamos que ΓZ también es igual a AZ, de modo que BZ también es igual a ZΓ; por tanto, el círculo descrito con el centro Z y como distancia una de las (rectas) ZA, ZB, ZΓ pasará también por los puntos restantes, y estará circunscrito en torno al triángulo ABΓ.

Por consiguiente, se ha circunscrito un círculo en torno al triángulo dado. Q. E. F.

(Porisma)

Y queda claro que, cuando el centro del círculo cae dentro del triángulo, el ángulo BAΓ que se encuentra en un segmento mayor que el semicírculo es menor que un recto; cuando el centro cae sobre la recta BΓ, el ángulo BAΓ que se encuentra en el semicírculo es recto; cuando el centro del círculo cae fuera del triángulo, el (ángulo) BAΓ que se encuentra en un segmento menor que el semicírculo es mayor que un recto [III, 31].

[De modo que cuando el ángulo dado es menor que un recto, AZ, EZ caerán dentro del triángulo, cuando es recto, sobre BΓ, cuando es mayor que un recto, fuera de BΓ. Q. E. F.] [98].

[98] Heiberg atetiza este último párrafo. Simson ya había observado: «Está manifiestamente viciado el texto griego en el corolario, en que se hace mención del ángulo dado, siendo así que en la proposición nada hay ni puede haber del ángulo dado» *(Los seis primeros libros, y el undécimo, y duodécimo de los Elementos de Euclides,* ed. cit., notas, pág. 307. Simson, por otra parte, también había reparado en un defecto de la demostración de la proposición IV, 5: la suposición no probada de que ΔZ, EZ habrán de coincidir (vid. *op. cit.*, pág. 98).

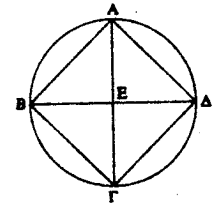
Proposición 6

Inscribir un cuadrado en un círculo dado.

Sea ABΓΔ el círculo dado.

Así pues, hay que inscribir un cuadrado en el círculo ABΓΔ.

Trácense dos diámetros AΓ, BΔ del círculo ABΓΔ formando ángulos rectos entre sí, y trácense AB, BΓ, ΓΔ, ΔA.



Y como BE es igual a EΔ: porque E es el centro, y EA es común y (forma) ángulos rectos, entonces la base AB es igual a la base AΔ [I, 4]. Por lo mismo, cada una de las (rectas) BΓ, ΓΔ es igual a cada una de las (rectas) AB, AΔ; por tanto, el cuadrilátero ABΓΔ es equilátero.

Digo además que también es rectangular.

Pues como la recta BΔ es un diámetro del círculo ABΓΔ, entonces BAΔ es un semicírculo; por tanto, el ángulo BAΔ es recto [III, 31]. Por lo mismo, cada uno de los (ángulos) ABΓ, BΓΔ, ΓΔA es también recto; por tanto, el cuadrilátero ABΓΔ es rectangular. Pero se ha demostrado que también es equilátero. Luego es un cuadrado [I, Def. 22] y está inscrito en el círculo ABΓΔ.

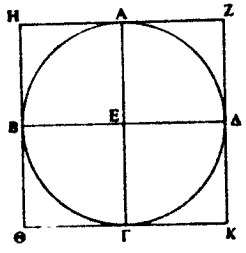
Por consiguiente, se ha inscrito el cuadrado ABΓΔ en el círculo dado. Q. E. F.

Proposición 7

Circunscribir un cuadrado en torno a un círculo dado.

Sea ΑΒΓΔ el círculo dado.

Así pues, hay que circunscribir un cuadrado en torno al círculo ΑΒΓΔ.



Trácense dos diámetros ΑΓ, ΒΔ del círculo ΑΒΓΔ formando ángulos rectos entre sí, y a través de los puntos Α, Β, Γ, Δ trácense ΖΗ, ΗΘ, ΘΚ, ΚΖ tangentes al círculo ΑΒΓΔ [III, 16, Por.].

Así pues, como ΖΗ toca el círculo ΑΒΓΔ, y ΕΑ se ha trazado desde el centro Ε hasta el punto de contacto Α, entonces los ángulos correspondientes a Α son rectos [III, 18]. Por lo mismo, los ángulos correspondientes a los puntos Β, Γ, Δ son rectos.

Y como el ángulo ΑΕΒ es recto, y el (ángulo) ΕΒΗ es también recto, ΗΘ es paralela a ΑΓ [I, 28]. Por lo mismo, ΑΓ es también paralela a ΖΚ. De modo que ΗΘ es también paralela a ΖΚ [I, 30]. De manera semejante demostraríamos que cada una de las (rectas) ΗΖ, ΘΚ es paralela a ΒΕΔ. Por tanto, ΗΚ, ΗΓ, ΑΚ, ΖΒ, ΒΚ son paralelogramos; luego ΗΖ es igual a ΘΚ y ΗΘ a ΖΚ [I, 34]. Ahora bien, como ΑΓ es igual a ΒΔ, mientras que ΑΓ es también igual a cada una de las (rectas) ΗΘ, ΖΚ y ΒΔ es igual a cada una de las (rectas) ΗΖ, ΘΚ [I, 34], entonces el cuadrilátero ΖΗΘΚ es equilátero.

Digo, además, que también es rectangular.

Pues como ΗΒΕΑ es un paralelogramo, y el (ángulo) ΑΕΒ es recto, entonces el (ángulo) ΑΗΒ es también recto [I, 34]. De manera semejante demostraríamos que los ángulos correspondientes a Θ, Κ, Ζ son rectos. Por tanto, ΖΗΘΚ es rectangular. Pero se ha demostrado que también es equilátero; luego es un cuadrado. Y está circunscrito en torno al círculo ΑΒΓΔ.

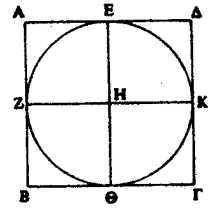
Por consiguiente, se ha circunscrito un cuadrado en torno al círculo dado. Q. E. F.

Proposición 8

Inscribir un círculo en un cuadrado dado.

Sea ΑΒΓΔ el cuadrado dado.

Así pues, hay que inscribir un círculo en el cuadrado ΑΒΓΔ.



Divídase en dos partes iguales cada una de las (rectas) ΑΔ, ΑΒ por los puntos Ε, Ζ, respectivamente [I, 10], y a través de Ε trácese ΕΘ paralela a una de las dos (rectas) ΑΒ, ΓΔ, y a través de Ζ trácese ΖΚ paralela a una de las dos (rectas) ΑΔ, ΒΓ [I, 31]; entonces cada una de las (figuras) ΑΚ, ΚΒ, ΑΘ, ΘΔ, ΑΗ, ΗΓ, ΒΗ, ΗΔ es un paralelogramo, y los lados opuestos evidentemente son iguales [I, 34]. Y como ΑΔ es igual a ΑΒ, y ΑΕ es la mitad de ΑΔ, mientras que ΑΖ es la mitad de ΑΒ, entonces ΑΕ es igual a ΑΖ; de modo que los lados opuestos también (son iguales); por tanto, ΖΗ es también igual a ΗΕ. De manera semejante demostraríamos que cada una de las (rectas) ΗΘ, ΗΚ es igual a cada una de las (rectas) ΖΗ, ΗΕ; por tanto, las cuatro (rectas) ΗΕ, ΗΖ, ΗΘ, ΗΚ son iguales entre sí. Luego el

círculo descrito con el centro H y como distancia una de las (rectas) (H)E, (H)Z, (H)Θ, (H)K pasará también por los puntos restantes; y tocará las rectas AB, BΓ, ΓΔ, ΔA por ser rectos los ángulos correspondientes a E, Z, Θ, K; porque si el círculo cortara las (rectas) AB, BΓ, ΓΔ, ΔA, la (recta) trazada formando ángulos rectos con el diámetro del círculo desde un extremo caería dentro del círculo; lo que se ha demostrado (que es) absurdo [III, 16]. Por tanto, el círculo trazado con el centro H y como distancia una de las (rectas) (H)E, (H)Z, (H)Θ, (H)K no corta las rectas AB, BΓ, ΓΔ, ΔA. Luego las tocará y estará inscrito en el cuadrado ABΓΔ.

Por consiguiente, se ha inscrito un círculo en el cuadrado dado. Q. E. F.

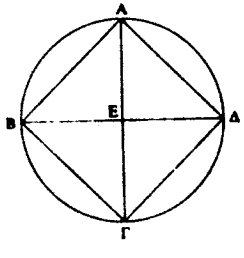
Proposición 9

Circunscribir un círculo en torno a un cuadrado dado.

Sea ABΓΔ el cuadrado dado.

Así pues, hay que circunscribir un círculo en torno al cuadrado ABΓΔ.

Pues una vez trazadas AΓ, BΔ, córtense entre sí en E.



Y como ΔA es igual a AB y AΓ es común los dos (lados) ΔA, AΓ son iguales a los dos (lados) BA, AΓ; y la base ΔΓ es igual a la base BΓ; por tanto, el ángulo ΔAΓ es igual al ángulo BAΓ [I, 8]; luego el ángulo ΔAB ha sido dividido en dos partes iguales por AΓ. De manera semejante demostraríamos que cada uno de los (ángulos) ABΓ, BΓΔ, ΓΔA ha sido dividido en dos partes iguales por las rectas

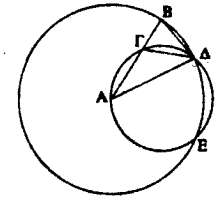
AΓ, ΔB. Y como el ángulo ΔAB es igual al (ángulo) ABΓ, y el (ángulo) EAB es la mitad del (ángulo) ΔAB, y el (ángulo) EBA es la mitad del (ángulo) ABΓ, entonces el (ángulo) EAB es igual al (ángulo) EBA; de modo que el lado EA es también igual al (lado) EB [I, 6]. De manera semejante demostraríamos que las (rectas) EA, EB son iguales respectivamente a las rectas EΓ, EΔ. Por tanto, las cuatro (rectas) EA, EB, EΓ, EΔ son iguales entre sí. Luego el círculo descrito con el centro E y como distancia una de las (rectas) (E)A, (E)B, (E)Γ, (E)Δ pasará también por los puntos restantes y estará circunscrito en torno al cuadrado ABΓΔ. Circunscríbase como ABΓΔ.

Por consiguiente, se ha circunscrito un círculo en torno al cuadrado dado. Q. E. F.

Proposición 10

Construir un triángulo isósceles cada uno de cuyos ángulos de la base sea el doble del restante.

Tómese una recta AB y córtese por el punto Γ de modo que el rectángulo comprendido por AB, BΓ sea igual al cuadrado de ΓA [II, 11]; y con el centro A y la distancia AB descríbase el círculo BΔE. Y adáptese al círculo BΔE la recta BΔ igual a la recta AΓ que no es mayor que el diámetro del círculo BΔE [IV, 1]; y trácense AΔ, ΔΓ y circunscríbase en torno al triángulo AΓΔ el círculo AΓΔ [IV, 5].



Ahora bien, como el (rectángulo comprendido) por AB, BΓ es igual al (cuadrado) de AΓ, y AΓ es igual a BΔ, entonces el (rectángulo comprendido) por AB, BΓ es igual al (cuadrado) de BΔ. Y como se ha tomado el punto B exterior al círculo AΓΔ, y desde B hasta el círculo AΓΔ han caído dos rectas BA, BΔ y una de ellas (lo) corta mientras que la otra cae (en él), y el (rectángulo comprendido) por AB, BΓ es igual al (cuadrado) de BΔ, entonces BΔ toca el círculo AΓΔ [III, 37]. Así pues, como BΔ lo toca, y ΔΓ ha sido trazada desde el punto de contacto Δ, entonces el ángulo BΔΓ es igual al ángulo ΔAΓ en el segmento alterno del círculo [III, 32]. Así pues, como el (ángulo) BΔΓ es igual al (ángulo) ΔAΓ, añádase a ambos el (ángulo) ΓΔA; entonces el (ángulo) entero BΔA es igual a los dos (ángulos) ΓΔA, ΔAΓ. Pero el (ángulo) externo BΓΔ es igual a los (ángulos) ΓΔA, ΔAΓ [I, 32]; luego el (ángulo) BΔA es también igual al (ángulo) BΓΔ. Pero el (ángulo) BΔA es igual al (ángulo) ΓBA, puesto que el lado AΔ es también igual al (lado) AB [I, 5]; de modo que el (ángulo) ΔBA es igual al (ángulo) BΓΔ. Por tanto, los tres (ángulos) BΔA, ΔBA, BΓΔ son iguales entre sí. Y como el ángulo ΔBΓ es igual al (ángulo) BΓΔ, el lado BΔ es también igual al lado ΔΓ [I, 6]. Pero BΔ es por hipótesis igual a ΓA; entonces ΓA es también igual a ΓΔ; de modo que el ángulo ΓΔA es también igual al ángulo ΔAΓ [I, 5]: por tanto, los (ángulos) ΓΔA, ΔAΓ son el doble del (ángulo) ΔAΓ. Pero el (ángulo) BΓΔ es igual a los (ángulos) ΓΔA, ΔAΓ; luego el (ángulo) BΓΔ es también el doble del (ángulo) ΓAΔ. Pero el (ángulo) BΓΔ es igual a cada uno de los (ángulos) BΔA, ΔBA; por tanto, cada uno de los (ángulos) BΔA, ΔBA es también el doble del (ángulo) ΔAB.

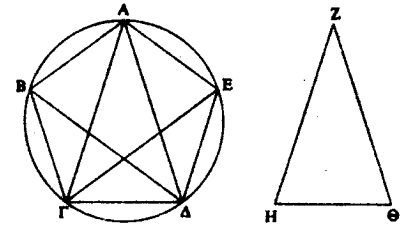
Por consiguiente, se ha construido un triángulo isósceles ABΔ cada uno de cuyos ángulos de la base ΔB es el doble del ángulo restante. Q. E. F. [99].

Proposición 11

Inscribir un pentágono equilátero y equiángulo en un círculo dado.

Sea ABΓΔE el círculo dado.

Así pues, hay que inscribir un pentágono equilátero y equiángulo en el círculo ABΓΔE.



[99] Existen razones para conjeturar que la construcción del triángulo de esta proposición y su relación con el pentágono regular fue un descubrimiento de los pitagóricos. Apoyan esta conjetura:

a. El escolio IV, núm. 2 (Stamatis, V, pág. 204) dice: «este libro es un descubrimiento de los pitagóricos».

b. En el resumen histórico de Eudemo recogido en Proclo, dice que Pitágoras descubrió «la construcción de las figuras cósmicas» (debemos entender los cinco sólidos regulares).

c. Jámblico *(Perí toû Pithagorikoû bíou,* cap. 18, s. 88) se hace eco de la leyenda en la que se atribuye a Hipaso la publicación de este descubrimiento secreto y su posterior muerte por impiedad. Para más detalles, *vid.* Heath, ed. cit., 2, págs. 97-100.

Tómese el triángulo isósceles ZHΘ cada uno de cuyos ángulos correspondientes a H, Θ sea el doble del correspondiente a Z [IV, 10], e inscríbase en el círculo ABΓΔE el triángulo AΓΔ de ángulos iguales a los de ZHΘ, de modo que el ángulo ΓAΔ sea igual al ángulo correspondiente a Z y los (ángulos) correspondientes a H, Θ sean iguales respectivamente a los (ángulos) AΓΔ, ΓΔA [IV, 2]; por tanto, cada uno de los (ángulos) AΓΔ, ΓΔA es también el doble de ΓAΔ. Entonces divídase en dos partes iguales cada uno de los (ángulos) AΓΔ, ΓΔA con las (rectas) ΓE, ΔB, respectivamente, y trácense AB, BΓ, ΔE, EA.

Así pues, como cada uno de los ángulos AΓΔ, ΓΔA es el doble de ΓAΔ, y ha sido dividido en dos partes iguales mediante las rectas ΓE, ΔB, entonces los cinco ángulos ΔAΓ, AΓE, EΓΔ, ΓΔB, BΔA son iguales entre sí. Pero los ángulos iguales están sobre circunferencias iguales [III, 26]; por tanto, las cinco circunferencias AB, BΓ, ΓΔ, ΔE, EA son iguales entre sí. Pero a las circunferencias iguales las subtienden rectas iguales [III, 29]; por tanto, las cinco rectas AB, BΓ, ΓΔ, ΔE, EA son iguales entre sí; luego el pentágono ABΓΔE es equilátero.

Digo además que es también equiángulo.

Pues como la circunferencia AB es igual a la circunferencia ΔE, añádase a ambas BΓΔ; entonces la circunferencia entera ABΓΔ es igual a la circunferencia entera EΔΓB. Ahora bien, el ángulo AEΔ está sobre la circunferencia ABΓΔ, y el ángulo BAE sobre la circunferencia EΔΓB; por tanto, el ángulo BAE es igual al (ángulo) AEΔ [III, 27]. Por lo mismo, cada uno de los ángulos ABΓ, BΓΔ, ΓΔE es igual a cada uno de los (ángulos) BAE, AEΔ; luego el pentágono ABΓΔE es equiángulo. Pero se ha demostrado que también es equilátero.

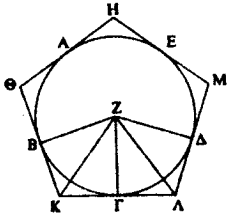
Por consiguiente, se ha inscrito un pentágono equilátero y equiángulo en el círculo dado. Q. E. F.

Proposición 12

Circunscribir un pentágono equilátero y equiángulo en torno a un círculo dado.

Sea ABΓΔE el círculo dado.

Así pues, hay que circunscribir un pentágono equilátero y equiángulo en torno al círculo ABΓΔE.



Considérense A, B, Γ, Δ, E los puntos de los ángulos del pentágono inscrito, de modo que las circunferencias AB, BΓ, ΓΔ, ΔE, EA sean iguales [IV, 11]; y a través de los puntos A, B, Γ, Δ, E trácense HΘ, ΘK, KΛ, ΛM, MH tangentes al círculo [III, 16, Por.], y tómese el centro Z del círculo ABΓΔE [III, 1], y trácense ZB, ZK, ZΓ, ZΛ, ZΔ.

Y como la recta KΛ toca el (círculo) ABΓΔE en Γ, y ZΓ ha sido trazada desde el centro Z hasta el punto de contacto Γ, entonces ZΓ es perpendicular a KΛ [III, 18]; por tanto, cada uno de los ángulos correspondientes a Γ es recto. Por lo mismo, los ángulos correspondientes a los puntos B, Δ son también rectos. Y como el ángulo ZΓK es también recto, entonces el (cuadrado) de ZK es igual a los (cuadrados) de ZΓ, ΓK [I, 47]. Por lo mismo, el (cuadrado) de ZK es igual a los (cuadrados) de ZB, BK; de modo que los (cuadrados) de ZΓ, ΓK son iguales a los (cuadrados) de ZB, BK, de los cuales el (cuadrado) de ZΓ es igual al (cuadrado) de ZB; por tanto, el (cuadrado) restante de ΓK es

igual al (cuadrado) de BK. Luego BK es igual a ΓK. Y como ZB es igual a ZΓ, y ZK es común, entonces los dos (lados) BZ, ZK son iguales a los dos (lados) ΓZ, ZK; y la base BK es igual a la base ΓK; entonces el (ángulo) BZK es igual al (ángulo) KZΓ [I, 8]; y el (ángulo) BKZ al (ángulo) ZKΓ; por tanto, el (ángulo) BZΓ es el doble del (ángulo) KZΓ y el (ángulo) BKΓ (el doble) del (ángulo) ZKΓ. Por lo mismo, el (ángulo) ΓZΔ es también el doble del (ángulo) ΓZΛ y el (ángulo) ΔΛΓ (el doble) del (ángulo) ZΛΓ.

Ahora bien, como la circunferencia BΓ es igual a ΓΔ, el ángulo BZΓ es también igual al (ángulo) ΓZΔ [III, 27]. Y el (ángulo) BZΓ es el doble del (ángulo) KZΓ, y el (ángulo) ΔZΓ (el doble) del (ángulo) ΛZΓ; entonces el (ángulo) KZΓ es también igual al (ángulo) ΛZΓ; pero el ángulo ZΓK es también igual al (ángulo) ZΓΛ. Entonces ZKΓ, ZΛΓ son dos triángulos que tienen dos ángulos iguales a dos ángulos y un lado igual a un lado, a saber: el común a ambos ZΓ; por tanto, también tendrán los lados restantes iguales a los lados restantes y el ángulo restante igual al ángulo restante [I, 26]; así pues, la recta KΓ es igual a la (recta) ΓΛ y el ángulo ZKΓ al (ángulo) ZΛΓ. Y como KΓ es igual a ΓΛ, entonces KΛ es el doble de KΓ. Por lo mismo, se demostraría que ΘK es también el doble de BK. Y BK es igual a KΓ; entonces ΘK también es igual a KΛ. De manera semejante se demostraría que cada una de las (rectas) ΘH, HM, MΛ también es igual a cada una de las (rectas) ΘK, KΛ; por tanto, el pentágono HΘKΛM es equilátero.

Digo, además, que también es equiángulo.

Pues como el ángulo ZKΓ es igual al (ángulo) ZΛΓ y se ha demostrado que el (ángulo) ΘKΛ es el doble del (ángulo) ZKΓ, y el (ángulo) KΛM el doble del (ángulo) ZΛΓ, entonces el (ángulo) ΘKΛ es igual al (ángulo) KΛM. De manera semejante se demostraría que cada uno de los (ángulos) KΘH, ΘHM, HMΛ es también igual a cada uno de los (ángulos) ΘKΛ, KΛM; por tanto, los cinco ángulos HΘK, ΘKΛ, KΛM, ΛMH, MHΘ son iguales entre sí. Luego el pentágono HΘKΛM es equiángulo. Pero se ha demostrado que también es equilátero y está circunscrito en torno al círculo ABΓΔE.

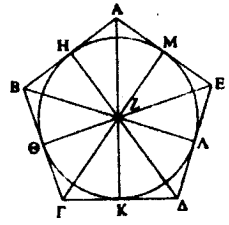
[Por consiguiente, se ha circunscrito un pentágono equilátero y equiángulo en torno al círculo dado.] Q. E. F. [100].

Proposición 13

Inscribir un círculo en un pentágono dado que es equilátero y equiángulo.

Sea ABΓΔE el pentágono dado equilátero y equiángulo.

Así pues, hay que inscribir un círculo en el pentágono ABΓΔE.



Divídanse, pues, en dos partes iguales los ángulos BΓΔ, ΓΔE con las rectas ΓZ, ΔZ, respectivamente; y a partir del punto Z, donde se encuentran las rectas ΓZ, ΔZ entre sí, trácense las rectas ZB, ZA, ZE. Y como BΓ es igual a ΓΔ, y ΓZ es común, los dos (lados) BΓ, ΓZ son iguales a los dos (lados) ΔΓ, ΓZ; y el ángulo BΓZ es igual al ángulo ΔΓZ; entonces la base BZ es igual a la base ΔZ, y el triángulo BΓZ es igual al triángulo ΔΓZ, y los ángulos restantes, subtendidos por los lados iguales, serán iguales respectivamente [I, 4]; por tanto, el ángulo ΓBZ es igual

[100] La conclusión de esta proposición se omite en los manuscritos.

al (ángulo) ΓΔZ. Y como el (ángulo) ΓΔE es el doble del (ángulo) ΓΔZ, y el (ángulo) ΓΔE es igual al (ángulo) ABΓ, mientras que el (ángulo) ΓΔZ es (igual) al (ángulo) ΓBZ, entonces el (ángulo) ΓBA es también el doble del (ángulo) ΓBZ; por tanto, el ángulo ABZ es igual al (ángulo) ZBΓ; luego el ángulo ABΓ ha sido dividido en dos partes iguales por la recta BZ. De manera semejante se demostraría que los ángulos BAE, AEΔ han sido divididos en dos partes iguales respectivamente por las rectas ZA, ZE. Ahora bien, trácense a partir del punto Z las (rectas) ZH, ZΘ, ZK, ZΛ, ZM perpendiculares a las rectas AB, BΓ, ΓΔ, ΔE, EA. Y como el ángulo ΘΓZ es igual al (ángulo) KΓZ, y el ángulo recto ZΘΓ es igual al ZKΓ, entonces ZΘΓ, ZKΓ son dos triángulos que tienen dos ángulos iguales a dos ángulos y un lado igual a un lado, a saber: ZΓ común a ambos que subtiende a uno de los ángulos iguales; por tanto, tendrá los lados restantes iguales a los lados restantes [I, 26]; luego la perpendicular ZΘ es igual a la perpendicular ZK. De manera semejante demostraríamos que cada una de las (rectas) ZΛ, ZM, ZH es igual a cada una de las (rectas) ZΘ, ZK; por tanto, las cinco rectas ZH, ZΘ, ZK, ZΛ, ZM son iguales entre sí. Luego el círculo descrito con el centro Z y como distancia una de las (rectas) (Z)H, (Z)Θ, (Z)K, (Z)Λ, (Z)M, pasará también por los demás puntos y tocará las rectas AB, BΓ, ΓΔ, ΔE, EA por ser rectos los ángulos correspondientes a los puntos H, Θ, K, Λ, M. Pues si no las tocara sino que las cortara, resultaría que la (recta) trazada formando ángulos rectos con el diámetro de un círculo desde un extremo cae dentro del círculo; lo cual se ha demostrado que es absurdo [III, 16]. Por tanto, el círculo descrito con el centro Z y como distancia una de las (rectas) (Z)H, (Z)Θ, (Z)K, (Z)Λ, (Z)M no corta las rectas AB, BΓ, ΓΔ, ΔE, EA; luego las tocará. Descríbase como HΘKΛM.

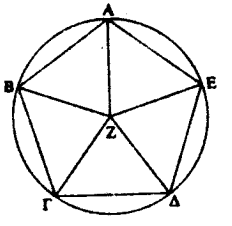
Por consiguiente, se ha inscrito un círculo en un pentágono dado que es equilátero y equiángulo. Q. E. F.

Proposición 14

Circunscribir un círculo en torno a un pentágono dado que es equilátero y equiángulo.

Sea ABΓΔE el pentágono dado que es equilátero y equiángulo.

Así pues, hay que circunscribir un círculo en torno al pentágono ABΓΔE.



Divídanse, pues, en dos partes iguales los ángulos BΓΔ, ΓΔE con las rectas ΓZ, ΔZ, respectivamente. Y desde el punto Z donde se encuentran las rectas, hasta los puntos B, A, E trácense las rectas ZB, ZA, ZE. De manera semejante al caso anterior se demostraría que los ángulos ΓBA, BAE, AEΔ han sido divididos en dos partes iguales por las rectas ZB, ZA, ZE, respectivamente. Ahora bien, como el ángulo BΓΔ es igual al (ángulo) ΓΔE, y el (ángulo) ZΓΔ es la mitad del (ángulo) BΓΔ, mientras que el (ángulo) ΓΔZ es la mitad del (ángulo) ΓΔE, entonces el (ángulo) ZΓΔ es también igual al ángulo ZΔΓ; de modo que el lado ZΓ es igual al lado ZΔ [I, 6]. De manera semejante se demostraría que cada una de las (rectas) ZB, ZA, ZE es igual a cada una de las (rectas) ZΓ, ZΔ; por tanto, las cinco rectas ZA, ZB, ZΓ, ZΔ, ZE son iguales entre sí. Luego el círculo descrito con el centro Z y como distancia una de las rectas ZA,

/B, ZG, ZD, ZE pasará también por los puntos restantes y estará circunscrito. Circunscríbase y sea ABΓΔE.

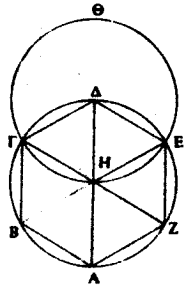
Por consiguiente, ha sido circunscrito un círculo en torno a un pentágono dado que es equilátero y equiángulo. Q. E. F.

Proposición 15

Inscribir un hexágono equilátero y equiángulo en un círculo dado.

Sea ABΓΔEZ el círculo dado.

Así pues, hay que inscribir un hexágono equilátero y equiángulo en el círculo ABΓΔEZ.



Trácese el diámetro AΔ del círculo ABΓΔEZ, y tómese el centro H del círculo, y con el centro Δ y la distancia ΔH descríbase el círculo EHΓΘ, y una vez trazadas EH, ΓH llévense hasta los puntos B, Z, y trácense AB, BΓ, ΓΔ, ΔE, EZ, ZA.

Digo que el hexágono ABΓΔEZ es equilátero y equiángulo.

Pues como el punto H es el centro del círculo ABΓΔEZ, HE es igual a HΔ. Como el punto Δ es a su vez el centro del círculo HΓΘ, ΔE es igual a ΔH. Pero se ha demostrado que HE es igual a HΔ; por tanto, HE es también igual a EΔ; luego el triángulo EHΔ es equilátero; entonces sus tres ángulos EHΔ, HΔE, ΔEH son iguales entre sí, porque los ángulos de la base de los triángulos isósceles son iguales entre sí [I, 5]; y los tres ángulos de un triángulo son iguales a dos rectos [I, 32]; por tanto, el ángulo EHΔ es la tercera parte de dos rectos. De manera semejante se demostraría que también el ángulo ΔHΓ es la tercera parte de dos rectos. Y como la recta ΓH levantada sobre EB hace los ángulos adyacentes EHΓ, ΓHB iguales a dos rectos, entonces el (ángulo) restante ΓHB es también la tercera parte de dos rectos; por tanto, los ángulos EHΔ, ΔHΓ, ΓHB son iguales entre sí; de modo que los ángulos BHA, AHZ, ZHE correspondientes a sus vértices son iguales [I, 15]. Luego los seis ángulos EHΔ, ΔHΓ, ΓHB, BHA, AHZ, ZHE son iguales entre sí. Pero los ángulos iguales están sobre circunferencias iguales [III, 26]; por tanto, las seis circunferencias AB, BΓ, ΓΔ, ΔE, EZ, ZA son iguales entre sí. Pero a las circunferencias iguales las subtienden rectas iguales [III, 29]. Por tanto, las seis rectas son iguales entre sí; luego el hexágono ABΓΔEZ es equilátero.

Digo, además, que es también equiángulo.

Pues como la circunferencia ZA es igual a la circunferencia EΔ añádase a ambos la circunferencia ABΓΔ; entonces la (circunferencia) entera ZABΓΔ es igual a la (circunferencia) entera EΔΓBA; y el ángulo ZEΔ está sobre la circunferencia ZABΓΔ, y el ángulo AZE sobre la circunferencia EΔΓBA; por tanto, el ángulo AZE es igual al (ángulo) ΔEZ [III, 27]. De manera semejante se demostraría que también los ángulos restantes del hexágono ABΓΔEZ son cada uno igual a uno de los ángulos AZE, ZEΔ; por tanto, el hexágono ABΓΔEZ es equiángulo. Pero se ha demostrado que es también equilátero; y ha sido inscrito en el círculo ABΓΔEZ.

Por consiguiente, se ha inscrito un hexágono equilátero y equiángulo en el círculo dado. Q. E. F. [101].

---

[101] Heiberg considera que Proclo se refiere a éste cuando alude a un porisma que sigue a un problema. El texto de Proclo dice: «el porisma

Porisma:

A partir de esto queda claro que el lado del hexágono es igual al radio del círculo.

De manera semejante al caso del pentágono, si trazamos tangentes al círculo a través de los puntos de división del círculo, se circunscribirá en torno al círculo un hexágono equilátero y equiángulo conforme a lo dicho sobre el pentágono; y además, por procedimientos semejantes a los expuestos en el caso del pentágono inscribiríamos y circunscribiríamos un círculo en un hexágono dado.

Proposición 16

Inscribir un pentadecágono equilátero y equiángulo en un círculo dado.

Sea ABΓΔ el círculo dado.

Así pues, hay que inscribir un pentadecágono equilátero y equiángulo en el círculo ABΓΔ.

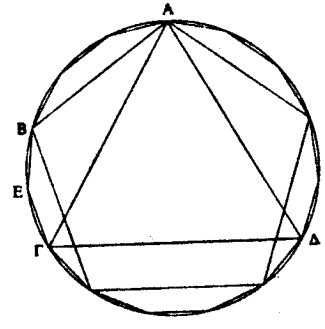
Inscríbase en el círculo ABΓΔ el lado AΓ del triángulo equilátero inscrito en él y el (lado) AB del pentágono equilátero; entonces, así como de los segmentos iguales hay 15 en el círculo ABΓΔ, así también habrá cinco en la circunferencia ABΓ que es la tercera parte del círculo, y habrá tres en la circunferencia AB que es la quinta parte del círculo; por tanto, en la (circunferencia) restante BΓ habrá dos. Divídase en dos partes iguales BΓ por el (punto) E

que se encuentra en el libro segundo *(tò dè en deutérōi biblíōi...)* es un porisma a un problema». Pero esto no podría aplicarse al único porisma que aparece en el libro II, el siguiente a II, 4. Para salvar esta referencia, Heiberg piensa que en lugar de *tò deutérōi biblíōi* debería leerse *tò d' biblíōi*, es decir, el presente libro IV.

[III, 30], entonces cada una de las circunferencias BE, EΓ es la quinceava parte del círculo ABΓΔ.

Por consiguiente, si después de trazar BE, EΓ, adaptamos al círculo ABΓΔ sucesivamente rectas iguales (a ellas), se habrá inscrito en él un pentadecágono equilátero y equiángulo. Q. E. F.

De manera semejante al caso del pentágono, si trazamos tangentes al círculo por los puntos de división del círculo, se circunscribiría en torno al círculo un pentadecágono equilátero y equiángulo. Además, mediante pruebas semejantes a las del caso del pentágono circunscribiríamos e inscribiríamos un círculo en el pentadecágono dado. Q. E. F. [102].



[102] Aquí, como en III, 10, usa el término *kyklon* («círculo») en el sentido inusual de «circunferencia de un círculo», en vez de atenerse a la definición 15 de I «figura plana comprendida por una línea...». Por otra parte, con esta proposición, Euclides termina la exposición de un procedimiento para construir polígonos regulares con 3, 4, 5, 6 y 15 lados. Ahora, dado cualquier polígono regular, podemos construir otro polígono regular con doble número de lados. Siguiendo este proceso, primero se describe un círculo en torno al polígono dado y luego se dividen en dos partes iguales todos los arcos subtendidos por los lados del polígono. Este método permite construir polígonos regulares con 3.2^n, 4.2^n, 5.2^n, 15.2^n lados, donde n es 0, o cualquier entero positivo.

ÍNDICE GENERAL

Págs.

INTRODUCCIÓN GENERAL 7

I. *«Euclides», un hombre para la geometría.* 7

1. Oscuro autor, incierto personaje 9
2. Obras 14
3. Los *Elementos* 17

II. *Los «Elementos», una escritura que hizo historia* 19

1. La tradición de los tratados de *Elementos.* 20
2. El lugar del tratado de Euclides en la tradición de *Elementos* 27
3. La institucionalización de los *Elementos.* 39

III. *La constitución de los «Elementos»* ... 48

1. El «pórtico axiomático» 48
2. Teoría elemental de la geometría plana (Libros I-IV) 65
3. La teoría generalizada de la proporción (Libros V-VI) 73

Págs.

4. Teoría de la aritmética (Libros VII-IX). 81
5. El libro X: «la cruz de los matemáticos». 88
6. Geometría del espacio (Libros XI-XIII). 96

IV. *Sobre la significación axiomática de los «Elementos»* 102

V. *El texto de los «Elementos». Versiones y ediciones* 123

1. El Euclides griego 125
2. El Euclides árabe 129
3. El Euclides latino 133
4. El Euclides de las lenguas vernáculas. 147

VI. *Los «Elementos» en España* 151

Una selección bibliográfica 182

ELEMENTOS

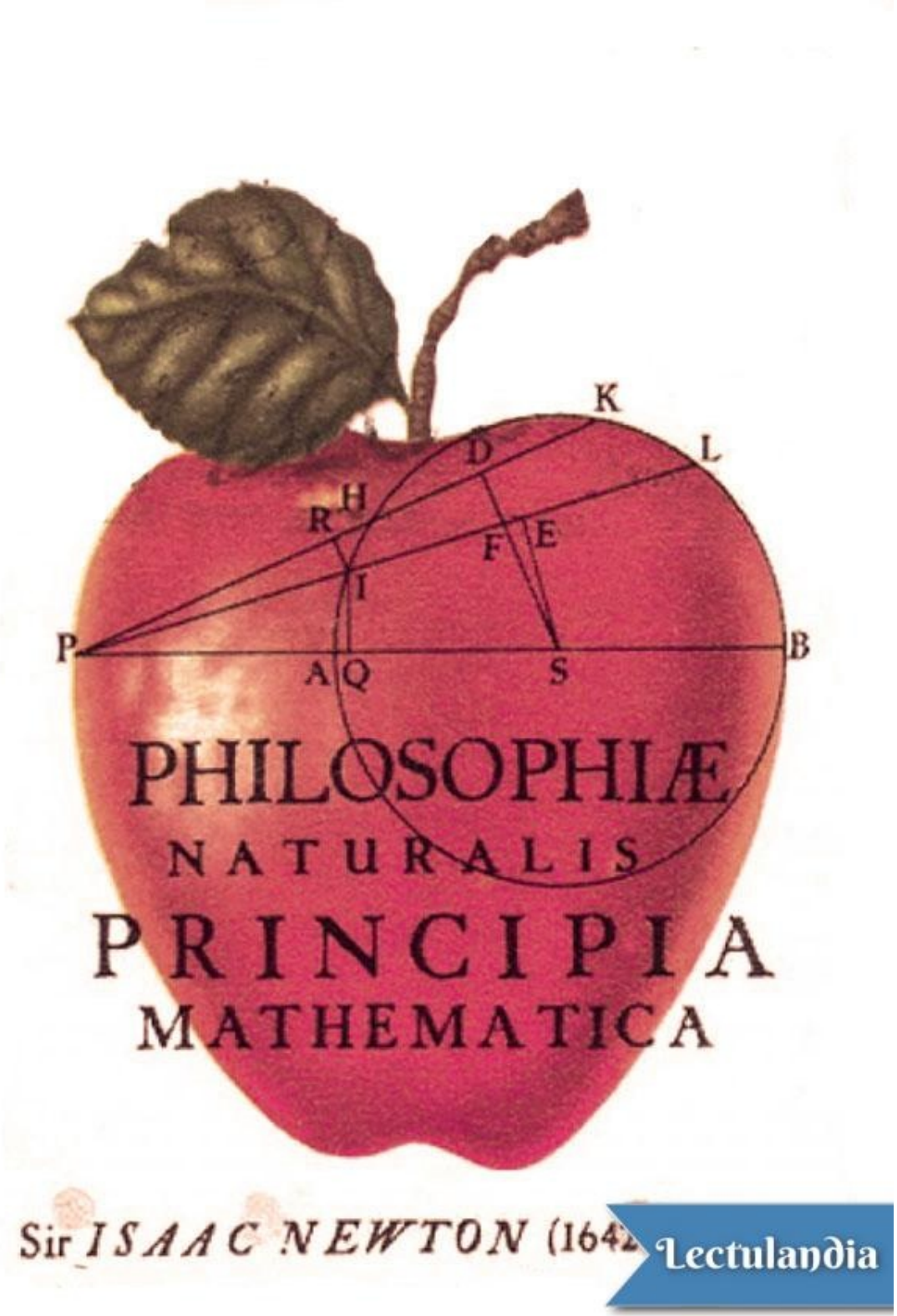
NOTA SOBRE LA PRESENTE TRADUCCIÓN 187

LIBRO PRIMERO 189

LIBRO SEGUNDO 265

LIBRO TERCERO 291

LIBRO CUARTO 341



Publicados en Londres en 1687, los principios matemáticos de la filosofía natural son uno de esos libros que todo el mundo cita pero muy pocos han leído; pues si el puesto que ocupa en la historia del pensamiento es tan principal como acreditado, su lectura presenta serias dificultades debidas a la complejidad propia de alguno de sus teoremas, junto a la sujeción deliberada del autor a las reglas del método geométrico en su demostración. Como es bien sabido, Newton resuelve aquí el teorema de los movimientos planetarios a la vez que los une a lo terrestres mediante una misma dinámica y una ley universal de gravitación; discute y explica fenómenos como el del movimiento de los cometas o las mareas; sienta las bases de la hidrostática, la hidrodinámica y la acústica; demuestra la imposibilidad de la hipótesis cartesiana de los vórtices; descubre, define por primera vez de modo no contradictorio y da reglas prácticas para la derivación e integración de funciones; y sistematiza un modo de estudio de la Naturaleza (a la que deben hacerse preguntas explícitas y cuantitativas mediante los experimentos) y de exposición de los conocimientos adquiridos mediante métodos matemáticos: lo que desde él se conoce propiamente como física.

Isaac Newton (1642-1727), recogiendo las aportaciones de Kepler y Galileo, consigue por vez primera construir un modelo matemático general que permite explicar tanto el movimiento de los cuerpos celestes como el de los terrestres.

Lectulandia

Sir Isaac Newton

Principios matemáticos de la filosofía natural (Principia)

ePub r1.0
casc 06.02.16

Título original: *Philosophiæ naturalis principia mathematica*
Sir Isaac Newton, 1687
Traducción: Eloy Rada
Retoque de cubierta: casc

Editor digital: casc
ePub base r1.2

más libros en lectulandia.com

I

Los mejores conocedores de la obra y de la figura de Newton suelen experimentar ante el personaje una extraña ambigüedad de sentimientos, desde la admiración más cálida a la repulsión casi inevitable. Las biografías más y mejor documentadas[1] dan también buena base para ello, por cuanto, hasta donde permite llegar la documentación, no parece el suyo un perfil amable y candoroso. Por el contrario, junto a la innegable categoría de genio, parece que se escondía tras la inescrutable máscara de *sir* Isaac Newton, un espíritu lleno de complejidades, cuya interpretación no ha dejado de tentar a psicoanalistas, biógrafos, estudiosos de humanos recovecos y hasta a pintores de retratos[2]. La única «descripción» que nos falta es la de un dramaturgo que hubiese teatralizado alguno de los aspectos más apasionantes de nuestro personaje, aunque quizá no se hubiera requerido para ello menos que un Shakespeare. A falta de ello y sin pretender suplir alguna de las biografías citadas más arriba, habrá que decir algo de su vida y, sobre todo, de su obra.

Nacido el día de Navidad de 1642 en Woolsthorpe, una granja de cierto rango, con unos cien acres de tierras de su pertenencia, y un par de meses después de la muerte de su padre, nadie podría haber predicho que aquel sietemesino que «cabía en una jarra de cuarto (de galón)» como, según decía él, había comentado su madre cuando nació, habría de ser un día *sir* Isaac Newton, *aeques auratus*, presidente de la Sociedad Real de Londres, etc. Las condiciones de su nacimiento eran casi de infortunio. Un padre recién muerto, un hijo sietemesino, débil, con una madre, Hanna Ayscoügh, perteneciente a una familia local ahora en declive económico apoyada por algunos parientes clérigos, pero sin medios ni para restaurar la vieja casa de piedra de la granja. Las dificultades de una época de guerra civil no constituían buenos presagios para el recién nacido. Pero, con todo, sobrevivió y con apenas tres años cumplidos vio a su madre partir, casada en segundas nupcias, hacia la casa rectoral de North Witham, cuyo viejo y hacendado titular, Barnabas Smith, llegó a convencer a Hanna mediante unas capitulaciones que incluían cierta dote a Isaac más el arreglo del edificio de la vieja granja. Los biógrafos registran este hecho como un momento crucial en la idiosincrasia de Newton. Rompe la relación del niño con su madre y origina encontrados sentimientos en su espíritu que, dicen, tendrán enormes

consecuencias. Una que tuvo importancia es que garantizó a Newton una cierta seguridad económica, incluso para los días de sus estudios en Cambridge. Las demás pertenecen a la intimidad de Newton y, que sepamos, apenas hay indicios de ellas; tan celoso fue de sus sentimientos. Newton quedó encomendado a los cuidados de su abuela materna y tampoco de ella dejó, que se sepa, ninguna semblanza, nada que permita adivinar sus recuerdos de esa temprana infancia.

Pero en agosto de 1653 el reverendo Barnabas falleció y la madre de Newton regresa a la granja de Woolsthorpe acompañada de un hijo de un año y de dos hijas algo mayores, tres hermanastros de Newton cuya presencia, al lado de la madre recuperada, no sabemos hasta dónde pudo afectar al pequeño Isaac.

Mientras tanto seguía acudiendo a una escuela local donde aprendió a leer y a escribir y muy poco más. Desde el regreso de su madre a casa hasta la primavera de 1655 todo son conjeturas, imágenes idílicas con su madre reencontrada, sus idas y vueltas a la escuela y acaso algún problema de celos con sus hermanastros Smith. Disponemos para estos detalles de dos «memoramdas» de Stukeley y de Conduitt[3], el segundo procedente en gran medida del primero, que fueron redactados con posterioridad a la muerte de Newton y que, por ello, hubieron de valerse de testimonios de terceros y de una memoria muy lejana e imprecisa para recordar anécdotas de un niño, entonces algo especial, pero no un prodigio todavía.

En marzo o abril de 1655 es enviado a la *Free Grammar School of King Edward VI* de [*Grantham*], donde quizá vio por vez primera al filósofo Henry More, con quien sin duda volverá a encontrarse en la Universidad de Cambridge. Grantham dista unas pocas millas de Woolsthorpe y Newton siguió en contacto frecuente con su familia, si bien se hallaba alojado en casa del boticario *mr*. Clark, cuya esposa, amiga de la madre de Newton, tenía de un primer matrimonio tres hijos de apellido Storer, uno de los cuales, con el tiempo, proporcionará a Newton (Libro III, Proposición XLI…) algunas observaciones desde Nueva Inglaterra sobre el cometa de 1680, mientras el otro hermano será su encargado en Woolsthorpe, y la tercera, «*miss* Storey» (en el «memoramdum» de Stukeley), será la encargada de darnos la única nota humana que conozcamos de Newton. Esta niña, después casada con *mr*. Vincent, dio a Stukeley alguna información de aquellos remotos días de 1655-1659 en que Newton estudiaba en Grantham. Sus hermanos mantendrán con Newton alguna relación personal (Arthur Storer como informador desde Nueva Inglaterra, Edward como administrador de Woolsthorpe). Pero ella fue la amiga de Newton en aquellos tiempos y posiblemente la única mujer de su vida. Newton sintió por ella un especial afecto, sin que sepamos mucho más, y quizá en aquellos años de semiadolescente, un romántico enamoramiento. Ella ya en la ancianidad lo recordaba bien como «un muchacho serio, silencioso, pensativo…» (*sober, silent, thinking lad*). Pero los comienzos no debieron ser fáciles. Es probable que la preparación de la escuela rural no fuera suficiente o quizá fuese otra la razón, mas el hecho es que Newton se encontró entre los últimos de la clase. Y ocurrió que un día camino de la escuela, el

alumno que estaba delante de el le propinó un golpe en el estómago. Y, aunque sólo jugaba con las niñas, en este caso sabía bien qué hacer. Esperó toda la larga jornada de clases y cuando salieron desafió a su contrincante. En el patio de la Iglesia pelearon, y aunque el rival era mayor que él, Newton tenía más genio y mayor rabia y obligó al contrario a pedir basta[4]. Después le agarró por las orejas y le metió la cabeza contra la valla de la Iglesia. No le bastó esto y a los pocos meses había humillado a todos colocándose en el primer puesto de la clase. Más tarde, en 1662, hizo una lista de sus «pecados» y en ella recuerda «pegar a Arthur Storer» («Beating Arthur Storer»).

Sus estudios en Grantham, bajo el magisterio de Henry Stokes, se centraron en el aprendizaje del latín, algo de griego y lecturas de los clásicos y de la Biblia. Por su cuenta leía y anotaba otras cosas, procedentes, quizá, de la pequeña biblioteca del boticario Clark[5] y dedicaba mucho de su tiempo a construir aparatos mecánicos de juguete poniendo en ello grandes dosis de observación y de habilidad mecánica. Al cabo de cuatro años y medio de estancia en Grantham, su madre le reclama en Woolsthorpe para que se vaya haciendo cargo de la granja y le asigna la compañía y asistencia de un criado. Newton fue un pésimo alumno agrícola. Su desinterés por la granja y por cuanto en ella ocurría acabó por convencer a su madre de que Isaac debía seguir por otros caminos. Se decidió su regreso a Grantham con el fin de preparar su ingreso en la Universidad de Cambridge. Quizá Hanna se vio ayudada en esta decisión por Stokes y por sus parientes clérigos, e inicialmente la decisión albergara la esperanza de dirigir ál joven Isaac hacia la carrera eclesiástica. De hecho la Iglesia ofrecía unas posibilidades mucho más seguras y atrayentes que la dura vida campesina. Y por otro lado la familia materna tenía relaciones en ese mundo. Sea como fuese, Newton regresó a Grantham durante otros nueve meses para completar su preparación. Stokes recibió a Newton con alegría y Stukeley nos asegura que le presentó ante los demás con lágrimas en los ojos en un discurso laudatorio, con el cual, al despedirle camino de la Universidad, le proponía como modelo a los demás escolares. Por lo visto, algunos de ellos también lloraban.

Y a buen seguro que Newton no dejaba entre ellos un mal recuerdo. A juzgar por las anécdotas que Stukeley llegó a escuchar de labios de sus informantes, tanto los Clark como los Storer y tal vez otros, debió encantar a los buenos habitantes de Grantham con su ingenio mecánico como constructor de cometas, relojes de agua, y otros varios juguetes mecánicos de general admiración. Pero quizá es más destacable el hecho de que en ellos puso a contribución su enorme capacidad de observación y de análisis del movimiento para hacerlos de tal modo que en ellos se diesen óptimos aprovechamientos de las fuerzas en juego[6]. Suele verse en este detalle una especie de premonición de su futura aplicación a la filosofía natural. Lo cierto es que sus habilidades mecánico-manuales le fueron de gran utilidad cuando hubo de preparar su instrumental de observación. Otro tanto cabe decir de sus habilidades como dibujante. Las paredes de su habitación en la casa de los Clark, parece ser que estaban llenas de

sus dibujos, entre ellos retratos del Rey, de Stokes, etc[7].

Por lo demás, los estudios de la escuela de Grantham debieron preparar a Newton para un dominio del latín y de los clásicos, sin ninguna referencia a problemas de filosofía ni matemáticas, salvo acaso algunas nociones elementales. A cambio proporcionaron a Newton, con la ayuda de las relaciones de su madre principalmente a través de su familia y de la amistad con la señora Clark, algunas conexiones en Cambridge, en especial la de Humphrey Babington, Fellow del Trinity y hermano de la señora Clark, cuya influencia benéfica, no sabemos si también los buenos oficios de Stokes, sirvió a Newton para cruzar por primera vez la Gran Puerta del Trinity College, el día 5 de junio de 1661. Aunque no era absolutamente pobre de fortuna (recuérdese que la dotación de Barnabas Smith llegaba a una renta de 50 libras por año), su situación de ingreso en el Trinity es de «subsizar», estudiante pobre que ha de ganar su sustento como sirviente en la mesa de los Fellows y estudiantes ricos. Los estatutos del Trinity daban cumplimiento con ello al precepto evangélico de socorrer a los pobres, para lo cual les dan el nombre de «Sizatores», esto es, «scolares pauperes». Y quizá Newton, heredero del señorío de Woolsthorpe y acostumbrado a tener criados, no encontró demasiado duro el oficio de «Sizar», puesto que tal vez estuvo desde el principio destinado al servicio de Babington como cuidador de los intereses del mismo en Cambridge, donde sólo venía a residir unas semanas por año. Si la situación social no era brillante, la comodidad y la autonomía debieron serlo, y los servicios a Babington a la postre resultaron beneficiosos para Newton.

Tampoco ahora tuvo amigos ni intimidad con nadie. Seguía siendo un estudiante pensativo y solitario. Al cabo de siete u ocho meses se encontró con John Wickins, que andaba pasando mal momento por causa de su compañero. Llegaron a entenderse y desde entonces compartieron la habitación, aunque esto no bastó para generar entre ellos una amistad profunda. Tal vez no pasó de ser una relación de buenos compañeros. Y no más amigos.

Pudiera pensarse que Newton fue un estudiante de los que hoy se suelen calificar de «expediente brillantísimo» o de «número uno» y que sus estudios tuvieron lugar en la Universidad más avanzada de aquel entonces. Nada de eso se corresponde con lo que sabemos de uno y de otra. La Universidad de Cambridge, como las demás de Inglaterra entonces, era un centro de poder real eclesiástico y el Trinity uno de los santuarios más conspicuos por su tradición, su importancia dentro del conjunto de los Colleges y por el espíritu puritano y cuasimonacal de sus estatutos, de sus estudios y de sus Fellows. Sus alumnos eran, por lo general, candidatos a ocupar puestos eclesiásticos, y si eran nobles, puestos políticos o en el aparato del estado. La enseñanza, en este contexto, era típicamente escolástica, con algunos toques renacentistas y basada en las antiguas disciplinas medievales que comprendían el bachillerato en Artes, o el magisterio en Artes o acaso en Teología. Gramática, retórica, dialéctica, un poco de física o de astronomía, algunas matemáticas o geometría, música y lecturas de los clásicos que empezaban a ser recomendadas por

los tutores con alguna profusión, en particular Plinio, Cicerón, Estrabón, etc. Al final, unos ejercicios de argumentación sobre la base silogística de Aristóteles y ello era suficiente para que un estudiante optase al bachillerato en Artes. Conocemos los textos que hubo de manejar Newton y nos quedan sus cuadernos de notas de estudio. Por ellos sabemos que Newton sintió poco interés por los programas que le proponía la Universidad y que en el momento en que aparecía ante su vista algo nuevo, abandonaba sus notas sobre los textos para internarse sólo por los nuevos horizontes. Esto ocurrió, de modo relevante, con la astronomía y con las matemáticas[8]. Sus progresos fueron siempre solitarios, nunca guiados por un maestro, quizá con la única excepción de Barrow en un momento, y siempre sin la posibilidad de discutirlos privadamente con colegas. Y si es cierto que a su llegada a Cambridge poseía aceptables conocimientos de latín, su preparación en las disciplinas relacionadas con la filosofía natural, matemáticas, astronomía, física, geometría, etc. era mucho menos que elemental. Sus «Notebooks» permiten afirmar que no concluyó la lectura de sus libros de texto y que muy pronto el análisis crítico de los mismos le empujó hacia la lectura de autores extracurriculares, como, por ejemplo, Descartes, Gassendi, Galileo (de quien parece que leyó el «Diálogo», aunque no los «Discorsi»), Boyle, Hobbes, K. Digby, J. Glanville, H. More, etc. En 1664 abre una entrada en su Notebook con el título de «Questiones Quaedam Philosophcae» (sic)[9] que en gran parte parece responder a sus lecturas privadas y no curriculares. El título va apostillado con una frase que expresa, al menos, su voluntad de investigar por su cuenta: «Amicus Plato, amicus Aristóteles magis amica veritas». Mas no hay que pensar que Newton deja de lado a Platón y a Aristóteles para adoptar sin más la filosofía cartesiana. Aunque ésta le sugiere más y mejores métodos de análisis, sus notas de la época muestran que leía planteando preguntas a los autores, proponiendo contraargumentos o contraejemplos, haciendo precisiones conceptuales u observacionales, etc. Los temas cartesianos que le llaman la atención desde el principio son la luz, el movimiento mecánico, la materia, el espacio, el sistema solar de vórtices y la geometría[10]. Cuando en la primavera de 1665 obtiene el grado de bachiller en Artes tiene también perfilado un programa personal de intereses científicos centrado en estos temas. La filosofía natural que descubre en 1664 no era demasiado ortodoxa en Cambridge, pero Newton era, a su vez, un tanto original dentro desella. Cuando hubo de ser examinado por Isaac Barrow sobre (Euclides) lo halló en blanco, y en cambio no le preguntó nada de la geometría de Descartes que dominaba ya bastante bien. Ello estuvo a punto de costar le la no elección para residente del Trinity, pero sus patrones, quizá otra vez Babington, le echaron una mano y su carrera quedó inicialmente garantizada, y lo que es más importante, se inicia su conocimiento y relación intelectual con Barrow.

Aproximadamente en los días en que se produjo su elección como escolar residente en el Trinity (abril de 1664), se inician sus notas y apuntes conservados de sus estudios matemáticos, siempre en solitario, relativos a la (geometría de Descartes) y al análisis moderno. Su guía inicial fue Ta segunda edición latina de la «Geometría»

de Descartes que había preparado Schooten, complementada con notas y textos de álgebra, sobre todo de Viéta. A la vez toma contacto con el cálculo infinitesimal; de Wallis. Y ya en 1664 logra un gran dominio de ambas disciplinas y también vislumbra algunas estrategias para hacer de ambas una síntesis, pues llega a percibir con claridad que la ecuación de una curva no es una mera representación de la misma sino que entraña la propia naturaleza de la curva, de tal modo que el estudió de la ecuación se convierte así en estudio *general* de una curva, de sus propiedades y de sus límites. Desde la primavera de 1664 hasta el otoño-invierno de 1666 en que redacta un tratado resumen de sus hallazgos, logra un dominio completo de los conocimientos matemáticos de la época y descubre además su propio método de fluxiones y también los teoremas que dan pie a dicho cálculo[11].

Pero estos logros matemáticos no son independientes por completo de sus estudios de mecánica. Más bien hay que pensar que, hasta cierto punto, están relacionados con ellos. Primero porque el estudio de Descartes le pone simultáneamente en relación con ambos campos, y en segundo lugar porque Newton conecta internamente unos problemas con otros. El análisis de curvas es un análisis de puntos en movimiento y análisis del movimiento es a su vez una reconsideración geométrica del mismo, de sus magnitudes espaciales y de las relaciones y proporciones matemáticas inherentes a esas curvas, rectas, áreas, ángulos, etc. El hecho de que ambas investigaciones fuesen paralelas y solapadas en el tiempo, más que una significación cronológica, tiene un valor epistemológico. Y un dato más lo puede confirmar: Newton en el análisis y construcción de curvas, de familias de curvas y de las propiedades de esas curvas (en particular las tónicas) utilizó como guía un cierto instrumental mecánico, artefactos compuestos de reglas y cuadrantes con hilos que determinaban un punto móvil, el cual, al moverse sobre un plano según le permitían las longitudes variables del hilo, describía curvas más o menos complejas. El movimiento de un punto aparece así vinculado a unos parámetros de naturaleza geométrica, y en definitiva matemática, que permiten establecer ecuaciones con incógnitas a despejar. Si se conoce la naturaleza de esas ecuaciones y el sistema de construcción de las figuras representativas del movimiento, entonces es posible una matemática de los puntos en movimiento. Eso se halla en el origen del método de fluxiones. En sus ensayos de mayo de 1666 lo afirma expresamente y muestra también que el cálculo fluxional no es un mero ejercicio de análisis de curvas[12].

Pero la razón última de esos progresos habría que situarla en el hecho de que, a la sazón, Newton estudia problemas cartesianos de impactos entre cuerpos perfectamente elásticos, de centros de gravedad y, por tanto, de lugares geométricos, de reflexión y refracción, etc. Esos estudios le llevan inevitablemente a plantearse problemas conceptuales, más allá de los puramente métricos, tales como el del concepto de «masa» (cantidad de materia), inercia, velocidad, fuerza… con los cuales ensaya múltiples estrategias. Una de ellas Te lleva a descubrir para los movimientos

circulares la ley del inverso del cuadrado del radio. En tal caso era obvia la comparación con los movimientos celestes de los planetas y de la Luna. Pero en este momento Newton aún no ha descubierto el concepto de «fuerza centrípeta» sino que sus análisis están guiados por la fuerza centrífuga, «endeavor to recede from the center». Y desde esta perspectiva no hay demasiado apoyo para asegurar, como hace la leyenda heroica, que en estos momentos se formuló para sí mismo la ley de la gravedad, ni siquiera como una intuición. Más bien parece que el análisis del movimiento circular en términos de leyes de impacto le llevó a formular el valor de la fuerza con la que se mueve un cuerpo en circulación uniforme por un círculo, en la hipótesis de que ese cuerpo es empujado contra la pared interior del círculo por una fuerza procedente del centro, fuerza que en esas condiciones resulta inversamente proporcional al cuadrado del radio. Esa hipótesis, en el caso de los planetas y de la Luna, es una hipótesis superponible a todas luces con la de los vórtices. Geométricamente admite un tratamiento inverso, el del desvío de la tangente. Newton lo consideró en el caso de la Luna, y la caída de los graves de Galileo le sugirió una analogía. Inicialmente no coincidió el resultado con el previsto. Newton no debió sentir sensación de fracaso por ello, ya que con toda probabilidad su idea de la gravedad era tan limitada por entonces como la de cualquier galileano. En cambio, obtuvo un resultado orientativo: la fuerza de la gravedad en la superficie de la Tierra es unas 4000 veces mayor que la fuerza de la Luna para apartarse del centro de la Tierra, y además, sustituyendo en su fórmula de la fuerza centrífuga los valores de longitudes y tiempos por la tercera Ley de Kepler (los cubos de los radios medios de los planetas son proporcionales a los cuadrados de los tiempos periódicos), las «fuerzas de receso desde el Sol serán inversamente como los cuadrados de las distancias desde el Sol[13]».

Durante buena parte de 1665 y de 1666 Newton vivió en Woolsthorpe como consecuencia de la peste que asoló Inglaterra y que acabó con el incendio de Londres; en marzo de 1666 regresa a Cambridge por breve tiempo y recibe el título de M. A. Para entonces ya ha iniciado la relación con Barrow y ha establecido con él un conocimiento y quizá una amistad que aparentemente duró hasta la muerte del viejo maestro impulsor del proyecto y construcción de la Biblioteca del Trinity. En esta época inician, aparentemente juntos, una serie de estudios matemáticos, ópticos y experimentos alquímicos[14], que Newton proseguirá por su cuenta años más tarde y que junto con los estudios matemáticos debieron servir al eminente Barrow para conocer a fondo las capacidades del «Scholar» que tenía ante él. Ello haría que Newton no encontrase mayores dificultades para ser elegido «minor Fellow» primero y «mayor Fellow» después, del Trinity.

Además de los estudios matemáticos, mecánicos y alquímicos, Newton aborda en estos años sus estudios de óptica. Su idea inicial sobre la naturaleza de la luz es, con certeza, igual de provisional y tentativa que las correspondientes de movimiento, materia y otras, entrelazadas en la filosofía mecánica cartesiana. Por ello no es fácil

sostener que los resultados de su investigación, más o menos contenidos en los ensayos publicados en los primeros años setenta en las «Philosophical Transactions» o incluso más tarde en las Lecciones de *Optica*, representen por completo el cuadro de su filosofía natural. Sus ideas sobre la naturaleza de la luz, sobre la estructura de la materia, sobre la naturaleza de la acción física pueden ser consideradas avanzadas y hasta aventuradas para sus contemporáneos, pero, para él mismo, eran aún indecisas y en buena medida imprecisas. En mi opinión, su larga dedicación a la alquimia está motivada por su deseo de llegar a penetrar en el secreto de la constitución de la materia y de las fuerzas que concurren en ella. Desde este punto de vista, las matemáticas, la mecánica o la óptica no son sino capítulos previos y periféricos para abordar el desafío filosófico que Newton recibió de la filosofía de su tiempo. Por lo demás, es bien conocido que, ante este desafío, Newton optó por responder en términos de «filosofía natural». Opción ésta que tiene tanto de filosófica como de metodológica, pues no es un mero arsenal experimental al que se propone acudir sino también una concepción de la verdad y, por supuesto, una concepción del mundo. Ninguno de estos temas fue objeto de tratados específicos, aunque sí dejó constancia de su opción en pasajes de la *Optica* o de los *Principia*. Igual que no fue explícito en materia filosófica tampoco lo fue con sus ideas tentativas en filosofía natural, y, por ello, los biógrafos y exégetas han de apelar a conjeturas a la hora de reconstruir sus pensamientos. Algo parecido a la reconstrucción de un antiguo edificio del que se conservasen muchos materiales pero no los planos[15].

Los años 1665-1666 suelen ser denominados «anni mirabiles» en el desarrollo intelectual de Newton por los hallazgos que logró en sus soledades de Cambridge y Woolsthorpe. Pero su programa de investigaciones se remontaba a la primavera de 1664 y se continuó hasta 1669. Y quizá lo admirable sea la coherencia y constancia infatigable con que lo prosiguió, avanzando alternativamente en cada frente. Mientras tanto, en su vida externa se producen unos acontecimientos importantes, de la mano ahora de Barrow, tales como su elección como Fellow del Trinity y después la renuncia de Barrow en su favor de la cátedra Lucasiana de matemáticas. Desde el 29 de octubre de 1669 Newton ocupará esa cátedra, con unas 100 libras de renta al año y convirtiéndose de paso en uno de los siete profesores de la Universidad. Barrow fue designado capellán del Rey y, tres años más tarde, maestro del Trinity. Pero para este momento, en que va a empezar su magisterio, el primer programa de investigación ya estaba prácticamente concluido. Durante los diez años siguientes, hasta 1679, se harán públicos algunos de sus resultados y se irán perfeccionando otros, a la vez que aparecerán intereses teológico-históricos de notable importancia en su vida. Lo más notable aquí será reseñar que, de nuevo gracias a las influencias de Barrow, consigue que el Rey descargue al titular de la cátedra Lucasiana de la obligación de tomar órdenes sagradas. Para entonces ya era secretamente arriano y no podía sufrir sin horror la idea de tener que volver a jurar el dogma trinitario. Pero esto es algo ajeno ahora a nuestro objetivo.

Dejando de lado su trabajo de profesor Lucasiano que aparentemente sólo representaba para Newton un quehacer residual, durante estos diez años siguientes a su nombramiento su atención se centró en una reelaboración del programa de investigación completado hasta 1669. Dejando a parte sus investigaciones teológico-históricas muy amplias y variadas y cargadas de pasión religiosa, podemos resumir su actividad en tres áreas de filosofía natural: la óptica, el análisis y la alquimia. En óptica completa su serie de experimentos, publica una parte de sus investigaciones y mantiene una extensa polémica con Hooke, Huygens, Linus, Pardies y Lucas. Ya he señalado que su óptica, incluso la obra completa[16], es un fragmento del *puzzle* total del pensamiento de Newton. Y ni siquiera estos ensayos de los años 70 son más que una parte de sus investigaciones. Por lo tanto, en este capítulo de su trabajo, su pensamiento andaba entrelazado con otras cuestiones, parte alquímicas, parte mecánicas, de las cuales solamente podemos inferir algunos aspectos por las respuestas que va dando o evitando dar a sus críticos.

Y aún eran más desconocidos sus pensamientos alquímicos. Pese a que en este período dedicó mucho esfuerzo y estudio al tema, solamente podemos saber de sus trabajos recurriendo a los manuscritos y apuntes que fue acumulando. Las pocas cosas que salieron de sus manos[17] apenas dan una idea de una mínima parte de sus investigaciones sobre la materia. Pero con todo, hay que concederles gran importancia, si se quiere llegar a comprender la génesis de algunos conceptos presentes en su *Optica* y en los *Principia*. No es fácil decidir si la teoría corpuscular de la luz y su correspondiente atomismo mecánico proceden de sus intuiciones alquímicas o si en sus investigaciones alquímicas busca una confirmación o una extensión de esas hipótesis. En todos los casos Newton da siempre la impresión de tratar de salvar los fenómenos y de indagar a través de ellos sobre algo más. Cuando tropieza con dificultades insalvables, a duras penas consigne disimular su contrariedad y, por lo general, evita decir algo que le comprometa con hipótesis poco sólidas. Pero la autocensura que se impone a la hora de escribir, no se la impone a la hora de investigar a solas en el laboratorio privado que tenía en su habitación del Trinity. Muchos de los conceptos que se manejan en el Libro II de los *Principia* como los de densidad, cohesión, resistencia, viscosidad, etc. han sido ensayados en el laboratorio. Y tratándose de un autodidacta, no es fácil determinar qué tipo de interacciones conceptuales iban ocurriendo en su mente al ensayar y contrastar ideas, hipótesis, experimentos y lecturas tan variadas. Hay que imaginar que su espíritu era una especie de laboratorio mental en el que determinadas experiencias y evidencias determinaban la aparición y prevalencia de unas ideas u otras según respondían mejor o peor a sus intereses científico-filosóficos.

En algunos casos el problema era más llevadero. Cuando aborda una revisión de sus trabajos matemáticos y se propone sistematizarlos, aunque no acabe el proyecto[18], las cosas permiten logros más satisfactorios, y su «método de series y fluxiones» queda suficientemente listo como para ser considerado prácticamente

completo por el momento. Y aquí es preciso destacar que su revisión le lleva a formular con entera precisión la estructura matemática de los problemas de velocidad, espacio, tiempo y movimiento: el método para analizar esos problemas de integración y diferenciación mediante cantidades fluentes (fluxiones). Si se compara este trabajo con el previo de 1668, «Enumeración de curvas de tres dimensiones» y se contemplan a la luz de un proyecto de investigación se puede observar que la estrategia es doble: por una parte, se mejora la estructura algebraica y, por tanto, se gana en generalidad y en rigor y, por otra, se conserva la perspectiva mecánica relacionada con la imagen de puntos en movimiento, incorporando con precisión variables de tiempo, velocidad y espacio. Diríase que las herramientas van estando listas. Desde 1672 a 1676 la polémica sobre la óptica le ocupa un tiempo que siempre lamentó. Y a partir de 1773 el regreso de Barrow como maestro del Trinity le empuja a preparar algunos textos matemáticos y, sobre todo, hacia la alquimia. Una vez conseguida la dispensa de tomar órdenes sagradas, sus trabajos teológicos y alquímicos ocupan su tiempo, junto con sus asuntos personales (muerte de su madre en 1679, arreglo de sus asuntos económicos en Woolsthorpe), de modo que para nuestra historia hemos de llegar hasta el invierno de 1679-1680 que constituye el primer paso directo hacia los *Principia*.

II

La Sociedad Real de Londres había elegido a Newton como miembro de la misma en los primeros años setenta y Newton había mantenido con ella algunas comunicaciones. De manera formal, había participado en sus debates enviando sendos ensayos sobre temas ópticos así como sobre un (telescopio de reflexión en el que aplicaba los principios de su teoría de los colores intentando eliminar los efectos de la aberración cromática. Pero, en un tono más personal y privado, mantuvo una correspondencia científica con Oldenburg, secretario de la Sociedad Real, que hacía de intermediario con Huygens, Leibniz, etc. A la muerte de Oldenburg, y tras algunas intrigas interiores, resultó elegido secretario de la S. R. Hooke, cuya presencia en ella era muy antigua y que había desempeñado el cargo de «curator» de los experimentos. Esto no facilitaba las relaciones de Newton con la institución, ya de por sí bastante deterioradas como consecuencia de la polémica pasada. Y pese a dos cartas, aparentemente conciliatorias[19], Newton se siente agraviado por la incomprensión de Hooke, y también por la pretensión de éste de que en su *Micrographia*[20] se podía hallar buena parte de todos esos fenómenos ópticos perfectamente salvados. La crítica de Hooke que era respetuosa y en alguna medida justa, indispuso de tal modo a Newton que, aparentemente, le decidió a no mantener nuevas relaciones epistolares con él. Pero el nuevo cargo de Hooke anima a éste a tratar de recuperar a Newton para la vida intelectual de la S. R. y aprovecha la nueva situación para convertir la

carta de cortesía comunicando su elección, en una invitación:

«… Por mi parte, estimaré como un gran favor el que tengáis a bien hacerme conocer por carta vuestras objeciones a cualquiera de mis hipótesis u opiniones: en especial si tenéis a bien hacerme saber qué es lo que pensáis de la hipótesis según la cual los movimientos celestes de los planetas se compondrían del movimiento rectilíneo por la tangente y del movimiento de atracción hacia un cuerpo central, así como vuestras objeciones a mi hipótesis sobre las leyes y causas de la elasticidad[21]»….

La primera de las hipótesis de Hooke se halla en su *Attempt to prove the Motion of the Earth by Observation* (Londres, 1674). La segunda en sus *Lectures de potentia resistendi* (Londres, 1678).

En esta misma carta Hooke da a Newton las noticias de «que en París se ocupan de un nuevo trabajo, la determinación de la longitud y latitud de las ciudades más importantes mediante los eclipses de los satélites de Júpiter… y han hallado que Brest está 18 leguas más cerca de París de lo señalado en los mapas… que Collins le ha mostrado un nuevo libro de de la Hire con un nuevo método de las cónicas y un tratado “de locis solidis”… que Flamsteed ha confirmado la paralaje de la órbita terrestre…, etc.» (paralaje que había sido predicha por Hooke).

Newton acababa de regresar de Woolsthorpe donde había estado unos meses con motivo de la última enfermedad y muerte de su madre y arreglando sus asuntos familiares. Con todo, responde inmediatamente (el 28 de noviembre de 1679[22]), excusándose por declinar la invitación a participar en asuntos filosóficos, por estar ahora dedicado a otras cosas, por haber dejado tales problemas de lado durante mucho tiempo y porque, en definitiva, «ha dicho su adiós a la filosofía». Le dice también que no ve cómo participar en algunos proyectos de medición similares a los franceses y, tras comentarle algunas de sus otras noticias, no tiene empacho en decirle que «con seguridad tendréis a bien creerme que antes de recibir vuestra carta yo no he oído hablar (en lo que recuerdo) sobre vuestra hipótesis de la composición de los movimientos celestes de los planetas y sobre el movimiento rectilíneo tangencial a la curva…», a pesar de que acaba de felicitarle por la confirmación de la paralaje predicha por Hooke precisamente en la misma obra donde formula las hipótesis en cuestión. Añade, generosamente, una oferta. Un experimento que serviría para probar experimentalmente el movimiento de rotación de la Tierra: «Supongámosmele *BDG* representa el globo terrestre que gira una vez cada día en torno a su centro C, moviéndose de Oeste a Este, según el orden de las letras *BDG*, que *A* es un cuerpo pesado, suspendido en el aire y que gira con la Tierra de tal modo que se encuentra constantemente sobre el punto *B*. A continuación imaginemos que se deja caer al cuerpo *A* y que su gravedad le comunica otro movimiento hacia el centro de la Tierra, sin que el anterior de Oeste a Este disminuya. Concedido que antes de su caída el cuerpo se hallaba más alejado del centro de la Tierra que el punto del suelo sobre el que está, resultará que el movimiento del cuerpo de Oeste a Este será más importante

que el movimiento de Oeste a Este del punto del suelo sobre el que estaba; por consiguiente, no descenderá según la perpendicular *AC*, sino que, rebasando las partes del suelo, caerá al Este de la perpendicular, describiendo en su caída una espiral *ADEC* contrariamente a la opinión común según la cual, si la Tierra se moviese, los cuerpos pesados en caída libre serían rebasados por la parte del suelo y caerían al Oeste de la perpendicular…». Y añade algunos detalles para que el experimento pueda ser bien ajustado y, por consiguiente, fiable, para acabar despidiéndose con una nueva protesta sobre su desconocimiento de las hipótesis de Hooke por lo que no puede hacerles objeciones.

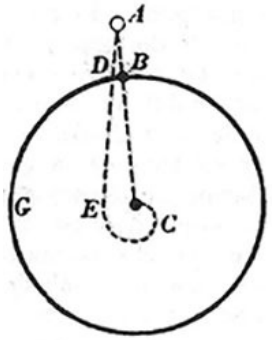


Fig. 1.

Hooke no se creyó casi nada de lo que Newton decía en su carta y, no obstante, persistió en sus intenciones de provocar a Newton para que se pronunciase más explícitamente.

Por otra parte, el problema de la caída libre de los graves ya disponía de una literatura científica notable y disponible para ambos: en primer lugar, y de manera innegable ambos conocían una presentación del problema y de sus protagonistas polémicos a través de un «Account» publicado por Gregory en las *Phil. Transactions*, en 1668[23]. Y de igual modo es de suponer que ambos conocieran los *Cogitata Physicomathematica* de M. Mersenne e incluso la recién publicada obra de Fermat[24].

El 4 de diciembre de 1679 Hooke lee en una reunión de la S. R. las partes no personales de la carta de Newton y en particular se debate el experimento propuesto. El día 14 lee en la reunión de la S. R. su propia respuesta a Newton en la que no trata de la forma de realizar el experimento propuesto, sino del resultado previsto por Newton, resultado que evidencia, para Hooke, una mala concepción del problema que se plantea y, como consecuencia, una solución insatisfactoria. No era necesario hacer las cosas tan explícitas, desde luego, y por ello Hooke se limita a mostrarlo. La carta está escrita con la caligrafía de algún amanuense algo descuidado y únicamente la firma es de mano de Hooke. Comienza ésta con tono amistoso haciendo unas

reflexiones sobre la esperanza que abriga de que Newton vuelva a la filosofía y agradeciendo su carta pasada así como las ideas contenidas en ella. Le dice que fue muy bien recibida por la comunidad de miembros de la S. R. a quienes leyó los pasajes de Mallament y del experimento propuesto «y nada más». Y continúa: «pero en cuanto a la línea curva por la que parece que suponéis que desciende el cuerpo (obsérvese que este problema es un problema nuevo que introduce Hooke aquí poniendo énfasis en lo que no era asunto principal en la carta anterior de Newton), es decir, en cuanto a que sea una especie de espiral que, después de unas cuantas revoluciones lo llevaría al centro de la Tierra, mi teoría del movimiento circular me hace pensar que esta línea sería muy diferente, que no se parecería casi nada a una espiral, sino que parecería más bien una especie de elipsoide. Sobre todo si se supone que el cuerpo en caída libre se halla en el plano equinoccial y que la Tierra se halla cortada en dos hemisferios separados por el espacio de una yarda dejando abierto un vacío para el paso del cuerpo que cae, mientras que la gravitación hacia el antiguo centro permanece como antes y que el globo terrestre se supone en giro diario en tomo a su eje y que el cuerpo en caída conserva impreso el movimiento de las partes superficiales de la Tierra desde las que fue lanzado; en estas condiciones yo imagino la línea por la que se movería este cuerpo parecida a una elipse. Supongamos (Fig. 2), por ejemplo, que *ABDE* representa el plano equinoccial de la Tierra limitado por la superficie, que C es el centro hacia el que tienden todas las líneas de gravitación. Si A representa al cuerpo grave que cae desde *A* atraído hacia C y movido mientras tanto por la revolución diurna de la Tierra desde *A* hacia *BDE*, etc., creo que la curva que entonces resultará descrita por ese cuerpo en caída libre será *AFGH* y que el cuerpo *A* no se aproximará nunca más cerca del centro C que el punto *G* (salvo que sufra la influencia de algún medio como el aire o cosa similar) y que continuará girando por la línea *AFGHAFG*…, etc. Pero si el medio en que se mueve tiene la fuerza de frenar o disminuir su movimiento, la curva que describirá será algo parecido a *AJKLMNOP* y que después de varias revoluciones el movimiento terminará en el centro *C*…».Y añade una corrección más: le hace ver que en Londres, donde se propone el experimento, a 51.º 32′ de latitud Norte la caída no será al Este sino al SSE, dada la inclinación del plano respecto al plano equinoccial[25].

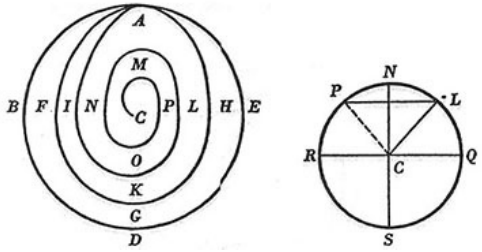


Fig. 2.

La respuesta de Newton es inmediata. El 13 de diciembre contesta y da por cerrada esta correspondencia con Hooke. Escribe Newton: «Estoy de acuerdo con usted en que en nuestra latitud el cuerpo caerá más bien al Sur que al Este siempre que a Altura desde la que caiga sea considerable. Creo también que si el peso se supone uniforme no caerá en espiral hacia el centro con alternativas subidas y bajadas provocadas por la fuerza centrífuga y la gravitación que le arrastran conjuntamente a la vez. Entonces yo imagino la trayectoria del cuerpo no elíptica, sino la curva *AFOGHJKL*, etc.». (Fig. 3).

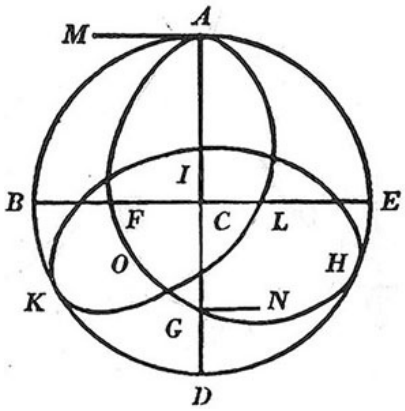


Fig. 3.

Aparece aquí un nuevo error de concepto por parte de Newton. Suponer a la gravedad constante es, por lo menos, no aceptar la ley del inverso del cuadrado. Una cosa es que la gravedad sea algo presente en todas partes y otra que su acción tenga el mismo grado de eficacia en cada punto del espacio desde el centro. Y Hooke lo señalará oportunamente. El 6 de enero de 1680 Hooke escribe a Newton: «vuestro cálculo de la curva que describirá un cuerpo atraído por una fuerza que se ejerce igualmente a todas las distancias del centro, como la de una bola que rueda en un cono cóncavo invertido es exacto, y los dos auges no se unirán más que tras un tercio de revolución; pero yo supongo que la atracción es siempre proporcional al inverso del cuadrado de la distancia al centro y, por consiguiente, la velocidad será en proporción subduplicada de la atracción y recíproca a la distancia».

Se ha destacado de esta carta de Hooke su error matemático[26] y su escasa preparación para desarrollar algebraicamente la cuestión que enuncia. Ello es, sin duda, verdad, como lo es el hecho de que Newton se dio perfecta cuenta de la debilidad de Hooke, pero el lector con perspectiva histórica puede, a su vez, dársela de que Hooke está formulando con mayor precisión que nadie hasta entonces la dinámica de los cielos o del sistema planetario. Y por otra parte, es posible intuir lo lejos que se halla Newton en este momento de su teoría de la gravitación universal.

Di ríase que la hipótesis formulada por Hooke parecería salida más bien de la mente de Newton. Y, sin embargo, no parece que cu este momento Newton hubiese llevado su teoría de la fuerza centrífuga o la «endeavor» más lejos de lo que lo había hecho en los años 66-69. El caso fue que no respondió a Hooke esta vez. Este vuelve a escribir el 17 de enero: «sólo falta ahora conocer las propiedades de una curva, ni circular ni concéntrica, generada por una fuerza de atracción central, que hace que las velocidades de caída respecto a la tangente, es decir, respecto a un movimiento rectilíneo uniforme, sean a todas las distancias inversamente proporcionales a los cuadrados de las distancias. Yo no dudo que gracias a vuestro excelente método, vos podáis establecer fácilmente cómo debe ser tal curva, lo mismo que sus propiedades, y sugerir la causa física de esta proporción».

Los errores matemáticos de Hooke en la formulación de su teoría le han hecho siempre excesivamente vulnerable a la crítica, empezando por la del propio Newton. Pero tanto éste como otros han tenido que ver en estas cartas el primer planteamiento nítido del problema a resolver. En la correspondencia de Newton con Halley en mayo-junio de 1686[27], Newton hace una reseña de esta correspondencia con Hooke y aunque reconoce que la propuesta de Hooke le llevó al descubrimiento del teorema de la elipse, da una imagen inexacta y desfavorable para Hooke de la propuesta de éste. Y lo sospechoso del caso es que Newton conservaba las cartas en su poder, por lo que no valdría argüir que hablaba de memoria.

Aún habría otros detalles para sospechar que Newton no poseía una idea adecuada del sistema del mundo en este momento y menos de la teoría de la gravedad (ni siquiera como fuerza central en los términos de Hooke). El cometa de 1679-1680 fue observado por Newton y la hipótesis que manejó es una hipótesis de trayectoria recta a través del sistema planetario. Flamsteed, Halley, Cassini y el propio Hooke se le adelantan en la sugerencia de que, tanto éste como el de 1682 que habían sido vistos dos veces, en diciembre antes del perihelio y en enero-febrero… después del perihelio eran un solo cometa cada vez y no dos cometas distintos cada vez[28].

No será fácil que los historiadores lleguen a un acuerdo sobre la importancia que tuvo Hooke en la gran realización de Newton. Este fue siempre hostil a la idea de reconocer a Hooke más de lo que era imposible negarle. Hooke siempre defendió como suyo más de lo que realmente había hecho. Y al tratar de reconstruir los hechos siempre nos faltará conocer algo que Newton ocultó: qué fue lo que realmente vio a través de las cartas de Hooke. De hecho cuando encontraba el buen camino avanzaba a grandes pasos. Eso ocurrió con el cálculo y con la óptica en 1664-1669, y quizá eso mismo ocurrió ahora cuando el planteamiento de Hooke le llevó «a descubrir el teorema de la elipse», y quizá algo más que eso, si uno se toma el trabajo de comparar sus concepciones de dinámica celeste anteriores y posteriores a este episodio.

Si Newton dio grandes pasos bajo el estímulo de esta correspondencia o por sí mismo es pura especulación. De hecho, desde este momento entra en una fase de

silencio epistolar, roto unas pocas veces para recabar de Flamnsteed algunos datos sobre el cometa de 1680 y poco más. No obstante compra el opúsculo titulado *Cometa* (otra vez la presencia de Hooke) del cual hace un examen concienzudo en 1681 y así lo registra en su *Waste Book*[29]. Pero a juzgar por los manuscritos de esta época, sus trabajos más importantes estuvieron relacionados con la alquimia, con la reorganización de sus anteriores trabajos matemáticos y con la historia teológico-bíblica. No obstante, también hace acopio de datos tanto sobre el cometa de 1680 como sobre el de 1682 (más larde «cometa Halley») con propósitos no conocidos y tampoco por razones aparentemente astronómicas. Quizá ciertos trabajos alquímicos estuviesen en el fondo de este interés y sólo más tarde, con ocasión de los *Principia*, encontraron una utilidad clara. Por otra parte, como veremos, tanto en la *Optica* como en los *Principia*, la materia y secuelas de estos cuerpos le sugiere ideas harto particulares.

El paso siguiente hacia los *Principia*, quizá el definitivo, fue una pregunta formulada por Halley en una entrevista que mantuvo con Newton probablemente en agosto de 1684. Newton recordaba aquella entrevista, según cuenta De Moivre, con las palabras siguientes: «En 1684 el doctor Halley vino a visitarle a Cambridge, y después de estar juntos algún tiempo el doctor le preguntó cuál creía él que debía ser la curva que debía ser descrita por los planetas, suponiendo la fuerza de atracción hacia el Sol inversamente proporcional al cuadrado de la distancia de ellos al Sol. *Sir* Isaac replicó inmediatamente que debería ser una elipse; el doctor se estremeció de alegría y asombro y le preguntó que cómo lo sabía, porque —respondió—, lo he calculado. Y entonces el doctor Halley le preguntó por sus cálculos inmediatamente; *sir* Isaac miró entre sus papeles y no pudo hallarlos, pero prometió rehacerlos y enviárselos…»[30].

La primera visita de Halley a Newton, probablemente en agosto, aunque otros la sitúan en mayo, parece que obedeció a una especie de decisión emotiva o «corazonada». En enero anterior se habían encontrado Hooke, Wren y el propio Halley en una conversación informal discutiendo sobre la ley del inverso del cuadrado. Los tres estaban convencidos de que esa ley debía ser la clave del arco de los movimientos celestes de los planetas. Y por otra parte tampoco era susceptible de ser establecida experimental u observacionalmente como lo pudieran ser la leyes de Kepler. La única forma de abordar el problema era tratar de demostrar su necesidad matemática. Hooke se comprometió a ello y Wren le ofreció una recompensa «y la parte correspondiente de gloria». Pero pasaron los meses y Hooke no cumplió a satisfacción de Wren. En la correspondencia de Halley con Newton de mayo-junio de 1686, se da cuenta de una conversación entre Wren y Newton de hacia 1677 en la cual Wren manifestó a Newton que había considerado el problema «de determinar los movimientos celestes sobre la base de principios filosóficos», ocasión que sirvió a Newton para constatar que Wren se hallaba convencido de la relación inversa del cuadrado, pero que era incapaz de demostrarla. Cuando Halley pregunta a Wren por

esto (a petición de Newton) contesta que «ha centrado sus pensamientos en establecer los movimientos planetarios sobre la composición de dos movimientos, uno de descenso hacia el Sol y otro impreso… pero ha sido incapaz de resolver el problema» y tampoco Hooke le ha ofrecido una demostración[31]. Este, por su parte, argumentaba que lo tenía resuelto pero guardaba en secreto la solución para probar a otros sabios.

El problema, pues, estaba planteado con toda virulencia en los medios próximos a la S. R. y un astrónomo como Halley tenía verdadera urgencia en poseer una respuesta a una cuestión que era central para sus trabajos. Ya en 1682 había conectado indirectamente con Newton a propósito de las observaciones del cometa de 1680 y de su estancia en París con Cassini observando el cometa de 1682[32]. Quizá conviene añadir que Halley, por su educación y su saber hacer y saber estar era una de las pocas personas indicadas para abordar a Newton y ganar su confianza, cosa que logró de modo permanente hasta el final de la vida de. Newton.

Otra cuestión, un tanto marginal, es cuál fue efectivamente la pregunta que Halley hizo a Newton. En el relato de De Moivre que hemos transcrito antes la pregunta fue «cuál debía ser la curva que describiría un planeta suponiendo la fuerza de atracción hacia el Sol inversamente proporcional al cuadrado de4a distancia hasta él». Pero Newton en otro fragmento[33] enuncia la pregunta de un modo más general y retrotrae la visita de Halley a «la primavera de 1684», mientras en otro lugar del mismo fragmento asegura que las Proposiciones I y XI las conocía desde diciembre de 1679 (fecha de la correspondencia con Hooke) en tanto que las Proposiciones VI, VII, VIII, IX, X, XII, XIII, XVII del Libro I y I, II, III, IV, del Libro II las había elaborado en junio-julio de 1684 (argumento a favor de la visita en mayo de Halley). El resto de la obra fue escrito desde diciembre de 1684 a mayo-junio de 1686. Todos estos recuerdos de Newton introducen elementos no siempre fiables en cuanto a fechas. En junio y quizá julio de 1684 estaba en Woolsthorpe según el registro del Trinity. Todo ello tiene una importancia secundaria para la comprensión del proceso generador de los *Principia*. Aceptando como probable que la visita de Halley se realizó en agosto, o menos probablemente en mayo, de 1684 y que Halley le preguntó por la curva descrita bajo la hipótesis del inverso del cuadrado por los planetas del sistema solar, la cuestión se puede restringir a saber qué es lo que Newton «había calculado».

I. B. Cohen cree que la pregunta pudo haberse formulado más o menos en los términos mencionados, pero, dado que se hallaba inserta en una conversación, podría ser entendida contextualmente por los interlocutores y no ser preciso, por tanto, haberla planteado más adecuadamente[34]. «Newton, dice I. B. Cohen, podría ser capaz de demostrar que si un objeto se mueve a lo largo de una elipse bajo la acción de una fuerza central dirigida hacia un foco, entonces dicha fuerza debe variar como el inverso del cuadrado de la distancia. Pero, tal como lo dice, la pregunta de Halley y la respuesta de Newton implican que éste podría probar la conversa. Esto es

seguramente erróneo, puesto que la conversa no es verdadera. Esto es, bajo la acción de una fuerza inversamente proporcional al cuadrado de la distancia, un objeto no se mueve necesariamente a lo largo de una elipse, sino que su trayectoria puede ser o una línea recta dirigida al centro de fuerza o una curva que puede ser un círculo, una parábola, una elipse e incluso una hipérbola. Es posible que la pregunta atribuida a Halley se haya presentado incorrectamente y que Halley preguntara a Newton realmente por la conversa: cuál habría se ser la fuerza si un planeta se mueve en una elipse…»[35].

Pero, por otra parte, si la explicación contextual es la más natural, dados los interlocutores, no cabe olvidar la posibilidad, recordada por Newton, de que el cálculo mencionado se hallase meramente esbozado (puesto que al rehacerlo «halló dificultades») y, por tanto, que hasta estuviese mal formulado inicialmente. La pregunta de Hooke era por la naturaleza de la curva y no por la fuerza generadora de la curva. Además, si ya lo había calculado, no se explica por qué tardó tres meses en enviarle a Halley unas páginas con el problema resuelto, si se considera que su ingenio resolviendo problemas no tenía igual.

Es preciso dar un rodeo ahora para acercarnos a la cuestión que habíamos planteado sobre los cálculos hechos por Newton. No sabemos, desde luego cuáles fueron los cálculos que realizó con motivo de la correspondencia con Hooke, pero, en cambio, algo sabemos sobre el estado de la cuestión en el momento.

La noción de una expansión de una acción física en razón inversa del cuadrado de la distancia tenía además de una buena apariencia para los astrónomos, dos antecedentes bien conocidos, ambos formulados y aceptados.

El primero de esos antecedentes era la ley de Kepler relativa a la *disminución de la intensidad luminosa* desde el foco de luz, precisamente en razón del inverso del cuadrado de la distancia. Es sabido que Kepler también propuso una atracción solar (de naturaleza no gravitatoria), pero no le atribuyó la misma relación sino que la cifró en un valor proporcional al inverso de la distancia. Los elementos conceptuales de este primer caso pueden sistematizarse un tanto simplemente: la acción física de iluminar, sea cual sea la naturaleza de la luz, tiene un origen central, una expansión prácticamente instantánea (dejando de lado la cuestión del medio) y esférica en torno al foco de luz y la ley de intensidad para cada punto del espacio esférico considerado es una ley de inverso del cuadrado de la distancia al centro, o foco de luz. El segundo de los antecedentes había sido publicado por Huygens en 1673 como apéndice de su *Horologium Oscilatorium* y es la *Regla de la fuerza centrífuga*. Para un movimiento circular uniforme esa *fuerza* es directamente proporcional al cuadrado de la velocidad e inversamente proporcional al cuadrado del radio del círculo[36].

Estos teoremas de Huygens (si bien los publicó sin demostración alguna) permitían calcular el valor de la fuerza planetaria para un movimiento circular uniforme que respondiese a la tercera ley de Kepler, suponiendo una fuerza central que retuviese a los planetas en sus órbitas. Dicho cálculo podría haber sido realizado

por Newton con toda facilidad y también por otros matemáticos, entre los cuales Wren, Halley, quizá Hooke, y otros más. Representemos por V la velocidad de un planeta en su órbita, por *r* el radio de la órbita circular, por T el tiempo periódico, por A la aceleración y por F la fuerza proporcional a ella. En tal caso,

$$A = \frac{V^2}{r} = \frac{(2\pi r/T)^2}{r} = \left(\frac{4\pi^2}{r^2}\right) x \left(\frac{r^3}{T^2}\right);$$

y puesto que $\left(\frac{r^3}{T^2}\right)$ es una constante (la tercera ley de Kepler) y F∞A, entonces F∞ $\frac{1}{r^2}$. Y Newton, en sus cálculos de la «endeavour» de la Luna hizo algo como esto, quizá, antes de 1673, y es cierto que algo parecido fue lo que hizo hacia 1666[37].

III

Y, sin embargo, no sería prudente afirmar que éste era el punto al que habría llegado Newton cuando respondió «lo he calculado». Con gran probabilidad había llegado algo más allá. Quizá había hallado que también en algunos puntos de una elipse se cumplía la ley del inverso, quizá en aquellos que no necesitan otro tratamiento que el del círculo para obtener ese mismo resultado, por ejemplo, en los extremos del diámetro de un círculo inscrito en la elipse y trazado con radio igual al semieje menor y otros casos parecidos. El caso es que, según cuenta De Moivre, «*sir* Isaac, tratando de cumplir su promesa, puso de nuevo manos a la obra, pero no logró llegar a la conclusión que creía haber obtenido antes con un examen cuidadoso. No obstante, ensayó una nueva vía, aunque más larga que la anterior, que le llevó de nuevo a su primitiva conclusión y entonces examinó con cuidado las causas por las que su primer cálculo resultó ser erróneo, y halló que al dibujar una elipse de prisa y a mano, había dibujado los dos ejes de la curva en lugar de dibujar dos diámetros un tanto inclinados uno hacia otro, de modo que posiblemente fijó su imaginación en dos diámetros conjugados, lo cual era un requisito imprescindible, y al darse cuenta, hizo que ambos cálculos coincidieran[38]».

El resultado de esta reconstrucción puesta en limpio para Halley llega a Londres en noviembre de 1684 llevado por Edward Paget, un Fellow del Trinity. Era un pequeño tratado de nueve páginas en la hipótesis de que fuera igual al de la S. R. En él se resuelve el problema planteado y se sientan las bases de una dinámica más general. Así debió de considerarlo Halley, que repite visita a Cambridge y arranca a Newton dos cosas, ambas importantes: permiso para dar noticia de la existencia de este tratado y el compromiso de presentar una copia en la S. R. para que sea

registrada. Todavía vuelve a Londres con tiempo para asistir a la reunión de la S. R. del 10 de diciembre y dar la noticia. La S. R., complacida, acepta y encarga la copia. Esta llega en enero siguiente y es anotada en su Registro (vol. VI, pág. 218). La gestión de Halley implicando a la S. R. significa un compromiso por parte de Newton para publicar su tratado *De motu*. Newton exigió reescribirlo y el resultado de esa reescritura fueron los *Principia*.

Dos cuestiones previas quedan pendientes. La primera es sobre el tránsito intelectual de Newton desde la correspondencia con Hooke hasta el «curioso tratado» que viera Halley y que presuntamente coincidía en gran medida con el conservado en el Registro de la S. R. La segunda, similar a la anterior, está relacionada con la transformación de ese mismo «curioso tratado» en los *Principia*.

No es fácil determinar el camino seguido por Newton desde 1679-1680 hasta noviembre de 1684 en lo referente a la transformación de sus ideas tanto dinámicas como astronómicas. Si tomamos como punto de partida las ideas que aparecen en su correspondencia con Hooke y como término de llegada el tratado enviado a la S. R., parece que hay que destacar una transformación radical de su concepto de «endeavor» («*conatus recedendi a centro*»), de los planetas y la Luna para mantenerse en su órbita hacia el concepto de «fuerza centrípeta». Este paso no está documentado en los manuscritos existentes y quizá estuviese determinado por la misma evolución del tratamiento matemático del problema del cuerpo que gira en un círculo uniformemente. La mera resolución de un problema de caída de un grave (como el propuesto por Hooke) en términos de un cuerpo que gira en torno a un centro bajo la urgencia de la gravedad como freno a su movimiento rectilíneo por la tangente, es una base conceptual suficiente para la transformación del concepto de «endeavor» en «vis centrípeta», siendo en tal caso, uno de los ejemplos matemáticamente equivalentes al de la gravedad. La primera de todas las aproximaciones, sólo explicitada en la correspondencia con Hooke, consiste en concebir la trayectoria de un grave en caída libre como una trayectoria en giro. Y ya desde hacía casi veinte años, había puesto a punto el método para determinar la «curvatura» (el grado de curvatura) de la línea descrita por un punto que se mueve continuamente en función de otros. Quizá la más fiable de las confirmaciones de esta hipotética vía newtoniana de aproximación al problema pueda hallarse en el propio «tratado» de la S. R. En él todas las consideraciones se hacen sin tener presente el concepto de cantidad de materia y sin el concepto de gravitación universal. En las «definiciones» se abordan los conceptos de «fuerza centrípeta», de «fuerza ínsita» y de «resistencia». En las 4 hipótesis se establece el marco: presencia o no de resistencia; la fuerza ínsita y el movimiento uniforme rectilíneo; relación de espacio con fuerzas simultáneas y sucesivas; y la relación entre el espacio y el cuadrado del tiempo, al inicio del movimiento, bajo la urgencia de la fuerza centrípeta[39].

Con esas bases teóricas se está completamente lejos de poder alcanzar el nivel general explicativo de los *Principia*, aunque se está en condiciones de abordar un

problema: el planteado por Hooke y eventualmente por Halley. Si seguimos considerando los teoremas, problemas y corolarios de este inicial tratado, nos encontramos con la solución a la pregunta de Halley en el Problema III, cuyo corolario dice: «Luego los planetas mayores giran en elipses que tienen su foco en el centro del Sol; y con radios trazados al Sol describen áreas proporcionales a los tiempos, exactamente como supuso Kepler…» (*«omnino ut supos-suit Keplerus»*)[40]. Pero es bien sabido que sólo unos meses más tarde, Libro III, Proposición XII, ya sabe que las leyes de Kepler son buenas aproximaciones empíricas pero no son matemáticamente exactas. En mi opinión, y por lo dicho hasta aquí, el primer paso de Newton consistió en hallar la mera solución matemática al problema debatido con Hooke, solución que daba también satisfacción a la pregunta de Halley. En otros términos, ello significa que Newton tomó esta solución matemática como punto de partida para desde ella avanzar hacia una explicación física tal y como su concepción de la filosofía natural lo exigía. Situados en esta perspectiva se puede mirar hacia atrás y comparar el nuevo enfoque con la primitiva visión del problema en 1679-1680 y mirando hacia adelante comprender la reelaboración conceptual que representan los *Principia*. En resumen, Newton ha incorporado el término de «fuerza centrípeta» para designar la noción de «fuerza atractiva» de Hooke y la ha aplicado con éxito al cálculo del cuerpo que gira en un círculo, en una elipse e incluso en otras cónicas. Si Newton era consciente de que la solución matemática del problema de «determinar la trayectoria de un cuerpo en caída libre sobre la Tierra en rotación era matemática y *físicamente* (s. m.) equivalente a hallar una órbita planetaria[41]» es algo más que problemático, incluso a la altura de su tratado de la S. R. Y ello por el hecho de que entre la equivalencia matemática y la física media toda la teoría dinámica que incluye la conceptualización de «masa», de «gravitación» y de fuerzas inerciales, así como la ley de la acción y reacción. El hecho de que en el pequeño tratado de la S. R. introduzca la novedad teórica de la «fuerza centrípeta» igual pero de dirección contraria a la «centrífuga» y, vinculada a ella, llegue a reformular la ley de las áreas de Kepler como propiedad matemática de las curvas y áreas generadas por los puntos en giro bajo esa condición, tanto si se supone que la curva generada es un círculo como otra cónica, ése viene a ser el punto máximo de distanciamiento aparente respecto a la concepción de 1679-1680. Esto no debe hacernos olvidar que Newton había seguido con interés los avances de la astronomía empírica y que en los primeros años 80 observó y recabó de Flamsteed observaciones con ánimo de calcular la trayectoria de los cometas de esos años. Y, con todo, en este tratado no se vislumbra ningún indicio de que la teoría de la gravitación universal ya estuviese perfilada. Por otra parte, quizá conviene ser cautos al respecto, puesto que Newton no era dado a hacer anuncios ni solía vender pieles de oso antes de cazarlos. Ello nos lleva a suponer que, aunque la idea estuviese ya en su mente, incluso con alguna claridad, no le satisfacía lo suficiente como para dar cuenta de ella. De hecho, al abordar el problema de los cometas en el Problema IV de este tratado de la S. R., la

hipótesis no es la de que estén sometidos a atracciones centrípetas, sino la de que los cometas se mueven uniformemente en línea recta, *juxta hipothesim quod cometa moveatur uniformiter in linea recta, determinanda est ejus via rectilínea*[42]. Fuese, pues, la que fuese la concepción general que Newton tenía del sistema del mundo, de las fuerzas y de los movimientos debidos a ellas, lo cierto es que en noviembre de 1684 no presenta más que la solución matemática al problema de Hooke y a la pregunta de Halley, previo cambio de «fuerza central» por «fuerza centrípeta». Y ello nos lleva a ponderar más exactamente la diferencia cualitativa que existe entre este primer intento y el resultado final, tras una larga reelaboración que duró casi año y medio y que no tituló «De Motu», como ha sido titulado después este primer ensayo y como él mismo tituló a otros ensayos posteriores pero previos al final y definitivo; a éste le dio un título seguramente más acorde con su propia idea de lo que realmente había construido: *Philosophiae Naturalis Principia Mathematica*.

Las *memorias* que *sir* Isaac Newton nos legó, directa o indirectamente, proceden todas de sus últimos años y adolecen de fragilidades memorísticas, cosa natural en un hombre que no llevó un diario de sus actividades y relaciones, pero, lo que es peor, adolecen también de una cierta distorsión óptica, no en vano era un maestro en el arte, mediante la cual su propia figura y originalidad quedan aumentadas, subrayadas o incluso mejoradas si hiciese falta. Y aunque es verdad que otros lo necesitan más, no obstante, esos «memoranda» han dado pie a exageraciones, con muy buena base erudita, que, sin embargo, no pueden entenderse en concienzudos investigadores de su vida y obra. Pemberton[43], con «buena» intención, sin duda, nos presenta a Newton empezando a construir sus *Principia* en la época de la gran plaga cuando se retiró a la granja de Woolsthorpe —historia de la manzana que no pasará, a buen seguro, de ser una historieta contada por el anciano Newton a la hora del chocolate—. Pudiera ser que Newton así lo presentara, pero el lector habrá de tener alguna indulgencia con las reconstrucciones de un anciano que siempre se creyó «único». En nuestra opinión, el nacimiento de los *Principia* hay que situarlo en el otoño-invierno de 1684-1685 y vincularlo directamente a la redacción del pequeño ensayo que envió a Halley y a la S. R. Esa redacción puso en marcha el proceso y no se completó hasta que envió el último pliego a Halley que, en nombre y por encargo de la S. R., actuaba como editor, corrector de pruebas, dibujante y quizá consultor del propio Newton[44].

Si al describir la naturaleza y propiedades de una curva, como le propusiera Hooke, o al demostrar para cuerpos que giran en elipses bajo la atracción de una fuerza central que ésta será necesariamente inversamente proporcional al cuadrado de la distancia desde el centro o foco y que cumplen las leyes de Kepler en respuesta a Halley, Newton articuló muy hábilmente, gracias a su «excelente método» para analizar curvas, un conjunto de ideas que con mayor o menor precisión y coherencia circulaban entre los astrónomos y filósofos naturales de la época, en cambio, cuando aborda la redacción de los *Principia* se sitúa en un plano de filósofo natural bien distinto del puramente matemático. Su instrumental matemático era un instrumento

necesario, pero nunca suficiente, para dar cima a su nuevo propósito. El nuevo y más ambicioso propósito es explicar el movimiento de los cuerpos en el mundo desarrollando el procedimiento encontrado para resolver un caso. Para ello no necesita tanto revisar su arsenal matemático, como las bases conceptuales de partida. Fue éste un esfuerzo notable, y en los diferentes borradores que se conservan, previos al manuscrito definitivo[45], todos con el título de «De Motu», se puede observar el proceso de análisis conceptual, de precisión y crítica rigurosa a que se autosometió hasta alcanzar formulaciones que le parecieron suficientes, consistentes y no redundantes tanto de las definiciones como de los Axiomas o leyes del movimiento. Newton no encontró ahora más antecedentes que la ley de la inercia cartesiana y la ley de la caída de los graves de Galileo. Todo lo demás era confusión, ficciones hipotéticas, o, a lo sumo, medias verdades intuidas, formuladas sin rigor y, lo que era peor, sin posibilidad de ser interconectadas sistemáticamente. El lector que se acerque a las variantes textuales recogidas en la edición Cohen-Koyré, todavía percibe el afán de Newton por mejorar la redacción de esas definiciones, por precisarlas más allá de cualquier equívoco. Y, lo que es más, diríase que Newton redactó estas *Definiciones* una última vez cuando ya había dado fin a la redacción del Libro I, puesto que la última frase del Escolio final de las Definiciones dice: «*Hunc enim in finem tractatum sequentem composui*» (para esto, pues, *COMPUSE* el tratado siguiente).

Por otra parte las Definiciones newtonianas fueron rápidamente revisadas y, en particular las relativas a *masa* y *fuerza*, sometidas a depuraciones conceptuales más finas y hasta distintas. En el Escolio con que cierra la Sección XI del Libro I afirma que términos como atracción o impulso tienen una dimensión cuantitativa, «de cantidades y proporciones matemáticas», como ya «he explicado en las definiciones». Y con esa dimensión matemática han de ser tomadas en esas Proposiciones en que habla de la gravedad. Pero añade que, «cuando se desciende a la física, hay que comparar estas razones matemáticas con los fenómenos para que quede claro cuáles condiciones de las fuerzas correspondan a cada clase de cuerpos atractivos». Y más tarde en la Proposición VII del Libro III, en la que enuncia la ley de la gravitación universal, «desciende a la física» precisamente desde la Sección XI del Libro I. Muchos ejemplos como éste podrían ilustrar el mecanismo autorreferencial por el que Newton va perfilando los conceptos básicos que introduce como Definiciones y Leyes. Pero, de paso, permiten entrever, con la ayuda de los borradores que se conservan, el trabajo de construcción y reconstrucción acumulativo con el que dio a luz su gran tratado. De las tres definiciones, fuerza centrípeta, fuerza ínsita y resistencia contenidas en el primer ensayo, a las ocho definiciones de los *Principia*, hay un salto cualitativo tan notable que sólo se explica aceptando la idea de un cambio de proyecto, fuese este cambio instantáneo o sucesivo y consecuencia del desarrollo del proyecto inicial. Y todo parece indicar que fue éste el caso. Quizá centrado en la redacción del primer tratado, llegó a ver la posibilidad de que no sólo Hooke, sino también los cartesianos y hasta el propio Huygens pidieran «basta»,

como antaño su compañero de escuela Storer.

Puesto en una situación semejante, eran muchas las ideas que podían concurrir para dar forma a semejante intento: nada digamos de sus ideas religiosas y de su concepción de Dios y de su presencia en el mundo, asunto éste que no se aclarará por recordar su arrianismo o su «muerte sin auxilios espirituales», ni tampoco por sus frecuentes maldiciones contra Atanasio o contra la «bestia» papal. Hay un trasfondo panteísta, impreciso, un «deísmo» muy peculiar, en su personal religiosidad, que operó sin explicitarse mucho (las cartas a Bentley o la polémica entre Clarke y Leibniz) y es posible y hasta probable que los *Principia* fuesen una pieza importante en esa actitud religioso-mística de Newton. Pero esto es mejor que lo diluciden los teólogos, dada la preternatural dificultad que reviste el problema. La otra fuente de ideas, ésta más constatable aunque no menos volátil, fue la alquimia. Sus denodados esfuerzos por desentrañar el misterio de la materia no fueron un fracaso tan absoluto como suele creerse. Le permitieron hacer explícitamente experimentos de densidades, y con ellos de cantidad de materia y, tras ellos, de movimientos pendulares que ya verá el lector cuán útiles resultaron a la hora de determinar valores gravitatorios y, comparativamente, oscilatorios de los satélites de Júpiter, por ejemplo. La naturaleza de los gases, de los fluidos, de los sólidos, etc. pudo ser incorporada como componente de problemas de mecánica. Lo mismo que sus conocimientos de óptica. Su concepción corpuscular de la materia, además de razones históricas externas, responde a sus trabajos de óptica y alquimia y no cuadraba mal con los «momentos» espaciales y temporales de sus fluxiones. Las fuerzas entre partículas, elasticidad, cohesión, etc., parecen exigir un soporte material para un filósofo mecánico poco dado a considerar a efectos tan tercos como la dureza de las piedras o de algunos metales meras propiedades espaciales o meras «ficciones» de la mente.

Finalmente, el lector hará por su cuenta un suma y sigue con los antecedentes matemáticos ya apuntados más arriba. Hasta qué punto este instrumental fue valioso ya es sabido. Newton mismo incorporó alguno de sus ensayos sobre cónicas, por ejemplo, casi tal cual estaba redactado, siendo éste uno de los ejemplos más claros de «integración» o de síntesis que se pueden ofrecer. Este, sin embargo, no descarta otros menos espectaculares[46], pero igualmente eficaces. Con todo, siempre habrá que recordar que la ensalada también llevaba consigo la pimienta del genio.

IV

Para simplificar un tanto la documentación disponible haremos de ella tres grandes grupos: el primero comprende las copias, las variantes en esas copias y las noticias que han quedado, relacionadas con el pequeño tratado enviado para ser registrado en la S. R. Aquí pueden agruparse, *a*) el enviado a Halley, vía Paget, en noviembre 1684; *b*) el «courious treatise de motu» que viera Halley con ocasión de su

segunda visita a Cambridge en diciembre; *c*) el enviado y registrado en la S. R. del cual quedan, además de esta copia, otras cuatro más y que fue editado por S. P. Rigaud[47], y, en nuestra opinión; *d*) la correspondencia con Hooke de 1679-1680.

Un segundo grupo, no bien determinado a causa de un cierto desorden en los papeles mismos y de su relativa dispersión, estaría constituido por todos los borradores, ensayos, correcciones, etc., posteriores al tratado de la S. R. y anteriores al manuscrito final (manuscrito M) enviado a la S. R. para que se editara con el título definitivo. Aquí aparecen algunos ensayos con títulos del siguiente tenor «de motu corporum in mediis regulariter cedentibus», «de motu corporum: definitiones», «de motu corporum líber primus», «de motu corporum liber secundus» (y una redacción inicial y después desechada del Libro III de los *Principia*), a lo cual hay que añadir el conjunto de observaciones, datos experimentales, etc., recolectados o efectuados por Newton como soporte empírico o confirmatorio de su construcción teórica. La correspondencia con Flamsteed es un buen ejemplo. Y aunque algunos de estos papeles, cartas, notas de experimentos con péndulos, con fluidos, con graves, etc., no sean habitualmente vistos como parte de los escritos previos o borradores, no deben olvidarse ni desconectarse del proceso general de construcción de los *Principia*. Las entradas del «Waste Book», algunas relacionadas con obras leídas en esta época, también ofrecen datos interesantes, que tenidos en cuenta por los historiadores, permiten intuir mejor lo que estaba pasando. Finalmente, el tercer gran grupo documental vendría dado por el manuscrito M sobre el cual se hizo la primera edición E_1[48], la correspondencia mantenida con Halley durante la impresión, algunos fragmentos desechados a última hora, las correcciones introducidas en los ejemplares de su propiedad de esta primera edición, así como por otras añadidas, vía Cottes, para la segunda, y otras menores para la tercera y final realizada en vida de Newton. Los casi cuarenta años, 1687-1726, que separan la primera edición de la última permitieron mejorar las observaciones y advertir deficiencias en la presentación, en el cálculo y en el orden lógico en que se presentaban algunas Proposiciones en la primera edición. Aunque no hubo ningún físico o matemático capaz de mejorar sustancialmente la concepción general formulada por Newton y, por consiguiente, la tercera edición no es más que un refinamiento en los detalles de la primera, no puede desconocerse el hecho de que los *Principia*, durante ese largo período de tiempo, generaron mecanismos empíricos y teóricos para la autocorrección, y también para la posterior evolución de esa misma concepción. Si los historiadores no pueden prescindir de Newton para entender la física, y aún la matemática, del siglo XVIII, pese a que hubiese sido reescrita por D'Alembert, Lagrange, Laplace, Euler, Gauss, los Bernouilli, o Boskovic, no es porque los *Principia* no fuesen susceptibles de una profunda revisión tanto de fondo como de forma, sino más bien porque aún no habían hecho nacer las estructuras teóricas derivadas de ellos, capaces de producir esa revisión. Hasta tal punto serían sólidas por un lado y fructíferas por otro las bases puestas por Newton en su obra física.

Lo primero que encuentra el lector, tras prólogos, prefacios y dedicatorias, es un pequeño grupo de definiciones conceptuales cuya formulación debió costar a Newton mucho esfuerzo de rigor y crítica puesto que esos conceptos, al menos en parte, circulaban de modo enmarañado y confuso, impregnados de ideas aristotélicas o escolásticas, ciertamente aligeradas de peso metafísico por la crítica de los siglos XIV y XV primero y por las depuraciones no siempre definitivas de Descartes, Galileo, Huygens, y algunos más. Así se ve Newton en la necesidad de empezar por definir la *cantidad de materia*, «quantitas materiae», la *cantidad de movimiento* y la *fuerza ínsita* (inercia), en términos cuantitativos o mensurables, de modo que resulten matematizables. Además, es preciso que estas definiciones tengan suficiente generalidad como para ser extensibles a toda materia, a todo movimiento con el fin de que esta «filosofía», fundada en principios matemáticos, alcance hasta donde alcancen materia y matemáticas, ya que no se trata de un sistema de cuerpos en movimiento sino que se trata de construir el sistema de todos los cuerpos, terrestres y celestes, en movimiento. Como él mismo dice en su prefacio «en este sentido la *mecánica racional* será la ciencia de los movimientos resultantes de cualesquiera fuerzas, y de las fuerzas requeridas para cualesquiera movimientos, ciencia propuesta y demostrada con precisión». Para ello era preciso poner en orden la relación que pudiera hacer posible cuantificar fuerzas y movimientos. Pero si la materia, en más o menos cantidad, mayor o menor, fuese igualmente indiferente al movimiento, entonces cualquier clase o cantidad de fuerza podría producir cualquier cantidad de movimiento, cosa, además de contraria a la experiencia y absurda para cualquier menestral de la mecánica vulgar, difícil de compaginar con la idea de cuantificar fuerzas y movimientos, toda vez que la indiferencia de la materia respecto al movimiento haría de éste, ante la presencia de una fuerza, algo indeterminable y, por así decirlo, libre. En tal caso ¿cómo cuantificar la fuerza? Si no se pudiera cuantificar mediante alguna relación necesaria, el movimiento existente en la cosa movida, no se podría cuantificar tampoco la fuerza «requerida» para generarlo. Pero en la cosa movida sólo se pueden distinguir dos cosas: la velocidad y la materia misma que es el objeto en movimiento. Y precisamente por eso es la cantidad de materia lo que requiere, más allá de las intuiciones comunes, una definición precisa: densidad y volumen conjuntamente, ambos mensurables, determinan la cantidad de materia en un cuerpo. Para el físico actual no es una noción absolutamente igual o equivalente a la de masa, pero ha sido la mejor durante largo tiempo, pese a adolecer de cierta posible circularidad, como otras de las aquí propuestas por Newton, toda vez que el concepto de densidad no es, ni mucho menos, un concepto primitivo sino vinculado a su vez a relaciones entre cantidades de materia y volumen. Que Newton lo entendiese vinculado al peso, en relación con el principio de Arquímedes, no salvaba la posible circularidad, pero tampoco disponía de mejores instrumentos conceptuales. Cantidad de materia, aunque siga siendo un concepto algo impreciso, es ahora un concepto susceptible de medida con valores proporcionales al peso, ya se llame cuerpo, ya

masa. Esto permite utilizar el término «cuerpo» en las definiciones siguientes sin equívocos engorrosos.

Tampoco se debe pasar por alto el esfuerzo realizado en las Definiciones III y IV para lograr una expresión libre de equívocos de las dos clases de fuerzas que van a entrar en consideración: la fuerza ínsita y la fuerza impresa. La primera «que puede llamarse muy adecuadamente fuerza de inercia» es proporcional al «cuerpo». La segunda «consiste en la mera acción» y no perdura en el cuerpo después de la acción. El resto de las definiciones, supuesta la pertinencia de las anteriores, se dedica a la fuerza centrípeta. También en estos casos cabe hacerse una cuestión del grado de simplicidad y primitivismo logrado en las definiciones y en los conceptos. De hecho, el lector hallará que en la Definición V se afirma con toda osadía que la gravedad es una fuerza centrípeta por la que los cuerpos tienden hacia el centro de la Tierra o, lo que quizá sea más atrevido, por la que los planetas son continuamente apartados de las trayectorias rectas y retenidos en giro por líneas curvas. Pero (sea la historia nuestro testigo), ésta iba a ser una de las mayores cruces del autor andando los años. Las fuerzas centrípetas, tal y como las presenta Newton, actúan sobre cuerpos y, por tanto, sus acciones pueden ser reducidas mentalmente a acciones instantáneas que se van produciendo en instantes sucesivos. Ya se consideren dichos instantes continuamente sucesivos o discretamente sucesivos, ello no hará al caso a la hora de tratar a dichas acciones como impulsos instantáneos, con la posibilidad de ser matemáticamente asociadas a infinitésimos de espacio o de tiempo o de movimiento. Ya Galileo había utilizado una estrategia similar en su análisis de la caída de los graves, si bien el aparato matemático era mucho menos sutil que el de Newton. Este parece que sufrió una reconversión a los ideales geométricos euclídeos en los años inmediatamente anteriores a la redacción de los *Principia*[49], menospreciando la poca elegancia de las construcciones analíticas de su tiempo. Esto le llevó, tal vez, a expresar las relaciones entre esas mínimas cantidades en términos de «razones nacientes» o «evanescentes». Esto no obstante, Newton conocía bien el tratamiento de esas cantidades «fluentes» mediante funciones continuas y en algunos casos no renunció a llevarlo a la práctica. Con todo, la concepción de las fuerzas impresas como «acciones» le ayudó también en la opción por el método de razones primeras y últimas. Por todo esto las definiciones primera, tercera y cuarta en las que se exponen los conceptos de fuerza inercial, fuerza impresa y sus relaciones con la masa y con el cambio de estado inercial tienen particular relevancia como exponentes de la concepción dinámica de Newton. No es, pues, extraño que hayan merecido tan amplia atención posterior. Y no sólo por el carácter arcaizante de esa concepción de fuerza (sino es también una continuación sofisticada de la estática arquimédea, cosa menos sorprendente en un cultivador de «priscae sapientiae»)[50], sino sobre todo por su carácter determinante de la dinámica newtoniana. Mientras tanto, las Definiciones V a VIII relativas a la fuerza centrípeta y sus clases (absoluta, aceleratriz, motriz) no adolecen menos que las otras de esos males, pero, por su carácter particular, el

análisis de las mismas se ha centrado en su particularidad. De ésta lo más notable es, sin duda, su aparente identificación con la gravedad y la no menos aparente inclinación de Newton a afirmar la realidad de la misma. Si la física actual acaba por darle la razón, después de haberle reprochado tantas veces que con esa identificación había reintroducido por la puerta trasera a las denostadas «cualidades ocultas» de la escolástica, no por ello habrían desaparecido los problemas mayores apuntados antes.

Sus definiciones se cierran con un Escolio famoso. En él Newton nos expone algo de su trasfondo metafísico: sus ideas de espacio absoluto, tiempo absoluto, movimiento absoluto. Estos conceptos se han utilizado en física de un modo parcial, el meramente necesario y suficiente para la física, pero en este texto parecen tener más alcance. No se trata de meros sistemas constantes de referencia, de un eje fijo de coordenadas, sino que el tiempo absoluto y el espacio absoluto aparecen aquí como referencias ontológicas y cosmológicas últimas y no como meros sistemas inmóviles de referencia. Aparecen concebidos como absolutamente independientes y como invariantes absolutas del mundo. El lector no dejará de recordar las discusiones derivadas de este planteamiento, desde la polémica Leibniz-Clarke hasta Mach, Einstein, etc., pero a la vez, es preciso recordar que, junto con el Escolio general y las «Queries» de la *Optica*, constituye uno de los textos más característicos y representativos del pensamiento de Newton, alejado en muchos puntos más de lo que pudiera parecer del puro pensamiento mecanicista.

En el segundo paso de la reconstrucción final de los diferentes tratados «de Motu» anteriores al manuscrito M, Newton reduce un largo conjunto de axiomas, leyes, y lo que acabó por denominar hipótesis a tres leyes, seis corolarios y un escolio. La primera de las leyes es la de la inercia, ya antes enunciada con mayor o menor precisión por Descartes, Galileo, Huygens, pero aceptada por los matemáticos, como dice Newton. En la segunda, establece la proporcionalidad entre los cambios de estado inercial (de reposo o movimiento) y las fuerzas impresas según la dirección recta de la acción de éstas. Pero necesitará de los corolarios para precisar los distintos casos de composición, descomposición, suma, resta, etc., en que se producen los cambios y las acciones de las fuerzas impresas. La tercera es la ley de la acción y reacción. Su origen en Newton se relaciona con sus estudios de los impactos y su formulación cartesiana que ahora encuentran una integración en una dinámica más general. Esta ley tiene aquí una particular aplicación («esta ley también tiene lugar en las atracciones»), como apostilla él mismo al tratar de justificarla. Su aplicación en el estudio de los impactos ya había dado sus frutos y ahora en su aplicación a la gravedad iba a dar más, pese a que la fuerza impresa en los impactos cuadre bien con «la mera acción que no permanece después en el cuerpo» y en cambio no cuadre tanto con la atracción, que es continua. En el Escolio, hace Newton repaso y homenaje de los predecesores en el uso de estas leyes y trata de dar carta de naturaleza y noble ascendencia a esta tercera. Por si las polémicas volviesen, no está demás una familia honrada como la de Galileo, Wren, Huygens, Mariotte. Pese a ello

la física ha revisado estas leyes de una forma paralela al desarrollo propio y casi como condición del mismo. La primera ha corrido la suerte de la crítica al concepto de «cantidad de materia» del cual es inmediatamente dependiente; la segunda ha sido tributaria del concepto newtoniano de «fuerza impresa» y, por tanto, de las reformulaciones del mismo, tentativas que aparecen ya en los *Marginalia leibnizianos*[51] y se van completando con D. Bernouilli en 1726 y D'Alembert en 1743[52]. En cuanto a la tercera, ha sido objeto de análisis de otro tipo: ¿es redundante y un mero corolario de las dos primeras? Es evidente que, desde un punto de vista cinemático, la física podía prescindir de ella, y desde un punto de vista dinámico, podía considerarla como una mera «suma» de las dos primeras. El esfuerzo de Newton por darle una justificación empírico-intuitiva, con ejemplos de mecánica empírica o de *trabajo* es un indicio de su fragilidad teórica. Además, y éste será un problema que no alcanza a la fuerza impresa en los impactos, ¿cómo ejercer acción y reacción a distancia?, esto es, ¿cómo ejercerla sin impacto, sin medio físico interpuesto que trasmita la acción o la reacción? Los vórtices cartesianos, materiales como eran, bien podían cumplir con este cometido mecánico, pero con la supresión del medio vorticial o etéreo u otro cualquiera no hay forma física de transportar la acción. En la primera definición aparece ya el problema. Cotes, en su prefacio, se hace eco de ello y parece abogar explícitamente por la «acción a distancia», naturalmente sin dar explicación mecánica, al igual que inclina demasiado la opinión del lector en favor de la consideración de la gravedad como «propiedad de la materia». Newton no se vinculó a una consideración tan fuerte[53].

Ya hemos dicho que los *Principia* resultan de la reelaboración y expansión del pequeño tratado de la S. R. En un primer plan esa reelaboración se debió concebir como dos libros: el primero sobre el movimiento de los cuerpos y el segundo sobre los movimientos del sistema solar según los principios establecidos en el primero. Este plan se realizó de hecho y se conserva en manuscritos entregados a la Biblioteca de la Universidad como si tales textos hubieran correspondido a las «Lectiones» de su cátedra Lucasiana, cosa que no ocurrió, al menos tal cual se entregaron y se conservan. Lo que sí fueron, desde luego, fue una primera redacción de ensayo, base de correcciones y de ampliaciones y, en suma, el antecedente inmediato de los *Principia*, del manuscrito M que sirvió para hacer la primera edición. Pero el primitivo libro sobre el movimiento acabó ampliado a dos, mientras que el segundo quedó desechado y sustituido por una nueva y algo más amplia redacción casi sobre el mismo esquema que el primitivo y que será el Libro III.

De este modo el Libro I vino a adquirir una notable extensión con 14 secciones, más que proposiciones contenía el tratado de la S. R. Este pequeño tratado se inicia con tres definiciones y cuatro hipótesis más dos lemas puramente instrumentales, y pasa inmediatamente a tres teoremas: el primero, igual a la Proposición I de los *Principia*, demuestra la igualdad de áreas en iguales tiempos en el caso de un cuerpo en giro alrededor de un centro, cuando esas áreas son descritas por un radio vector

trazado desde el cuerpo al centro de giro. El segundo se corresponde con la Proposición IV de los *Principia* y define las fuerzas centrípetas de los cuerpos que giran en las condiciones del movimiento circular uniforme. (En la segunda y tercera ediciones de los *Principia* se suprimen los diagramas). El tercero —correspondiente a la Proposición VI de los *Principia*— expresa la fórmula general de la proporción de la fuerza centrípeta para un cuerpo que gire en una curva. Siguen unos problemas, el primero determina, para un punto de la circunferencia, la fuerza central que permite a un cuerpo girar por ella —Proposición VII de los *Principia*— y el segundo y el tercero plantean el mismo problema para el centro y el foco de una elipse —Proposiciones X y XI—. Sigue luego un cuarto teorema —Proposición XV— en el que demuestra que cuando la fuerza centrípeta es inversamente proporcional al cuadrado de la distancia al centro, los cuadrados de los tiempos periódicos son proporcionales a los cubos de los ejes transversos de las elipses. No olvida recordar aquí que «con este método es posible determinar las órbitas de la Tierra, Marte, Júpiter y Saturno…». Un cuarto problema —Proposición XVIII de los *Principia*— investiga el caso de que «supuesto que la fuerza centrípeta sea inversamente proporcional al cuadrado de la distancia al centro, y conocida la cantidad de dicha fuerza, hállese la elipse que describe un cuerpo lanzado desde un lugar dado y con una velocidad dada siguiendo una línea recta». Caso este que bien pudiera ser el de la Luna tratado en los años 65-66 y que responde también al problema debatido con Hooke. Todo lo contenido hasta aquí va a ser tema de las secciones segunda y tercera de los *Principia*, lo cual muestra la magnitud de la ampliación a que sometió Newton su primera redacción. El resto del tratado de la S. R. son tres problemas sobre la caída de graves en medios no resistentes —Proposición XXXII— suponiendo a la gravedad como una fuerza centrípeta de ley igual al inverso del cuadrado; y los problemas sexto y séptimo en que trata de la resistencia de los medios a la caída de los cuerpos —Corolarios de las Proposiciones II y III del Libro II— y un Escolio cierra el tratado (Proposición IV del Libro II) sobre los proyectiles lanzados paralelamente al horizonte. Eso es todo lo que Newton incluyó en el pequeño tratado de la S. R[54]. ¿Cómo entender el paso, en dieciséis o diecisiete meses, hasta una obra de tan gran complejidad como los *Principia*? Si repasamos los títulos de las catorce secciones del Libro I, hallamos que la primera es de contenido puramente matemático. Sus once Lemas responden a una reelaboración «*ad hoc*» de los estudios sobre tratamiento de cuadraturas y su tratamiento mediante límites de cantidades fluentes. En esta sección se halla implícita una opción metodológica vinculada no sólo a sus preferencias, en ese momento, por las formulaciones geométricas en lugar de las analíticas, sino también a su modo de entender las relaciones espacio-tiempo-movimiento, que se hallaba vinculado a imágenes geométricas o a las representaciones geométricas de las mismas, cosa muy evidente cuando aborda la construcción de problemas de movimiento, etc. Sus generalizaciones son siempre, incluso en el desarrollo de sus estudios de análisis,

generalizaciones a partir de casos discutidos hasta el detalle. Para hacer lo mismo con los problemas y teoremas de movimiento opta por su representación y ésta es más comprensible y directa en un modelo geométrico que en un modelo analítico. De hecho, él mismo dio una versión de los hechos congruente con esta interpretación[55]. Las secciones II-XII-XIII se dedican a temas de fuerzas. Pero, como no puede menos de ocurrir, las magnitudes de semejantes fuerzas han de entenderse como condiciones necesarias de las propiedades métricas de algunos movimientos. Además, la sección tercera, pese al enunciado «sobre el movimiento de cuerpos en secciones cónicas excéntricas», se dedica a la determinación de las leyes de las fuerzas centrípetas tendentes a algún centro o foco. Y ambas, segunda y tercera, tienen una gran parte de proposiciones problemáticas. Esta característica, junto con la introducción de los Lemas necesarios para resolver los problemas, hace de estas secciones iniciales un buen exponente del método heurístico newtoniano: los fenómenos no son aquí tales fenómenos sino hipótesis ideales de carácter geométrico, como por lo demás concierne a «principios matemáticos». Solamente los corolarios introducen aplicaciones al mundo real, o meramente las insinúan, pero sin identificar absolutamente los fenómenos de referencia con las hipótesis formuladas en los problemas[56].

Cuando el lector inicia la sección cuarta se encuentra con el Lema XV y con la Proposición XVIII, Problema X. En resumen: mucho instrumental matemático y problemas resueltos gracias a ese instrumental. Y cuando concluya la sección quinta habrá pasado por XXVII Lemas y por XXI Problemas y solamente habrá llegado a la Proposición XXIX. Habrá recorrido un elenco de cuestiones relativas al hallazgo de puntos, curvas, lugares geométricos, descripción de curvas tangentes a rectas dadas, etc., cuestiones todas que Newton había trabajado y sistematizado mucho antes[57]. Y algo parecido vale para la sección VI, aunque en ella aparezca un Lema tan notable como el XXVIII y su aplicación en el Problema XXIII siguiente. De hecho estas seis primeras secciones son sobre todo una brillante aplicación o una extensión de sus conocimientos matemáticos a los problemas ya enunciados en el pequeño tratado de la S. R., junto con una organización de la geometría asociada a esos problemas y en algún caso, como el del último Lema mencionado, del análisis algebraico merecido por determinadas ecuaciones[58]. En la sección VII se detendrá con la solución del problema de ascenso y descenso rectilíneo de los cuerpos bajo la condición del inverso del cuadrado de la distancia. Es el último problema del pequeño tratado de la S. R. y empieza a dar por supuestas las cuadraturas de las curvas[59].

Podríamos resumir el contenido de las siete secciones restan les del Libro I diciendo que en ellas amplía los casos, tratados o insinuados hasta aquí, mediante una sucesiva introducción de variables procedentes de las diferentes perspectivas que pueden entrar en consideración, aunque las más notables sean las de la astronomía empírica. De hecho, la VIII se centra en cómo hallar órbitas de cuerpos que giran bajo

la condición de fuerzas centrípetas, la IX del movimiento de órbitas y ápsides, la X del movimiento pendular, la XI del movimiento de cuerpos que se atraen mutuamente, y así se llega a la sección XII, completamente original (requiere un nuevo Lema) en la que estudia la estructura de las fuerzas atractivas en cuerpos esféricos y que le permitirá descubrir, entre otras cosas, que la atracción de la Tierra en su interior desde la superficie al centro no decrece en razón del cuadrado de la distancia, sino sólo de la distancia. Es toda esta sección autónoma en tal medida que no depende en su tratamiento más que de la Ley tercera (en la Proposición LXXV) y del Corolario 3 del Lema VII (en las Proposiciones LXX y LXXI) La sección XIII es de nuevo una aplicación de la anterior, mientras la XIV debe entenderse en relación con su teoría corpuscular de la luz y no exenta de reivindicación de esa misma teoría frente a argumentos utilizados en su contra en la polémica de los años setenta. Pero, con todo, vale la pena recordar aquí que en este Libro I se encuentran algunos de los problemas cuya solución merece ser recordada por su novedad: así, en la sección VI aborda el llamado problema de Kepler (el cálculo de áreas para sectores focales de la elipse), en la Proposición LXIX calcula la Corrección que es preciso introducir a la tercera ley de Kepler en función de las masas de los cuerpos en giro, en la Proposición LXVI aborda el célebre problema de los tres cuerpos, etc. El esfuerzo matemático acumulado a lo largo de este Libro I puede parecer desproporcionado con respecto a los problemas «reales» que trata, pero nada tiene de trivial si se considera en su conjunto Sean lo que fueren las fuerzas centrípetas, en torno a ellas se constituye por vez primera un sistema capaz de calcular espacios, velocidades, tiempos, masas, trayectorias, etc. con aproximaciones jamás conocidas. Añádase la generalidad alcanzada en esta teoría del movimiento (ideal o matemático si se quiere) y pese a las restricciones que el propio Newton concede de buena gana tendremos una primera mecánica «la más perfecta de todas», como dice en el prólogo, con que se tropieza la historia, un tratado sobre el movimiento de los cuerpos de alcance teórico general.

Si el Libro I puede describirse como teoría general del movimiento de los cuerpos en condiciones ideales (en medios sin resistencia, con masa pero sin forma ni volumen, sin problemas de elasticidad ni de viscosidad, etc.), en el Libro II la perspectiva cambia notablemente. Newton, probablemente al redactar algunas Proposiciones finales sobre los movimientos en medios resistentes, comprendió que el tema era más amplio de lo que en un principio sospechó y quizá vislumbró también que sólo una aplicación a un mundo más real del aparato teórico construido en el Libro I merecería el calificativo de «Filosofía Natural», y que explorar el movimiento de la materia en interacción con la materia era una ocasión pintada para acercarse a profundas y oscuras intuiciones sugeridas por sus estudios ópticos y alquímicos. El caso es que cerró el Libro I y abrió un segundo Libro de naturaleza mucho más concreta, con aplicaciones a problemas más específicos en los cuales introduce las complejidades de la resistencia, la viscosidad, la densidad, la elasticidad, etc. Aunque

siempre se ha puesto de relieve la aparición en este Libro II de los indicios y noticias primeras publicadas sobre el cálculo fluxional —Lema II y su Escolio, Escolio que varía notablemente de la primera y segunda ediciones a la tercera— nos parece mucho más relevante llamar la atención del lector sobre la naturaleza misma de los problemas que se tratan en él. Las tres primeras secciones las consagra Newton al estudio del movimiento en medios resistentes —resistencias en razón de la velocidad, en razón del cuadrado de la velocidad y en razón compuesta de ambas— resistencias que, por otra parte, muy bien pudieran no darse en la naturaleza, pero cuyo estudio proporcionará el modelo para otras cualesquiera que de hecho se dieren. Ya advierte en el Escolio de la sección primera que «la resistencia en razón de la velocidad es más una hipótesis matemática que una hipótesis natural». Pero, con todo, en corolarios posteriores encuentra casos de aplicación inmediata. Y en otros casos, como aduce en el mismo Escolio, media una aplicación de las Leyes II y III, junto con evidencias empíricas de su propia experiencia de laboratorio, en las que ha tratado de medir viscosidades, elasticidades, etc. De ello informa al lector en el Escolio final de la sección tercera, aunque no le informa de por qué atribuye los valores de resistencia a cada una de las partes. En este orden de cosas Newton se mueve con menos seguridad y menor base experimental u observacional que en temas astronómicos y se ve forzado a recurrir a estrategias suplementarias, en buena parte analogías y extrapolaciones de situaciones ideales estudiadas en el Libro I. Al igual que un buen número de Proposiciones y hasta de secciones del Libro I no tienen aparentemente vinculación directa con el tema-guía de su trabajo (la dinámica celeste) y, sin embargo, merecieron su atención por diferentes razones, de igual modo en este Libro II hay algunas cuestiones aparentemente desconectadas del tema principal. La sección IV dedicada al análisis del movimiento circular en medios resistentes o la sección V, por ejemplo, que se consume en el análisis de la densidad y de la compresión de los fluidos y de la hidrostática. Y sin embargo, las aplicaciones van apareciendo, poco a poco, principalmente en los corolarios, en los cuales tiene presentes problemas debatidos, en el momento en que escribe, por otros filósofos naturales, de algunos de los cuales da cuenta, como en el caso del Escolio de la Proposición XXIII sobre el magnetismo. Y aún es más explícita su intención de tratar un problema concreto en la sección VI siguiente, dedicada al estudio del movimiento pendular en medios resistentes. Y si el lector se detiene en una lectura minuciosa de la Proposición XL y del Escolio subsiguiente, comprobará en qué medida la ampliación de todo este Libro II está vinculada a experimentos conocidos o directamente concebidos y realizados por Newton, no siempre con ocasión de la redacción del libro y, por lo tanto, la medida en que Newton trataba de tener presentes en estos momentos otros problemas procedentes de otras investigaciones y de otros contextos, para los cuales ve ahora una vía de solución en el contexto general en que se mueve. La sección VIII, en la que aborda una antigua cuestión sobre la que no vuelve directamente, está dedicada al

movimiento ondulatorio. En ella aparece de nuevo el tratamiento corpuscular de las ondas en medos fluidos, aparentemente suficiente para explicar fenómenos ondulatorios. Y el lector avisado hará por su cuenta un repaso de las cuestiones debatidas sobre la naturaleza de la luz. Y, aunque breve, no será menos polémica la sección IX en la que desmonta con gran facilidad toda la teoría vorticial cartesiana. Le bastará aplicar su batería matemática, como ya había sugerido a Hoofe respecto a la teoría de Mallement, y mostrar que no es compatible con los cálculos que sean a su vez compatibles con las observaciones. Aparece así un universo vacío y libre de materias más o menos sutiles dotadas de actividad mecánica. Pero ¿cómo pueden, entonces, realizarse todos estos movimientos en espacios libres, sin vórtices? Eso podrá entenderlo el lector, dice Newton, a partir del Libro I, y se expondrá más completamente en seguida en el Sistema del Mundo.

Este Libro II viene, por tanto, a introducir algunos nuevos elementos relacionados con las fluxiones, aunque ya hubiesen aparecido insinuados en el Libro I (Lemas I, II y Escolio del Lema XI así como en las Proposiciones XXXIX, Corolario 3, XLV, Casos 2 y 3) relativos a sus ideas sobre las cantidades infinitamente pequeñas. Pero no es seguro que las necesidades del proyecto inicial exigiesen un rodeo como el que el lector se encuentra al principio de este libro, salvo que se llegue a la convicción de que Newton, cosa no improbable, se tropezase con problemas tanto conceptuales como instrumentales más complejos de lo previsto. De hecho la temprana aparición (Proposición II) de la hipótesis de trabajo sobre la proporcionalidad de la resistencia respecto a la velocidad le va llevando a cálculos más complejos y a introducir el célebre Lema II. Estos complejos problemas no obedecen únicamente a un deseo de Newton de mostrar su competencia matemática y de alejar de las polémicas a inexpertos, sino que es más que probable también una ampliación del horizonte filosófico de Newton durante el proceso de redacción de la primera parte, ampliación que, por otra parte, no contaba siempre con suficiente base experimental. Pero además, algunas de sus brillantes deducciones estaban de antemano orientadas por otros trabajos (la ley de Boyle, trayectorias de proyectiles, sonidos, comportamiento de la luz, cicloides y péndulos isócronos, etc.), pero no se hallaban sometidos, ni esos ni otros fenómenos, a un tratamiento unificado. Esa es la primera característica de este Libro II. De modo más sutil, el intento de organizar la dinámica de los cuerpos en cualesquiera clase de medios, resistentes o no resistentes, en torno a una ley profunda de la naturaleza —la ley de la gravedad— va adquiriendo carta de naturaleza paso a paso, hasta que el lector se ve irremediablemente en el dilema de elegir entre un universo gravitacional y otro universo lleno de materias vorticiales cuyo comportamiento dinámico sería absolutamente incompatible con los fenómenos. Una vez alcanzada esta perspectiva unificadora, solamente faltaba mostrar el grado de aplicación de estos principios al universo. De ahí el título mismo del Libro III: «Sobre el Sistema del Mundo».

Con el rótulo anterior como título, el Libro III es resultado de la reescritura de una redacción anterior, adoptando una forma retórica en la exposición —«more mathematico» dice Newton— para alejar a posibles objetantes sin base matemática y quizá pensando en Hooke cuyas reclamaciones, vía Halley, ya le zumbaban en los oídos cuando se decide a hacer el cambio de ropaje[60], pero reteniendo con algunos cambios y mejoras los tópicos de la primitiva versión en lenguaje «popular». En general se trata de reconstruir el sistema copernicano de acuerdo con las leyes de Kepler y las correcciones computacionales posibles gracias al nuevo método, incorporando como centro de la reconstrucción la ley de la gravedad, que le permite explicar también las mareas, predecir irregularidades en el movimiento de la Luna y los planetas y explicar la trayectoria de los cometas.

El lector es advertido desde la breve introducción, antepuesta por Newton, de la relativa independencia y autonomía de este libro, pero no debiera, a mi entender, creer demasiado en ella, lo que tampoco significa que se trate aquí de una mera extensión de alguno de los problemas tratados anteriormente, o para ser más exactos, no es posible prescindir de los dos anteriores como patrón explicativo del tercero, aunque sea posible prescindir de una gran parte de las Proposiciones contenidas en ellos a la hora de exponer «matemáticamente» las consecuencias que aquí se exponen. Cuando en la Proposición VII Newton propone su ley de la gravitación universal y afirma que «la gravedad acontece en todos los cuerpos y es proporcional a la cantidad de materia existente en cada uno», no está obteniendo una consecuencia matemática de una proposición o de varias, sino que está expresando como inferencia de todo lo anterior la condición necesaria para que en el caso de los planetas, y del mundo en general, se cumplan las proposiciones matemáticas sobre el movimiento del Libro I (supuestas las restricciones que haya que suponer, entre otras las del Libro II). Por ello, todo este libro puede ser visto como un ejercicio de aplicación de los otros dos a un caso, el del sistema del mundo. Los presupuestos necesarios, además de los dos libreas anteriores considerados como marco teórico, aparecen explicados en las diez primeras proposiciones junto con las Reglas y Fenómenos incluidos o las Hipótesis y Lemas intercalados a lo largo del libro. En ese ejercicio Newton se centró en unos pocos casos en los que consideraba que su tratado podría mostrar su poder de resolución: el cálculo de órbitas y cuerpos celestes, una teoría de los movimientos y perturbaciones de la Luna, una explicación de las mareas como efecto de la gravedad hacia el Sol y hacia la Luna y una aproximación a la teoría de los cometas como cuerpos que gravitan en torno al Sol.

Pero hay un aspecto en este tercer libro que debemos resaltar aquí para que el lector lo compruebe bajo la óptica que crea oportuna: lo motiva inicialmente el amplio aparato observacional reunido por Newton para ser incorporado como datos observables de cada uno de los problemas. Si notamos que no era «profesionalmente» un astrónomo ni un observador dotado de instrumental adecuado que dependía de las tablas de observación y de los datos publicados o comunicados por otros y además

tenemos en cuenta las nachas tablas y datos que logró reunir ya desde la primera edición (si bien esta labor continuó para las siguientes y dio lugar a múltiples correcciones y cambios de datos), tendremos una mejor idea del alcance del proyecto abordado por Newton en este libro. Pero para comprender el alcance de su estudio no conviene contextualizar sus resultados entre los logros de la astronomía actual. Esto nos llevaría a considerar a este libro como un mero tratado de ciencia positiva, como un producto tópicamente baconiano sin mayor trascendencia que la búsqueda y acumulación de resultados. Si esto hiciéramos, habríamos dejado de lado el carácter *filosófico* que Newton pretendió para su Sistema del Mundo. Y no se trata solamente del aspecto metodológico vinculado frecuentemente a las «hipótesis» (nueve en la primera edición) con que inicia la exposición ni a la posterior transformación de las mismas, en las dos ediciones siguientes, en las «Regulae Philosophandi» (tres reglas, una hipótesis y seis fenómenos en la segunda; cuatro reglas, una hipótesis y seis fenómenos en la tercera), porque esto meramente tiene importancia respecto a la idea que Newton tenía de cómo hacer filosofía, sino que más bien ha de tenerse presente qué filosofía hace o pretende hacer con ese u otro método cualquiera. Y tampoco es excesivamente relevante aunque dé una pista el contenido de la Hipótesis III de la primera edición (suprimida en las otras dos) pese a su compromiso ontológico («todo cuerpo puede transformarse en otro cuerpo de otro género cualquiera y asumir todos los grados intermedios de cualidades»), puesto que Newton oficia de filósofo natural en el sentido histórico del término en ese momento y, como hicieran Descartes, Gassendi, More, etc. aspira a desvelar los secretos de la naturaleza hasta donde sea posible, sin importar que alguno de ellos se resista al intento. Saber cómo es el mundo, cómo se mueve y por qué, descubrir que solamente fuerzas proporcionales a la distancia y proporcionales al inverso del cuadrado de la distancia permiten movimientos armónicos y estables en círculos o en elipses, etc. es alcanzar a entrever el plan divino, entrar en el secreto de la naturaleza y, como había soñado la metafísica, contemplar al universo «sub especie divinitatis». Sólo desde esta perspectiva se puede comprender que un libro de mecánica, por celeste que ésta sea, acabe con un Escolio General de amplio sentido teológico (aunque incluido en la segunda edición y prefigurado en las «Qeries» de la *Optica* y en parte en la conclusión general redactada para la primera pero no incluida finalmente). Todo lo cual sugiere que el conjunto de la obra, y más explícitamente este tercer libro, no es sino un retazo de una más amplia y «religiosa» concepción del mundo a cuyo descubrimiento dedicó no sólo estos esfuerzos sino también los que dedicó a la óptica, a la historia sagrada y profana, a los estudios bíblicos y, con mucha mayor extensión y menos éxito, a la alquimia. Si se supone que para Newton el concepto de naturaleza incluía a Dios como esencia o parte de la misma y que las manifestaciones, «fenómenos», de la naturaleza en sus diferentes estados y «grados de cualidades» eran también fenómenos de la divinidad, quizá se esté en una perspectiva cercana a la que, al menos en momentos de su madurez, poseyó el propio Newton. Y, desde esta

perspectiva, las leyes de la mecánica, las de los astros y, en definitiva, las del mundo no serían meramente leyes naturales sino, sobre todo, leyes divinas «matemáticamente» demostradas. Newton, que sepamos, no formuló jamás estos pensamientos en términos explícitos, pero algunas expresiones del Escolio General y otras de la polémica con Leibniz vía Clarke inducen a suponerle cercano a ellos. No menos se infiere esa cercanía de la lectura que los teólogos afines a él hicieron por su cuenta. Y aunque su sentido del rigor científico le llevase a cautelas, con la gravedad lo mismo que con la divinidad, no por ello logra ocultar por completo el último contexto de su pensamiento como filósofo natural. Pero, prescindiendo ahora de estas conjeturas filosóficas, vale la pena subrayar algunas de las aplicaciones concretas en las que Newton ensaya su poderoso arsenal teórico y que el lector encontrará en este libro: entre ellas el Problema III en la Proposición XIX, en el que discute la relación entre los diámetros polar y ecuatorial de la Tierra, logrando por primera vez una imagen muy aproximada a la realidad[61] y la consecuencia inmediata de la Proposición siguiente según la cual los pesos de cuerpos iguales varían en razón de la distancia al centro de la Tierra con el conjunto de mediciones aportadas y comparadas entre sí realizadas a diferentes latitudes con péndulos isócronos. De no menor interés resulta su análisis de los movimientos de la Luna, tanto de los observados como de los meramente inferidos, que se extiende desde la Proposición XXV hasta la XXXVIII, con un extenso Escolio a la Proposición XXXV en el que, por si el lector no se hubiera percatado por sí mismo, le advierte de los méritos teóricos de la teoría de la gravitación y de sus posibilidades heurísticas, especialmente en casos tan complejos como el de los movimientos de la Luna (un problema de tres cuerpos) aunque hasta para él mismo «el cálculo de este movimiento sea difícil» y haya de recurrir a aproximaciones tolerablemente cercanas a las observaciones. Tampoco escapará a la consideración del lector la aplicación de la teoría de la gravedad al cálculo de la precesión de los equinoccios en la Proposición XXXIX, dando una explicación a un fenómeno observado desde los tiempos de Hiparco. Más misteriosos eran los movimientos y naturaleza de los cometas, a los que Newton había prestado atención ya desde sus días de estudiante y que, pese a su frecuencia y a las múltiples observaciones y conjeturas de que habían sido objeto, todavía en los años 82 y 83 eran ampliamente desconocidos por el mismo Newton, como refleja su correspondencia con Flamsteed. No obstante, gracias a la nueva teoría que acaba de conformar, a las observaciones recopiladas y a la recuperación de su técnica de estudio de lugares y puntos de rectas y curvas (Lemas V, VI, VII, VIII y IX), llega al brillante Problema XXI (Proposición XLI), con el ejemplo del cometa de 1680 en el cual reúne amplia información observacional y construye con gran arte una representación a escala de su trayectoria, de su luminosidad aparente y de las velocidades y tiempos de su paso por las inmediaciones de la Tierra. Y aunque el ejemplo acabe con curiosas reflexiones de Newton sobre el papel restaurador de la

materia cósmica que podrían jugar los cometas y sus enormes colas de partículas gaseosas, todo bajo un aparente sabor alquímico-naturalista especial, no es menos evidente que con esta Proposición y la siguiente sobre la forma de corregir la trayectoria hallada de un cometa logra por fin dar una ley para sus movimientos y su misterioso comportamiento que aventa para siempre los «terrores» de la humanidad. No olvida en esta última Proposición del libro incluir las observaciones disponibles de otros cometas (en particular el del año 64-65 y del año 82-83 de los cuales él mismo había sido testigo también), tanto si dispone de observaciones recientes y más o menos fiables como si se trata de meras noticias que vienen de la antigüedad.

Y para dar fin a su aventura hallará el lector el célebre *Escolio General*, incluido en la segunda edición, aunque en la primera ya había redactado una especie de *conclusión* no incluida al fin, Escolio que, tras resumir muy brevemente los datos principales del Sistema del Mundo recién descrito, se adentra en una consideración sobre el papel de Dios en ese sistema y sobre la acción y presencia del mismo en su funcionamiento. Este sorprendente final tendrá el lector que ponerlo en relación con otros acontecimientos externos, tales como el inevitable choque con la filosofía cartesiana y cor la metafísica racionalista, especialmente de Leibniz, aunque el obispo Berkeley, en casa, ya había arremetido contra su concepción del espacio, del tiempo y del universo mecánico. No menor interés hallará el lector en los dos últimos párrafos del Escolio. En el primero de ellos Newton se confiesa desconocedor de la naturaleza de la gravedad y se niega a proponer hipótesis imaginarias y a reconocer que las haya propuesto —«hypotheses non fingo»— pero afirma que la gravedad, sea lo que sea, existe de hecho «satis est quod revera existat», afirmación que para él no dispone de menor base experimental que las demás afirmaciones sobre propiedades de los cuerpos tales como la impenetrabilidad, la movilidad, etc. En resumen que si Descartes propuso imaginarias hipótesis como la de los vórtices, tal no es su caso con la gravedad ni con otros temas (quizá menos fáciles de recordar, como el de la naturaleza de los colores), cosa que no le impide pasar inmediatamente a mencionar a «cierto, espíritu sutilísimo» que atraviesa a todos los cuerpos y gracias al cual, fenómenos como la cohesión, la interacción química o eléctrica, la emisión de luz y calor y hasta la sensación y movimientos musculares humanos podrían ser explicados, pero por desgracia «no disponemos de suficiente número de experimentos mediante los cuales deben mostrarse y determinarse con exactitud las leyes de las acciones de este espíritu», al cual en la versión inglesa de esta obra se adjetiva de «elástico y eléctrico», sin que sepamos todavía si no merecería quizás otros cuantos calificativos más.

V

Quizá sea hora de retomar alguno de los hilos que hemos dejado sueltos más

arriba y rematar el cuadro histórico en que se acabaron los *Principia*. Sabemos que Newton se concentró sobre los iniciales teoremas que envió a Halley en noviembre de 1684 y que no mucho después, quizá en febrero ya habían alcanzado una expansión aceptable, aunque por esos días hubo de viajar por asuntos privados a sus posesiones de Woolsthorpe, pero no más tarde de abril ya estaba de nuevo sobre ellos, ampliándolos continuamente, sin que en estos meses se prodigase en otra cosa que alguna muy breve comunicación con el exterior. Por estos tiempos tiene a su servicio un amanuense, Humphrey Newton (sin parentesco entre ellos), que va poniendo en limpio las sucesivas versiones ampliadas de un primer original con el título de «De motu corporum», sobre el cual vuelve una y otra vez *sir* Isaac hasta que en un momento considera completo lo que hoy tenemos como Libro I de los *Principia* y se lo envía a Halley y a la S. R. para su publicación. Esto ocurre en los últimos días de abril de 1686 y únicamente envía el Libro I. Halley, en medio de grandes apuros económicos toma a su cargo la publicación, puesto que las finanzas de la S. R. estaban, como siempre, en pésima situación y la promesa hecha a Newton, vía Halley, era muy explícita. Si ahora no se cumplía con ella, sólo Dios sabe cómo iba a reaccionar el hermético Newton. Halley, además, era consciente de la joya que tenía entre manos, pues hasta parece que había colaborado con Newton haciendo correcciones y observaciones sobre un borrador al que *sir* Isaac le dio acceso y, por último, si su dinero no se lo permitía, su cuna le obligaba a cumplir como caballero en un asunto en que, por si fuera poco, él era el primer apasionado, rareza ésta de la nobleza inglesa que ya Fontenelle subrayó en su *Elogio* fúnebre de Newton.

Con la llegada de este primer libro se produce una correspondencia entre Halley y Newton en la que aparecen de soslayo las pretensiones de Hooke sobre la paternidad de las ideas centrales contenidas en el libro. Las iras de Newton no se hacen esperar y entre las consecuencias podemos señalar el retraso en la entrega de los otros dos libros y, quizá, la introducción de complejidades matemáticas, confesadas en el tercero, para mostrarle a aquel sempiterno reclamante quién era cada uno. Es claro que Newton se sintió perturbado y que trató de, sin faltar materialmente a la verdad, mostrar a Halley dónde estaban cada uno de ellos en 1678-1679. Pero curiosamente las cartas más explícitas sólo aparecieron a principios de este siglo[62]. Si fueron retiradas de la S. R. o del propio archivo de Newton por una mano bien guiada o temerosa de que dieran pie a mayor polémica es cosa un tanto discutida pero no imposible. En todo caso, por esto, junto con algunos otros pequeños incidentes más, retrasa Newton la entrega del Libro II hasta los primeros días de marzo de 1687 y el tercero hasta un mes más tarde, con lo cual sólo el 5 de julio de 1687 puede Halley comunicar a Newton que «por fin he concluido vuestro libro y espero que le agrade. Las últimas correcciones llegaron justo a tiempo para ser incluidas. Enviaré de su parte los ejemplares que desee a la S. R., a *mr*. Boyle, a *mr*. Paget, a *mr*. Flamsteed y a algún otro, si hay en la ciudad, que usted indique; y le he enviado para regalar a sus amigos de la Universidad 20 copias, las que encarecidamente le ruego que acepte»;

reparto este que hizo Humphrey siguiendo instrucciones de *sir* Isaac y alguno de los receptores (el doctor Babington del Trinity, por ejemplo) dijo que «tendría que estudiar siete años antes de comprender algo de él[63]». Pero donde sí fue rápidamente comprendido fue en los centros intelectuales del continente que dieron cuenta de su publicación en sendas recensiones o presentaciones. La primera, como no podía ser menos, es debida a Halley y aparece en las *Philosophical Transactions* de enero-marzo 1687 escrita antes de la aparición del mismo libro, si bien el número 186, que es el correspondiente a ese trimestre, se retrasó por estar el editor (Halley) ocupado en la impresión de la obra recensionada. No es necesario advertir que los términos de esta recensión son, además de esclarecedores para el lector, sumamente precisos y justos en la exposición del contenido de la obra. En segundo lugar apareció una recensión menos técnica, quizá debida a Locke en el vol. 8 de la *Bibliothéque Universelle*, de marzo de 1688, donde el autor traduce al francés los títulos de las secciones de los Libros I y II, pero al carecer de secciones el tercero se detiene en reseñar los hallazgos «más curiosos». Aunque también es elogiosa esta recensión no permite a los «matemáticos» comprender el alcance de la obra. Otra cosa bien diferente es la recensión publicada en las «*Acta Eruditorum*», en junio de 1688, de cuyo autor no se tiene noticia pero que sin duda captó muy completamente el contenido y valor de la obra, y es seguro que Leibniz conoció antes esta recensión que la obra misma, aunque era él con Huygens y pocos más los únicos que tal vez no necesitasen ayudas de terceros lectores para comprender y juzgar sobre sus méritos. De hecho en el *Journal des Savants* de agosto de 1688 aparece una nueva noticia de los *Principia*, con algunas críticas esta vez, orientadas hacia el carácter hipotético de las soluciones newtonianas, en especial en el «Sistema del Mundo», críticas que pudieron ser un factor para que Newton se decidiese a cambiar las «Hipótesis» de la primera edición por las «Reglas» en las siguientes, y acaso le ayudaron a pronunciar la enfática frase del Escolio General de «Hipotheses non fingo». Como quiera que ello sea, lo cierto es que el continente, tanto como las islas, se dieron por enterados de los *Principia*, y, como es natural, admiraciones y elogios aparte, empezó la crítica. Y no sólo la crítica menor de erratas o incluso errores, sino también la crítica de fondo, primero la procedente del cartesianismo dominante y luego la sistemática. De la primera fue Newton testigo, mientras que de la segunda bien se puede decir que fue, en parte, causante con haber puesto a punto la herramienta del análisis y del cálculo. Pero él no alcanzó a conocer los grandes avances del siglo XVIII desde Clairaut, D'Alembert y Lagrange a Laplace, Euler, etc.

Después de la publicación de los *Principia* su vida intelectual se hace poco a poco cansina y oficialista. En ese mismo año vive un episodio de enfrentamiento a Jaime II, rey católico, por asuntos de la Universidad, y empieza una especie de culto a su persona que se extiende por el Reino Unido y le convierte en una especie de Papa científico. Premia a sus amigos y persigue a sus adversarios o meros competidores, abandona Cambridge en 1696 por un puesto de director de la fábrica de moneda de

Londres, persigue a los falsificadores con saña incontenible, adula y cultiva a la nobleza, es hecho caballero por la reina Ana, enloquece unos meses en 1693-1694, preside la S. R. durante casi 20 años, responde a cartas y a desafíos matemáticos, polemiza con Leibniz sobre el descubrimiento del cálculo y otras cosas, es indiferente con la desgracia de algún amigo como Whiston, Fatio, etc. se hace más reservado en su credo unitario y más sociable con los que le interesa, etc., cosas todas que, en una vida tan larga y cargada de honores, poderes y odios, no es demasiado difícil ir acumulando[64]. Y entre todas esas cosas hay alguna que debemos recordar aquí: reeditó dos veces los *Principia*.

Apenas tres años después de la aparición de la primera edición ya aparecen correcciones, arreglos y sustituciones hechas por Newton en sus ejemplares o en hojas sueltas intercaladas, al igual que ensayos alternativos de algunas Proposiciones o pruebas de las mismas. Estos tanteos, debidos a su propia iniciativa o a la de algunos amigos del círculo newtoniano, como Fatio de Doullier, David Gregory, que hasta pensaron en convertirse en editores de la segunda edición, fueron conformando a lo largo de los años noventa un corpus de «adenda et corrigenda» que hizo a Newton pensar en una segunda edición. De hecho tal vez lo mencionó alguna vez, porque tanto Fatio como Gregory se sintieron indicados para ello. El caso fue que sólo hacia 1708 se decide a esta tarea, bajo la presión de Richard Bentley a la sazón Maestro del Trinity y casi seguro antiguo alumno de las «Lectiones» de Newton en la cátedra Lucasiana. Por qué Newton no designó primero a Fatio y luego a Gregory es cosa algo misteriosa y que sólo se puede adivinar por otros episodios con ambos. Quien vino a merecer su confianza fue finalmente un joven matemático que resultó un perspicaz conocedor de los *Principia*, y tan valioso ahora como Halley en la primera edición: Roger Cotes[65].

La importancia de los cambios introducidos es insinuada por Newton en el breve prefacio a la segunda edición pero Ball[66] hace un balance más explícito: «Tengo una lista manuscrita de las adiciones y variantes introducidas en la segunda edición; los cambios son muy numerosos, de hecho encuentro que de las 494 (i. e. 510-516) páginas de la primera edición, 397 se han modificado más o menos en la segunda edición». No se piense que todos estos cambios son significativos, ni mucho menos. Muchos son meros retoques estilísticos, quizá por el hecho de que Bentley era un magnífico latinista y en su entorno Cotes también se viese impulsado a ello. Para Ball «los cambios más importantes son el nuevo prefacio debido a Cotes, las Proposiciones sobre la resistencia de los fluidos, sección VII, Libro II, Proposición XXXIV-XL, la Proposición sobre la precesión de los equinoccios, Proposición XXXIX del Libro III, y las Proposiciones XLI-LXII del Libro III sobre la teoría de los cometas». Curiosamente Ball olvida en esta lista el Escolio General, el cambio en el inicio del Libro III de las Hipótesis por las Reglas, además de algunas sustituciones de redacciones anteriores por otras nuevas, reordenaciones de Proposiciones y

supresiones o inclusiones de algunos Corolarios, junto con el índice alfabético que ahora aparece por vez primera. Como resultado de la meticulosa lectura que Cotes hizo del original para imprenta y de las «pruebas de imprenta» que Newton previamente iba revisando, Cotes puede decirle a Newton en carta de 22 de diciembre de 1712 «Creo que la edición será muy correcta». No exageraba Cotes y bien puede decirse que el texto de los *Principia*, salvo algunas cosas de importancia secundaria, queda fijado con esta edición, en la que Newton intervino activamente y, en cierto modo, creativamente, tratando de asumir todos los problemas que con su propio examen de la primera y algunos procedentes de otros había llegado a considerar relevantes. El nombre de Cotes aparece, junto con el de Halley, Whiston, Clarke, vinculado al de Newton en otros problemas, tales como el informe sobre la forma de determinar la longitud en el océano, etc., pero desgraciadamente murió joven, a los 34 años, y mereció de Newton, cuando éste lo supo, la célebre frase «si Cotes hubiera vivido algo hubiésemos aprendido». Por otra parte, junto con Clarke, lo hallamos al lado de Newton en sus querellas filosóficas con Leibniz, al cual va dirigida una buena parte de su prefacio a la segunda edición, texto que Leibniz encontró «pleine d'aigreur». La correspondencia entre ambos, relativa a esta edición, no sólo sirvió a Newton para mejorar el texto que envió a imprenta, sino que le permitió entrever algunas de las cosas que anotó en sus ejemplares de la segunda edición y que sirvieron para mejorar la tercera. Valga como ejemplo la Regla IV añadida en la tercera edición y que responde a objeciones de Cotes a la redacción de la Proposición V del Libro III.

De la segunda edición aparecida en Cambridge en julio de 1713 se hicieron unos 750 ejemplares que desaparecieron del mercado con tal rapidez que en 1714 una asociación de libreros holandeses se decidió a hacer por su cuenta una reimpresión clandestina o no autorizada del texto de Cambridge, cosa repetida en 1724 añadiendo algunos tratados matemáticos de Newton, ya publicados antes en diferentes formas, y algunas cartas.

La manía correctora de Newton tampoco cesó con la segunda edición y prosiguió en su tarea de corregir, intercalar y precisar sus palabras. Hacia 1723, ya con 81 años, se decide a encomendar a H. Pemberton, un médico formado en Leyden y en París, y estudioso matemático, la tarea de una tercera edición. Pemberton resultó un editor cuidadoso y meticuloso pero carecía del vigor matemático de Halley y de Cotes, limitándose a hacer nimias sugerencias de estilo[67], a cuidar de las pruebas de imprenta y a introducir las modificaciones que Newton proponía. Cuando la edición estuvo completada en 1726, las diferencias con la segunda apenas son señaladas en cuanto a mejoras sustantivas. Se puede notar que aparecen incluidos algunos nuevos experimentos al final del Libro III (el 14 por ejemplo, que ya había sido publicado en las *Phil. Trans*. en 1719), algunos nuevos datos astronómicos, la Regla IV incluida ahora y el cambio del *Escolio Leibniziano* subsiguiente al Lema II del Libro II por otro en el que ni siquiera después de muerto concede a Leibniz un recuerdo a su

contribución a la puesta en circulación del cálculo. ¡Sus iras llegaban tan lejos! Quizá las 200 guineas con que obsequió a Pemberton por sus servicios como editor se debían tanto a su generosidad como a su agradecimiento por haberle dado la ocasión de borrar de su libro y de la gloria el nombre de aquel rival a quien pudo ver muerto, pero no humillado como había visto a Hooke. Hecho esto podía morir tranquilo. Cosa que no dejó para muy tarde. Apenas un año y el 20 de marzo de 1727 en su casa de la calle Martin Street, esquina a Leicester Square en el barrio londinense de Chelsea, murió aquel hombre a quien «se concedió llegar más cerca de los dioses que ningún mortal». Sepultado en la Abadía de Westminster inevitablemente presente para quien la visite[68], mereció que en la Torre de Londres se le fundiese una medalla apenas cuatro años después de su muerte con su cabeza en busto y una inscripción, «Félix cognoscere causas», en torno a ella. También el Trinity erigió su estatua en mármol blanco frente a la Capilla cuya inscripción, un verso de Virgilio, reza «*Qui genus humanum ingenio superávit*».

Aquí ofrecemos al lector una traducción de la obra que hizo prorrumpir a sus contemporáneos en tales cantos. Hemos utilizado como texto principal el presentado por Koyré-Cohen, también como texto de la tercera edición. No hemos intentado castellanizar por completo un texto latino tan profundamente idiosincrático como es éste, y por ello el lector hallará algunas expresiones algo violentas para el buen decir actual. Pero traducir a Newton, nos ha parecido, exigía respetar lo que de newtoniano y a la vez de su tiempo pudiera haber en su texto. Hemos procurado, pues, que sus palabras y su estilo se pudieran entrever, también en castellano.

E. Rada.

APÉNDICE BIBLIOGRÁFICO

I. LAS EDICIONES DE LOS *PRINCIPIA*

Las ediciones latinas:

Newton realizó tres ediciones de los *Principia*. La primera (E_1 en la clave establecida por I. B. Cohen) en 1687 en Londres. La segunda (E_2) 1713, en Cambridge. La tercera (E_3) 1726, en Londres. Pero, puesto que cada una dio lugar a varias reimpresiones y alcanzó una difusión más amplia que la edición original misma, las enunciaremos por familias, entendiendo que cada una constituye una familia de textos discrepante de las otras en varios puntos. No es necesario advertir que las discrepancias desde la primera a la tercera también muestran cierta evolución, de mayor o menor interés según los casos, de la opinión y enfoque de Newton en esos puntos. Por otra parte conviene tener presente que no todos los textos alternativos que Newton bosquejó para algunos de los puntos que le parecieron dignos de retoque fueron luego incluidos en el texto impreso.

Familia 1.ª:

1.0 (1687). Editor E. Halley. Londres. 250 a 350 ejemplares. Primera impresión.

1.1 (1687). Editor E. Halley. Londres. 50 ejemplares. Segunda impresión.

1.10 (1954). Reprint de 1.1. Londres, Dawson and Sons.

1.1 (1965). Reprint de 1.0. Bruselas. Ed. Culture et Civilisation.

Familia 2.ª:

2.0 (1713). Editor Roger Cotes. Cambridge. Aproximadamente 750 ejemplares.

2.1 (1714). Amsterdam. Sociedad de libreros (sobre 2.0).

2.2 (1723). Amsterdam. Sociedad de libreros. Con algunos tratados matemáticos y geométricos (2.1).

Familia 3.ª:

3.0 (1726). Editor H. Pemberton. Londres. 1000 + 250 + 50 ejemplares.

3.10 (1739-1742). Editores PP. Th. Le Seur y F. Jacquier, Ginebra, con comentarios (3.00). Llamada la «edición de los jesuitas» aunque los editores eran franciscanos «ex Gallicana Minimorum Familia».

3.11 (1760). En Colonia. Reprint de (3.10) con correcciones.

3.20 (1779-1792). Editor S. Horsley, *Opera quae extant omnia*. En 5 vol., Londres. Incluye también el «de Mundi Systemate», la «Theoria Lunae» y las «Lectiones opticae». Reprints recientes y disponibles.
3.12 (1780-1785). Editor J. Tessanek. Nuevas notas sobre (3.11) del editor.
3.13 (1822). En Glasgow. Reprint de (3.10) con revisiones de J. M. Wright.
3.13.1 (1833). En Glasgow. Reprint segunda reimpresión de (3.10).
3.1 (1871). En Glasgow. Reimpresión del texto sin notas ni comentarios con lo cual el texto vuelve a ser (3.00) aunque la tipografía y algunos detalles son de las ediciones (3.13). Fue preparada por *sir* William Thomson (Lord Kelvin) y Hugh Blackburn.
10.20.30 (1971-1972). Editores Alexander Koyre y I. B. Cohen. Cambridge Un. Press. Aparecen las variantes textuales producidas por Newton desde la primera (1.0) hasta la tercera (3.00) no sólo las impresas sino también las que escribió en sus copias y en hojas sueltas intercaladas dentro de las páginas en los respectivos lugares. Editan como texto principal el correspondiente a (3.00) y en notas a pie de página las variantes textuales. Se halla precedida de una Introducción y seguida de Apéndices. Tres volúmenes de formato in fol. 380 + XL, 547 + 548 a 916 págs. respectivamente.

Traducciones

Los *Principia* han sido objeto de traducciones varias a distintos idiomas hasta nuestros días. No entraremos en la descripción de las mismas. Sólo advertir que unas lo han sido desde el original latino de la tercera edición y otras desde la versión inglesa de 1729 o de 1934.

Versiones inglesas

1. (1729). Londres, por Andrew Motte. John Machin (vide Escolio a la Proposición XXXIII del Libro III) añade las leyes del movimiento lunar.
2. (1777). Londres. Robert Thorp. Solamente se publica la traducción del Libro I. Por alguna razón no encontró editor para los otros dos, aún sin publicar.
3. (1802). Londres. Reimpresión de (2). Tampoco se publican los Libros II y III.
4. (1803). Segunda edición de (1). Londres. Incluye, además de una traducción del «Sistema del Mundo», una breve biografía de Newton.
5. (1819). Londres. Reimpresión de (4).
6. (1848). Nueva York. Primera edición americana sobre (4) con algunos cambios. De esta edición se hacen hasta tres reimpresiones hasta 1850. Quizá una cuarta para exportar al mundo de habla inglesa.
7. (1934). California, Berkeley, U. of C. Press. Florian Cajori revisa la traducción de Motte (1). Retiene el «Sistema del Mundo» y añade un apéndice histórico. Se reedita en 1934, 1946, 1947, 1952, 1960, 1962.
8. (1964). Nueva York. Foto-reprint de (6) y de (6-1850) con introducción de A. del Vecchio, eliminando los prefacios de Newton y Cotes.
9. (1968). Londres. Foto-reprint de (1), con introducción de I. B. Cohen.
10. (1969). Londres. Foto-reprint de (2), con introducción de I. B. Cohen.

Existen, además, una treintena larga de ediciones parciales, repertorios y extractos, etc., generalmente preparados para estudiantes en el siglo XIX, ya como libros de texto, ya como ilustración de otras cuestiones. Es curioso que breves textos de los *Principia* aparezcan desde 1715 apoyando panfletos de variada índole.

Versiones francesas

1. (1756). Debida a *madame* la Marquise du Chastellet. París. Con la probable ayuda de Voltaire y de Clairaut. Se conservan algunas, impresiones de prueba de 1756 y 1757-1759. Por diferentes dificultades en la realización de las planchas, no se hizo la edición definitiva para venta hasta 1759.
2. (1966). París. Foto-reprint de (1) con múltiples alteraciones procedentes de otras copias. Carece de aclaraciones sobre los cambios y sobre los originales utilizados. Librerie Se. et Tecnique A. Blanchard.

Versiones alemanas

1. (1872). Berlín. Debida a J. Ph. Wolfers, con introducción y notas.
2. (1963). Darmstadt. Foto-reprint de (1).

Italiana

(1966). Turín. Coordinada por Alberto Pala. Con una introducción y una biografía de Newton.

Japonesa

(1930). Tokyo. Debida a Kunio Oka.

Rusas

(1915-1916). San Petersburgo. Debida al lugarteniente general de la Marina A. N.

Kriloff. Reeditada en 1936 en Moscú y Leningrado.

Sueca

(1927-1931). Debida a C. V. L. Charlier. Lund. Con apéndices analíticos.

Rumana

(1956). Debida a Víctor Mariam y revisada por Víctor Valcovici. Academia.

Españolas

(1982). Debida a Antonio Escohotado y otros, con introducción, incluye el «Sistema del Mundo»; parte del texto Motte-Cajori y tiene a la vista el texto latino. Madrid. Editora Nacional.

Omitimos la reseña de fragmentos, repertorios y florilegios de los *Principia*. En traducciones directas o indirectas son demasiados para incluirlos aquí. Además, muchas veces, son fragmentos que no indican el origen de donde han sido tomados. Otra cosa sería evaluar el grado de conocimiento que, merced a ellos, se ha llegado a tener de los *Principia*.

II. LA OBRA DE NEWTON

1. Ensayos y obras menores publicados en su época

En las *Philosophical transactions*:

1672. Febrero, núm. 80, vol. 6, 3075-3087. «*A Letter… Containing his New Theory about Light and Colors*…». Primero publicado sobre el tema.
1672. Marzo, núm. 81, vol. 7, 4004-4010. «An accompt of a New Catadioptrical Telescope…».
1672. Abril, núm. 82, vol. 7, 4032-4034. «*Letter… containing some more suggestions about his New Telescope*».
1672. Abril, núm. 82, 4034-4035. «Answer to… an ingenious French Philosopher». (A. Auzout).
1672. Mayo, núm. 83, vol. 7, 4056-4059. «*Considerations upon part of a Letter… concerning… Telescope*». Respuesta a M. de Berce.
1672. Mayo, núm. 83, 4059-4062. «Some experiments proposed by… (MORAY) … Observations… by the Author of that Theory April 13-1672».
1672. Junio, núm. 84, vol. 7, 4091-4093. «… Letter… being an Answer to Latin Letter… of» (PARDIES).
1672. Julio, núm. 85, vol. 7, 5004-5007. «A serie of Quere's… to be determined by Experiments… concluding his new Theory of Light and Colors…».
1672. Julio, núm. 85, vol. 7, 5014-5018. «Answer… to A second Letter from Paris…» (PARDIES).
1672. Noviembre, núm. 88, vol. 7, 5084-5103. «Answer to some considerations upon his doctrine…». (A Hooke cuya crítica había sido leída, pero no publicada, en la S. R.).
1673. Julio, núm. 96, vol. 8, 6087-6092. «Answer to the foregoing Letter (de Huygens) further explaining…».
1673. Octubre, núm. 97, vol. 8, 6108-6111. «Letter from Cambridge April 3-1673 … concerning the number of Colors…».
1674. Enero, núm. 110, vol. 9, 219-225. «An Answer to (Francis Hall) A Letter…».
1675. Noviembre, núm. 121, vol. 10, 501-502. «… Considerations on (Linus) A Letter… from Liege…».
1676. Enero, núm. 121, vol. 10, 503-504. «Another Letter (de J. Gasconies) relating to te same argument».
1676. Marzo, núm. 123, vol. 11, 556-561. «A particular Answer to… (Linus) …from Cambridge… Febr. 29-1676».
1676. Septiembre, núm. 128, vol. 11, 698-705. «Answer to…» (Lucas).
1676. Diciembre. Se lee su segundo ensayo «on light and colors» en la S. R., pero no se publica en las Phil. Trans.
1701. Marzo, núm. 270, vol. 22, 824-829. «Scala Graduum Caloris. Calorum Descriptiones et Signa» (publicado anónimamente).
1715. Enero, núm. 342, vol. 29, 173-224. «An account of the Book entiruld *Commercium Epistolicum* Collini et…» (publicado anónimamente).
1721. Enero, núm. 368, vol. 31, 172. «Tabula Refractionum» (anónimo).

Póstumamente

1742. Enero, núm. 465. «An Instrument for observing the Moon's Distance from the Fixt Stars at Sea».

2. Publicaciones menores debidas a otros

Harris, J. (editor) (1710): «De Natura accidorum» y «Some Thoughts about nature of Acids», en *Lexicón Tecnichon*, vol. II, Londres.

Conduitt, J. y Smith, B. (1733): «Observations upon the prophecies of Daniel and the Apocalypse of St. John». Londres.

Greaves, John (1738): «The second Cubit of the Jews», en *Miscellaneus Works*. Londres.
Birch, Th. (1744): «Letters to Boyle», en *Works of Boyle*, vol. I, Londres.
Dodsley, R. y J. (1752): «Four Letters… to Dtr. Bentley…». Londres.
Birch, Th. (1757): «Second Paper on Color and Light, read at the R. S. 1675-1676», en *The History of the R. S. of London*, vol. III, Londres.

3. Obras mayores

3.1

1687. 1.ª ed. *Philosophiae Naturalis Principia Mathematica*. Londres. Edit. Halley, Edmond. Jusu Regiae Societatis.
1713. 2.ª ed. Cambridge, Univ. Press. Ed. Cotes, Roger.
1726. 3.ª ed. Londres. Ed. Pemberton, Henry.

3.2

1704. 1.ª ed. *Optiks*: Or, a Treatise of the Reflexions, Refractions, Inflexions and Colors of Light. Also Two Treatises of the Species and Magnitude of Curvilinear Figures. Londres. Se reimprime en 1704, 1717 (con nuevas Cuestiones) 1718, 1721 y 1730.
1706. 1.ª ed. latina. *Optice*: Sive de Reflexionibus, Refractionibus, Inflexionibus et Coloribus Lucís *Libri Tres*. (Incluye «Ennumeratio» y «Tractatus»). Traducción de Samuel Clarke, Londres. La edición estuvo al cuidado de A. de Moivre. Se reimprime en 1706 y (sin tratados) en 1719.

3.3

1707. 1.ª ed. *Arithmetica Universalis*: Sive de compositione et Resolutione Arithmetica Líber, cui accessit Halleiana… Methodus. Univ. Press, Cambridge. (Editado por W. Whiston).
1722. 2.ª ed. Revisión de Newton. Aunque se atribuye a J. Machín. Se omite el artículo de Halley.
1720. 1.ª ed. inglesa. *Universal Arithmetic.*…, etc. Trad. de Jos. Raphson, Londres. Con el artículo de Halley.
1728. 2.ª ed. inglesa. Con correcciones. Sin el artículo de Halley, Londres…
1736 (†postum). *The method of fluxions and infinite series with i'ts aplication to the geometry of curve-lines*… Traducción de W. Jones, Londres.

1737. *A treatise of the method of fluxions and infinite series.*…, etc. Londres (anónimo). (Ambas traducciones de un mismo manuscrito incompleto).

3.4

1712. 1.ª ed. *Commercium epistolicum D. Johnnis Clinns et aliorum de analysi promota; Jusu Societatis Regiae in lucem editum*. Londres. (La comisión de redacción estaba presidida por J. Keill. Los materiales proporcionados por Newton). Se reimprime en 1712, en 1722 (con adiciones de Newton) seis veces, y en 1725, tres veces.

3.5

1728. *A treatise of the System of the World*. Versión inglesa anónima, Londres.
1728. *De mundi systemate liber*. Londres.
1728. Optical lectures read in the Publick Schools of Cambridge, Anno Domini 1669. Traducción anónima, Londres.
1729. The mathematical principies of natural philosophy. Traducción de Andrew Motte, Londres.
1779. *Opera quae existant omnia*, 5 vol., Londres.

3.6

1725. *Abridged chronology… et observados.*…, por Nicolás Freret. (Traducido al francés y obtenido de modo fraudulento provocó la respuesta de Newton en la siguiente), París.
1725. *Remarks apon the observations made… translated into French ly*… en. Ph. Transact. 389, vol. 33, 315-321.
1728. *The chronology of ancient kingdoms amended…* (Editado por Conduitt, Londres. Otras tres reimpresiones en ese año).

4. Newton editor

Barrow, I. (1669-1670): *Lections XVIII Cantabrigiae in scholis… in qubus opticorum phaenomenon genuinae rationes investigantur*. Londres.
Varen, Bernhard (1672): *Geographia generalis… aucta et illustrata ab Isaaco Newton*. Cambridge.

5. Manuscritos. Localización actual de… (Los más importantes).

Los dividimos en seis grandes grupos, el último subdividido en subgrupos.

1. U. C. L. (Univ. of Cambridge Library) con tres bloques:

a) Manuscritos depositados por Newton como textos de sus lecciones de cátedra Lucasiana.

b) Manuscritos donados por la familia Portsmouth.

c) Manuscritos adquiridos de otras fuentes.

2. Colección *Keynes*. Depositada en el King's College de Cambridge. Procede de la subasta de 1936 realizada por el vizconde de Lymington. Contiene algunas cartas, notas de Conduitt, algunas notas de Newton sobre los problemas universitarios de 1687-1688 y amplia documentación alquímica.

3. Yahuda Col. En la Universidad Nacional de Jerusalén. Escritos teológicos y de historia religiosa.

4. Mint Col. Devueltos al Mint y hoy depositados en el Public Record Office. Relacionados con la actividad en el Mint.

5. Bodleiana Col. En el New College de Oxford. Principalmente escritos teológicos, exégesis y cronología.

6. Otras colecciones:

6.1. Royal Society Col. Fondos institucionales, cartas, y relacionados con los *Principia*.

6.2. Trinity College. Documentos institucionales.

6.3. Babson College (Mass.).

6.4. Dibner Col. en el Instituto Smithsoniano.

6.5. Conde de Macclesfield. Separada antes de la muerte de Newton por W. Jones. No abierta al público.

6.6. Fitzwilliam Museum. Cambridge. Un Note Book.

6.7. Documentos sueltos en Banco de Londres, British Library, Clitton College de Bristol. Iron and Steel Institute, G. L. Keynes. Lrd. Lymington, Magdalen College de Oxford, Univ. de St. Andrews.

6.8. En Europa continental.

6.8.1. Colección Fabius en París. No abierta al público.

6.8.2. Colección Bodmer en Ginebra. No abierta al público.

6.8.3. Academia Soviética. Disponible.

6.9. En EE. UU. de América y otros.

6.9.1. Schaffner Col. Universidad de Chicago.

6.9.2. Pierpont Morgan Library. Nueva York.

6.9.3. Standord University Library.

6.9.4. University of Texas Library. Humanities Center.

6.9.5. Yale University.

6.9.6. A. Ph. S. Philadelphia, Clark Memorial de Los Ángeles, Columbia University, Countway Medical Center de Boston, Historical Society de Pensilvania, Huntington Library de California, Lenigh University de Pensilvania, M. I. T. Library Mass. Mount Wilson Observatory de Pasadena, Bibliot. Publi. de Nueva York, Universidad de Princeton… y alguno más en Australia.

III. EDICIONES RECIENTES DE LA OBRA DE NEWTON (Generalmente disponibles en el mercado).

Edleston, J. (editor) (1850): *Correspondence of Sir Isaac Newton and Professor Cotes, including letters of other eminent men*… Londres, John W. Parker y Cambridge, John Deighton.

Cohen, I. B. (editor) (1958): *Isaac Newton's papers and letters on Natural Philosophy*. Mass. Cambridge U. Press.

Turnbull, J. F. et allii (editores) (1959-1977): *The correspondence of Isaac Newton* (7 vol.). Cambridge U. Press. For the Royal Society.

Hall, A. R. y Boas M. (editores) (1962): *Unpublished scientific papers of Isaac Newton*. Cambridge U. Press.

Whiteside, D. T. (editor) (1964-1967): *The mathematical works of I. Newton*. N. Y. Johnson Reprint Corporation.

Whiteside, D. T. (editor) (1967-1980): *The mathematical papers of Sir Isaac Newton* (8 vol.). Cambridge U. Press.

Cohen, I. B. (editor) (1969): *A treatise of the System of the World*. Londres, Revisión de la edición de Conduitt J. de 1728.

Koyre, A. y Cohen, I. B. (editores) (1972): *Philosophiae Naturalis Principia Mathematica* (2 vol. + 1 vol. *Introducción*, de Cohen). Cambridge U. Press.

Whiteside, D. T. (editor) (1973): *Isaac Newton's Cambridge lectures on optics* 1670-1672. Facsímil del manuscrito 4002 de C. U. L. Cambridge U. Library.

McGuire, J. E. y Tamny Mertin (editores) (1983): *Certain philosophical questions: Newton's Triniiy Notebook*. Cambridge U. Press.

Shapiro Alan, E. (editor) (1984-1985): *The optical papers of Isaac Newton* (2 vol.). Cambridge U. Piess.

Traducciones castellanas:

Solís, C. (1977): *Optica o tratado de las Reflexiones, Refracciones, Inflexiones y Colores de la luz*. Madrid, Alfaguara. Con Introducción y notas.

Escohotado, A. y Saenz de Heredia, M. (1982): *Principios matemáticos de Filosofía Natural* (edición preparada por A. Escohotado). Madrid, Editora Nacional. Con Introducción.

Rada, E. (1983): *El Sistema del Mundo*. Trad., Introducción y notas. Madrid, Alianza Editorial (2.ª reimpresión, 1986).

IV. REPERTORIOS Y BIBLIOGRAFÍAS

Babson Catalog (1950): *A descriptive catalogue of the Grace K. Babson Collection of the works of Sir I. Newton* (y Suplemento 1955). Mass.

Cohen, I. B. (1960): «Recent scholarship». (*ISIS*), 51, 489-514.

Gray, George (1907): *A bibliography of the works of Sir I. Newton, together with a list of books ilustrating his works*. Cambridge U. Press.

Pighetti, Clelia (1960): «Cinquant'anni di studi newtoniani», *Rivista critica di Storia della Fia*, 15, 181-203.

Portsmouth Ctal (1888): *Catalogue of the Portsmouth Collection*. Cambridge.

Sotheby Catal (1936): *Catalogue of the Newton Papers sold by order of the Viscount Lymington*. Londres.

Thackray, A. (editor) (1985): *ISIS Cumulative Index*. De 1953 a 1982 (vols. 44-73). Ultima salida. Univ. de Pennsylvania y History of S. S.

Wallis, Peter y Ruth (1977): *Newton and newtoniana* 1672-1975. Falkestone. Kent.

Westfall, R. S. (1976): «The changing world of the newtonian industry», *Jour of the hist. of ideas*, 37, 175-184.

Whiteside, D. T. (1962): «The expanding World of newtonian research», *History of Science*, 1, 16-29.

Whitrow, Magda (editor) (1971): *ISIS Cumulative Bibliography.*… [para la] History of S. Society. Vide entrada *Newton*. Londres y Univ. de Pennsylvania.

V. BIOGRAFÍAS Y ESTUDIOS CONEXOS

Andrade, E. N. da C. (1950): *Isaac Newton*. Londres.

Andrade, E. N. da C. (1954): *sir Isaac Newton*. Londres.

Brewster, D. (1855): *Memoirs of the Life Writings and Discoveries of Sir Isaac Newton* (2 vol.). Edimburgo. Reprint 1965; Johnson Rep. Corp. Londres y Nueva York.

Christianson, G. E. (1984): *In the presence of the Creator: Isaac Newton and his time*. Londres, Nueva York.

De Morgan Aug (1885): *Newton: his friend: and his niece*. Londres.

De Morgan Aug (1914): *Essays on the Ufe and work of Newton*. Chicago (edit. Ph. E. B. Jourdain).

Fontenelle Bernard le Bovier de (1728): «The elogium of Isaac Newton». Londres. La versión francesa apareció en 1733 en la *Histoire de L'Academie.*…

Foster, C. W. (1928): «*Sir Isaac Newton's family*», en *Reports and pap. of the architect. societies*…, 39, I, 1-62. Londres.

Harrison, J. R. (1978): *The library of Isaac Newton*. Cambridge.

Manuel Frank, E. (1968): *A portrait of Isaac Newton*. Cambridge, Mass.

Manuel Frank, E. (1974): *The Religión of Isaac Newton*. Oxford.

More Louis Trenchard (1934): *Isaac Newton: A biography*. Nueva York.

Stukeley, William (1936): *Memoirs of Isaac Newton's life*. A Hastings White, Londres.

Sullivan, J. W. N. (1938): *Isaac Newton*; 1642-1727. Londres.

Turnor, Edmund (editor) (1806): *Collections for the history of the town and soke of Grantham*. Londres.

Villamil, Richard (1931): *Newton: the man*. Londres.

Westfall, R. S. (1980): *Never at rest: a biography of Isaac Newton*. Cambridge, Nueva York.

Whiteside, D. T. (1966): «Newton's marvellous year: 1966 and all that», en *Not. and Rec. of Royal Society*, 21, 32-41.

VI. OBRAS GENERALES

Aiton, E. J. (1972): *The vortex theory of planetary motions. Londres y Nueva York*.

Ball. W. W. Rouse (1893): *An essay on Newton's Principia*. Londres (reimpreso por Johnson Rep. Corp., 1972, Londres y Nueva York).

Baylu, Francis (1835): *An account of the Rvd. John Flamstee.*… (Reproducción fotográfica en Dawson of Pall Malí, 1966, Londres).

Bechler, Zev (1982): *Contemporary newtonian research*. Reidel P. C. Dordrecht, Boston.

Birch, Th. (1757): *The history of the Royal Society of London for improving of natural knowledge*. Londres (reimpreso por Johnson Reprint Corp., 1968, Londres y Nueva York).

Bloch, León (1908): *La philosophie de Newton*. París.

Boyer, Carl B. (1939): *Concepts of the calculus: A critical and historical discussion of the derivative and the integral*. Colmb. U. Press, Nueva York.

Boyer, Carl B. (1956): *History of analytic geometry*. Scrip. Math., Nueva York.

Broad, C. D. (1927): *Sir Isaac Newton*. Londres.

Brougham, H. y Routh, E. J. (1855): *Analytical view of Sir Isaac Newton's Principia*. Londres (reimpreso por Johnson Reprint Corp., 1972, Londres y Nueva York).

Brunet, Pierre (1931): *L'introduction des theories de Newton en Trance au XVIII siécle*. A Blanchard, París, vol. 1.

Buchdahl, Gerd (1969): *Metaphysics and the philosophy of Science: the clasical origins. Descartes to Kant*. B. Blackwell, Oxford.

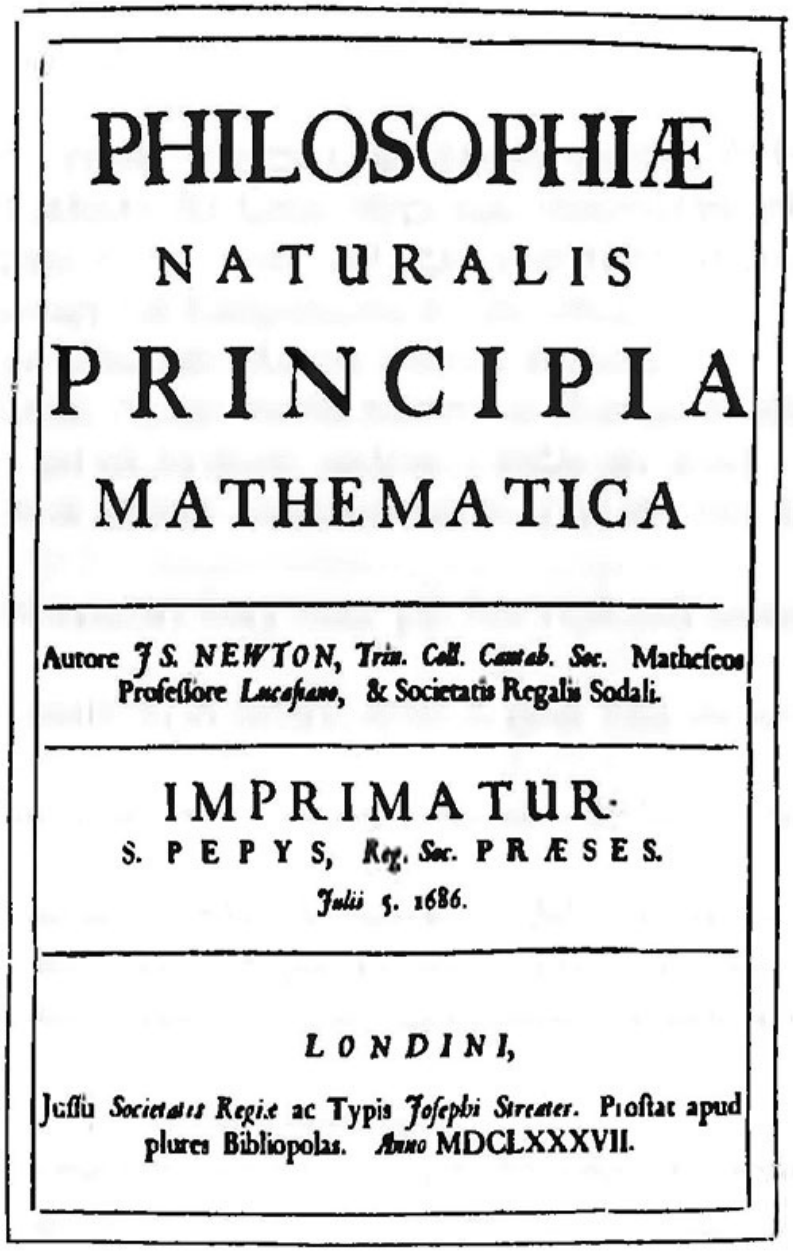
Burke, J. G. (editor) (1983): *The uses of Science in the age of Newton*. U. Calif. Press, Los Ángeles.

Burtt, E. A. (1925): *The metaphysical foundations of modern physical Science*. Nueva York (versión castellana en Ed. Sudamericana, Buenos Aires, 1950).

Castillejo, David (1981): *The expanding force in Newton's cosmos*. Ediciones de Arte y Cultura, Madrid.

Cohen, I. B. (1956): *Franklin and Newton; an inquiry…* Philadelphia, 2.ª (1966).
Cohen, I. B. (1960): *The birth of a new physics*.
Cohen, I. B. (1971): *Introduction to Newton's Principia. Cambridge*.
Cohen, I. B. (1980): *The newtonian revolution*. Cambridge U. Press, Cambridge (hay trad. cast. en Alianza Edit., Madrid, 1984).
Collins, John y otros (1856): *Commercium epistolicum J. Collins et aliorum de analysi promota, etc*. Edición de Biot y Lefort, Mallet-Bachelier, París.
Craig, J. (1946): *Newton at the mint*. Cambridge.
D'Alembert, Jean le Rond (1743): *Traite de Dynamique…* David l'aine, París. Reimpreso 1967. Cult. et Civilisation, Bruselas.
Dobbs, B. J. T. (1975): *The foundations of Newton's Principia; the hunting of the geene Lyon*. Cambridge.
Dugas, Rene (1955): *A history of machanics* (trad. ing. de J. R. Maddox). Grifton, Neuchatel.
Edwards (jnr.), C. H. (1979): *The historical development of the calculus*. Springer-Verlag, Nueva York, nueva edición, 1982.
Elkana, Y. (1974): *The Interaction between Science and philosophy*. N. Jersey.
Espinase, Margaret (1956): *Robert Hooke*. W. Heinemann, Londres. Euler, Leonhard (1736): *Mechanica sive motus scientia analytice expósita*. 2 vol. San Petersburgo. Reeditada en Opera Omnia, 1911. Sociedad Suiza de Ciencias Naturales. Teubner, Leipzig y Berlín.
Fatio de Duillier, Nicholas (1949): *De la cuase de la pesanteur*. Reconstruida y editada por B. Gag nebin en Notes and Records de la R. S., vol. 6.
Gillispie, Charles C. (1960): *The edge of objectivity; an essay in the history of scientific ideas*. Princeton U. Press.
Greenstreet, W. J. (editor) (1927): *Isaac Newton* 1642-1727. Colectivo memorial. Bell and Sons, Londres.
Guerlac, H. (1963): *Newton et Epicure*. París.
Guerlac, Henry (1981): *Newton on the Continent*. Cornell U. Press, Ithaca y Londres.
Hahn, Roger (1971): *The anatomy of a scientific institution; the París Academy of Sciences* 1666-1803. U. of Calif. Press, Los Ángeles.
Hall, A. R. (1954): *Scientific revolution*. Londres.
Hall, A. R. (1963): *From Galileo to Newton*. Londres.
Hall, A. R. (1980): *Philosophers at war*. Cambridge U. Press.
Halley, Edmond (1932): *Correspondence and papers of Edmond Halley*. Editado por E. F. MacPike. Clarendon Press, Oxford.
Herivel, J. (1965): *The background to Newton's Principia. A study of Newtorís dynamical resarches in the years* 1664-1684. Clarendon Press. Oxford.
Hesse, M. B. (1965): *Forces and fields*. Littlefield, Adams & Co., Totowa, New Jersey.
Hessen, Boris (1971): *The social and economic roots of Newton's Principia*. Reimp. del texto de 1931 con nueva introducción de Cohen I. B. Howard Fertig, Nueva York.
Hill, Christopher (1961): *The century of revolution* 1603-1714. Nelson and Sons, Londres (hay trad. castellana en Ayuso, Madrid, 1972). Holton, Gerald (1978): *The scientific imagination: case studies*. Cambridge U. Press.
Jacob, Margaret (1976): *The newtonians and the english revolution* 1689-1720. Ythaca, Nueva York.
Jammer, Max (1957): *Concepts of forcé*. Harvard U. Press, Mass.
Knight, David (1975): *Sources for the history of Science*, 1660-1914. Cornell U. Press, Ithaca.
Koyre, A. (1957): *From the closed world to the infinite universe*. Johns Hopkins U. P. (versión castellana en Siglo XXI, Madrid, 1979).
Koyre, A. (1965): *Newtonian studies*. Cambridge, Mass. (edición francesa, Gallimard, París, 1968).
Koyre, A. (1973): *Chute des corps et mouvement de la terre de Kepler a Newton*. (Trad. del inglés, Vrin, París. Primera publicación en Transactions de *Amer. Phil. Soc*., 1955).
Lagrange, Joseph Louis (1788): *Mechanique analytique*. Desaint Libraire, París.
Leibniz, G. W. (1973): *Marginalia in Newtoni Principia Mathematica*. (Editado por E. A. Fellman y traducción francesa de J. F. Courtine, Vrin, París).
Leibniz, G. Q. y Clarkc, S. (1980): *La polémica Leihniz-Clarke*. Taurus, Madrid (ed. E. Rada).
MacPike, E. F. (1937): *Hevelius, Flamsteed and Halley three contemporary astronomers and their mutual relations*. Taylor y Francis, Londres.
MacPike, E. F. (1939): *Dtr. Edmond Halley* (1656-1742). Taylor y Francis, Londres.
Marcolongo, R. (1919): *Il problema dei tre corpi da Newton* (1686) *ai nostri giorni*. Ulrico Hoepli, Milán.
McMullin, E. (1977): *Matter and activity in Newton*. Notre Dame U., Indiana.
Palter, Robert (editor) (1970): *The annus mirabilis of Sir Isaac Newton* 1666-1966. M. I. T. Press, Cambridge, Mass. y Londres.
Patterson, Louise Diehl (1949): «Hooke's gravitation theory and its influence on Newton», en *ISSIS*, vols. 40 y 41 (1949 y 1950).
Pemberton, Henry (1728): *A view of Sir Isaac Newton's philosophy*. S. Palmer, Londres.
Rigaud, S. P. (1838): *Historical essay on the first publication of Sir Isaac Newton's Principia*. Oxford.
Rigaud, Stephen Peter (1841): *Correspondence of scientifie men… in the collection of… earl of macclesfield*, 2 vol. Oxford U. Press.
Rosenberger, Ferdinand (1895): *Isaac Newton und seine physicalishen principien. Leipzig*.
Royal, Society: *Notes and records of the Royal Society*. Londres.

Royal, Society (1947): *Newton tercentenery celebrations*, 15-19 *july* 1946. Cambridge U. Press.
Sabra, A. I. (1967): *Theories of light from Descartes to Newton*. Londres.
Scott, Wilson L. (1970): *The conflict between atomism and conservartion theory 1644 to 1860*. MacDonald, Londres.
Snow, A. L. (1926): *Matter and gravity in Newton's physical philosophy*. Londres.
Solís, Carlos (1982): *La filosofía de la ciencia de Newton*. En Seminario de Teoría de la Ciencia. Ed. Univ., Salamanca, págs. 77-112.
Thackray, A. (1970): *Atoms and powers; an essay on newtonian matter-theory and the developement of chemistry*. Cambridge, Mass.
Thomas, E. R. (1934): *Newton and the origin of colors; a study of one of the earliest examples of scientific method*. Londres.
Truesdell, Clifford (1968): *Essays in the history of mecanics*. Springer-Verlag, Nueva York (hay traducción castellana en E. Tecnos, 1975, Madrid).
Turnbull, Herbert W. (1945): *The mathematical discoveries of Newton*. Blackie and Son, Londres.
Westfall, Richard S. (1971): *Force in Newton's physics: the Science of dynamics in the seventheenth century*. Macdonald, Londres.
Westman, R. S. y McGuire, J. E. (1977): *Hermeticism and the scientific revolution*. U. of California, Los Ángeles.



PHILOSOPHIÆ
NATURALIS
PRINCIPIA
MATHEMATICA

Autore *J S. NEWTON, Trin. Coll. Cantab. Soc.* Matheſeos Profeſſore *Lucaſiano*, & Societatis Regalis Sodali.

IMPRIMATUR.
S. PEPYS, *Reg. Soc.* PRÆSES.
Julii 5. 1686.

LONDINI,
Juſſu *Societatis Regiæ* ac Typis *Joſephi Streater*. Proſtat apud plures Bibliopolas. *Anno* MDCLXXXVII.

Cabecera y título (Primera edición)
Original en el Trinity College, Cambridge (C. 15. 158).

A esta obra físico-matemática
del muy ilustre varón Isaac Newton,
honra insigne de nuestro siglo
y de nuestro pueblo[1]

He aquí la Ley del Universo, las divinas medidas de la masa,
He aquí el cálculo del Cielo; leyes que, mientras establecía
Los principios de las cosas, el Creador de todo no quiso violar,
Y así establecer los fundamentos de las obras.
Se abren del cielo vencido los últimos arcanos,
Y no se oculta ya qué fuerza mueve los últimos círculos.
Sentado el Sol en su trono ordena a todas las cosas
Dirigirse hacia El con rápido descenso, y ya no deja a los carros
Celestes moverse en línea recta por los inmensos espacios vacíos;
Sino que, siendo El el centro, atrae a cada cosa en giros inmutables.
Ya está claro cual sea el tortuoso camino de los horribles cometas;
Ni ya nos causa asombro la aparición del astro con cabellera.
Al fin aquí sabemos por qué avanza la plateada Luna
Con pasos desiguales; por qué, hasta ahora rebelde a los astrónomos,
Rechaza el freno de los números,
Por qué regresan los nodos, por qué los auges se adelantan.
Y también podemos saber cuán grande es la fuerza
Con la que la errante Luna empuja el flujo del mar
Cuando con quebradas olas abandona las Ovas
Y desnuda las arenas, peligro de los navegantes,
Lanzándolas una y otra vez a la cima de las costas.
Cosas que tantas veces han torturado a los Sabios antiguos
Y que en vano torturan a las Escuelas con ronca contienda
Las vemos claras ahora matemáticamente desveladas.
Ya el error con su niebla no aplasta a quienes
La sublime agudeza del genio concedió
Entrar en la morada de los dioses y escalar las alturas del Cielo.
Levantaos mortales, desechad los terrenos cuidados
Y distinguid desde ahora las fuerzas de la mente divina
Tan amplia y largamente distante de la vida de las bestias.
Quien ordenó en tablas escritas castigar las muertes,
Robos, adulterios y crímenes de perjurio y fraude,
quien había aconsejado a los pueblos errantes

Rodear las ciudades de altas murallas, era un sabio;
O quien alegró a las gentes con el don de Ceres,
O sacó de las uvas consuelo en las penas,
O enseñó a juntar diferentes sonidos Pintados en una caña del Nilo,
Y a transformar en signos visibles las voces distintas,
Explicó menos la suerte de los hombres; de modo que
Sólo consideró unas pocas necesidades de la vida.
Pero ya somos admitidos en convite a la mesa de los dioses,
Ya podemos manejar las leyes superiores del Universo
Y ya se abren los ocultos misterios de la oscura Tierra,
El orden inmóvil de las cosas y los secretos
Que ocultaron los siglos pasados.
Vosotros, los que gozáis del néctar celeste,
Celebrad conmigo a quien tales cosas nos muestra,
A Newton que abre el cerrado cofre de la verdad,
A Newton, amado de las musas, en cuyo limpio pecho
Habita Febo, de cuya mente se apoderó con todo su Numen;
Pues no está permitido a un mortal tocar más de cerca a los dioses.

Edmundo Halley

Prefacio del autor al lector

Habiendo los antiguos, en la investigación de la naturaleza, practicado sobre todo la Mecánica (como lo hizo *Pappo*), y como los más modernos, desechadas ya las formas sustanciales y las cualidades ocultas, hubiesen intentado reducir los fenómenos de la naturaleza a leyes matemáticas, nos parece oportuno tratar en esta obra la parte Matemática que se relaciona con la Filosofía. Los antiguos establecieron dos mecánicas: la Racional, que procede por demostraciones exactas, y la Práctica. A la práctica pertenecen todas las artes manuales de las que propiamente toma el nombre de Mecánica. Pero como los artesanos suelen proceder con escasa exactitud, ocurre que la Mecánica entera se distingue de la Geometría de tal modo que lo que se hace con exactitud se asimila a la Geometría y lo que se hace con poca exactitud a la Mecánica. Sin embargo, los errores no pertenecen a las artes, sino a los artífices. El que construye con menor exactitud es un mecánico más imperfecto, y si alguien pudiese construir con absoluta exactitud, este tal sería el más perfecto mecánico. Así pues, los trazados de las líneas rectas y curvas en que se apoya la Geometría pertenecen a la Mecánica. La Geometría no enseña a trazar estas líneas, sino que lo postula. Postula que el aprendiz procure trazarlas exactamente antes de alcanzar el límite de la Geometría; después enseña cómo se resuelven los problemas mediante estas operaciones, puesto que trazar rectas y círculos es cuestión problemática pero no geométrica. Desde la Mecánica se postula su solución, mientras en Geometría se enseña el uso de las soluciones. Y bien puede gloriarse la Geometría de que de tan pocos principios postulados de otro sitio logre tan grandes resultados. Se funda, pues, la Geometría en la práctica mecánica y no es otra cosa que aquella parte de la Mecánica universal que propone y demuestra con exactitud el arte de medir. Mas, como las artes manuales se cifran ante todo en mover los cuerpos, ocurre que comúnmente se asocie la Geometría con la magnitud y la Mecánica con el movimiento. En este sentido la Mecánica racional será la Ciencia, propuesta y demostrada exactamente, de los movimientos que resultan de cualesquiera fuerzas y de las fuerzas que se requieren para cualesquiera movimientos. Esta parte de la Mecánica fue cultivada por los antiguos en las cinco fuerzas relativas a las artes manuales, los cuales apenas tuvieron en cuenta a la gravedad (pues no es fuerza manual) más que en los pesos a mover por dichas fuerzas. En cambio nosotros, cultivando no las artes, sino la filosofía, y escribiendo no de las fuerzas manuales,

sino de las naturales, tratamos sobre todo lo relativo a la gravedad, levedad, elasticidad, resistencia de los fluidos y fuerzas por el estilo, ya sean de atracción o de repulsión; y por ello proponemos estos nuestros como principios matemáticos de filosofía. Pues toda la dificultad de la filosofía parece consistir en que, a partir de los fenómenos del movimiento, investiguemos las fuerzas de la naturaleza y después desde estas fuerzas demostremos el resto de los fenómenos. A esto se refieren las proposiciones generales que tratamos en los Libros primero y segundo. En el libro tercero proponemos un ejemplo de esto con la explicación del sistema del mundo. Pues allí, a partir de los fenómenos celestes, por medio de proposiciones demostradas matemáticamente en los libros anteriores, se deducen las fuerzas de la gravedad por las que los cuerpos tienden hacia el Sol y a cada uno de los planetas. Después, a partir de estas fuerzas, también por proposiciones matemáticas, se deducen los movimientos de los planetas, cometas, Luna y mar. Ojalá que fuera posible deducir los demás fenómenos de la naturaleza a partir de principios mecánicos con el mismo género de argumentación, pues muchas cosas me mueven a sospechar que puedan depender todos ellos de ciertas fuerzas con las que las partículas de los cuerpos, por causas aún desconocidas, bien se atraen unas a otras uniéndose según figuras regulares, bien huyen y se separan unas de otras; y, siendo estas fuerzas desconocidas, en vano los filósofos hasta ahora intentaron acercarse a la naturaleza. Espero, sin embargo, que con este modo de filosofar o con otro mejor, los principios aquí enunciados añadan alguna luz.

En la edición de ellos tuvo capital importancia Edmundo Halley, varón muy sutil y erudito en toda clase de saberes, quien no solamente corrigió las pruebas de imprenta y supervisó el dibujo de las figuras, sino también fue causante de que me decidiera a su edición. Puesto que habiéndome pedido la figura demostrada por mí de las órbitas celestes, no cesó de rogarme que la comunicase a la Sociedad Real, la cual, después, con sus peticiones y benigna protección, hizo que yo empezara a pensar en darlo a la luz. Sin embargo, después de comprobar las desigualdades de los movimientos de la Luna e intentar otras cosas referentes a las leyes y medidas de la gravedad y otras fuerzas, así como a las figuras que habrían de describir cuerpos en mutua atracción según determinadas leyes, referentes también a movimientos de muchos cuerpos entre sí, a los movimientos de cuerpos en medios resistentes, a las fuerzas, a densidades y movimiento de los medios, referentes, por último, a las órbitas de los cometas y cosas por el estilo, pensé que debía aplazar la edición hasta más tarde para corregir todas estas cosas y darlas a la publicidad juntas. Por lo que respecta a los movimientos lunares (aunque imperfectos), los he reunido en los Corolarios de la Proposición LXVI, por no sentirme obligado a estudiar cada uno más prolijamente de lo que el caso requiere y demostrar uno a uno e interrumpir la serie de las demás proposiciones. Algunas cosas descubiertas más tarde decidí incluirlas en lugares menos idóneos antes que cambiar el número de las proposiciones y de las citas. Ruego encarecidamente que todo esto sea leído rectamente y que los defectos

en materia tan ardua no tanto sean vituperados cuanto investigados con nuevos intentos de los lectores y suplidos con benevolencia.

Cambridge,
colegio de la S. Trinidad.
8 de mayo de 1686[2]

Prefacio del autor a la segunda edición

En esta segunda edición de los Principios se han corregido muchas cosas aquí y allá y se han añadido otras. En la Sección II del Libro primero se hace más fácil y extenso el procedimiento de descubrir las fuerzas que permiten a los cuerpos girar en órbitas dadas. En la Sección VII del Libro segundo se indaga con más rigor y se confirma con nuevos experimentos la teoría de la resistencia de fluidos. En el Libro tercero se deduce de modo más completo de sus propios principios la teoría de la Luna y la precesión de los equinoccios, así como la teoría de los cometas aparece confirmada por muchos, y más exactamente calculados, ejemplos de trayectorias.

Isaac Newton

Londres, marzo, 28, 1713

Prefacio del editor a la segunda edición

Te ofrecemos, amable lector, una nueva y largamente deseada edición de la filosofía newtoniana ampliamente corregida y aumentada. Por los índices adjuntos podrás conocer los principales temas contenidos en esta famosa obra: los cambios y añadidos te los puede reseñar casi el propio Prefacio del Autor. Sólo falta que añadamos algo sobre el método de esta filosofía.

Podemos reducir a tres las clases de tratamientos que han abordado la Física. Hubo quienes atribuyeron a las diversas especies de cosas cualidades ocultas específicas, de las que hacían después depender las diferentes operaciones de cada cosa por una razón oculta y desconocida. En esto consistió toda la doctrina escolástica procedente de Aristóteles y los peripatéticos. Sostienen, en efecto, que los efectos particulares se siguen de las naturalezas particulares de los cuerpos; no enseñan, en cambio, de dónde proceden tales naturalezas y, por tanto, nada enseñan. Al centrarse únicamente en los nuevos nombres de las cosas y no en las cosas mismas, hay que pensar que han encontrado un cierto lenguaje filosófico, pero en cambio no han ofrecido filosofía alguna.

Otros hay que han creído conseguir, una vez desechado el fárrago inútil de palabras, un reconocimiento de su esfuerzo. Sostuvieron, pues, que toda la materia es homogénea y que toda la variedad de formas que se descubre en las cosas procede de ciertas propiedades simples y de fácil comprensión de las partículas componentes. Y correctamente se establece una progresión de lo más simple a lo más complejo si a dichas propiedades primarias de las partículas no se atribuyen otros modos que aquellos que la propia Naturaleza atribuye. Pero cuando se permiten establecer cualesquiera figuras o magnitudes desconocidas de las partes, lugares y movimientos inciertos, y hasta imaginar determinados fluidos ocultos que penetren libremente a través de los poros de los cuerpos, dotados de una omnipotente sutilidad, agitados de ocultos movimientos, entonces derivan hacia los sueños y abandonan la verdadera constitución de las cosas, la cual, en verdad, raramente puede obtenerse de falaciosas conjeturas, cuando apenas si es posible investigarla por medio de las más seguras observaciones. Quienes toman hipótesis como fundamento de sus especulaciones, aun cuando después procedan del modo más meticuloso de acuerdo con las leyes de la Mecánica, hay que decir que con seguridad componen una fábula elegante y graciosa quizá, pero fábula al fin.

Queda todavía una tercera clase, la de los que profesan la filosofía experimental. Estos pretenden que las causas de todas las cosas han de derivarse de los principios más simples que sea posible; y en el lugar de un principio no asumen jamás cosa alguna que todavía no se haya comprobado a partir de los fenómenos. No se imaginan hipótesis, ni las aceptan en física, a no ser como cuestiones cuya verdad se disputa. Proceden, pues, con un doble método: analítico y sintético. Se deducen las fuerzas de la naturaleza y sus leyes más simples mediante análisis a partir de ciertos fenómenos elegidos y después a partir de aquí se ofrece, mediante síntesis, la constitución de los demás. Esta es la manera de filosofar, con mucho la mejor, que, por encima de todas las demás, nuestro autor creyó que era preciso seguir. Y sólo a ésta creyó digna de ser cultivada y honrada con su propia obra. De ésta dio, pues, un magnífico ejemplo, cual es la explicación del sistema del mundo deducida con todo éxito de la teoría de la gravedad. Otros hubo que sospecharon o imaginaron que la fuerza de la gravedad está presente en todos los cuerpos: pero sólo él, y el primero, pudo demostrarlo a partir de las apariencias, y establecer un fundamento seguro por medio de las más brillantes especulaciones.

Sé que algunas personas, incluso de gran renombre, ocupados más bien con otros prejuicios, difícilmente han podido dar su conformidad a este nuevo principio y por lo mismo han preferido lo incierto a lo cierto. No es mi intención poner en tela de juicio su fama, sino que más bien procede, lector benévolo, que exponga unas pocas cosas de las que tú mismo puedes extraer las conclusiones adecuadas.

Así pues, con objeto de hallar un principio argumental que parta de las cosas más simples e inmediatas, examinaremos un poco cuál es la naturaleza de la gravedad en las cosas terrestres para poder después acercarnos con más seguridad a las cosas celestes enormemente distantes de nosotros. Todos los filósofos están ya de acuerdo en que todos los cuerpos circunterrestres gravitan hacia la Tierra. Una amplia y múltiple experiencia confirma que no se dan cuerpos absolutamente leves. La llamada levedad relativa no es verdadera levedad, sino tan sólo aparente, y que surge de la gravedad más fuerte de los cuerpos contiguos.

Por tanto, dado que todos los cuerpos gravitan hacia la Tierra, del mismo modo la Tierra gravita igualmente hacia todos los cuerpos; que la acción de la gravedad es mutua e igual uno respecto a otro se muestra como sigue: distingamos la masa total de la Tierra en dos partes cualesquiera, iguales o en cualquier forma desiguales; si los pesos de las partes no fuesen mutuamente iguales, el peso menor cedería al peso mayor y las partes juntas empezarían a moverse en línea recta hasta el infinito hacia la región del espacio hacia la que tendiese el peso mayor, lo que está en absoluto contra toda experiencia. Y por tanto, habrá que decir que los pesos de las partes están constituidos en estado de equilibrio; esto es, que la acción de la gravedad es mutua y recíprocamente igual[3].

Los pesos de los cuerpos equidistantes del centro de la Tierra son como las cantidades de materia en ellos. Esto se sigue efectivamente de la igualdad de

aceleración de todos los cuerpos que caen en virtud de la fuerza de su peso y a partir del reposo; pues las fuerzas por las que cuerpos desiguales se aceleran igualmente deben ser proporcionales a las cantidades de materia a mover. Ahora bien, puesto que todos los cuerpos que caen se aceleran igualmente, es evidente que en el vacío *Boyleano* recorren al caer espacios iguales en tiempos iguales, una vez suprimida la resistencia del aire; y todavía se comprueba con más exactitud por medio del experimento de los péndulos.

Las fuerzas de atracción de los cuerpos a distancias iguales son como las cantidades de materia de los cuerpos. Dado que los cuerpos gravitan hacia la Tierra y ésta hacia los cuerpos con fuerzas iguales, el peso de la Tierra hacia cada cuerpo, esto es, la fuerza con la que un cuerpo atrae a la Tierra será igual al peso de dicho cuerpo hacia la Tierra. Dicho peso era como la cantidad de materia del cuerpo y, por tanto, la fuerza con la que cada cuerpo atrae a la Tierra o la fuerza absoluta de un cuerpo será como su cantidad de materia.

Por tanto, se origina y se compone la fuerza de atracción de los cuerpos totales de las fuerzas atractivas de las partes, puesto que se ha demostrado que si se aumenta o disminuye la cantidad de materia se aumenta o disminuye proporcionalmente su fuerza. Por eso, la acción de la Tierra ha de atribuirse a la fusión de las acciones de todas las partes, y por tanto es necesario que todos los cuerpos terrestres se atraigan entre sí mutuamente con fuerzas absolutas que estén en razón a la materia atrayente. Esta es la naturaleza de la gravedad en la Tierra; veamos ahora cuál será en los cielos.

Todo cuerpo persevera en su estado de reposo o movimiento uniforme en línea recta, salvo en la medida en que le obliguen a cambiar dicho estado fuerzas impresas en él; es ésta una ley de la naturaleza aceptada por todos los filósofos. En consecuencia, se sigue que los cuerpos que se mueven en círculos, y que por tanto se separan continuamente de las tangentes a sus órbitas, son retenidos dentro de su trayectoria curva por una fuerza constantemente actuante. Por tanto, necesariamente los planetas que giran en órbitas curvilíneas han de entrañar alguna fuerza en cuya virtud permanente se aparten continuamente de las tangentes.

Y justo es conceder lo que se deduce mediante razones matemáticas y se demuestra con toda certeza, que todos los cuerpos que se mueven según una curva descrita en un plano y que, trazando su radio a un punto en reposo o en movimiento, describen áreas en torno a dicho punto proporcionales a los tiempos, están sometidos a fuerzas que tienden a dicho punto. Puesto que hay acuerdo entre los astrónomos en que los planetas primarios describen áreas proporcionales a los tiempos en torno al Sol y los secundarios en torno a sus primarios, se debe concluir que la fuerza por la que constantemente son apartados de las rectas tangentes y obligados a girar en las órbitas curvas, se dirige hacia los cuerpos que ocupan los centros de las órbitas. De este modo, esta fuerza puede llamarse adecuadamente centrípeta respecto al cuerpo que gira; respecto al cuerpo central, en cambio, puede llamarse de atracción, sea cual sea el origen causal que se le atribuya.

Efectivamente, no sólo hay que conceder esto, sino que se puede demostrar matemáticamente: si varios cuerpos distintos giran con movimiento uniforme en círculos concéntricos y los cuadrados de los tiempos periódicos son como los cubos de las distancias al centro común, las fuerzas centrípetas serán inversamente como los cuadrados de las distancias. O, si los cuerpos giran en órbitas semejantes a círculos y no cambian los ejes de las órbitas, las fuerzas centrípetas de los cuerpos que giran serán inversamente como los cuadrados de las distancias. Todos los astrónomos están de acuerdo en alguna de estas fórmulas en todos los planetas. Por tanto, las fuerzas centrípetas de todos los planetas son inversamente como los cuadrados de las distancias a los centros de las órbitas. Si alguien objetara que los ejes de las órbitas de los planetas, sobre todo el de la Luna, no están en reposo, sino que son arrastrados por un cierto movimiento lento, se le podría responder que, aunque concedamos que ese movimiento mínimo hubiese desplazado el eje de su sitio y por ello la proporción de la fuerza centrípeta se desvíe un tanto del cuadrado, podría calcularse dicho desvío y ser prácticamente insensible. La propia razón de la fuerza centrípeta lunar, que debía ser la más perturbada de todas, supera bien poco al cuadrado; pues se aproxima casi sesenta veces más a éste que al cubo. Pero la respuesta sería más exacta si dijéramos que ésta deriva de los ejes, no procede de los desvíos de la proporción del cuadrado, sino que tiene otros orígenes, como muy bien se muestra en esta filosofía. Queda, pues, sentado que las fuerzas centrípetas por las que los planetas primarios tienden al Sol y los secundarios a sus respectivos primarios, muy exactamente son inversas a los cuadrados de las distancias.

De lo dicho se sigue que los planetas son mantenidos en sus órbitas por alguna fuerza que constantemente actúa en ellos; se sigue que dicha fuerza se dirige siempre hacia los centros de las órbitas; se sigue que su fuerza aumenta en dirección al centro y disminuye en la dirección contraria; y aumenta precisamente en la proporción en que disminuye el cuadrado de la distancia al igual que disminuye exactamente en la proporción en que aumenta el cuadrado de la distancia. Veamos ahora, una vez efectuada la comparación entre las fuerzas centrípetas de los planetas y la fuerza de gravedad, si tal vez son fuerzas del mismo género. Y serán del mismo género si hallásemos aquí y allá las mismas leyes y los mismos efectos. Veamos, pues, primero la fuerza centrípeta de la Luna, que es la más próxima a nosotros.

Los espacios rectilíneos que son recorridos en un tiempo dado por cuerpos que parten del reposo, a la vez y bajo el impulso de distintas fuerzas, son proporcionales a dichas fuerzas; y esto también se deduce de razonamientos matemáticos. La fuerza centrípeta de la Luna que gira en su órbita será, pues, a la fuerza de la gravedad en la superficie de la Tierra como sería el espacio más pequeño que la Luna podría describir cayendo hacia la Tierra en virtud de la fuerza centrípeta, si la imagináramos desprovista de todo movimiento circular, respecto al espacio que en el mismo tiempo mínimo describiría un grave en las inmediaciones de la Tierra y que cayera por la fuerza de su propia gravedad. El primero de dichos espacios es igual al seno verso del

arco descrito por la Luna en dicho tiempo, puesto que equivale a la desviación de la Luna de la tangente producida por la fuerza centrípeta; y por tanto, puede calcularse tanto a partir de los datos de la Luna en un tiempo periódico como de su distancia del centro de la Tierra. Se puede hallar el otro espacio por medio del experimento de los péndulos, tal como muestra Huygens. Hecho el cálculo, el espacio del primero será al espacio del segundo, o la fuerza centrípeta de la Luna que gira en su órbita será a la fuerza de la gravedad en la superficie de la Tierra como el cuadrado del semidiámetro de la Tierra al cuadrado del semidiámetro de la órbita. La fuerza centrípeta de la Luna girando en su órbita está, por lo que se ha dicho, en la misma razón que la fuerza centrípeta de la Luna en las proximidades de la superficie terrestre. Por consiguiente, la fuerza centrípeta en la proximidad de la Tierra es igual a la fuerza de la gravedad. No son, por tanto, dos fuerzas distintas, sino una y la misma: pues si fuesen distintas, los cuerpos, bajo ambas fuerzas juntas, caerían hacia la Tierra doblemente más deprisa que bajo la sola acción de la gravedad. Se sigue, pues, que la fuerza centrípeta por la que la Luna continuamente se ve apartada de la tangente y retenida en su órbita es la propia fuerza de gravedad terrestre que alcanza hasta la Luna. Y es, por otra parte, conforme a la razón el que dicha fuerza se extienda a tan grandes distancias, dado que ninguna disminución de dicha fuerza es posible detectar ni siquiera en las más altas cimas de las montañas. Gravita, pues, la Luna hacia la Tierra sin que, y en acción mutua, la Tierra a su vez deje de gravitar igualmente hacia la Luna. Cosa ampliamente confirmada en este tratado, donde se trata del movimiento de las mareas y de la precesión de los equinoccios, originados tanto de la acción de la Luna como de la del Sol sobre la Tierra. Por todo esto vemos, finalmente, de acuerdo con qué ley decrece la fuerza de la gravedad a mayores distancias de la Tierra. Puesto que la gravedad no es distinta de la fuerza centrípeta lunar y ésta es a su vez inversamente proporcional al cuadrado de la distancia, también la gravedad disminuirá según esa misma razón.

Pasemos ahora a los demás planetas. Puesto que las revoluciones de los primarios en torno al Sol y de los secundarios alrededor de Júpiter y de Saturno son fenómenos del mismo género que la revolución de la Luna en torno a la Tierra y puesto que está demostrado que las fuerzas centrípetas de los primarios se dirigen hacia el centro del Sol y las de los secundarios hacia los centros de Júpiter y Saturno, al igual que la fuerza centrípeta de la Luna se dirige hacia el centro de la Tierra y además, puesto que todas las susodichas fuerzas son inversamente como los cuadrados de las distancias a los centros, al igual que la fuerza de la Luna lo es respecto al cuadrado de la distancia a la Tierra, se habrá de concluir que es la misma la naturaleza de todas ellas. Y así del mismo modo que la Luna gravita hacia la Tierra y la Tierra a su vez hacia la Luna, del mismo modo gravitarán todos los secundarios hacia sus primarios y a su vez los primarios hacia sus secundarios, al igual que todos los primarios hacia el Sol y el Sol hacia los primarios.

Por consiguiente, el Sol gravita hacia todos los planetas y todos hacia el Sol.

Puesto que los secundarios, al acompañar a sus primarios, giran mientras tanto también en torno al Sol a la vez que los primarios. Por la misma razón, pues, los planetas de una y otra clase gravitan hacia el Sol y el Sol hacia ellos. Además, se sigue que los planetas secundarios gravitan hacia el Sol por más razones, tales como las desigualdades lunares; de ellas tenemos una exposición exacta, realizada con admirable sutileza, en el Libro tercero de esta obra.

Que la fuerza atractiva del Sol se propaga por doquier hasta las más enormes distancias y que se difunde por cada rincón del espacio circundante se deduce con toda claridad del movimiento de los cometas, los cuales, partiendo de distancias inmensas, llegan hasta las inmediaciones del Sol, incluso algunas veces llegan tan cerca del mismo al alcanzar sus perihelios, que parecería que iban a caer en su globo. La teoría de éstos se la debemos a nuestro ilustre autor, en vano buscada hasta ahora por los astrónomos, descubierta con todo éxito en este siglo y demostrada con toda exactitud por los fenómenos. Se sigue, pues, que los cometas se mueven en secciones cónicas que tienen sus focos en el centro del Sol y que con los radios que llegan al Sol describen áreas proporcionales a los tiempos. De estos fenómenos se deduce, y matemáticamente se comprueba, que las fuerzas por las que los cometas son retenidos en sus órbitas están dirigidas al Sol y son inversamente como los cuadrados de las distancias al centro del mismo. Gravitan, por tanto, los cometas respecto al Sol; y por tanto, la fuerza de atracción del Sol no sólo alcanza a los cuerpos de los planetas situados a unas distancias dadas y casi en el mismo plano, sino que también alcanza a los cometas situados en los lugares más distintos y a las más diversas distancias. Tal es, pues, la naturaleza de los cuerpos que gravitan, que sus fuerzas llegan a todas las distancias hasta todos los cuerpos en gravitación. De aquí se sigue que todos los planetas y cometas se atraen mutuamente y gravitan mutuamente entre sí; esto también se halla confirmado por la perturbación de Júpiter y Saturno, no desconocida de los astrónomos, y originada de las acciones mutuas de estos planetas, y también por el lento movimiento, mencionado antes, de los ápsides y cuya causa es muy similar.

Por ello, finalmente tenemos que llegar a decir que la Tierra, el Sol y todos los cuerpos celestes que acompañan al Sol se atraen mutuamente. Por consiguiente, aún las más mínimas partículas de cada uno tendrán sus propias fuerzas de atracción según la cantidad de materia que tengan, tal como se vio antes de las partículas terrestres. Para distancias distintas también serán sus fuerzas inversamente como el cuadrado de las distancias; pues se demuestra matemáticamente que de partículas que se atraen, según dicha ley deben formarse los globos que, a su vez, se atraen conforme a la misma ley.

Las conclusiones que anteceden se basan en el siguiente axioma que todos los filósofos aceptan: las causas y las propiedades de los efectos que aún no se conocen, y que son del mismo género que los que se conocen, son causas y propiedades iguales a las de los efectos que se conocen. ¿Quién duda de que si la gravedad es la causa de

la caída de una piedra en *Europa*, también será la causa de la caída en *América*? Si se diese mutua gravitación entre una piedra y la Tierra en *Europa*, ¿quién negará la mutua gravitación en *América*? Si la fuerza de atracción de la piedra y la Tierra se compone en *Europa* de las fuerzas atractivas de las partes, ¿quién negará que en *América* la composición será semejante? Si la atracción terrestre alcanza a toda clase de cuerpos y a cualquier distancia en *Europa*, ¿por qué no diremos que se propaga de igual modo en *América*? Toda la ciencia se basa en esta regla, puesto que si la suprimimos nada podríamos afirmar universalmente. La constitución de las cosas singulares se hace patente por medio de las observaciones y los experimentos, y, por tanto, sólo mediante esta regla podemos hablar de la naturaleza de todas las cosas.

Puesto que son graves todos los cuerpos que se hallan en la Tierra o en el firmamento y sobre los cuales es posible hacer experimentos u observaciones, habrá que decir de una vez por todas que la gravedad afecta a todos los cuerpos. Y al igual que no se debe concebir cuerpo alguno que no sea extenso, móvil e impenetrable, del mismo modo tampoco debe concebirse un cuerpo que no sea grave. La extensión, movilidad e impenetrabilidad de los cuerpos no se revelan si no es experimentalmente; exactamente del mismo modo ocurre con la gravedad. Todos los cuerpos de los que tenemos observaciones son extensos, móviles e impenetrables; de ello concluimos que todos los cuerpos, incluso aquellos de los que no tenemos observación ninguna, son extensos, móviles e impenetrables. De igual modo, todos los cuerpos de los que tenemos observación son graves; de aquí concluimos que todos los cuerpos, incluso aquellos de los que no tenemos observación, son graves. Si alguien dijera que no son graves los cuerpos de las estrellas fijas, dado que su gravedad aún no ha sido observada, por la misma razón sería posible decir que ni son extensos, ni móviles, ni impenetrables, dado que tales propiedades aún no han sido observadas en las estrellas fijas. ¿Qué necesidad hay de más palabras? O la gravedad tiene un lugar entre las cualidades primarias de todos los cuerpos o no la tienen ni la extensión, ni la movilidad, ni la impenetrabilidad. Y por tanto, o se explica correctamente la naturaleza de las cosas por medio de la gravedad de los cuerpos o no se explica correctamente por medio de la extensión, la movilidad y la impenetrabilidad de los cuerpos.

Estoy oyendo ya a algunos rechazar esta conclusión y cuchichear no sé qué sobre las cualidades ocultas. Suelen argumentar siempre que la gravedad es algo oculto; y hay que apartar de la filosofía las causas ocultas. A éstos fácilmente se responde: no son causas ocultas aquellas cuya existencia se demuestra claramente por medio de observaciones, sino sólo aquellas cuya existencia es oculta, imaginaria y aún no demostrada. La gravedad no será, pues, causa oculta de los movimientos celestes, dado que se ha demostrado a partir de los fenómenos que tal fuerza realmente existe. Estos tales son más bien los que acuden a causas ocultas y establecen como determinantes de dichos movimientos a no sé qué vórtices de una ficticia materia completamente desconocida para los sentidos[4].

¿Acaso la gravedad se llama causa oculta, y por esta razón se desecha de la ciencia, simplemente por ser oculta la causa de dicha gravedad y no haber sido aún descubierta? Los que tal dicen cuídense, no sea que digan algo absurdo que venga a arruinar los fundamentos de toda filosofía. Puesto que las causas suelen proceder mediante encadenamiento continuo de las más complejas a las más simples, cuando lleguemos a la causa más simple ya no nos será posible ir más allá. Por tanto, de la causa más simple no se podrá dar una explicación mecánica, pues si se pudiera dar no se trataría de la causa más simple. ¿Llamaríamos ocultas a estas causas simplicísimas y las suprimiríamos también? Entonces suprimiríamos también a las inmediatamente siguientes y a las siguientes, hasta que la filosofía entera resultase vacía y limpia de toda causa.

Hay quienes afirman que la gravedad es algo preternatural y la consideran un milagro perpetuo. Por tanto, también quieren rechazarla, dado que las causas preternaturales no tienen cabida en la física. Apenas vale la pena detenerse en responder a esta burda objeción que mina además a toda la filosofía. O bien niegan la gravedad inherente a todos los cuerpos, cosa que no se puede decir; o bien, por la misma razón, sostendrán que es preternatural lo que no tiene su origen en otras propiedades corpóreas y, por tanto, en causas mecánicas. Ciertamente, se dan las propiedades primarias de los cuerpos, las cuales, puesto que son primarias, no dependen de otras. Observen, pues, si todas éstas son también preternaturales o no, y, por consiguiente, recusables, y piensen en consecuencia qué filosofía podría seguirse después.

Otros hay para los que toda esta física celeste es tanto menos grata cuanto parece oponerse a los dogmas cartesianos o apenas puede casarse con ellos. En su derecho están si les gusta esta opinión; pero justo será que se conduzcan consecuentemente y no nieguen a los demás la libertad que para sí piden. Por tanto, será lícito defender y abrazar la filosofía *newtoniana*, que nos parece más verdadera y mantener las causas comprobadas por medio de fenómenos, más bien que las ficticias y aún no comprobadas. Compete a la verdadera filosofía derivar la naturaleza de las cosas de causas realmente existentes; esto es, buscar aquellas leyes con las que el mismo artífice quiso establecer el maravilloso orden de este mundo, no aquéllas con las que pudo hacerlo si así le hubiese parecido. Pues no es contradictorio el que muchas causas, distintas entre sí, produzcan los mismos efectos. Pero aquella será la causa verdadera que realmente y de hecho produce el efecto, mientras las demás no tienen sitio en la verdadera filosofía. En los relojes mecánicos el mismo movimiento del índice horario puede provenir bien de una pesa, bien de un muelle incorporado. Pero si un reloj dado, de hecho está dotado de una pesa, caerá en ridículo el que suponga un muelle y, de acuerdo con la hipótesis, pretenda explicar así el movimiento del índice horario; es preciso investigar previamente con más atención la construcción interna de la máquina para poder determinar el verdadero origen del movimiento propuesto. Y no merecen un juicio muy distinto los filósofos aquellos que han

considerado a los cielos repletos de una materia sutil y en perpetua acción en los vórtices. Pues, aunque pudieran dar cuenta de los fenómenos a partir de sus hipótesis del modo más exacto, aún no se podría decir que han ofrecido una verdadera filosofía ni que han hallado las causas verdaderas de los movimientos celestes, a no ser que hayan demostrado que tales causas existen realmente, o que no existen otras. Por tanto, si se demostrase que la atracción de todos los cuerpos tiene verdaderamente lugar en la naturaleza de las cosas, y además se hubiese demostrado por qué razón todos los movimientos celestes pueden explicarse a partir de ella, entonces sería vana y en verdad ridícula la objeción del que dijese que estos mismos movimientos deberían explicarse por medio de vórtices, si esto fuese posible o incluso lo concediésemos. Pero no lo hemos concedido; pues de ningún modo pueden explicarse los fenómenos por los vórtices, cosa que nuestro autor demuestra ampliamente y con claros argumentos, para que, los que sostienen tal desafortunada opinión, tanto reconstruyendo el inútil engendro como adornándolo con nuevos comentarios, sean más dignos de indulgencia por sus sueños.

Si los planetas y los cometas fuesen arrastrados en torno al Sol por los vórtices, sería preciso que los astros arrastrados y las partes de los vórtices circundantes se moviesen con la misma velocidad y con la misma dirección y que tuvieran la misma densidad y la misma fuerza de inercia que la masa de la materia. Sin embargo, nos consta que los planetas y cometas, mientras caminan por las mismas regiones del cielo, se mueven con distintas velocidades y distintas direcciones. Se sigue, pues, necesariamente que aquellas partes del fluido celeste que están a la misma distancia del Sol, se mueven a la vez en direcciones distintas y con velocidades distintas, puesto que unas serán la velocidad y dirección necesarias para que puedan circular los planetas y otras para que puedan hacerlo los cometas. Pero al no poder explicar esto, o se habrá de reconocer que ningún cuerpo celeste es transportado por la materia de los vórtices, o se habrá de sostener que sus movimientos son producidos, no por uno solo e idéntico vórtice, sino por muchos que son distintos entre sí y ocupan el mismo espacio alrededor del Sol.

Si se suponen muchos vórtices contenidos en el mismo espacio penetrándose mutuamente y girando con diferentes movimientos, puesto que tales movimientos deben ser conformes con los movimientos de los cuerpos transportados, movimientos que son altamente regulares y se producen en secciones cónicas, ora muy excéntricas, ora muy próximas a un círculo, con toda razón uno se habrá de preguntar a qué se debe el que se conserven intactos y no se hayan, en cambio, desajustado ni un poco a lo largo del tiempo por obra de la materia que les opone resistencia. En realidad, si estos movimientos ficticios son más compuestos y se explican más difícilmente que los verdaderos movimientos de los planetas y de los cometas, me parece inútil aceptarlos en filosofía, pues toda causa debe ser más simple que su efecto. Concediendo venía a las fantasías, alguien afirmaría que todos los planetas y cometas están rodeados de atmósferas similares a la de nuestra Tierra; hipótesis que, sin

embargo, parecerá más racional que la de los vórtices. Afirmaría después que tales atmósferas por su naturaleza se mueven en torno al Sol y describen secciones cónicas, movimientos que en realidad pueden concebirse mucho más fácilmente que el movimiento similar de los vórtices mutuamente permeables. Por último, establecería que es preciso creer que los propios planetas y cometas son arrastrados alrededor del Sol por sus respectivas atmósferas y cantará victoria por haber encontrado las causas de los movimientos celestes. Pero quien crea que esta fábula debe rechazarse, también rechazará la otra, puesto que la hipótesis de las atmósferas y la de los vórtices no se distinguen más que un huevo de otro huevo.

Galileo mostró que el desvío de la trayectoria recta de una piedra lanzada y que se mueve parabólicamente tiene su origen en la gravedad de la piedra hacia la Tierra, esto es, en una cualidad oculta. Sin embargo, podría ocurrir que otro filósofo cualquiera de más fino olfato proponga otra causa. Imagina, pues, el tal filósofo que cierta materia sutil, que no puede percibirse ni por la vista, ni por el tacto, ni por sentido alguno, actúa en zonas inmediatas a la superficie de la Tierra. Pero sostiene que esta materia es llevada hacia direcciones diversas por movimientos distintos y muchas veces contrarios a la vez que describe líneas parabólicas. Y así describirá a partir de aquí con pulcritud el desvío de la piedra e incluso merecerá el aplauso del vulgo. La piedra, dirá, flota en aquel Huido sutil y siguiendo el curso del mismo no puede dejar de describir a la vez la misma senda. Pero como el fluido se mueve en líneas parabólicas, entonces también será necesario que la piedra se mueva parabólicamente. ¿Quién no sentirá admiración ante el sutil ingenio de nuestro filósofo, que deduce con claridad y al alcance del vulgo los fenómenos de la naturaleza de causas mecánicas tales como la materia y el movimiento? ¿Quién no se reirá del bueno de Galileo viéndole sostener con un gran aparato matemático la necesidad de recuperar á las cualidades ocultas excluidas ya afortunadamente de la filosofía? Pero me sonroja gastar más tiempo en bagatelas.

El resultado finalmente viene a ser: el número de cometas es enorme, sus movimientos son altamente regulares y cumplen las mismas leyes que los movimientos de los planetas. Se mueven en curvas cónicas, curvas que son muy excéntricas. Recorren cualquier lugar del firmamento en cualquier dirección y atraviesan libremente los espacios planetarios y frecuentemente caminan en sentido contrario al orden del Zodíaco. Las observaciones astronómicas confirman sin duda todos estos fenómenos que no pueden explicarse por los vórtices e incluso no pueden coexistir con los vórtices de los planetas. A no ser que la tal materia imaginaria desaparezca del cielo, no habrá lugar para los movimientos de los cometas.

Si los planetas son arrastrados en torno al Sol por los vórtices, las partes de éstos que rodean inmediatamente a cada uno de los planetas tendrá la misma densidad que éstos, como se ha dicho más arriba; y por tanto, toda aquella materia que está inmediatamente contigua al perímetro del orbe magno tendrá una densidad semejante a la de la Tierra, mientras que la que se halla dentro del orbe magno o del de Saturno

será igual o mayor. Pero para que la estructura del vórtice pueda constituirse y permanecer es preciso que las partes menos densas ocupen el centro y las más densas se alejen del mismo. Mas como los períodos de los planetas están en razón de la potencia $^{3}/_{2}$ de las distancias al Sol, es preciso que los períodos de las partes de los vórtices conserven la misma proporción. Y de aquí se sigue que las fuerzas centrífugas de estas partes habrán de estar en razón inversa a los cuadrados de las distancias. Por tanto, las que distan más del centro pugnan con menos fuerza por alejarse de él; por ende, si fuesen menos densas es necesario que cedan ante la mayor fuerza con que las partes más cercanas al centro pugnan por alejarse. Ascenderán, por tanto, las más densas y descenderán las menos densas y cambiarán entre sí de sitio hasta que la materia toda del vórtice esté ordenada y dispuesta de tal modo que pueda reposar en equilibrio. Si dos fluidos de densidades distintas se colocan en el mismo vaso ocurrirá que el fluido de mayor densidad se dirigirá al lugar más bajo, dada su mayor gravedad; y por una razón semejante hay que decir que las partes más densas de un vórtice se irán hacia los lugares más alejados por su mayor fuerza centrífuga. Por consiguiente, toda la parte del vórtice, que es con mucho la mayor, que se extiende fuera del globo terrestre, tendrá una densidad y fuerza inercial por cantidad de materia no menor que la densidad o inercia de la materia del mismo. De ahí surgirá una resistencia tan grande y ostensible a las trayectorias de los cometas que yo diría que es capaz de frenar y hasta detener sus movimientos. Nos consta, sin embargo, por el movimiento altamente regular de los cometas, que no sufren resistencia alguna y por tanto no atraviesan materia alguna que ofrezca resistencia y que, por tanto, tenga densidad o fuerza de inercia alguna, pues la resistencia de los medios procede bien de la inercia de la materia fluida, bien de la falta de lubricidad. La que procede de la falta de lubricidad es muy pequeña y apenas puede detectarse en los fluidos conocidos comúnmente, salvo que fuesen de una viscosidad similar a la del aceite o la miel. La resistencia observable en el aire, en el agua, en el mercurio o en fluidos semejantes no viscosos, es casi toda del primer tipo y apenas puede disminuirse por cualquier grado ulterior de enrarecimiento si permanece la densidad y la fuerza de inercia del fluido a las que siempre es proporcional esta resistencia, como demostró claramente nuestro autor en su brillante teoría de las resistencia, que ahora, en esta segunda edición, se expone un poco más rigurosamente a la vez que se confirma con experimentos de cuerpos que caen.

Los cuerpos al moverse comunican poco a poco su movimiento al fluido circundante, y al comunicarlo lo van perdiendo, y al perderlo se desaceleran. La desaceleración es proporcional al movimiento comunicado, y éste, cuando se da la velocidad del móvil, es como la densidad del fluido; por tanto, la desaceleración o la resistencia será como la densidad del fluido; y esto no hay modo de evitarlo, a no ser que el fluido que llene la parte posterior del móvil restituya el movimiento perdido. Pero esto sería imposible de mantener salvo que el empuje del fluido sobre la parte posterior fuese igual al empuje de la parte delantera sobre el fluido, esto es, salvo que

la velocidad relativa con que el fluido irrumpe contra el cuerpo por detrás fuese igual a la velocidad con que el cuerpo irrumpe contra el fluido o, lo que es lo mismo, que la velocidad absoluta del fluido que irrumpe por detrás del móvil sea el doble que la velocidad absoluta del fluido impactado por el móvil, lo que es imposible. No hay, pues, modo de evitar la resistencia de los fluidos procedente de su densidad e inercia. Hay que concluir, por tanto, que el fluido celeste no tiene fuerza inercial alguna dado que no ofrece resistencia alguna, que no hay fuerza alguna que comunicar a móvil alguno, dado que no hay inercia ninguna; que no hay fuerza alguna que produzca cambio alguno en los cuerpos ni singulares ni en conjunto, puesto que no hay fuerza alguna que comunique movimiento a los cuerpos; que no existe la más mínima capacidad de obrar al no existir la menor facultad de producir cualquier tipo de mutación. Por qué, pues, no llamar inútil e indigna de un sabio a una hipótesis que, sobre carecer de fundamento, no sirve en lo más mínimo para explicar la naturaleza de las cosas. Los que quieren ver el cielo lleno de materia inerte suprimen el vacío sólo de palabra, pero en la realidad lo mantienen. Puesto que no puede hallarse razón alguna que permita distinguir semejante materia del espacio vacío, toda la polémica se reduce a cuestión de nombres y no de cosas. Pero si hay además algunos tan adictos a la materia, que de ningún modo podrían admitir espacio vacío de cuerpos, veamos hasta qué punto es obligado hacerlo.

Y ello porque, o sostienen que esta constitución que imaginan del mundo lleno por todas partes procede de la voluntad de Dios con el fin de dar apoyo a las operaciones de la naturaleza mediante un éter sutil que todo lo llena y en todo está presente, cosa que no se puede sostener, puesto que, como se ha mostrado por los fenómenos de los cometas, la eficacia de tal éter es nula, o sostienen que procede de la voluntad de Dios para algún fin desconocido, cosa que no debe decirse, ya que semejante argumento llevaría igualmente a establecer otra constitución cualquiera del mundo, o, finalmente, sostienen que no procede de la voluntad de Dios, sino de cierta necesidad natural. Y así, finalmente es preciso venir a parar a las filas de una grey de indeseables. Son tales los que creen que el hado y no la providencia lo gobierna todo, que la materia necesariamente ha existido siempre y en todas partes, que es infinita y eterna. Supuesto esto también, será uniforme en todo lugar, puesto que la diversidad de formas no cuadra en absoluto con la necesidad. También será inmóvil, puesto que si se moviera necesariamente en una dirección dada con una determinada velocidad, con igual necesidad se movería en dirección distinta y con velocidad distinta; pero al no ser posible moverse en direcciones distintas y con velocidades distintas, es preciso que sea inmóvil. Por tanto, este mundo lleno de las más bellas formas y de la mayor variedad de movimientos, no ha podido tener otro origen que la libre voluntad de un dios providente y gobernante.

De esta fuente, salieron todas las así llamadas leyes de la naturaleza, en las que tantas muestras de sabiduría y no de necesidad aparecen. Por tanto, hay que encontrarlas observando y experimentando y no a partir de conjeturas inciertas.

Quien cree que puede encontrar por su sola razón y con la ayuda de su sola capacidad mental los principios de la Física y las leyes de la naturaleza, necesita, o bien establecer que el mundo procede de la necesidad y que sigue las leyes nacidas de ella, o bien, si el orden del mundo ha sido creado por la voluntad de Dios, que él, humana miseria, ha comprendido qué es lo mejor que puede ser creado. La verdadera y auténtica filosofía se basa en los fenómenos, los cuales, si nos inducen a nosotros, o a otros menos dispuestos, a aceptar tales principios en los que se trasluce el gran saber y la suprema potestad de un ser sabio y todopoderoso, no deben ser rechazados bajo el pretexto de que tal vez sean menos aceptables para otros hombres. Ya llamen milagros o cualidades ocultas a los principios que rechacen, no deben atribuirse a las cosas nombres maliciosamente puestos, a no ser que al fin se desee confesar que efectivamente la filosofía debe descansar en el ateísmo. Por causa de estos hombres no debe degradarse la filosofía si no se quiere cambiar el orden de las cosas.

Gozará de crédito, pues, ante los más exigentes y equitativos jueces aquella manera de hacer filosofía que se basa en experimentos y observaciones. Apenas podemos decir en qué grado ilumina y cuánto dignifica a este modo de hacer filosofía la meritísima obra de nuestro ilustre autor, cuyo talento al resolver los más difíciles problemas, hasta el punto de que no era dado esperar su solución de la mente humana, con razón admiran y alaban quienes conocen con cierta profundidad estos temas. Rotos los arcanos, nos abrió paso hacia los más bellos misterios de la naturaleza. De tal modo nos esclareció la armoniosa belleza del sistema del mundo, que ni el propio rey Alfonso, si resucitase, desearía en él mayor simplicidad y graciosa armonía. De tal modo, pues, que es ya más fácil comprender la majestad de la naturaleza, gozar de la más dulce contemplación, venerar y dar culto sin esfuerzo al fundador y señor del universo, cosas todas que son, con mucho, el fruto más logrado de la filosofía. Es preciso estar ciego para no ver al instante a través de las óptimas y sabias estructuras de las cosas la sabiduría y bondad infinitas de un autor omnipotente; es preciso estar loco para no reconocerlo.

Se erguirá, pues, la admirable obra de Newton como un formidable castillo contra los ataques de los ateos y en ningún otro sitio se hallarán más fácilmente dardos contra la caterva impía que en esta aljaba. Así lo vio ya el primero, y lo publicó tanto en latín como en inglés el ilustre y admirable en todos los géneros del saber Ricardo Bentley, eximio protector de las artes, gloria de su siglo y de nuestra Universidad, digno y conspicuo maestro de nuestro colegio de la *Santísima Trinidad*. A él me siento obligado por muchas razones. Incluso tú, amable lector, no le negarás tu agradecimiento; pues habiendo disfrutado durante mucho tiempo de la íntima amistad del ilustre autor (amistad que piensa que no debe tenerse en menos por la posteridad que los propios escritos que brillan en el mundo literario) se ocupó a la vez de la fama del amigo y del crecimiento de las ciencias. Y así, al ser ya raros y sumamente caros los ejemplares que quedaban de la anterior edición, persuadió con sus lamentaciones e incluso sin violencia empujó finalmente a aquel ilustre varón, insigne no menos por

su modestia que por su enorme erudición, a que autorizase esta nueva edición de su obra, revisada y aumentada con añadidos importantes, a sus expensas y bajo su cuidado; a mí me pidió, era su derecho, algo no demasiado ingrato: que cuidase de hacerlo sin erratas en lo que pudiese.

Rogerio Cotes
Miembro del colegio de la Santísima Trinidad,
Profesor Plumiano de Astronomía
y Filosofía experimental.

Cambridge, 12 de mayo de 1712

En esta tercera edición, debida a los cuidados de *Henry Pemberton*, muy docto en estas cuestiones, se explican con alguna mayor extensión ciertos experimentos del Libro II sobre la resistencia a los graves que caen en la atmósfera hacia la Tierra. En el Libro III se expone con más amplitud el argumento con que se prueba que la Luna permanece en su órbita a causa de la gravedad y se añaden nuevas observaciones sobre la proporción mutua de los diámetros de Júpiter realizados por *D. Pound*. Se añaden también algunas observaciones del cometa de 1680, hechas por *Kirk* en noviembre en Alemania y recién llegadas a nuestras manos, con las que se muestra cuán cerca están las órbitas parabólicas de las trayectorias de los cometas. También se precisa un poco más que antes la elipse de la órbita del cometa que reseñó *Halley*. Y se muestra que dicha órbita elíptica a través de los nueve signos celestes no realizó su giro con menor exactitud que lo suelen hacer los planetas a través de sus órbitas elípticas descritas en astronomía. Además, se añade la órbita del cometa de 1723, observado por *Bradley*, profesor de astronomía de *Oxford*.

Isaac Newton

Londres, 12 de enero de 1725-26.

Definición Primera

La cantidad de materia es la medida de la misma originada de su densidad y volumen conjuntamente.

El aire dos veces más denso, en también doble espacio, es cuádruple, en triple espacio, séxtuple. Lo mismo se debe entender de la nieve o del polvo condensados por compresión o licuefacción. Y esta misma es la razón para todos los cuerpos que por cualquier causa se condensan de modos diversos. No tengo, en cambio, aquí razón ninguna para un medio, si hubiere alguno, que atraviese libremente los intersticios de las partes. A esta cantidad llamo en lo sucesivo cuerpo o masa. Se hace manifiesta por el peso de cualquier cuerpo, pues, por medio de experimentos muy exactos con péndulos, hallé que era proporcional al peso, como después se mostrará[5].

Definición II

La cantidad de movimiento es la medida del mismo obtenida de la velocidad y de la cantidad de materia conjuntamente.

El movimiento del todo es la suma de los movimientos de cada parte y, por tanto, es el doble en un cuerpo el doble mayor con igual velocidad y cuádruple con doble velocidad[6].

Definición III

La fuerza ínsita de la materia es una capacidad de resistir por la que cualquier cuerpo, por cuanto de él depende, persevera en su estado de reposo o movimiento uniforme y rectilíneo.

Esta es siempre proporcional al cuerpo y no se diferencia en nada de la inercia de la masa, a no ser en el modo de concebirla. Por la inercia de la materia ocurre que todo cuerpo difícilmente se aparta de un estado de reposo o movimiento. De donde la fuerza ínsita también puede llamarse con toda propiedad fuerza de inercia[7]. Ejerce el cuerpo esta fuerza solamente en el cambio de su estado hecho por otra fuerza impresa en él y es su ejercicio, bajo diverso aspecto, tanto resistencia como ímpetu: resistencia, en tanto en que el cuerpo, para conservar su propio estado, se opone a la fuerza impresa; ímpetu, en tanto que el mismo cuerpo, cediendo difícilmente a la fuerza del obstáculo resistente, intenta cambiar el estado de éste. El vulgo atribuye la resistencia a los cuerpos en reposo y el ímpetu a los cuerpos en movimiento; pero el movimiento y el reposo, tal como el vulgo los concibe, sólo se distinguen relativamente uno de otro, y no siempre verdaderamente reposan las cosas que el vulgo considera en reposo.

Definición IV

La fuerza impresa es la acción ejercida sobre un cuerpo para cambiar su estado de reposo o movimiento uniforme y rectilíneo.

Consiste esta fuerza en la sola acción y no permanece en el cuerpo después de ella, pues el cuerpo permanece en el nuevo estado únicamente por inercia. La fuerza impresa tiene diferentes orígenes, tales como un golpe, una presión, la fuerza centrípeta.

Definición V

La fuerza centrípeta es aquella en virtud de la cual los cuerpos son atraídos, empujados, o de algún modo tienden hacia un punto como a un centro.

De esta clase es la gravedad por la que los cuerpos tienden al centro de la Tierra; el magnetismo por el que el hierro tiende hacia el imán; y la fuerza, cualquiera que sea, por la que constantemente los planetas se ven apartados de las trayectorias rectilíneas y se ven obligados a permanecer girando en líneas curvas. Una piedra volteada en una honda intenta escapar de la mano del hondero y su intento hace estirarse la honda y más cuanto más rápidamente gira, y en cuanto se suelta se aleja. Llamo centrípeta a la fuerza contraria al mencionado intento; por ella, la honda retiene constantemente la piedra hacia la mano y la mantiene en el círculo, y, por tanto, se dirige hacia la mano o hacia el centro del círculo. Igual ocurre con todos los cuerpos que giran en círculo. Todos intentan alejarse del centro dél círculo y, a no ser

por una fuerza contraria a este intento, que los cohíba y los obligue en sus órbitas y a la que por ello llamo centrípeta, se alejarían todos en línea recta con movimientos uniformes. Si se despojase de la fuerza de gravedad, un proyectil no caería a tierra, sino que se alejaría por los cielos en línea recta y esto con movimiento uniforme si además suprimimos la resistencia del aire. La gravedad lo aparta continuamente de la trayectoria rectilínea y lo doblega continuamente hacia la Tierra, y esto más o menos según sea su gravedad y su velocidad de movimiento. Cuanto menor sea su gravedad por la cantidad de materia o la velocidad con que se dispara, tanto menos se desviará de la trayectoria rectilínea y más lejos llegará. Si una bola de plomo disparada con la fuerza de la pólvora desde lo alto de un monte a una determinada velocidad y en la dirección de una línea horizontal recorriese, describiendo una curva, dos millas antes de caer al suelo, con velocidad doble llegaría casi el doble más lejos y con diez veces la velocidad llegaría también casi diez veces más lejos, supuesto que no haya resistencia del aire. Y, aumentando la velocidad, puede aumentarse a voluntad la distancia de lanzamiento y disminuir la curvatura de la trayectoria hasta el punto de alcanzar diez, treinta o noventa grados e incluso circundar la Tierra y hasta alejarse hacia los cielos y con el movimiento de ida escapar al infinito. Y por la misma razón por la que un proyectil podría girar en un círculo por la fuerza de la gravedad y dar vuelta a toda la Tierra, asimismo la Luna por la fuerza de la gravedad, supuesto que sea grave, u otra fuerza cualquiera por la que sea empujada hacia la Tierra, puede ser desviada hacia la Tierra continuamente de la trayectoria rectilínea y girar en su órbita; y sin tal fuerza, la Luna no podría retenerse en su órbita. Esta fuerza, si fuese sólo un poco menor, no podría apartar a la Luna de su curso rectilíneo, y si fuese un poco mayor, la apartaría demasiado y la precipitaría de su órbita hacia la Tierra. Se requiere, pues, que sea de una magnitud exacta. Y es cometido de los matemáticos calcular la fuerza con la que un cuerpo en una órbita determinada y a una velocidad dada podría mantenerse exactamente, y a la inversa, determinar la trayectoria curva a que es empujado un cuerpo que es lanzado con una fuerza dada y desde un punto dado. Además, la magnitud de esta fuerza centrípeta es de tres clases: absoluta, acelerativa y motriz.

Definición VI

Magnitud absoluta de la fuerza centrípeta es la medida mayor o menor de la misma según la eficacia de la causa que la expande desde un centro en todas las direcciones en torno.

Como ocurre con la fuerza magnética, según la masa del imán o según la intensidad de la fuerza magnética, que es mayor en unos imanes que en otros.

Definición VII

La magnitud acelerativa de la fuerza centrípeta es su medida proporcional a la velocidad que genera en un tiempo dado.

Así, la fuerza del imán es mayor a menor distancia y menor a mayor distancia, o la fuerza gravitacional es mayor en los valles y menor en las cimas de los montes muy altos, y menor todavía (como luego se mostrará) a mayores distancias de la superficie terrestre, mientras permanece igual a distancias iguales, puesto que acelera por igual a todos los cuerpos que caen (graves o leves, grandes o pequeños) si se suprime la resistencia del aire.

Definición VIII

La magnitud motriz de la fuerza centrípeta es la medida de la misma proporcional al movimiento que genera en un tiempo dado.

Así, el peso es mayor en un cuerpo mayor y menor en un cuerpo menor; y en el mismo cuerpo resulta mayor cerca de la Tierra y menor en el cielo. Dicha cantidad es la centripetencia o tendencia al centro de todo el cuerpo y (por así decirlo) su peso; se pone siempre de manifiesto por medio de una fuerza contraria a ella y equivalente por la que sea posible impedir la caída del cuerpo.

Es conveniente, para ser breves, llamar a estas magnitudes fuerzas motrices, aceleratívas y absolutas; y para distinguirlas, referirlas a los cuerpos que tienden a un centro, a los lugares de los cuerpos y al centro de fuerzas: a saber, la fuerza motriz a un cuerpo, como si se tratara del impulso de un todo hacia un centro, impulso compuesto de los impulsos parciales de todas las partes; y la fuerza aceleratriz al lugar del cuerpo, como si se tratara de cierta eficacia difundida desde el centro por cada punto en torno para mover los cuerpos situados en él; la fuerza absoluta al centro, como si estuviera dotado de una causa sin la que las fuerzas motrices no se propagarían por las regiones entorno, ya sea tal causa un cuerpo central (como el imán en el centro de la fuerza magnética o la Tierra en el centro de la fuerza de gravitación) u otra causa oculta. Tal concepto es meramente matemático, puesto que no considero aquí las causas y las bases físicas de las fuerzas.

La fuerza acelerativa es a la fuerza motriz como la velocidad al movimiento. Puesto que surge la cantidad de movimiento de la velocidad y de la cantidad de materia y la fuerza motriz surge de la fuerza acelerativa y de la misma cantidad de materia conjuntamente. Pues la suma de los impulsos de la fuerza aceleratriz sobre cada parte representa la fuerza motriz del todo. Por consiguiente, en la superficie de la Tierra, donde la gravedad aceleratriz o fuerza de gravitación es la misma para

todos los cuerpos, la gravedad motriz o peso es idéntico al cuerpo; pero si se sube por el espacio, donde la gravedad aceleratriz es menor, el peso disminuirá también y será siempre igual a la gravedad aceleratriz y al cuerpo conjuntamente. Así, en sitios donde la gravedad aceleratriz es dos veces menor, el peso de un cuerpo dos o tres veces menor será cuatro o seis veces menor.

Por cierto, que llamo en el mismo sentido fuerzas aceleratívas y motrices a las atracciones y a los impulsos. Utilizo unas por otras, e indiferentemente, las palabras atracción, impulso, tendencia de cualquier tipo a un centro, y lo hago considerando a tales fuerzas, no en su aspecto físico, sino sólo en el matemático. De ahí que cuide el lector de no creer que con estas palabras yo esté definiendo algún género o modo de acción o causa o propiedad física, o que estoy atribuyendo a los centros (que son puntos matemáticos) verdaderas fuerzas físicas, si me hallare diciendo que los centros atraen o que las fuerzas son centrales.

Escolio[8]

Nos ha parecido oportuno explicar hasta aquí los términos menos conocidos y el sentido en que se han de tomar en el futuro. En cuanto al tiempo, espacio, lugar y movimiento, son de sobra conocidos para todos. Hay que señalar, sin embargo, que el vulgo no concibe estas magnitudes si no es con respecto a lo sensible. De ello se originan ciertos prejuicios para cuya destrucción conviene que las distingamos en absolutas y relativas, verdaderas y aparentes, matemáticas y vulgares.

I. El tiempo absoluto, verdadero y matemático en sí y por su naturaleza y sin relación a algo externo, fluye uniformemente, y por otro nombre se llama duración; el relativo, aparente y vulgar, es una medida sensible y externa de cualquier duración, mediante el movimiento (sea la medida igual o desigual) y de la que el vulgo usa en lugar del verdadero tiempo; así, la hora, el día, el mes, el año.

II. El espacio absoluto, por su naturaleza y sin relación a cualquier cosa externa, siempre permanece igual e inmóvil; el relativo es cualquier cantidad o dimensión variable de este espacio, que se define por nuestros sentidos según su situación respecto a los cuerpos, espacio que el vulgo toma por el espacio inmóvil: así, una extensión espacial subterránea, aérea o celeste definida por su situación relativa a la Tierra. El espacio absoluto y el relativo son el mismo en especie y en magnitud, pero no permanecen siempre el mismo numéricamente. Pues si la Tierra, por ejemplo, se mueve, el espacio de nuestra atmósfera que relativamente y respecto a la Tierra siempre permanece el mismo, ahora será una parte del espacio absoluto por la que pasa el aire, después otra parte y así, desde un punto de vista absoluto, siempre cambiará.

III. Lugar es la parte del espacio que un cuerpo ocupa y es, en tanto que espacio, absoluto o relativo. Digo parte del espacio, no situación del cuerpo ni superficie

externa. Pues los sólidos iguales siempre tienen lugares iguales; las superficies, en cambio, por la desemejanza de las figuras son muchas veces desiguales. La situación, hablando propiamente, no tiene cantidad y no es tanto un lugar cuanto una propiedad del lugar. El movimiento del todo es el mismo que la suma de los movimientos de las partes, esto es, la traslación del todo de su lugar es la misma que la suma de las traslaciones de sus lugares de las partes, y por tanto, el lugar del todo es igual a la suma de los lugares de las partes y, por consiguiente, interno y solidario con el cuerpo.

IV. Movimiento absoluto es el paso de un cuerpo de un lugar absoluto a otro lugar absoluto, el relativo de un lugar relativo a otro lugar relativo. Así, en una nave empujada por las velas desplegadas, el lugar relativo de un cuerpo es aquella región de la nave en que está el cuerpo, o sea, la parte de la cavidad total que llena dicho cuerpo y que, por consiguiente, se mueve a la vez que la nave: mientras que el reposo relativo es la permanencia del cuerpo en la misma región de la nave o en la misma parte de su cavidad. Pero el reposo verdadero es la permanencia del cuerpo en la misma parte del espacio inmóvil en que se mueve la nave misma junto con su cavidad y todos sus contenidos. De donde si la Tierra verdaderamente está en reposo, el cuerpo que en la nave permanece relativamente en reposo se moverá verdadera y absolutamente con la misma velocidad con que la nave se mueve sobre la Tierra. Si la Tierra también se mueve, constará el verdadero y absoluto movimiento del cuerpo, parte del verdadero movimiento de la Tierra en el espacio inmóvil, parte de los movimientos relativos de la nave sobre la Tierra: y si el cuerpo también se mueve relativamente a la nave, constará su verdadero movimiento, parte del verdadero movimiento de la Tierra en el espacio inmóvil, parte de los movimientos relativos, tanto de la nave respecto a la Tierra como del cuerpo respecto a la nave, y de estos movimientos relativos constará el total movimiento relativo del cuerpo respecto a la Tierra. Así, si la parte de Tierra ocupada por la nave se mueve verdaderamente hacia Oriente con velocidad de 10 010 partes y la nave es empujada hacia Occidente por el viento y las velas con velocidad de 10 partes y un marino camina por la nave hacia Oriente con velocidad de 1 parte, el marino se moverá absolutamente en el espacio inmóvil hacia Oriente con la velocidad de 10 001 partes, y relativamente a la Tierra se moverá hacia Occidente con la velocidad de 9 partes.

El tiempo absoluto se distingue del relativo en Astronomía por la ecuación del tiempo vulgar. Pues desiguales son los días naturales, que son tenidos por iguales por el vulgo al medir el tiempo. Los astrónomos corrigen esta desigualdad al medir con tiempos más exactos los movimientos celestes. Es posible que no haya ningún movimiento igual con el que medir exactamente el tiempo. Todos los movimientos pueden acelerarse y retardarse, pero el flujo del tiempo absoluto no puede alterarse. La duración o permanencia de las cosas en la existencia es la misma, tanto si los movimientos son rápidos, como si son lentos, como si no los hubiese; por tanto, la duración se distingue claramente de sus medidas sensibles, a la vez que de ellas se

deduce por la ecuación astronómica. La necesidad de esta ecuación para la determinación de los fenómenos se patentiza tanto por el experimento del reloj oscilatorio como por los eclipses de los satélites de Júpiter.

Del mismo modo que el orden de las partes del tiempo es inmutable, así lo es el orden de las partes del espacio. Si éstas se movieran de sus lugares, se moverían (por así decirlo) de sí mismas. Pues el tiempo y el espacio son los cuasilugares de sí mismos y de todas las cosas. Todas las cosas se sitúan en el tiempo en cuanto al orden de la sucesión y en el espacio en cuanto al orden de lugar. Es de su esencia el ser lugares y es absurdo pensar que los lugares primeros se muevan. Por tanto, estos son lugares absolutos y únicamente las traslaciones desde estos lugares son movimientos absolutos.

Mas como estas partes del espacio no pueden verse y distinguirse unas de otras por medio de nuestros sentidos, en su lugar utilizamos medidas sensibles. Por las posiciones y distancias de las cosas a un cierto cuerpo que consideramos inmóvil, definimos todos los lugares; posteriormente interpretamos todos los movimientos por respecto a los antedichos lugares, en tanto que los concebimos como pasos de los cuerpos por estos lugares. Así, usamos de los lugares y movimientos relativos en lugar de los absolutos y con toda tranquilidad en las cosas humanas: para la Filosofía, en cambio, es preciso abstraer de los sentidos. Pues es posible que en la realidad no exista ningún cuerpo que esté en total reposo, al que referir lugar y movimiento.

Se distinguen el reposo y movimiento absolutos y relativos entre sí por sus propiedades, causas y efectos. Es propiedad del reposo que los cuerpos verdaderamente quietos están en reposo entre sí. Por tanto, al ser posible que un cuerpo cualquiera en la región de las estrellas fijas, o más lejos, permanezca en reposo absoluto y no se pueda saber por las situaciones respectivas de los cuerpos entre sí en nuestras cercanías si alguno de ellos conserva su posición constante respecto al cuerpo lejano, por ende no se puede definir el reposo verdadero por las posiciones relativas de estos cuerpos.

Es propiedad del movimiento que las partes que conservan su posición dada respecto al todo participan de los movimientos de los mismos todos. Pues todas las partes de los cuerpos que giran tienden a separarse del eje del movimiento y la fuerza de los móviles que se desplazan surge de la fuerza conjunta de las partes singulares. Así que, al mover los recipientes de los cuerpos, se mueven también las cosas que reposan relativamente dentro de esos recipientes. Y por tanto, el movimiento verdadero y absoluto no puede definirse por la traslación respecto a las cercanías del cuerpo que son consideradas como en reposo. Porque los cuerpos exteriores deben, no sólo ser considerados en reposo, sino también reposar verdaderamente. Pues de lo contrario, todo lo incluido, además de participar del de traslación de las cercanías de los recipientes, participará también de los movimientos verdaderos de los recipientes, y suprimida aquella traslación, no reposará verdaderamente, sino que solamente será considerado como en reposo; pues son los recipientes respecto a los contenidos como

la parte exterior del todo a la parte interior, o como la corteza al núcleo. Movida la corteza se mueve el núcleo también, como parte del todo, sin traslación de las cercanías de la corteza.

Es afín a la anterior propiedad el que, movido el lugar, se mueva también lo contenido en él: por tanto, el cuerpo que se mueve de un lugar movido, participa también del movimiento de su lugar. Por consiguiente, todos los movimientos, que surgen del movimiento de sus lugares, son partes solamente de movimientos totales y absolutos, y todo movimiento completo se compone del movimiento del cuerpo de su lugar primero, y del movimiento de este lugar del suyo, y así sucesivamente hasta que se llegue al lugar inmóvil, como en el ejemplo del navegante propuesto más arriba. De donde los movimientos completos y absolutos no pueden definirse si no es por lugares inmóviles y por eso más arriba los relacioné con los lugares inmóviles, y los relativos en cambio con los lugares móviles. Lugares inmóviles no son otra cosa que las posiciones constantes que conservan entre sí todas las cosas desde el infinito hasta el infinito y que, por tanto, siempre permanecen inmóviles y constituyen el espacio que llamo inmóvil.

Las causas, por las que los movimientos verdaderos y los relativos se distinguen mutuamente, son fuerzas impresas en los cuerpos para producir el movimiento. El movimiento verdadero ni se engendra ni se cambia, a no ser por fuerzas impresas en el mismo cuerpo movido; en cambio, el movimiento relativo puede generarse y cambiarse sin fuerzas impresas en tal cuerpo. Basta con imprimirlas solamente en los otros cuerpos respecto a los que se da la relación para que, cediendo éstos, cambie la relación dada en que consiste el reposo o movimiento relativo de aquel cuerpo. Por otra parte, el movimiento verdadero siempre se cambia por fuerzas impresas en el cuerpo movido, mientras que el movimiento relativo no se cambia necesariamente por estas fuerzas impresas. Pues si dichas fuerzas se aplican de tal modo hacia los demás cuerpos respecto a los que se da la relación que se conserve el lugar relativo, se conservará la relación en que consiste el movimiento relativo. Puede, pues, cambiarse todo el movimiento relativo mientras se conserva el verdadero, y conservarlo mientras se cambia el verdadero y absoluto; por tanto, el movimiento verdadero en absoluto puede consistir en tales relaciones.

Los efectos por los que los movimientos absolutos y los relativos se distinguen mutuamente son las fuerzas de separación del eje de los movimientos circulares. Pues en el movimiento circular meramente relativo estas fuerzas son nulas, pero en el verdadero y absoluto son mayores o menores según la cantidad de movimiento. Si se cuelga un cubo de un hilo muy largo y se gira constantemente hasta que el hilo por el torcimiento se ponga muy rígido y después se llena de agua y se deja en reposo a la vez que el agua, y entonces con un empujón súbito se hace girar continuamente en sentido contrario y, mientras se relaja el hilo, persevera durante un tiempo en tal movimiento, la superficie del agua será plana al principio, al igual que antes del movimiento del vaso, pero después, al transmitir éste su fuerza poco a poco al agua,

hace que ésta también empiece a girar sensiblemente, se vaya apartando poco a poco del centro y ascienda hacia los bordes del vaso, formando una figura cóncava (como yo mismo he experimentado) y con un movimiento siempre creciente sube más y más hasta que efectuando sus revoluciones en tiempos iguales que el vaso, repose relativamente en él. Muestra este ascenso el intento de separarse del centro del movimiento, y por tal intento se manifiesta y se mide el movimiento circular verdadero y absoluto del agua, y aquí contrario totalmente al movimiento relativo. Al principio, cuando mayor era el movimiento relativo del agua en el vaso, ese movimiento no engendraba ningún intento de separación del eje; el agua no buscaba el borde subiendo por los costados del vaso, sino que permanecía plana, y por tanto su movimiento circular verdadero no había aún empezado, pero después cuando decreció el movimiento relativo del agua, su ascensión por los costados del vaso indicaba el intento de separarse del eje y este conato mostraba su movimiento circular, verdadero y siempre creciente y al final convertido en máximo cuando el agua reposaba relativamente en el vaso. Por tanto, este conato no depende de la traslación del agua respecto de los cuerpos circundantes y, por tanto, el movimiento circular verdadero no puede definirse por tales traslaciones. Único es el movimiento circular verdadero de cualquier cuerpo que gira, y responde a un conato único como un verdadero y adecuado efecto; los movimientos relativos, en cambio, por las múltiples relaciones externas, son innumerables, pero como las relaciones carecen por completo de efectos verdaderos, a no ser en tanto que participan de aquel único y verdadero movimiento. De donde, incluso en el sistema de los que quieren que nuestro cielo gire bajo el cielo de las estrellas fijas y arrastre consigo a los planetas, los planetas y cada una de las partes del cielo que reposan relativamente a sus cercanías celestes, se mueven verdaderamente. Pues cambian sus posiciones relativas (al revés de lo que ocurre con las verdaderamente en reposo) y a la vez que son arrastrados con sus cielos participan de sus movimientos y, como partes de todos que giran, intentan alejarse de sus centros.

Así pues, las cantidades relativas no son las cantidades mismas que los nombres indican, sino las mediciones sensibles de ellas (verdaderas o erróneas) de las que se sirve el vulgo en lugar de las cosas medidas. Pero si los significados de las palabras se han de definir por el uso, por los nombres de tiempo, espacio, lugar y movimiento, más propiamente hay que entender estas mediciones sensibles; y el discurso será insólito y puramente matemático si se entienden aquí las cosas medidas. Por tanto, violentan las palabras consagradas los que interpretan aquí estos vocablos refiriéndolos a las cantidades medidas. Y no menos dañan a la Matemática y a la Filosofía quienes confunden las verdaderas cantidades con sus relaciones y las medidas vulgares.

Es muy difícil conocer los movimientos verdaderos de cada cuerpo y distinguirlos de hecho de los aparentes; además, porque las partes de aquel espacio inmóvil, en que los cuerpos se mueven verdaderamente, no se captan por los sentidos. Sin embargo,

no es el caso desesperado. Pues surgen argumentos, parte de los movimientos aparentes, que son diferencias de los movimientos verdaderos, parte de las fuerzas, que son causas y efectos de los movimientos verdaderos. Así, si a dos esferas, unidas entre sí por un hilo de determinada longitud, se las hace girar en torno al común centro de gravedad, aparecerá por la tensión del hilo el conato de las esferas de alejarse del eje de giro, y de ello se puede calcular la cantidad de movimiento circular. Después, si se aplican a la vez dos fuerzas iguales en las caras alternas de las esferas para aumentar o disminuir el movimiento circular, aparecerá, por el aumento o disminución de la tensión del hilo, el aumento o disminución del movimiento; y después, por fin, se podrían hallar las caras de las esferas en que deberían imprimirse las fuerzas para que el movimiento aumentase al máximo, esto es, las caras posteriores, o las que siguen en el movimiento circular. Pero, conocidas las caras que siguen y las caras opuestas que preceden, se conocerá la determinación del movimiento. De este modo se podría averiguar la cantidad y la determinación de este movimiento circular en un cierto vacío inmenso, donde nada hubiese externo y sensible con lo que se pudiesen comparar las esferas. Si ahora se establecen en dicho espacio algunos cuerpos lejanos que guarden entre sí cierta posición dada, tales como las estrellas fijas en nuestro firmamento, entonces no es posible saber a partir de la traslación relativa de las esferas entre los cuerpos si es a éstos o a aquéllos a quienes hay que atribuir el movimiento. Pero si se atiende al hilo y se encuentra que la tensión del mismo es la misma que la requerida por el movimiento de las esferas, será lícito concluir que el movimiento es de las esferas y entonces también deducir la determinación de este movimiento de la traslación de las esferas entre los cuerpos. A inferir, sin embargo, los movimientos verdaderos de sus causas, de sus efectos y diferencias con los aparentes y, al revés, sus causas y efectos a partir de los movimientos ya verdaderos, ya aparentes, se enseñará más extensamente en lo que sigue. Pues para este fin compuse el tratado siguiente.

AXIOMAS O LEYES DEL MOVIMIENTO[9]

Ley Primera

Todo cuerpo persevera en su estado de reposo o movimiento uniforme y rectilíneo a no ser en tanto que sea obligado por fuerzas impresas a cambiar su estado.

Los proyectiles perseveran en sus movimientos a no ser en cuanto son retardados por la resistencia del aire y son empujados hacia abajo por la gravedad. Una rueda, cuyas partes en cohesión continuamente se retraen de los movimientos rectilíneos, no cesa de dar vueltas sino en tanto en que el aire la frena. Los cuerpos más grandes de lo$ cometas y de los planetas conservan por más tiempo sus movimientos, tanto de avance como de rotación, realizados en espacios menos resistentes.

Ley II

El cambio de movimiento es proporcional a la fuerza motriz impresa y ocurre según la línea recta a lo largo de la cual aquella fuerza se imprime.

Si una fuerza cualquiera produce un movimiento dado, doblada producirá el doble y triplicada el triple, tanto si se aplica de una sola vez como si se aplica gradual y sucesivamente. Este movimiento (dado que se determina siempre en la misma dirección que la fuerza motriz) si el cuerpo se movía antes, o bien se añade sumándose a él, o se resta si es contrario, o se añade oblicuamente, si es oblicuo, y se compone con él según ambas determinaciones.

Ley III

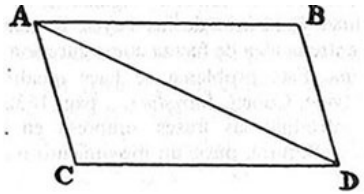
Con toda acción ocurre siempre una reacción igual y contraria: O sea, las acciones mutuas de dos cuerpos siempre son iguales y dirigidas en direcciones opuestas.

El que empuja o atrae a otro es empujado o atraído por el otro en la misma medida. Si alguien oprime una piedra con el dedo, también su dedo es oprimido por la piedra. Si un caballo arrastra una piedra atada con una soga, el caballo es retroarrastrado (por así decirlo) igualmente, pues la soga estirada en ambas direcciones y con el propio impulso de contraerse tirará del caballo hacia la piedra y de la piedra hacia el caballo y tanto se opondrá al progreso de uno cuanto ayude al avance del otro. Si un cuerpo cualquiera golpeando sobre otro cuerpo cambiara el movimiento de éste de algún modo con su propia fuerza, él mismo a la vez sufrirá el mismo cambio en su propio movimiento y en sentido contrario por la fuerza del otro cuerpo (por la igualdad de la presión mutua). A tales acciones son iguales los cambios de movimientos, no de velocidades, y siempre que se trate de cuerpos no fijados por otra parte. Igualmente los cambios de velocidad en sentido contrario, puesto que los movimientos cambian igualmente, son inversamente proporcionales a los cuerpos. Se cumple esta ley también para las atracciones como se comprobará en un escolio próximo.

Corolario Primero

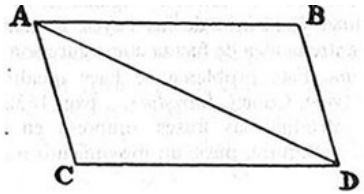
Un cuerpo recorre la diagonal de un paralelogramo bajo dos fuerzas conjuntas en el mismo tiempo en que los dos lados bajo las dos acciones por separado.

Si un cuerpo, en un tiempo dado, con la sola fuerza M impresa en el punto A es transportado con movimiento uniforme de A a B y con la sola fuerza N impresa en el mismo punto es transportado de A a C, complétese el paralelogramo ABDC y con ambas fuerzas el cuerpo será transportado en el mismo tiempo en diagonal de A a D. Porque, puesto que la fuerza actúa según la línea AC paralela a BD, esta fuerza, por la Ley II, en nada modificará la velocidad de acercamiento a la línea BD generada por otra fuerza. Por tanto, un cuerpo llegará en el mismo tiempo hasta la línea BD, tanto si se imprime la fuerza N como si no se imprime; por tanto, al cabo de tal tiempo se hallará al cuerpo en algún lugar de la línea BD. Por la misma razón al final de dicho tiempo se hallará en algún punto en la línea CD y, por tanto, es necesario encontrarla en la unión D de ambas líneas al cabo del mismo tiempo. Seguirá, pues un movimiento rectilíneo de A a D por la Ley I[10].

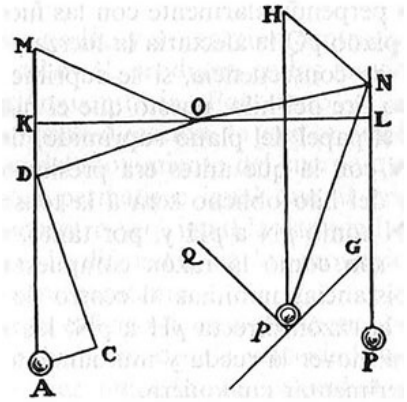


Corolario II

Así se evidencia la composición de la fuerza directa AD de las fuerzas oblicuas AB y BD, y a la vez la resolución de cualquier fuerza directa como AD en fuerzas oblicuas como AB y BD. Tales composición y resolución se confirman ampliamente por la mecánica.



Supongamos los radios desiguales OM y ON que parten del centro O de una rueda y sostienen los pesos A y P mediante los hilos MA y NP y se trata de hallar las fuerzas de los pesos para mover la rueda: Trácese la recta KOL por el centro O, que corte perpendicularmente los hilos en K y en L. Trácese un círculo con centro en O y con la distancia mayor OL de las de los intervalos OK y OL, círculo que cortará al hilo MA en D; y que sean AC paralela a la recta trazada OD, y DC perpendicular. Puesto que nada importa si los extremos de los hilos están fijos o no al plano de la rueda en K, L y D, los pesos tendrán el mismo valor si se suspenden de los puntos K y L o D y L. Exprésese ahora toda la fuerza del peso A por la longitud de toda la línea AD y ésta se resolverá en las fuerzas AC y CD de las cuales AC, atrayendo el radio OD directamente hacia el centro, en nada influye para mover la rueda; mientras la otra fuerza DC, atrayendo el radio DO perpendicularmente, equivale a atraer perpendicularmente al radio OL igual al propio OD; esto es, igual al peso P, dado que dicho peso es al peso A como la fuerza DC a la fuerza DA, es decir (por la semejanza de los triángulos ADC, DOK) como OK a OD o a OL. Por tanto, los pesos A y P, que son inversamente como los radios OK y OL, puesto uno a continuación de otro, tendrán el mismo valor y permanecerán en equilibrio lo que es una propiedad conocida en la balanza, en la palanca y en la polea. Si el peso de uno de los dos es mayor que el del otro, según esta razón su fuerza para mover la rueda será tanto mayor.



Puesto que si el peso *p* igual a P se suspende parcialmente de un hilo N*p* y de otra parte se apoya en el plano oblicuo *p*G y se trazan *p*H y NH, la primera perpendicular a la horizontal y la segunda perpendicular al plano *p*G, la fuerza del peso *p* hacia abajo puede representarse por la línea *p*H y puede resolverse en las fuerzas *p*N, HN. Si un plano *p*Q fuese perpendicular al hilo *p*N cortando al plano *p*G según una línea paralela al horizonte y el peso *p* solamente reposara sobre estos planos *p*Q, *p*G presionará a estos planos perpendicularmente con las fuerzas *p*N, HN; de tal modo que al plano *p*Q le afectaría la fuerza *p*N y al plano *p*G la fuerza HN. Y en consecuencia, si se suprime el plano *p*Q de modo que el peso tire del hilo, puesto que el hilo sosteniendo al peso hace ahora el papel del plano suprimido, tirará de él con la misma fuerza *p*N con la que antes era presionado el plano. De donde la tensión del hilo oblicuo será a la tensión del otro hilo perpendicular PN como *p*N a *p*H y, por tanto, si las razones de los pesos *p* y A son como la razón compuesta de las razones inversas de las distancias mínimas al centro de la rueda de sus hilos *p*N, AM y la razón directa *p*H a *p*N, los pesos tendrán el mismo valor para mover la rueda y mutuamente se equilibrarán, como puede experimentar cualquiera.

El peso *p*, por otra parte, al caer sobre dichos dos planos oblicuos tiene el carácter de cuña entre las dos caras internas de un cuerpo hendido y de este modo se ponen de manifiesto las fuerzas de la cuña y del mazo: Así ocurre que la fuerza con la que el peso *p* presiona al plano *p*Q es a la fuerza con que él mismo es empujado contra los planos según la línea *p*H por la gravedad o por el golpe del mazo, como *p*N a *p*H, y a la fuerza con que presiona al otro plano *p*G, como *p*N a NH. Incluso se deduce de aquí, por pura división semejante de fuerzas, la fuerza del tornillo sin fin, salvo que la cuña es empujada por una palanca. La utilización de este corolario es muy clara y pone de manifiesto su verdad, al depender de lo dicho toda la mecánica demostrada por los autores de diversos modos. Fácilmente también se derivan de lo dicho las fuerzas de las máquinas que suelen hacerse con ruedas, tomos, poleas, palancas, correas y pesas que suben directa u oblicuamente u otras energías mecánicas, tales

como las fuerzas de los tendones para mover los esqueletos de los animales.

Corolario III

La cantidad de movimiento que se obtiene tomando la suma de los movimientos hechos en una dirección y la diferencia de los realizados en sentido contrario, no cambia por la acción de los cuerpos entre sí.

Puesto que una acción y su reacción contraria son iguales por la Ley III y, por la Ley II, producen en los movimientos cambios iguales en dirección contraria. Por tanto, si los movimientos ocurren hacia la misma dirección, lo que se añade al cuerpo que se separa se detrae del movimiento del cuerpo que le sigue, de tal modo que la suma permanece igual que al principio. Y si los cuerpos van al encuentro será igual la cantidad detraída de cada móvil y, por ende, la diferencia de los movimientos en sentido opuesto permanecerá constante.

Así, si un cuerpo esférico A es el triple mayor que un cuerpo esférico B, y tiene dos unidades de velocidad, y B le sigue en la misma recta con diez unidades de velocidad y, por tanto, el movimiento total de A será con respecto a B como seis a diez: supongamos que tienen seis y diez unidades de movimiento, la suma total será de dieciséis unidades. Por tanto, en el encuentro de ambos si el cuerpo A obtiene tres o cuatro o cinco unidades, el cuerpo B pierde otras tantas y saldrá el cuerpo A después del choque con nueve, diez u once unidades y el cuerpo B con siete, seis o cinco, permaneciendo siempre la suma total de dieciséis como al principio. Si el cuerpo A obtuviese nueve o diez u once o doce unidades y, por ende, después del choque sale con quince o dieciséis o diecisiete o dieciocho, el cuerpo B al perder tantas unidades como aumenta A, o bien sale con una unidad al haber perdido otras nueve, o bien reposa al haber perdido sus diez unidades de movimiento, o bien regresa al haber perdido su movimiento (por así decirlo) con una parte más, o bien regresa con dos unidades por haber perdido doce unidades de movimiento hacia adelante. Así las sumas de los movimientos concurrentes 15 + 1 ó 16 + 0 y las diferencias de los contrarios 17 - 1 y 18 - 2 siempre darán dieciséis unidades de movimiento como antes del choque y la reflexión. Conocidos, por tanto, los movimientos con que salen los cuerpos después del choque se halla la velocidad de cualquiera de ellos suponiendo que es a la velocidad anterior al choque como el movimiento posterior es al anterior. Como en el ejemplo anterior, donde el movimiento del cuerpo A era de seis unidades antes del choque y de dieciocho después y la velocidad de dos unidades antes del choque: se hallará que su velocidad será de seis después del choque, diciendo que el movimiento de seis unidades es al movimiento de dieciocho unidades después del choque como la velocidad de dos unidades antes del choque es seis unidades después del choque.

Porque si se trata de cuerpos no esféricos o que se mueven según rectas distintas que se inciden oblicuamente y se pregunta por sus movimientos después del choque, hay que conocer la situación del plano en el que concurren los cuerpos en el punto de reunión; después (por el Corolario II) hay que distinguir en dos los movimientos de cada cuerpo, uno perpendicular a ese plano y otro paralelo al mismo; los movimientos paralelos, dado que los cuerpos actúan entre sí mutuamente según una línea perpendicular a dicho plano, han de conservarse idénticos antes y después de la reflexión. Y a los movimientos perpendiculares hay que atribuirles cambios iguales en sentido contrario, de tal modo que la suma de los movimientos concurrentes y la diferencia de los contrarios permanezca igual que antes. De tales choques suelen originarse también movimientos circulares de los cuerpos en torno a su centro, pero no consideraré en lo que sigue estos casos, ya que sería demasiado prolijo demostrar todo lo relativo a estas cuestiones[11].

Corolario IV

El centro común de gravedad de dos o más cuerpos no cambia su estado de movimiento o reposo por las acciones de los cuerpos entre sí; por tanto, el centro de gravedad común de los cuerpos en interacción (excluidas las acciones o impedimentos externos) o reposa o se mueve uniformemente en línea recta.

Pues si los puntos se mueven con movimiento uniforme en línea recta, y la distancia entre ellos se divide según una razón dada, el punto de división o reposa o se mueve uniformemente en línea recta. Después en el Lema XXIII y en su Corolario se demostrará esto, si el movimiento de los puntos ocurre en el mismo plano; del mismo modo se puede demostrar si tales movimientos no ocurren en el mismo plano. Luego si cualesquiera cuerpos se mueven uniformemente en líneas rectas, el centro común de gravedad de dos cualesquiera o está en reposo o se mueve en línea recta con movimiento uniforme; ello por el hecho de que la línea que une los centros de los cuerpos que se mueven uniformemente en línea recta es dividida según una determinada razón desde el centro común. De modo semejante el centro común de estos dos y un tercer cuerpo cualquiera o reposa o se mueve uniformemente en línea recta y ello porque desde dicho centro se divide la distancia entre el centro común de los dos cuerpos y el centro del tercero, según una razón dada. Del mismo modo el centro común de estos tres y el de un cuarto cualquiera o reposa o se mueve uniformemente en línea recta, puesto que desde dicho centro se divide la distancia entre el centro común de los tres y el centro del cuarto según una razón dada y así hasta el infinito. Por tanto, en un sistema de cuerpos que ejercen acciones mutuamente entre sí y carecen por completo de otros influjos exteriores y, por tanto, se mueve cada uno uniformemente en una línea recta el centro común de gravedad de

todos o está en reposo o se mueve uniformemente en línea recta.

Además en un sistema de dos cuerpos que actúan entre sí mutuamente, siendo las distancias de los centros de cada uno respecto al centro común de gravedad inversas a los cuerpos, los movimientos relativos de los propios cuerpos tanto al acercarse a dicho centro como al separarse serán iguales entre sí.

Por tanto, tal centro ni se adelanta ni se atrasa, ni sufre cambio alguno en su estado de movimiento o reposo por causa de cambios iguales de movimiento realizados en sentidos contrarios, ni tampoco por interacciones de esos cuerpos. Pero en un sistema de muchos cuerpos, dado que el centro común de gravedad de dos cuerpos cualesquiera en mutua interacción no se ve afectado en absoluto en su estado por dicha acción, y el centro común de gravedad de los demás, con los que dicha acción no tiene relación alguna, nada sufre por ella, en cambio la distancia de estos dos centros se divide desde el centro común de todos los cuerpos en partes inversamente proporcionales a las sumas totales de los cuerpos de los que son centros; y en consecuencia, al conservar aquellos dos centros su estado de movimiento o reposo, el centro común de todos también conserva el suyo: es evidente que el centro común de todos nunca cambia su estado de movimiento o reposo por la acción de dos cuerpos entre sí. En semejante sistema, pues, todas las acciones de cuerpos entre sí, o son acciones entre dos o son compuestas de acciones entre dos cuerpos y, por tanto, nunca producen cambio en el estado de movimiento o reposo del centro común de todos. Por tanto, como aquel centro, cuando los cuerpos no se influyen mutuamente, o reposa o discurre uniforme y rectilíneamente, persevera, pues, idéntico, sin que lo impidan las acciones de los cuerpos entre sí, o siempre en reposo o siempre en movimiento uniforme y rectilíneo, a no ser que sea apartado de dicho estado por fuerzas impresas extrínsecamente al sistema. Así pues, la ley de un sistema de muchos cuerpos es la misma que la de un solo cuerpo en lo que se refiere a la permanencia en estado de movimiento o reposo, pues el movimiento progresivo, tanto de un cuerpo aislado como de un sistema de cuerpos, debe apreciarse siempre por el movimiento del centro de gravedad.

Corolario V

Los movimientos entre sí de los cuerpos incluidos en un determinado espacio son los mismos, ya esté dicho espacio en reposo, ya se mueva recta y uniformemente sin movimiento circular.

Pues las diferencias de los movimientos tendentes a un lado y las sumas de las tendentes al lado contrario, son las mismas desde el principio en ambos casos (por hipótesis) y de tales sumas y diferencias proceden los choques y fuerzas con los que los cuerpos mutuamente se afectan. Por tanto, en virtud de la Ley II, serán iguales los

efectos de los choques en ambos casos; y por tanto los movimientos de los cuerpos permanecerán iguales entre sí en un caso a los movimientos entre sí en el otro. Esto se comprueba con un experimento clarísimo: todos los movimientos se comportan de modo igual entre sí en una nave tanto si se halla en reposo como si se halla en movimiento uniforme y rectilíneo.

Corolario VI

Si los cuerpos se moviesen entre sí de cualquier modo y fuesen empujados por fuerzas aceleratívas iguales según líneas paralelas, todos ellos se seguirán moviendo entre sí del mismo modo que si no estuviesen empujados por tales fuerzas.

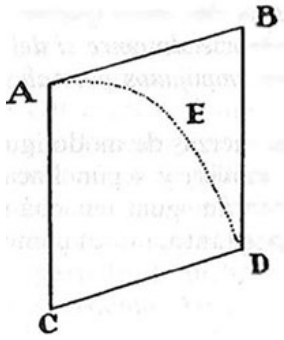
Pues al actuar tales fuerzas de modo igual (según las magnitudes de los cuerpos a mover) y según líneas paralelas, todos los dichos cuerpos se moverán igual (en cuanto a la velocidad) en virtud de la Ley II y, por tanto, nunca cambiarán sus posiciones y movimientos entre sí.

Escolio

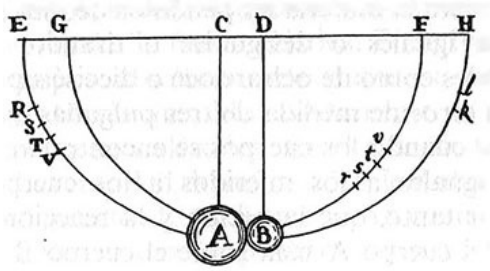
Hasta ahora he ofrecido los principios aceptados por los matemáticos y confirmados por muy amplia experiencia. Por las dos leyes primeras y los dos Corolarios primeros, *Galileo* descubrió que la caída de los graves ocurre según la razón cuadrada del tiempo y que el movimiento de los proyectiles ocurre en parábola, de acuerdo con la experiencia, a no ser en la medida en que tales movimientos se retardan un poco por la resistencia del aire. Para un cuerpo que cae la gravedad uniforme, actuando de modo igual en cada unidad de tiempo sobre partículas iguales, imprime fuerzas iguales en dicho cuerpo y genera velocidades iguales. Y en la totalidad del tiempo imprime toda la fuerza y genera la velocidad total proporcional al tiempo. Y los espacios recorridos en tiempos proporcionales son como las velocidades y los tiempos conjuntamente; esto es como la razón cuadrada de los tiempos. Y en un cuerpo lanzado hacia arriba la gravedad uniforme imprime fuerzas y disminuye la velocidad en proporción al tiempo; y los tiempos de ascensión hasta el punto más alto son como son las velocidades a disminuir, y la altura máxima es como los tiempos y las velocidades conjuntas o en razón cuadrada de las velocidades. Y el movimiento de un cuerpo proyectado según una línea recta cualquiera se compone del movimiento procedente de la proyección con el procedente de la gravedad. Así si un cuerpo A con el solo movimiento de proyección pudiese en un cierto tiempo recorrer la recta AB y con el movimiento de caída en el mismo tiempo pudiese recorrer la línea de caída AC; complétese el paralelogramo ABCD y dicho cuerpo al

final del tiempo dado con el movimiento compuesto se hallará en el punto D; y la línea curva AED que describe dicho cuerpo será una parábola a la que la recta AB es tangente en A y cuya ordenada BD es como AB cuadrado[12]. De dichas leyes y corolarios dependen las demostraciones acerca de los péndulos oscilantes, de acuerdo con la experiencia diaria de los relojes. A partir de estos principios y de la tercera Ley hallaron *sir Christopher Wren*, *John Wallis* S. T. D. y el caballero *Christian Huygens*, príncipes de los geómetras de la época última, las reglas del choque y reflexión mutua de dos cuerpos, y casi a la vez la comunicaron a la *Sociedad Real* colaborando entre ellos plenamente (en cuanto a estas leyes); primero, en verdad, *Wallis* y después *Wren* y *Huygens* ofrecieron su hallazgo. Pero además esta verdad fue corroborada por *Wren* ante la *Royal Society* por medio del experimento de los péndulos; también el preclaro *Mariotte* se dignó exponer enseguida todo esto en un libro completo[13]. Pero es cierto que para que este experimento concuerde con las teorías de modo aceptable, hay que tener en cuenta tanto la resistencia del aire cómo la fuerza elástica de los cuerpos concurrentes. Suspendamos dos cuerpos esféricos AB de hilos paralelos iguales, AC, BD de los centros C, D. Desde dichos centros y con tales intervalos tracemos los semicírculos EAF, GBH bisecados por los radios CA, CB. Empujemos el cuerpo A hasta el punto R del arco EAF y (apartando el cuerpo B) soltémosle desde allí y tras una oscilación regresará al punto V. RV es el retardo debido a la resistencia del aire. Sea ST la cuarta parte puesta en medio de este RV, de tal modo que RS sea igual a TV y RS sea a ST como 3 a 2. Por otra parte dicho ST representará muy aproximadamente el retardo en el descenso desde S a A. Pongamos B en su lugar. Dejemos caer el cuerpo A desde el punto S y su velocidad en el punto de reflexión A será sin gran error muy similar a la que tendría si cayese en vacío desde el punto T. Representemos pues dicha velocidad por la cuerda del arco TA, puesto que es de sobra conocido para los geómetras el enunciado de que la velocidad de un péndulo en el punto más bajo es como la cuerda del arco que describe al caer. Después de la reflexión el cuerpo A alcanzará el punto *s*, y el cuerpo B el punto *k*. Apartemos el cuerpo B y hallemos el punto *v*. Si el cuerpo A se lanza desde dicho punto y tras una oscilación vuelve al punto *r*, sea *st* la cuarta parte del propio *rv*, situada en medio, de tal modo que *rs* y *tv* sean iguales y esté representada por la cuerda del arco *t*A la velocidad alcanzada por el cuerpo A en el punto A inmediatamente después de la reflexión. Pues *t* será el punto verdadero y correcto al que debería llegar el cuerpo A, suprimida la resistencia del aire. Del mismo modo hay que corregir el punto *k*, hasta el que llega el cuerpo B, y hallar el punto *l*, hasta el cual debió ascender dicho cuerpo en el vacío. De este modo podría experimentarse todo como si estuviésemos en un medio vacío. Finalmente habría que multiplicar (por así decirlo) al cuerpo A por la cuerda del arco TA que representa su velocidad para obtener su movimiento en el punto A inmediatamente antes de la reflexión; después por la cuerda del arco *t*A para obtener su movimiento en el punto A inmediatamente después de la reflexión. Y del mismo modo el cuerpo B habría de ser multiplicado

por la cuerda del arco BI para obtener su movimiento inmediatamente después de la reflexión. Y de modo semejante cuando dos cuerpos son lanzados a la vez desde lugares diversos, hay que hallar los movimientos de cada uno, tanto antes como después de la reflexión, y sólo entonces se comparan entre sí los movimientos y se deducen los efectos de la reflexión. Con experimentos sobre la materia en péndulos de diez pies, lo mismo que con pesos iguales o desiguales o tirando cuerpos desde distancias grandes como de ocho, doce o dieciséis pies, siempre he encontrado sin error de medida de tres pulgadas que los cambios de movimiento, cuando los cuerpos se encontraban directamente, eran siempre iguales a los inferidos a los cuerpos en sentido contrario y, por tanto, que la acción y la reacción siempre eran iguales. Así, si el cuerpo A caía sobre el cuerpo B en reposo con nueve unidades de movimiento y, perdidas siete, partía después del choque con dos, el cuerpo B retrocedía con estas otras siete. Si estos cuerpos iban al encuentro, A con doce unidades y B con seis.



A retornaba con dos, B regresaba con ocho, efectuada la resta de catorce entre uno y otro. Quítense doce unidades del movimiento del cuerpo A y no quedará nada; quítense otras dos más y se obtendrá un movimiento de dos unidades en la otra dirección; y del mismo modo del movimiento de B que tiene seis unidades quítense catorce y se obtendrán ocho en sentido contrario. Y cuando los cuerpos iban en la misma dirección, A más rápido con catorce unidades de velocidad, B más lento con cinco, después del choque salía A con cinco y B con catorce al haberse transferido nueve unidades de A a B. Y lo mismo en el resto. Jamás se alteraba en la reunión o colisión de cuerpos la cantidad de movimientos concurrentes y de las diferencias de los movimientos contrarios pues el error de una pulgada o dos en las medidas lo atribuía a la dificultad de llevar a cabo cada una de modo suficientemente preciso. Era difícil tanto soltar a la vez los péndulos de tal modo que los cuerpos se encontraran en los puntos más bajos AB, como registrar los puntos *s*, *k* a que ascendían los cuerpos después del encuentro. Y por otra parte, los propios cuerpos de los péndulos con densidades desiguales en diversos puntos y con textura irregular por otras causas, llevaba a error.



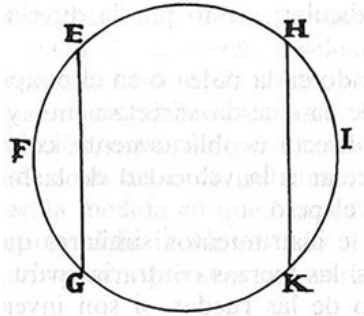
Pero no sea que alguien objete que la regla para cuya prueba se ha aducido este experimento presupone que los cuerpos son absolutamente duros o por lo menos perfectamente elásticos, casos que no se hallan en lo más mínimo entre las cosas naturales; añado que los experimentos descritos acontecen tanto en cuerpos blandos como duros, sin depender en lo más mínimo del grado de dureza. Pues si dicha regla hubiese de ensayarse en cuerpos no completamente duros, la reflexión deberá disminuir sólo en cierta cantidad proporcional a la fuerza elástica. En la teoría de *Wren* y *Huygens* los cuerpos absolutamente duros se reflejan con la misma velocidad del choque. Mas ciertamente se puede afirmar esto de los perfectamente elásticos. En el caso de los imperfectamente elásticos la velocidad de retorno ha de deducirse junto con la fuerza elástica; por cuanto tal fuerza (salvo cuando las partes se doblan con el choque, o se expanden como lo hacen bajo la acción de un martillo) es cierta y determinada (según creo) y hará que los cuerpos se separen con velocidad relativa que estará en una razón dada con la velocidad relativa de choque. Esto lo ensayé con pelotas de lana bien prensadas. Primero dejando caer el péndulo y midiendo la reflexión hallé la fuerza elástica; después mediante esta fuerza calculé las reflexiones en otros ejemplos de choque y los experimentos concordaban. Retrocedían siempre las pelotas con una velocidad relativa tal que se relacionaba con la velocidad relativa de choque, aproximadamente como 5 a 9.

Las de acero retrocedían casi con la misma velocidad, las de madera con un poco menos, mientras que en las de vidrio la proporción era casi 15 a 16. Y así de este modo se ha comprobado la tercera Ley, en cuanto a las acciones y reacciones, por medio de una teoría que se adecúa plenamente con los experimentos.

En las atracciones muestro esto brevemente como sigue: Imaginemos dos cuerpos cualesquiera A, B que se atraen mutuamente; imaginemos un obstáculo cualquiera interpuesto entre ellos que impida su choque. Si uno de los dos cuerpos A es atraído hacia B más que B hacia A, el obstáculo será empujado más por la presión del cuerpo A que por la presión del cuerpo B y, en consecuencia, no permanecerá en equilibrio. Prevalecerá la presión más fuerte y hará que el sistema de los dos cuerpos y el obstáculo se mueva hacia B en línea recta y, con un movimiento continuamente acelerado, en espacios libres iría hasta el infinito. Lo cual es absurdo y contrario a la primera Ley, pues por la primera Ley el sistema deberá permanecer en su estado de

reposo o de movimiento uniforme y rectilíneo y, por tanto, ambos cuerpos empujarán igualmente al obstáculo y, por tanto, se atraerán mutuamente con igualdad. Ensayé esto con hierro e imán. Si se colocan éstos aparte en vasijas adecuadas y contiguas y flotan juntas en un baño de agua, ninguno de ellos impulsará al otro sino que mutuamente con iguales atracciones cada uno sostendrá los impulsos del otro y finalmente puestos en equilibrio reposarán.

Asimismo la gravedad entre la Tierra y sus partes es mutua. Divídase la Tierra FI con un plano cualquiera EG en dos partes EGF y EGI: Los pesos de estas dos partes serán mutuamente iguales. Pues si con otro plano HK, paralelo al anterior EG, se divide la parte mayor EGI en dos partes EGKH y HKI de las cuales AKI es igual a la primeramente separada EFG, es evidente que la parte intermedia EGKH por su peso no se inclinará a ninguna de las dos partes, sino que, por así decirlo, quedará suspendida y en equilibrio entre ambas y reposará. Pero la parte extrema HKI con todo su peso caerá sobre la parte media y la empujará hacia el otro extremo EGF y, por tanto, la fuerza con la que EGI suma de las dos partes HKI y EGKH, tiende hacia la tercera zona EGF, es igual al peso de la parte HKI, esto es al peso de la parte tercera EGF. Y por tanto los pesos de las dos partes EGI, EGF, son iguales entre sí como pretendí demostrar. Y a no ser que dichos pesos fuesen iguales toda la Tierra flotando en el éter libre cedería al mayor peso y huyendo de él se alejaría perdiéndose en el infinito[14].



Al igual que son equivalentes en los choques y reflexiones los cuerpos cuyas velocidades son inversamente como las fuerzas ínsitas, del mismo modo son equivalentes y se sostienen mutuamente con empujes contrapuestos en el uso de instrumentos mecánicos aquellos agentes cuyas velocidades calculadas según la determinación de las fuerzas son inversamente como las fuerzas. Así los pesos son equivalentes al mover los brazos de la balanza, los cuales cuando se mueve la balanza son inversamente como sus velocidades hacia arriba o hacia abajo; esto es, los pesos, si suben y bajan en línea recta, son equivalentes y son inversamente como las distancias desde los puntos de los que se suspenden hasta el eje de la balanza; pero si suben o bajan oblicuamente impedidos por planos inclinados u otros obstáculos

interpuestos, entonces son inversamente como los ascensos y descensos realizados de acuerdo con la perpendicular; y esto por la dirección de la gravedad hacia abajo.

Del mismo modo en la polea o en el polipasto la fuerza de la mano que tira de la cuerda directamente y que sea al peso, ascendiendo éste directa u oblicuamente, como la velocidad del ascenso perpendicular a la velocidad de la mano que tira de la cuerda, sostendrá el peso.

En los relojes e instrumentos similares que estén hechos de ruedas conectadas, las fuerzas contrarias para promover o impedir el movimiento de las ruedas, si son inversamente como las velocidades de las partes de las ruedas en que actúan, se sostendrán mutuamente. La fuerza de un torno para apretar un cuerpo es a la fuerza de la mano que maneja la manivela como la velocidad circular de la manivela en el lugar donde se aplica la mano a la velocidad de avance del torno hacia el cuerpo oprimido. Las fuerzas con las que una cuña obliga a las dos partes de un madero fijo son a la fuerza del mazo en la cuña como el avance de la cuña según la dirección de la fuerza impresa en ella por el mazo es a la velocidad con que las partes del madero ceden a la cuña según líneas perpendiculares a las caras de la cuña. E igual es la explicación de todas las máquinas.

La eficacia y utilidad de éstas consiste únicamente en que disminuyendo la velocidad aumentamos la fuerza y viceversa; de donde se resuelve para todo tipo de máquinas apropiadas el problema *de mover un determinado peso con una determinada fuerza* o de superar con una fuerza dada otra resistencia también dada. Pues si se hicieran las máquinas de tal modo que las velocidades del agente y del resistente fuesen inversamente como las fuerzas, el agente sostendrá la resistencia; y la vencerá con una mayor disparidad de velocidades. Hasta el punto de que si la disparidad de velocidades es tan grande que quede vencida toda resistencia, tanto la procedente de la contigüidad o del rozamiento de los cuerpos como de la cohesión de los cuerpos continuos o que han de ser separados o de los pesos que han de ser elevados, superada toda esa resistencia, la fuerza sobrante producirá una aceleración del movimiento proporcional a ella misma, parte en la máquina parte en el cuerpo resistente. Por lo demás no es el intento presente tratar de mecánica. Hasta aquí solamente pretendí mostrar en qué medida es evidente y cuán cierta es la tercera Ley del movimiento. Pues si se considera la acción de un agente como el producto de su fuerza y velocidad, y, de modo semejante, la reacción del resistente, como el producto de las velocidades de sus partes singulares y de las fuerzas de resistencia procedentes de su fricción, cohesión, peso y aceleración, acción y reacción en el uso de toda clase de instrumentos siempre serán iguales entre sí. Y, en la medida en que la acción se propaga por medio del instrumento y al fin incide en todo el cuerpo resistente, su determinación última siempre será contraria a la determinación de la reacción.

Libro primero
DEL MOVIMIENTO DE LOS CUERPOS

Sección primera
DEL MÉTODO DE LAS RAZONES PRIMERAS
Y ÚLTIMAS POR CUYO MEDIO
SE DEMUESTRA LO QUE SIGUE

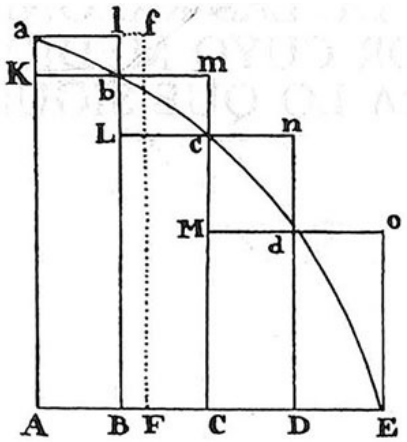
Lema Primero

Las cantidades, así como las razones de cantidades, que tienden a la igualdad constantemente en un cierto tiempo finito y antes del límite de dicho tiempo se aproximan mutuamente más que una diferencia dada, al final se hacen iguales.

Si lo niegas, sean al final desiguales y sea su diferencia final D. Luego no pueden acercarse a la igualdad más que hasta una diferencia dada D. Contra la hipótesis.

Lema II

Si en una figura AacE comprendida entre las rectas Aa, AE, y la curva acE se inscriben varios paralelogramos Ab, Bc, Cd, etc. construidos sobre bases iguales AB, BC, CD, etc. y con lados Bb, Cc, Dd paralelos al lado Aa de la figura; y se completan los paralelogramos aKbl, bLcm, cMdn, etc., si entonces se disminuye la anchura de estos paralelogramos y se aumenta infinitamente el número de ellos: digo que las razones últimas que se dan entre la figura inscrita AKbLcMdD, la circunscrita AalbmcndoE y la curvilínea AabcdE son razones de igualdad.

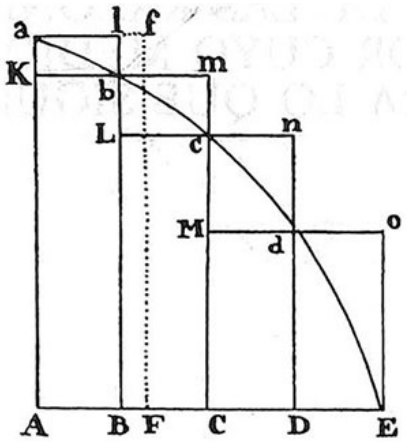


Pues la diferencia de las figuras inscritas y circunscritas es la suma de los paralelogramos K*l*, L*m*, M*n*, D*o*, esto es (debido a la igualdad de las bases de todos ellos) al rectángulo construido sobre la base de uno de ellos K*b* y la suma de las alturas A*a*, esto es, el rectángulo AB*la*. Pero este rectángulo, dado que su anchura AB disminuye *in infinitum*, deviene menor que uno dado cualquiera. Luego (por el Lema I) la figura inscrita y la circunscrita, y mucho más la curvilínea intermedia, al final se hacen iguales. Q. E. D.

Lema III

Las mismas razones últimas son también razones de igualdad cuando las anchuras AB, BC, CD, etc., de los paralelogramos son desiguales y todas disminuyen in infinitum.

Sea AF igual a la anchura máxima, y complétese el paralelogramo FAaf. Este será mayor que la diferencia de la figura inscrita y de la figura circunscrita; pero al disminuir su anchura AF infinitamente se hará menor que un rectángulo cualquiera dado. Q. E. D.



Corolario 1. De aquí que la suma última de paralelogramos evanescentes coincide en todo punto con la figura curvilínea.

Corolario 2. Y mucho más la figura rectilínea, comprendida por las cuerdas de los arcos evanescentes *ab*, *bc*, *cd*, etc., coincidirá al fin con la figura curvilínea.

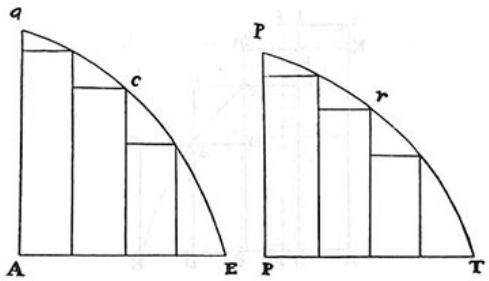
Corolario 3. Lo mismo que la figura rectilínea circunscrita comprendida por tangentes de los mismos arcos.

Corolario 4. Por ello, estas figuras últimas (en cuanto a los perímetros *ac*E) no son rectilíneas, sino límites curvilíneos de figuras rectilíneas.

Lema IV

Si en dos figuras AacE, PprT se hallan inscritas (como antes) dos series de paralelogramos y el número es idéntico en ambas y al ir disminuyendo las anchuras in infinitum las razones últimas de los paralelogramos en una figura respecto a los paralelogramos de la otra, uno a uno, permanecen iguales; digo que las dos figuras AacE y PprT están mutuamente en esa misma razón.

Pues tal como los paralelogramos son uno a uno, así (por composición) es la suma de todos a la suma de todos, y así será la figura a la figura; pues están (por el Lema III) la primera figura a la suma primera, y la figura posterior a la suma posterior, en razón de igualdad. Q. E. D.



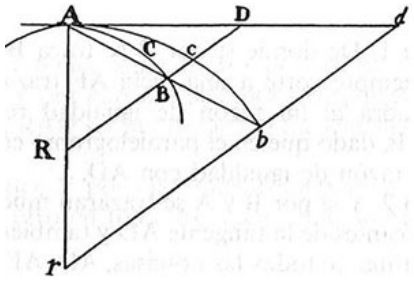
COROLARIO. De aquí que si dos cantidades de cualquier clase se dividen de algún modo en el mismo número de partes; y tales partes, al ir aumentando su número y disminuyendo su tamaño *in infinitum*, alcanzan una razón dada entre sí, la primera con la primera, la segunda con la segunda, y las demás por su orden con las otras, estarán todas entre sí en aquella razón dada. Pues si en las figuras de este Lema se toman los paralelogramos entre sí como partes, las sumas de las partes siempre serán como la suma de los paralelogramos; y por esto al ir aumentando el número de paralelogramos y de partes y disminuyendo el tamaño de los mismos *in infinitum*, estarán en razón última de paralelogramo a paralelogramo, esto es (por hipótesis) en razón última de parte a parte.

LEMA V

Todos los lados homólogos de figuras semejantes son proporcionales entre sí tanto si son rectilíneas como si son curvilíneas; y las áreas son como el cuadrado de los lados.

LEMA VI

Si un arco ACB dado en posición es subtendido por la cuerda AB, y en el punto A, en medio de la curvatura continua, es tocado por la recta tangente AD prolongada hacia ambos lados y después los puntos A y B se acercan y coinciden, digo que el ángulo BAD, contenido entre la cuerda y la tangente, disminuye in infinitum y finalmente desaparece.

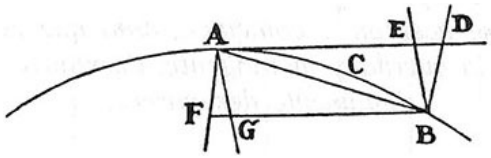


Pues si dicho ángulo no desapareciera, el arco ACB contendría junto con la tangente AD un ángulo igual a un recto y, por tanto, la curvatura en el punto A no sería continua, contra la hipótesis.

LEMA VII

Con los mismos supuestos, digo que la razón última entre la cuerda, el arco y la tangente entre sí, es la razón de igualdad.

Pues mientras el punto B se aproxima al punto A, imagínese a AB y AD prolongadas a los puntos lejanos *b* y *d*, mientras se traza *bd* paralela a la secante BD. Sea el arco A*cb* siempre semejante al arco ACB. Si los puntos A y B se aproximan, por el Lema anterior el ángulo *d*A*b* se desvanece, y las líneas rectas finitas A*b* y A*d* y el arco intermedio A*cb* coincidirán y, por tanto, serán iguales. De donde se sigue que las rectas AB y AD y el arco ACB siempre proporcionales a aquéllas también desaparecerán y tendrán como razón última la igualdad. Q. E. D.



COROLARIO 1. De donde si por B se traza BF paralela a la tangente, que siempre corte a una recta AF trazada por A, esta BF siempre tendrá al fin razón de igualdad respecto al arco evanescente ACB, dado que en el paralelogramo completo AFBD siempre tendrá razón de igualdad con AD.

COROLARIO 2. Y si por B y A se trazaran muchas rectas, BE, BD, AF, AG, secantes de la tangente AD y también de su paralela BF, la razón última de todas las abscisas, AD, AE, BF, BG, de la cuerda y del arco AB será entre sí la razón de igualdad.

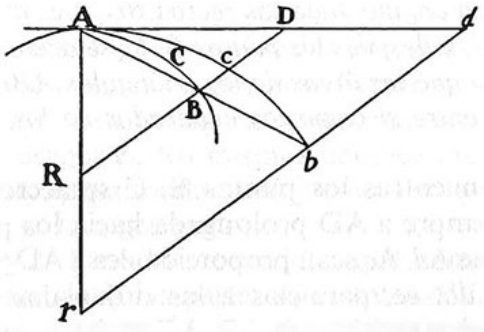
COROLARIO 3. Y, por tanto, todas estas líneas pueden sustituirse entre sí en todos

los razonamientos sobre razones últimas.

Lema VIII

Si las rectas dadas AR, BR, junto con el arco ACB, la cuerda AB y la tangente AD, constituyen tres triángulos RAB, RACB, RAD y los puntos A, B se aproximan mutuamente, digo que la forma última de los triángulos evanescentes es la de semejanza y la razón última de igualdad.

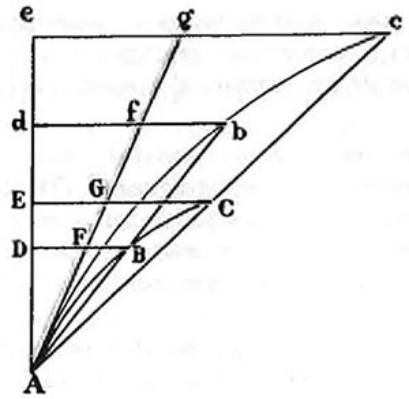
Pues mientras se aproxima B a A imagínese siempre que AB, AD y AR se prolongan hasta los puntos más lejanos *b*, *d* y *r*, y que se traza *rbd* paralela a RD y que el arco A*cb* permanece siempre semejante al arco ACB. Al coincidir los puntos A y B el ángulo *b*A*d* desaparecerá y, por tanto, los triángulos siempre finitos *r*A*b*, *r*A*cb* *r*A*d* coincidirán y por eso mismo serán semejantes e iguales. Y por ello, sus siempre semejantes y proporcionales RAB, RACB, RAD serán también al fin semejantes e iguales entre sí. Q. E. D.



Corolario. Y por ello, tales triángulos pueden ser sustituidos uno por otro en toda argumentación sobre razones últimas.

Lema IX

Si la recta AE y la curva ABC de posición dada se cortan mutuamente en un ángulo dado A y se aplican ordenadamente a dicha recta en un ángulo dado las rectas BD, CE, que a su vez tocan la curva en B, C, y después los puntos B, C, se acercan a la vez hacia el punto A, digo que las áreas de los triángulos ABD, ACE serán al final entre sí como los cuadrados de los lados.



En efecto, mientras los puntos B, C se acercan al punto A, imaginemos siempre a AD prolongada hacia los puntos lejanos *d* y *e*, de modo que A*d*, A*c* sean proporcionales a AD y AE y tracemos las ordenadas *dh*, *ec*, paralelas a las ordenadas DB, EC y que corten a las prolongaciones de AB, AC en *b* y *c*. Supóngase que se traza la curva A*bc* semejante a la curva ABC así como la recta A*g* tangente a las curvas en A y que corta a las ordenadas DB, EC, *db*, *ec* en F, G, *f*, *g*. Entonces, manteniendo la longitud A*e*, que los puntos B, C se aproximen al punto A y al desaparecer el ángulo *c*A*g* las áreas curvilíneas A*bd*, A*ce* coincidirán con las rectilíneas A*fd*, A*ge*; y por ello (por el Lema V) serán entre sí como el cuadrado de los lados A*d*, A*e*. Pero a dichas áreas siempre serán proporcionales las áreas ABD, ACE y a tales lados siempre serán proporcionales los lados AD, AE. Luego también las áreas ABD, ACE serán al final como el cuadrado de los lados AD, AE. Q. E. D.

Lema X

Los espacios que un cuerpo recorre empujado por una fuerza cualquiera finita tanto si dicha fuerza es determinada y siempre la misma, como si es continuamente creciente o decreciente, son como los cuadrados de los tiempos al comienzo del movimiento.

Represéntense los tiempos por las líneas AD, AE y las velocidades originadas por las ordenadas DB, EC; los espacios descritos con estas velocidades serán como las áreas ABD, ACE descritas por estas ordenadas, esto es (por el Lema IX) al comienzo del movimiento como el cuadrado de los tiempos AD, AE. Q. E. D.

Corolario 1. Y de aquí se deduce fácilmente que los recorridos de los cuerpos que describen partes semejantes de figuras semejantes en tiempos proporcionales, recorridos que son originados por fuerzas iguales aplicadas de modo semejante a los cuerpos y que son medidos por las distancias de los cuerpos respecto a los lugares de

las figuras semejantes a los que hubiesen llegado dichos cuerpos en los mismos tiempos proporcionales sin tales fuerzas, son aproximadamente como los cuadrados de los tiempos en que se han generado.

COROLARIO 2. Pero los recorridos que son originados por fuerzas proporcionales a las partes semejantes de figuras semejantes aplicadas semejantemente, son como las fuerzas multiplicadas por el cuadrado de los tiempos.

COROLARIO 3. Lo mismo ha de entenderse de cualesquiera espacios que describan los cuerpos bajo la acción de distintas fuerzas. Serán, respecto al comienzo del movimiento, como las fuerzas multiplicadas por el cuadrado de los tiempos.

COROLARIO 4. Y por tanto, las fuerzas están, al comienzo del movimiento, en razón directa a los espacios descritos y en razón inversa a los cuadrados de los tiempos.

COROLARIO 5. Y los cuadrados de los tiempos son directamente como los espacios descritos e inversamente como las fuerzas.

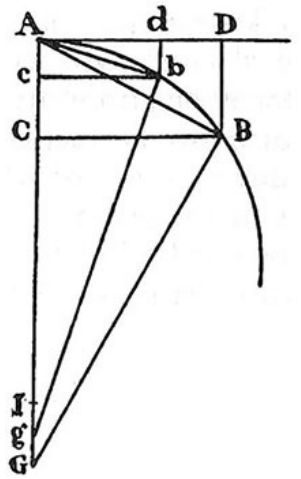
ESCOLIO

Si se comparan entre sí cantidades indeterminadas de diversos géneros y una de ellas se dice que es como otra, directa o inversamente, esto significa que la primera crece o disminuye según la misma razón que la segunda o según la inversa. Y si una de ellas se dice que es como otras dos o más directa o inversamente, esto significa que la primera aumenta o disminuye según la razón compuesta de las razones, según las cuales las otras o las inversas de las otras aumentan o disminuyen. Así, si se dice que A es directamente como B y directamente como C e inversamente como D, esto significa que A aumenta o disminuye en la razón de B x C x 1/D, esto es, que A y BC/D son mutuamente según una razón dada.

LEMA XI

La subtensa evanescente del ángulo de contacto, en todas las curvas que tienen curvatura finita en el punto de contacto, es al fin como el cuadrado de la subtensa del arco contiguo.

CASO 1. Sea dicho arco AB, la tangente AD, sea BD la subtensa del ángulo de contacto perpendicular a la tangente, y AB la subtensa del arco. Trácense AG y BG perpendiculares a la tangente AD y a la subtensa AB y que se encuentren en G.



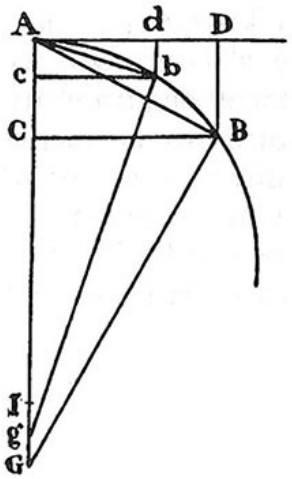
Muévanse después los puntos D, B, G hacia *d, b, g*, y sea I la última intersección de BG y AG cuando D y B se acercan hacia A. Es evidente que la distancia GI puede ser más pequeña que una distancia dada. Por otra parte (por la naturaleza de los círculos que pasan por los puntos ABG, A*bg*) AB^2 es igual a AG x BD y Ab^2 es igual a A*g* x *bd*; y por tanto la razón de AB^2 a Ab^2 se compone de las razones de AG a *Ag* y BD a *bd*. Pero como GI puede tomarse como menor que cualquier longitud dada, puede hacerse que la razón AG a A*g* diste menos de la igualdad que una cantidad dada y, por tanto, que la razón AB^2 a Ab^2 diste menos de la razón de BD a *bd* que una cantidad dada. Luego, por el Lema I, la razón última entre AB^2 y Ab^2 es la misma que la razón última de BD a *bd*. Q. E. D.

CASO 2. Inclínese BD sobre AD en un ángulo dado cualquiera y la razón última de BD a B*d* siempre será la misma que antes y, por tanto, la misma que AB^2 a Ab^2. Q. E. D.

CASO 3. Y aunque no esté dado el ángulo D, sino que la recta BD converja en un punto dado o bien se construya con otra regla cualquiera, sin embargo los ángulos D, *d* se construirán con una ley común y siempre tenderán a la igualdad y se acercarán entre sí más que una diferencia dada y, por tanto, al fin serán iguales, por el Lema I, y por tanto las líneas BD, *bd* están entre sí en la misma razón que antes. Q. E. D.

COROLARIO 1. De donde, como las tangentes AD, A*d*, de los arcos AB, A*b* y sus senos BC, *bc* sean al final iguales a las cuerdas AB, A*b*; sus cuadrados serán al fin como las subtensas BD, *bd*.

COROLARIO 2. Tales cuadrados serán asimismo al final como las sagitas de los arcos que bisecan a las cuerdas y convergen en un punto dado. Puesto que dichas sagitas son como las subtensas BD, *bd*.



COROLARIO 3. Por tanto, la sagita es como el cuadrado del tiempo en que un cuerpo describe el arco a una velocidad dada.

COROLARIO 4. Los triángulos rectilíneos ADB, A*db* están al fin en razón cúbica de los lados AD, A*d* y en razón de la potencia $\frac{3}{2}$ (sesquiplicada) de los lados DB, *db*; por cuanto que se da la existencia de la razón compuesta de los lados AD y DB, A*d* y D*b*. Igualmente los triángulos ABC, A*bc* están al final en razón cúbica de los lados BC, *bc*. Llamo razón sesquiplicada a la subduplicada de la triplicada, que se compone, pues, de la simple y la subduplicada[15].

COROLARIO 5. Y puesto que DB, *db* son al final paralelas y en razón cuadrada de las mismas AD, A*d*, las áreas curvilíneas últimas ADB, A*db* serán (por la naturaleza de la parábola) dos tercios de los triángulos rectilíneos ADB, A*db*, y los segmentos AB, A*b* terceras partes de dichos triángulos. Y en consecuencia, dichas áreas y dichos segmentos estarán en razón triplicada, ya de las tangentes AD, A*d*, ya de las cuerdas y arcos AB, A*b*.

ESCOLIO

Por lo demás, en todo esto suponemos que el ángulo de contacto no es ni infinitamente mayor que los ángulos de contacto contenidos por los círculos y sus tangentes, ni infinitamente menor que los mismos; esto es, que la curvatura en el punto A no es ni infinitamente pequeña, ni infinitamente grande, o también, que el intervalo AI es de magnitud finita. Puede, en efecto, tomarse a DB como AD^3: en tal caso ningún círculo puede trazarse por el punto A entre la tangente AD y la curva AB y por consiguiente el ángulo de contacto será infinitamente menor que los de los círculos. Y por un argumento semejante si DB se hace sucesivamente como AD^4,

AD^5, AD^6, AD^7, etc., se tendrá una serie de ángulos de contacto que va hasta el infinito y de los cuales cualquiera de los siguientes es infinitamente menor que su antecesor. Y si DB se hiciera sucesivamente como AD^2, $AD^{3/2}$, $AD^{4/3}$, $AD^{5/4}$, $AD^{6/5}$, $AD^{7/6}$, etc., se tendrá otra serie infinita de ángulos de contacto de los cuales el primero es del mismo género que los de los círculos, el segundo infinitamente mayor, y cualquier siguiente infinitamente mayor que el anterior. Además, entre dos ángulos cualesquiera de éstos puede intercalarse una serie, también tendente al infinito, de ángulos intermedios de los cuales cualquier posterior será infinitamente mayor o menor que el anterior. Como si entre los términos AD^2 y AD^3 se intercala la serie $AD^{13/6}$, $AD^{11/5}$, $AD^{9/4}$, $AD^{7/3}$, $AD^{5/2}$, $AD^{8/3}$, $AD^{11/4}$, $AD^{14/5}$, $AD^{17/6}$, etc. Y de nuevo entre dos cualesquiera ángulos de esta serie se puede intercalar una nueva serie de ángulos intermedios, distantes entre sí infinitos intervalos. No conoce la naturaleza un límite.

Lo que se ha demostrado de las líneas curvas y de las superficies comprendidas, puede aplicarse fácilmente a las superficies curvas de los sólidos y a los contenidos. He adelantado estos Lemas para evitar tediosas y largas deducciones *ad absurdum* al estilo de los antiguos geómetras. Pues las demostraciones se hacen más breves por el método de los indivisibles. Pero como la hipótesis de los indivisibles es más difícil y además, tal método se considera menos geométrico, he preferido reducir las demostraciones de las cosas que siguen a las sumas y razones últimas de cantidades evanescentes, y a las sumas y razones primeras de cantidades nacientes, esto es, a los límites de las sumas y de las razones y, por tanto, he preferido anteponer, con la brevedad que he podido, las demostraciones de dichos límites. Por este medio se consigue lo mismo que con el método de los indivisibles y así podremos utilizar con mayor seguridad principios ya demostrados. Por tanto, en lo que sigue, cuantas veces considere cantidades como si constaran de partículas, o cuantas veces tome pequeñas curvas por líneas rectas, no quiero entender nunca que se trata de indivisibles, sino de divisibles evanescentes, ni tampoco de sumas o razones de partes determinadas, sino de los límites de las sumas y de las razones; y la fuerza de tales demostraciones debe atribuirse siempre al método de los Lemas precedentes.

Pudiera objetarse que no hay proporción última alguna entre cantidades evanescentes, ya que antes de que desaparezcan no son últimas, y, después de desaparecidas, no puede darse ninguna. Pero por la misma razón podría decirse que un cuerpo que llega a un punto en el que se acaba su movimiento no tendría una velocidad última, puesto que dicha velocidad no sería última antes de que dicho cuerpo alcance el punto final de su movimiento, y cuando le haya alcanzado ya no tendrá velocidad alguna. La respuesta es fácil: por velocidad última se entiende aquella con la que el cuerpo se mueve, no antes de alcanzar el punto final y cesa, por consiguiente, el movimiento, ni tampoco después de haberlo alcanzado, sino aquella con la que se mueve cuando lo alcanza, esto es, aquella velocidad con la que el cuerpo alcanza el punto final y aquella con la que cesa el movimiento. De igual

manera ha de entenderse por razón última de cantidades evanescentes la razón de cantidades, no antes de que desaparezcan, ni después de desaparecidas, sino aquella con la que desaparecen. De igual modo ocurre con la razón primera de cantidades nacientes, que es aquella con la que nacen. Y la suma primera y última es aquella con la que empiezan o acaban la existencia (o también a aumentar o disminuir). Existe un límite que puede alcanzar la velocidad al final del movimiento, pero no puede traspasarlo. Esta es la velocidad última. Y semejante es la razón del límite de todas las cantidades y proporciones nacientes y evanescentes. Y dado que tal límite es cierto y definido, el problema de determinarlo es puramente geométrico. Por lo demás es legítimo utilizar medios geométricos para determinar y demostrar cosas también geométricas.

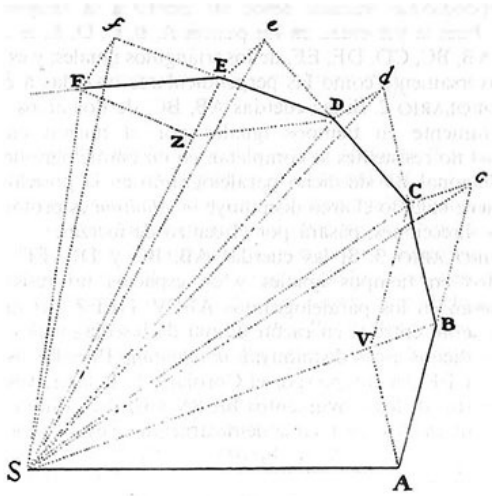
También pudiera objetarse que si se dan razones últimas de cantidades evanescentes, también se darán magnitudes últimas; y así toda cantidad constará de indivisibles, contra lo que demostró Euclides sobre los inconmensurables en el libro décimo de los elementos. Pero esta objeción se apoya en una hipótesis falsa. Las razones últimas con las que tales cantidades desaparecen en realidad no son razones de cantidades últimas, sino límites a los que tienden a acercarse siempre las razones de cantidades continuamente decrecientes, límites a los que pueden acercarse más que una diferencia dada, pero nunca traspasarlo, ni tampoco alcanzarlo antes de que las cantidades disminuyan *in infinitum*. El tema se ve mejor en las cosas infinitamente grandes. Si dos cantidades con una diferencia dada aumentan infinitamente, habrá una razón última de dichas cantidades, a saber, la razón de igualdad, sin que por ello se den también las cantidades últimas o máximas de las cuales era esta la razón. Por tanto, cuando en lo que sigue diga, con el fin de facilitar la comprensión de las cosas, cantidades mínimas o evanescentes o últimas, no se entienda que son cantidades de determinada magnitud sino que debe pensarse siempre en cantidades infinitamente decrecientes.

Sección II[16]
SOBRE EL DESCUBRIMIENTO DE LAS FUERZAS CENTRÍPETAS

Proposición I. Teorema I

Las áreas, descritas por cuerpos que giran sujetos a un centro de fuerzas inmóvil por radios unidos a dicho centro, están en el mismo plano inmóvil y son proporcionales a los tiempos.

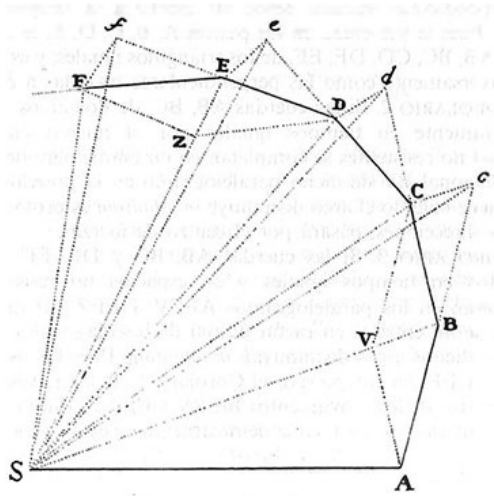
Divídase el tiempo en partes iguales y en la primera parte de tiempo el cuerpo por su fuerza ínsita recorra la recta AB. En la segunda parte de tiempo, si nada lo impide, la recta llegará hasta *c* (por la Ley I) describiendo la línea B*c* igual a la misma AB; de modo que las áreas ASB, BS*c* descritas por el trazado de los radios AS, BS, *c*S, al centro habrán de ser iguales. Pero cuando el cuerpo llega a B, supóngase que una fuerza centrípeta produce un único y gran impulso y haga que el cuerpo se separe de la línea B*c* y siga la recta BC. Trácese la paralela *c*C a BS que toca a BC en C; y al completarse la segunda parte de tiempo el cuerpo (por el Corolario I de las Leyes) se hallará en C, en el mismo plano que el triángulo ASB. Únase SC, y el triángulo SBC, por ser paralelas SB, C*c*, será igual al triángulo SB*c* y, por tanto, también al triángulo SAB. Y, por argumentos semejantes, si una fuerza centrípeta actúa sucesivamente en C, D, E, etc., haciendo que el cuerpo en cada parte de tiempo describa cada recta CD, DE, EF, etc., estarán todas ellas en el mismo plano y el triángulo SCD será igual al triángulo SBC, y SDE al propio SCD, y SEF a SDE. Por tanto, en tiempos iguales se describen áreas iguales en un plano inmóvil; y, por composición, las sumas de cualesquiera áreas SADS y SAFS son entre sí como los tiempos en que se describen. Auméntese ahora el número de triángulos y disminúyase su altura *in infinitum* y su perímetro último ADF será una línea curva[17] (por el Corolario 4 del Lema III); y por tanto la fuerza centrípeta, por la que un cuerpo es continuamente separado de la tangente de dicha curva, actúa continuamente; y áreas cualesquiera descritas SADS, SAFS proporcionales siempre a los tiempos en que se describen, serán, en este caso, proporcionales a los mismos tiempos. Q. E. D.



COROLARIO 1. La velocidad de un cuerpo atraído hacia un centro inmóvil en un espacio no resistente es inversamente como la perpendicular trazada desde tal centro a la tangente de la curva. Pues la velocidad en los puntos A, B, C, D, E, es como las bases AB, BC, CD, DE, EF, de los triángulos iguales; y estas bases son inversamente como las perpendiculares trazadas a ellas.

COROLARIO 2. Si las cuerdas AB, BC, de dos arcos descritos sucesivamente en tiempos iguales por el mismo cuerpo en espacios no resistentes se completan en un paralelogramo ABCV, y la diagonal BV de dicho paralelogramo en la posición que al final tiene cuando el arco disminuye *in infinitum* es prolongada en ambas direcciones, pasará por el centro de fuerzas.

COROLARIO 3. Si las cuerdas AB, BC y DE, EF de arcos descritos en tiempos iguales y en espacios no resistentes se completan en los paralelogramos ABCV, DEFZ, las fuerzas en B y E serán entre sí en razón última de las diagonales BV, EZ, cuando dichos arcos disminuyan *in infinitum*. Pues los movimientos BC y EF del cuerpo (por el Corolario I de las Leyes) estarán compuestos de los movimientos Be, BV y E*f*, EZ. Ahora bien, BV y EZ, iguales a C*c* y F*f*, en la demostración de esta Proposición se originaban por impulsos de la fuerza centrípeta en B y E, y por consiguiente, son proporcionales a tales impulsos.



COROLARIO 4. Las fuerzas, por las que cualesquiera cuerpos en espacios no resistentes son desviados de trayectorias rectilíneas y constreñidos a trayectorias curvas, son entre sí como las sagitas convergentes al centro de fuerzas; sagitas de los arcos descritos en tiempos iguales y bisecantes de las cuerdas de dichos arcos mientras éstos disminuyen *in infinitum*. Pues tales sagitas son las mitades de las diagonales de las que hemos hablado en el Corolario 3.

COROLARIO 5. Y, por tanto, estas fuerzas son a la fuerza de la gravedad como tales sagitas son a las sagitas perpendiculares al horizonte de los arcos parabólicos que describen los proyectiles en los mismos tiempos.

COROLARIO 6. Lo mismo vale (por el Corolario V de las Leyes) cuando los planos en que se mueven los cuerpos junto con el centro de fuerzas situado en los mismos no están en reposo, sino que se mueven uniformemente y en línea recta.

PROPOSICIÓN II. TEOREMA II

Todo cuerpo que se mueve en una curva descrita en un plano, y con un radio trazado a un punto, inmóvil o en movimiento rectilíneo y uniforme, describe áreas en torno a dicho punto proporcionales a los tiempos, es empujado por una fuerza centrípeta tendente a dicho punto.

CASO 1. Pues todo cuerpo que se mueve en una curva es desviado de su trayectoria recta por alguna fuerza actuante sobre él (por la Ley I). Y esta fuerza que aparta al cuerpo de la trayectoria rectilínea y le obliga a describir en torno al punto inmóvil S los triángulos iguales e infinitamente pequeños SAB, SBC, SDC, etc.,

actúa en el punto B según una línea paralela a *c*C (por la Proposición XL del Libro I de los *Elementos* y la Ley II), esto es, según la dirección de la línea BS; y en el punto C actúa según la línea paralela a *d*D, esto es, según la línea SC, etc. Por tanto, actúa siempre según líneas tendentes al mencionado punto S inmóvil. Q. E. D.

CASO 2. Y, por el Corolario V de las Leyes, es indiferente que la superficie en que un cuerpo describe una línea curva esté en reposo o que se mueva, junto con el cuerpo, la figura descrita y su punto S, uniformemente y en directo.

COROLARIO 1. En espacios o medios no resistentes, si las áreas no son proporcionales a los tiempos, las fuerzas no tienden al punto donde concurren los radios sino que se adelantan hacia la región hacia la que se dirige el movimiento, caso de que la descripción de las áreas se acelere, o se atrasan, caso de que se retarde.

COROLARIO 2. Incluso en medios resistentes, si se acelera la descripción de las áreas, las direcciones de las fuerzas se apartan del punto de encuentro de los radios hacia la dirección hacia la que va el movimiento.

ESCOLIO

Un cuerpo puede ser urgido por una fuerza centrípeta compuesta de muchas fuerzas. En tal caso el sentido de la Proposición es que dicha fuerza, compuesta de todas las demás, tiende al punto S. Además si alguna fuerza actúa constantemente según una línea perpendicular a la superficie descrita, obligará al cuerpo a separarse del plano de su movimiento; pero no aumentará ni disminuirá la cantidad de superficie descrita y, por tanto, ha de despreciarse al componer las fuerzas.

PROPOSICIÓN III. TEOREMA III

Todo cuerpo que, unido por un radio al centro de otro cuerpo que se mueve de cualquier modo, describe alrededor de dicho centro áreas proporcionales a los tiempos es urgido por una fuerza compuesta de la fuerza centrípeta tendente a ese otro cuerpo y de toda la fuerza aceleratriz con la que es urgido ese otro cuerpo.

Sea L el primer cuerpo, y sea T el otro; y (por el Corolario VI de las Leyes) si ambos cuerpos son urgidos según líneas paralelas por una fuerza nueva, igual y contraria a aquella que urge al otro cuerpo T, el primer cuerpo L describiría alrededor del otro cuerpo T las mismas áreas que antes: en cambio la fuerza por la que era urgido el otro cuerpo T será ahora destruida por una fuerza igual y contraria y, por tanto (por la Ley I), ese otro cuerpo T abandonado ya a sí mismo o bien quedará en reposo o bien en movimiento uniforme y directo: y el cuerpo primero L urgido por la fuerza restante pasará a describir en torno al otro cuerpo T áreas proporcionales a los

tiempos. Por tanto (por el Teorema II) la diferencia de fuerzas tiende hacia el otro cuerpo T como a su centro. Q. E. D.

COROLARIO 1. Por tanto, si un cuerpo L, unido por un radio a otro T, describe áreas proporcionales a los tiempos y de la fuerza total (tanto si es simple como si está compuesta de muchas fuerzas según el Corolario II de las Leyes) por la que es urgido el primer cuerpo L se resta (por el mismo Corolario de las Leyes) toda la fuerza aceleratriz por la que es urgido el otro cuerpo, toda la fuerza restante por la que es urgido el primer cuerpo tenderá al otro cuerpo T como a su centro.

COROLARIO 2. Y si esas áreas son aproximadamente proporcionales a los tiempos, la fuerza restante tenderá aproximadamente al otro cuerpo T.

COROLARIO 3. Y viceversa, si la fuerza restante tiende aproximadamente al otro cuerpo T, esas áreas serán aproximadamente proporcionales a los tiempos.

COROLARIO 4. Si un cuerpo L, unido por un radio a otro cuerpo T, describe áreas que, comparadas con los tiempos, son muy desiguales y el otro cuerpo T o reposa o se mueve uniformemente en directo, la acción de la fuerza centrípeta tendente hacia ese otro cuerpo T o es nula o se mezcla y se compone con acciones muy potentes de otras fuerzas, y la fuerza total compuesta de todas ellas, si son más de una, tiende hacia otro centro (móvil o en reposo). Lo mismo ocurre cuando el otro cuerpo se mueve con cualquier clase de movimiento si se toma la fuerza centrípeta que resulta después de restar toda la fuerza que actúa sobre el otro cuerpo T.

ESCOLIO

Puesto que la igual descripción de áreas es índice del centro hacia el que tiende la fuerza por la que más afectado es un cuerpo y por la que es apartado del movimiento rectilíneo y mantenido en su órbita, ¿por qué no podemos en lo sucesivo tomar la igualdad en la descripción de áreas como un índice del centro en torno al que ocurre todo movimiento circular en los espacios libres?

PROPOSICIÓN IV. TEOREMA IV

Las fuerzas centrípetas de los cuerpos que describen círculos distintos con movimientos iguales tienden a los centros de los mismos círculos; y son entre sí como los cuadrados de los arcos descritos en tiempos iguales divididos respectivamente por los radios de los círculos.

Estas fuerzas tienden a los centros de los círculos, por la Proposición II y por el Corolario 2, Proposición I, y son entre sí como los senos versos de los arcos descritos en tiempos iguales muy pequeños, por el Corolario 4 de la Proposición I, esto es,

como los cuadrados de los mismos arcos divididos por los diámetros de los círculos, por el Lema VII; y por tanto, como estos arcos son como los arcos descritos en unos tiempos iguales y los diámetros son como los radios, las fuerzas serán como los cuadrados de los arcos descritos en el mismo tiempo divididos por los radios de los círculos. Q. E. D.

COROLARIO 1. Puesto que estos arcos son como las velocidades de los cuerpos, las fuerzas centrípetas estarán en razón compuesta, directamente de la razón cuadrada de las velocidades e inversamente de la razón simple de los radios.

COROLARIO 2. Y, como los tiempos periódicos están en razón compuesta directamente de la razón de los radios e inversamente de la razón de las velocidades, las fuerzas centrípetas están en razón compuesta directamente de la razón de los radios e inversamente de la razón cuadrada de los tiempos periódicos.

COROLARIO 3. De donde, si se igualan los tiempos periódicos y, por tanto, las velocidades son como los radios, también las fuerzas centrípetas serán como los radios; y al contrario.

COROLARIO 4. Si tanto los tiempos periódicos como las velocidades están en razón de la raíz cuadrada de los radios; las fuerzas centrípetas serán iguales entre sí; y al contrario.

COROLARIO 5. Si los tiempos periódicos son como los radios y, por tanto, las velocidades iguales, las fuerzas centrípetas serán inversamente como los radios; y al contrario.

COROLARIO 6. Si los tiempos periódicos están en razón de la potencia $^{3}/_{2}$ de los radios y, por tanto, las velocidades están en razón de la raíz cuadrada de los radios, las fuerzas centrípetas serán inversamente como los cuadrados de los radios; y al contrario.

COROLARIO 7. Y, en general, si el tiempo periódico es como una potencia R^{n} cualquiera del radio R y, por tanto, la velocidad es inversamente como la potencia R^{n-1} del radio, la fuerza centrípeta será inversamente como la potencia R^{2n-1} del radio; y al contrario.

COROLARIO 8. Con respecto a los tiempos, las velocidades y las fuerzas con los que los cuerpos describen partes semejantes de cualesquiera figuras semejantes y que tienen los centros colocados en posición semejante dentro de dichas figuras se siguen las mismas cosas de la demostración de los casos precedentes aplicada a estas otras. Se aplica sustituyendo la descripción igual de las áreas por los movimientos iguales y tomando las distancias de los cuerpos a los centros como radios.

COROLARIO 9. Se sigue también de esta demostración que el arco que describe un cuerpo que gira uniformemente en círculo con una fuerza centrípeta dada en un tiempo cualquiera es media proporcional entre el diámetro del círculo y el descenso realizado por el cuerpo con la misma fuerza dada y cayendo durante el mismo tiempo[18].

ESCOLIO

El caso del Corolario 6 se da en los cuerpos celestes (como ya observaron por su parte nuestros *Wren, Hooke* y *Halley*); por ello, en lo que sigue he decidido exponer más ampliamente lo relativo a la fuerza centrípeta que decrece en razón del cuadrado de las distancias a los centros.

Además, como resultado de la Proposición precedente y de sus Corolarios, se puede deducir también la proporción entre una fuerza centrípeta y otra fuerza cualquiera conocida, tal como la de la gravedad, pues si un cuerpo gira en círculo concéntrico a la Tierra por la fuerza de su gravedad, esta gravedad es la fuerza centrípeta del mismo. Pues, el tiempo de una revolución, así como el arco descrito en un tiempo dado, vienen dados a partir del descenso de los graves, por el Corolario 9 de dicha Proposición. Y por medio de tales proposiciones. *Huygens* comparó la fuerza de la gravedad con las fuerzas centrífugas de los cuerpos que giran, en su célebre tratado *de Horologio Oscillatorio*.

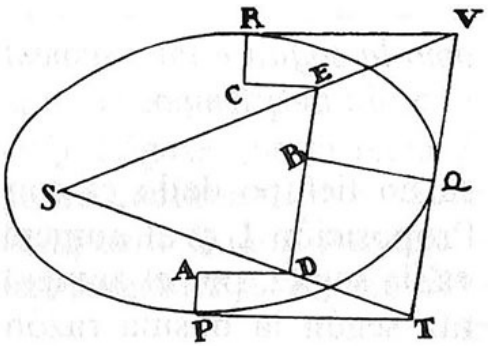
Lo que antecede puede también demostrarse de este modo: imagínese descrito dentro de un círculo a un polígono de los lados que se quiera. Si un cuerpo moviéndose con una velocidad dada por los lados del polígono en cada ángulo del mismo es rechazado por el círculo, la fuerza con la cual en cada ocasión choca contra el círculo será como su velocidad; y por tanto la suma de las fuerzas en un tiempo dado será como dicha velocidad y el número de rechazos conjuntamente; esto es (si se da la clase del polígono) como la longitud descrita en el tiempo dado aumentada o disminuida en razón de la misma, longitud al radio del círculo predicho; es decir, como el cuadrado de dicha longitud dividido por el radio; y por tanto, si el polígono, acortando sus lados *in infinitum*, coincidiese con el círculo, como el cuadrado del arco descrito en un tiempo dado dividido por el radio. Esta es la fuerza centrífuga con la que el cuerpo empuja contra el círculo; e igual a ella es la fuerza contraria con la cual el círculo repele continuamente al cuerpo hacia el centro.

PROPOSICIÓN V. PROBLEMA I[19].

Hallar el centro de una figura dada, descrita por un cuerpo con fuerzas tendentes a dicho centro común y con una velocidad dada en un lugar cualquiera.

Sean las tres rectas PT, TQV, VR tangentes a la figura descrita en los puntos P, Q, R y que se corten en T y V. Trácense PA, QB, RC perpendiculares a las tangentes e inversamente proporcionales a las velocidades del cuerpo en los dichos puntos P, Q, R desde los que parten; esto es, de tal modo que PA sea a QB como la velocidad en Q a la velocidad en P y QB sea a RC como la velocidad en R a la velocidad Q. Trácense en ángulo recto por los extremos A, B, C de las perpendiculares las líneas AD, DBE,

EC que se cortan en D y E. Y trazadas TD, VE se cortarán en el centro buscado S.

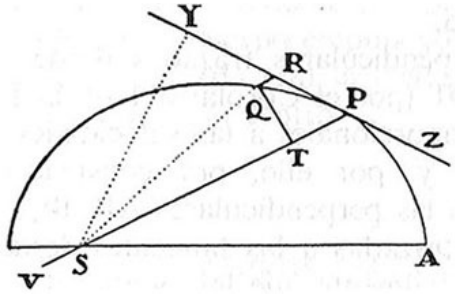


Pues las perpendiculares trazadas desde el centro S a las tangentes PT, QT (por el Corolario 1 de la Proposición I) son inversamente proporcionales a las velocidades del cuerpo en los puntos P y Q y, por ello, por construcción, directamente proporcionales a las perpendiculares AP, BQ, esto es, como las perpendiculares trazadas a las tangentes desde el punto D. De donde se sigue fácilmente que los puntos S, D, T están en una misma recta y por un argumento semejante los puntos S, E, V están también en la misma recta; y por tanto el centro S se halla en el cruce de las rectas TD, VE. Q. E. D.

Proposición VI. Teorema V[20]

Si en un espacio sin resistencia un cuerpo gira en una órbita alrededor de un centro inmóvil y describe en un tiempo muy pequeño un arco naciente en ese instante e imaginamos trazada la sagita del arco que a su vez divida en dos la cuerda y, prolongada, pase por el centro de fuerzas: la fuerza centrípeta en el punto medio del arco será directamente como la sagita e inversamente como el cuadrado del tiempo.

Pues la sagita en un tiempo dado es como la fuerza (por el Corolario 4 de la Proposición I) y al aumentar el tiempo según una razón cualquiera, la sagita, por el aumento del arco según la dicha razón, aumenta según la misma razón cuadrada (por los Corolarios 2 y 3 del Lema XI) y, por tanto, es como la fuerza y el cuadrado del tiempo. Divídase por el cuadrado del tiempo a ambos y la fuerza será directamente como la sagita e inversamente como el cuadrado del tiempo. Q. E. D.



También se puede demostrar esto fácilmente por el Corolario 4 del Lema X.

Corolario 1. Si un cuerpo P girando en torno a un centro S describe la curva APQ; hágase que la recta ZPR sea tangente a dicha curva en un punto cualquiera P y trácese hasta la tangente desde otro punto cualquiera Q una paralela a SP con distancia QR y trácese QT perpendicular a la línea SP; entonces la fuerza centrípeta será inversamente como el sólido $\frac{SP^2 \times QP^2}{QR}$; siempre que se considere la magnitud del sólido alcanzada en el último momento, cuando coinciden los puntos P y Q. Pues QR es igual a la sagita del doble del arco QP en cuya mitad se halla. P, y el doble del triángulo SQP o SP x QT es proporcional al tiempo en que es descrito dicho doble arco; por tanto, puede tomarse como representación del tiempo.

Corolario 2. Por un argumento semejante la fuerza centrípeta es inversamente como el sólido $\frac{SP^2 \times QP^2}{QR}$, siempre que SY sea perpendicular desde el centro de fuerza a PR tangente a la órbita. Pues los rectángulos SY x QP y SP x QT son iguales.

Corolario 3. Si una órbita es, o un círculo, o toca concéntricamente un círculo, o corta concéntricamente a un círculo, esto es, contiene un ángulo muy pequeño de contacto o de sección con el círculo y tiene la misma curvatura y el mismo radio de curvatura en el punto P; y si la cuerda PV del círculo es trazada desde el cuerpo por el centro de fuerzas; la fuerza centrípeta será inversamente como el sólido SY^2 x PV. Pues PV es $\frac{QP^2}{QR}$.

Corolario 4. Supuesto esto, la fuerza centrípeta es como el cuadrado de la velocidad directamente, y dicha cuerda inversamente. Pues la velocidad es inversamente como la perpendicular SY, por el Corolario 1 de la Proposición I.

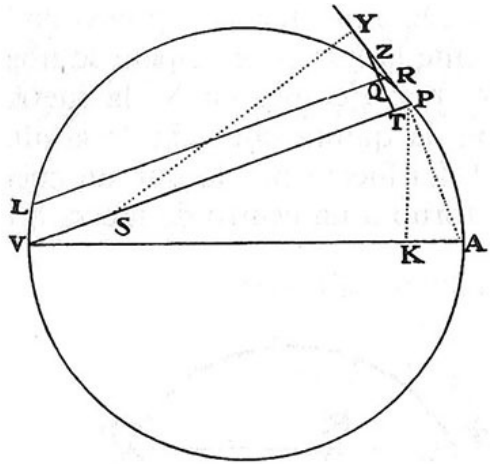
Corolario 5. De aquí que, si se da una figura curvilínea APQ y se da también en ella

el punto S al que se dirige continuamente la fuerza centrípeta se puede hallar la ley de la fuerza centrípeta por la que un cuerpo P es retenido en el perímetro de dicha figura, desviado continuamente de la trayecto ria recta, figura que resultará descrita al girar. Efectivamente se puede hallar calculando, o bien el sólido $\frac{SP^2 \text{ x } QT^2}{QR}$, o bien el sólido SY^2 x PV inversamente proporcional a dicha fuerza. Daremos ejemplos de esto en los problemas siguientes.

Proposición vii. Problema ii[21].

Si un cuerpo gira en la circunferencia de un círculo, hállese la ley de la fuerza centrípeta tendente a un punto dado cualquiera.

Sea VQPA la circunferencia del círculo; y sea S el punto dado hacia el que tiende la fuerza como hacia su centro; sea P el cuerpo que se mueve en la circunferencia y Q el punto inmediato hacia el cual se mueve y sea PRZ la tangente del círculo en el punto precedente. Trácese por el punto S la cuerda PV; y trazado el diámetro del círculo VA únanse AP y trácese sobre SP la perpendicular QT que prolongada cortará a la tangente PR en Z y, finalmente, trácese por el punto Q la recta LR paralela a SP que corte el círculo en L y a la tangente PZ en R. Por la semejanza de los triángulos ZQR, ZTP, VPA tendremos que RP^2, esto es, QRL será a QT^2 como AV^2 a PV^2. Y, por tanto, $\frac{QRL \text{ x } PV^2}{AV^2} = QT^2$. Multiplíquense estos dos miembros iguales por $\frac{SP^2}{QR}$ y al coincidir los puntos P y Q sustitúyanse PV por RL. Tendremos asi que $\frac{SP^2 \text{ x } PV^3}{AV^2} = \frac{SP^2 \text{ x } QT^2}{QR}$. Luego (por los Corolarios 1 y 5 de la Proposición vi) la fuerza centrípeta es inversamente como $\frac{SP^2 \text{ x } PV^3}{AV^2}$; esto es (por estar dado AV^2) inversamente como el producto del cuadrado de la distancia o de la altura SP por el cubo de la cuerda PV. Q. E. I.

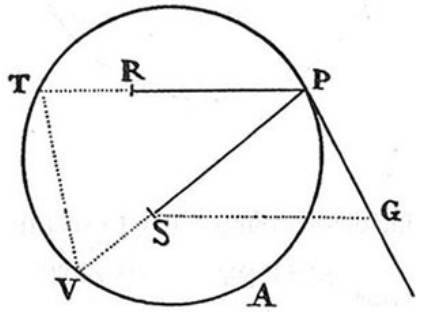


Lo mismo de otro modo

Trácese la perpendicular SY sobre la prolongación de la tangente PR; y, por semejanza de los triángulos SYP, VPA, tendremos que AV será a PV como SP a SY: por tanto $\frac{SP \text{ x } PV}{AV} = SY$, y $\frac{SP^2 \text{ x } PV^3}{AV^2} = SY^2$ x PV. Y por tanto (por los corolarios 3 y 5 de la Proposición vi) la fuerza centrípeta es inversamente como $\frac{SP^2 \text{ x } PV^3}{AV^2}$ esto es, por estar dada AV, inversamente como SP^2 x PV^3 Q. E. I.

Corolario 1. De aquí que si el punto dado S hacia el que tiende continuamente la fuerza centrípeta se ubica en la circunferencia del círculo, por ejemplo en V, la fuerza centrípeta será inversamente como la quinta potencia de la altura SP.

Corolario 2. La fuerza por la que un cuerpo P gira en un círculo APTV en torno a un centro de fuerza S es a la fuerza por la que el mismo cuerpo P puede girar en el mismo círculo y en el mismo tiempo periódico en torno a otro centro de fuerza como RP^2 x SP al cubo de la recta SG trazada desde el centro de fuerza primitivo S a la tangente PG de la circunferencia y paralela a la recta que une el cuerpo P con el segundo centro de fuerza. Pues por la construcción de esta proposición la fuerza primera es a la segunda como RP^2 x PT^3 es a SP^2 x PV^3, esto es como SP x RP^2 SP^3 x PV^3 es a $\frac{SP^3 \text{ x } PV^3}{PT^3}$ también (por semejanza de los triángulos PSG, TPV) a SG^3.

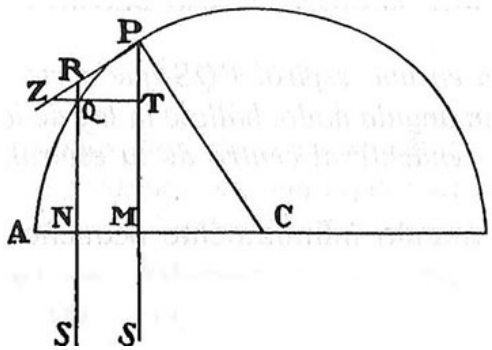


COROLARIO 3. La fuerza por la que gira un cuerpo P en una órbita en torno a un centro de fuerzas S es a la fuerza por la que el mismo cuerpo P giraría en la misma órbita en el mismo tiempo periódico en torno a otro centro de fuerzas R, como el sólido SP x RP^2, contenido a su vez por la distancia del cuerpo al primer centro de fuerzas S y por el cuadrado de su distancia al segundo centro de fuerzas R, es al cubo de la recta SG trazada desde el primer centro de fuerzas S a la tangente a la órbita PG y paralela a la recta RP que une el cuerpo P con el segundo centro de fuerzas R. Pues las fuerzas en esta órbita para un punto P son iguales que en un círculo de la misma curvatura.

PROPOSICIÓN VIII. PROBLEMA III.

Si un cuerpo se mueve en el semicírculo PQA: hállese la ley de la fuerza centrípeta tendente a un punto S tan lejano que todas las líneas PS, RS dirigidas a él puedan ser consideradas como paralelas.

Desde el centro C del semicírculo trácese el semidiámetro CA que corte perpendicularmente dichas paralelas en M y N y únase CP. Por semejanza de los triángulos CPM, PZT y RZQ tenemos que CP^2 es a PM^2 como PR^2 a QT^2 y por la naturaleza del círculo PQ^2 es igual al rectángulo QR x (RN x QN) o, cuando los puntos P y Q tienden a juntarse, igual al rectángulo QR x 2PM. Luego CP^2 es a PM^2 como QR x 2PM a QT^2 y por tanto $\frac{QT^2}{QR} = \frac{2PM^3}{CP^2}$, y $\frac{QT^2 \times SP^2}{QR} = \frac{2PM^3 \times SP^2}{CP^2}$. Así pues (por los Corolarios 1 y 5 de la Proposición VI) la fuerza centrípeta es inversamente como $\frac{2PM^2 \times SP^2}{SP^2}$, esto es $\left(\text{despreciando la razón dada } \frac{2SP^2 \times SP^2}{CP^2}\right)$ inversamente como PM^2. Q. E. I.



Esto mismo se deduce fácilmente también de la Proposición precedente.

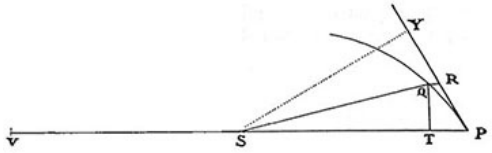
ESCOLIO

Y por un argumento no muy distinto se hallará que un cuerpo se movería en una elipse o también en una hipérbola o en una parábola, por una fuerza centrípeta que sería a su vez inversamente como el cubo de la ordenada dirigida a un centro de fuerzas infinitamente lejano.

PROPOSICIÓN IX. PROBLEMA IV.

Si un cuerpo gira en una espiral PQS que corta a todos los radios SP, SQ, etc., en un ángulo dado, hállese la ley de la fuerza centrípeta tendente al centro de la espiral.

Sea dado el ángulo infinitamente pequeño PSQ y por estar dados todos los ángulos se obtendrá una figura de la forma SPRQT. Luego está dada la razón $\frac{QT}{QR}$, y $\frac{QT^2}{QR}$ es como QT, esto es (por la forma de la figura dada) como SP. Altérese cuanto sea el ángulo PSQ y la recta que subtiende al ángulo de contacto QRT se alterará (por el Lema XI) en razón cuadrada del propio PR o QT. Luego $\frac{QT^2}{QR}$ permanecerá igual que antes, esto es como SP. Por lo tanto $\frac{QT^2 \times SP^2}{QR}$ es como SP^3 y por lo mismo (por los Corolarios 1 y 5 de la Proposición VI), la fuerza centrípeta es inversamente como el cubo de la distancia SP. Q. E. I.



Lo mismo de otro modo

La perpendicular SY trazada a la tangente y la cuerda PV del círculo concéntrico secante de la espiral están en razones dadas con respecto a la altura SP. Por tanto SP^3 es como SY^2 x PV, esto es (por los Corolarios 3 y 5 de la Proposición VI) inversamente como la fuerza centrípeta.

Lema XII

Todos los paralelogramos descritos en torno a cualesquiera diámetros conjugados de una elipse o de una hipérbola dadas son iguales entre sí.

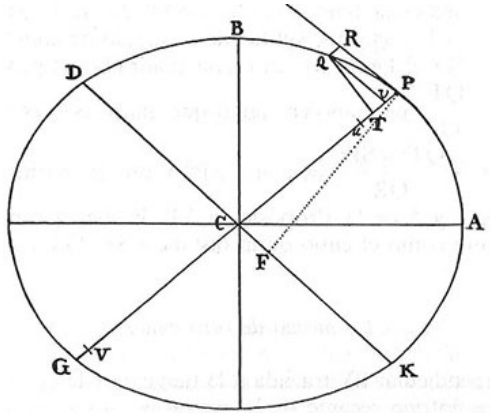
Consta por las cónicas.

Proposición X. Problema V

Si un cuerpo gira en una elipse, hallar la ley de la fuerza centrípeta tendente al centro de la elipse.

Sean CA, CB los semiejes de la elipse: GP, DK diámetros conjugados; PF, QT las perpendiculares a los diámetros, QV una ordenada del diámetro GP, y si se completa el paralelogramo Q*v*PR, el rectángulo P*v*G será (por las cónicas) a Qv^2 como PC^2 A CD^2 Y (por semejanza de los triángulos Q*v*T, PCF) Qv^2 es a QT^2 como PC^2 a PF^2 y multiplicando las razones, el rectángulo P*v*G a QT^2 como PC^2 a CD^2, y PC^2 a PF^2, esto es *v*G a $\frac{QT^2}{Pv}$ como PC^2 a $\frac{CD^2 \text{ x } PF^2}{PC^2}$. Sustitúyase QR por P*v* y (por el Lema XII) BC x CA por CD x PF, así como (al acercarse los puntos P y Q) 2PC por *v*G tendremos que, multiplicando extremos y medios entre sí, $\frac{QT^2 \text{ x } PC^2}{QR} = \frac{2BC^2 \text{ x } CA^2}{PC}$. Así pues (por el Corolario 5 de la Proposición VI) la fuerza centrípeta es

inversamente como $\frac{2BC^2 \text{ x } CA^2}{PC}$; esto es (por estar dados $2BC^2$ x CA^2) inversamente como $\frac{1}{PC}$; o lo que es lo mismo directamente como la distancia PC. Q. E. I.



Lo mismo de otro modo

En la recta PG y en la otra parte del punto T tómese el punto *u* de tal modo que T*u* sea igual al propio T*v*; después tómese *u*V de modo que sea a *v*G como DC^2 a PC^2. Y puesto que por las cónicas Qv^2 es a P*v*G como DC^2 a PC^2, Qv^2 será igual a P*v* x *u*V. Añádase a ambos el rectángulo *u*P*v* y el cuadrado de la cuerda del arco PQ resultará igual al rectángulo VP*v* y, por tanto, el círculo tangente a la sección cónica en el punto P, y que pase por el punto Q, pasará también por el punto V. Acérquense los puntos P y Q, y la razón *u*V a *v*G que es la misma que la de DC^2 a PC^2, vendrá a ser la misma que la de PV a PG o la de PV a 2PC; y por tanto PV será igual a $\frac{2DC^2}{PC}$. Por ello la fuerza por la que un cuerpo P gira en una elipse será inversamente como $\frac{2DC^2}{PC}$ x PF^2 (por el Corolario 3 de la Proposición VI), esto es (por estar dado $2DC^2$ x PF^2) directamente como PC. Q. E. I.

Corolario 1. La fuerza es, por tanto, como la distancia del cuerpo al centro de la elipse; y viceversa, si la fuerza es como la distancia, el cuerpo girará en una elipse que tiene su centro en el centro de fuerzas, o si no, tal vez en un círculo en el que también puede convertirse la elipse.

Corolario 2. Y serán iguales también los tiempos periódicos de las revoluciones efectuadas en torno al mismo centro en todas las elipses. Pues los tiempos mencionados en las elipses semejantes serán iguales (por los Corolarios 3 y 8 de la Proposición IV), en cambio en las elipses que tienen el eje mayor común están entre sí directamente como el área total de las elipses e inversamente como las partes de área descritas en el mismo tiempo; esto es directamente como sus ejes menores e inversamente como las velocidades de los cuerpos en los vértices principales; o lo que es lo mismo, directamente como los ejes menores e inversamente como las ordenadas al mismo punto del eje común; en consecuencia (por la igualdad de las razones directas e inversas) en razón de igualdad.

Escolio

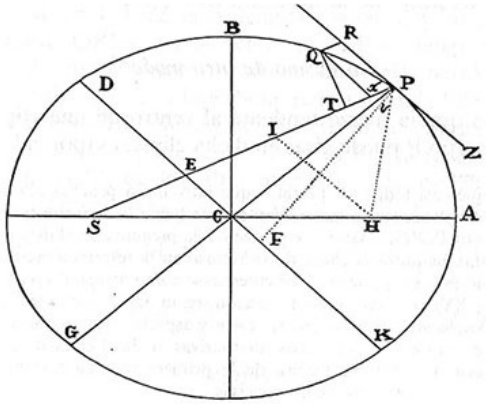
Si una elipse, por alejarse su centro infinitamente, se convierte en una parábola, el cuerpo se moverá en esta parábola; y la fuerza ahora tendente a un centro infinitamente distante resultará constante. Este es el teorema de *Galileo*. Y si la sección parabólica de un cono (al cambiar la inclinación del plano secante del cono) se convierte en una hipérbola, el cuerpo se moverá en el perímetro de ésta con una fuerza centrípeta transformada en centrífuga. Y del mismo modo que si en un círculo o una elipse las fuerzas tienden al centro de la figura situado en la abscisa, estas fuerzas, aumentando o disminuyendo las ordenadas según una razón dada o también alterando el ángulo de inclinación de las ordenadas a la abscisa, siempre aumentan o disminuyen en razón a las distancias al centro, mientras los tiempos periódicos permanezcan iguales, del mismo modo en todas las figuras si las ordenadas aumentan o disminuyen en una razón dada, o el ángulo de inclinación se altera en cualquier sentido permaneciendo igual el tiempo periódico, las fuerzas tendentes a cualquier centro situado en la abscisa aumentan o disminuyen en cada ordenada en razón a las distancias al centro.

Sección III
DEL MOVIMIENTO DE LOS CUERPOS EN SECCIONES CÓNICAS EXCÉNTRICAS

Proposición XI. Problema VI

Si un cuerpo gira en una elipse, hallar la ley de la fuerza centrípeta tendente al foco de la elipse.

Sea S el foco de la elipse. Trácese SP secante tanto del diámetro DK de la elipse en E como de la ordenada Q*v* en *x* y complétese el paralelogramo Q*x*PR. Es evidente que EP es igual al semieje mayor AC dado que trazada desde el otro foco H la línea HI paralela a la propia EC, por ser iguales CS, CH, son iguales ES, El, igual que EP es la semisuma de PS, PI, esto es (por ser paralelas HI, PR e iguales los ángulos IPR, HPZ) de PS, PH que conjuntamente son iguales a todo el eje 2AC. Trácese la perpendicular QT sobre SP y llamando L al «latus rectum» principal de la elipse (o sea $\frac{2BC^2}{AC}$), L·QR será a L·P*v* como QR a P*v*, esto es como PE o AC a PC; y L·P*v* será a G*v*P como L a G*v*; y G*v*P será a Qv^2 como PC^2 a CD^2 y (por el Corolario 2 del Lema VII), al coincidir Q y P, ocurrirá que Qv^2 será igual a Qx^2; y Qx^2 o Qv^2 es a QT^2 como EP^2 a PF^2, esto es, como CA^2 a PF^2 o también (por el Lema XII) como CD^2 a CB^2, y multiplicando todas estas razones L·QR será a QT^2 como AC·L PC^2·CD^2 o también como 2CB^2·PC^2·CD^2 a PC·G*v*·CD^2·CB^2, o bien, como 2PC a G*v*. Pero al coincidir los puntos Q y P ocurre que 2PC es igual a G*v*. Luego también se igualan L·QR y QT^2 por ser proporcionales a aquéllos. Multiplicando ahora estos iguales por $\frac{SP^2}{QR}$ tendremos que L·SP^2 es igual a $\frac{SP^2 \cdot QT^2}{QR}$. Luego (por los Corolarios 1 y 5 de la Proposición VI) la fuerza centrípeta es inversamente como L·SP^2, esto es, inversamente como el cuadrado de la distancia SP. Q. E. I.



Lo mismo de otro modo[22].

Puesto que la fuerza tendente al centro de una elipse, por la que un cuerpo P puede girar en dicha elipse, es (por el Corolario 1 de la Proposición X) como la distancia CP del cuerpo al centro de la elipse; trácese CE paralela a la tangente de la elipse PR; y la fuerza por la que el cuerpo P puede girar en torno a otro punto cualquiera S, si CE y PS se cruzan en E, será como $\frac{PE^3}{SP^2}$ (Por el Corolario 3 de la Proposición VII), esto es, si el punto S es foco de la elipse y, por tanto, PE está dado, como el inverso de SP^2. Q. E. I.

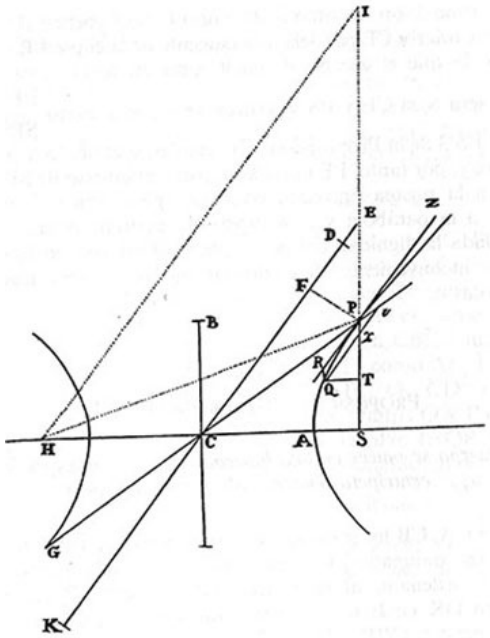
Con la misma brevedad con que aplicábamos el problema quinto a la parábola y a la hipérbola podríamos hacerlo aquí. Pero dada la dignidad del problema y su utilidad en lo sucesivo no hay inconveniente en confirmar mediante demostración los demás casos.

Proposición XII. Problema VII

Si un cuerpo se mueve en una hipérbola, hállese la ley de la fuerza centrípeta tendente al foco de la figura.

Sean CA, CB los semiejes de la hipérbola; sean PG, KD otros diámetros conjugados; PF una perpendicular al diámetro KD; y Q*v* una ordenada al diámetro GP. Trácese SP secante del diámetro DK en E y de la ordenada Q*v* en *x*, y complétese el paralelogramo QRP*x*. Es evidente que EP es igual al semieje transverso Ac, puesto que trazada la línea HI, paralela a EC, desde el otro foco H de la hipérbola, por la

igualdad de CS y CH serán iguales ES y EI; además porque EP es la semidiferencia de PS, PI esto es (por el paralelismo de IH, PR y la igualdad de los ángulos IPR, HPZ) de PS, PH, cuya diferencia es igual a todo el eje 2AC. Trácese QT perpendicular a SP. Llamando L al «latus rectum» principal de la hipérbola $\left(\text{o } \frac{2BC^2}{AC}\right)$ tendremos $\frac{L \times QR}{L \times Pv} = \frac{QR}{Pv} = \frac{Px}{Pv}$ esto es (por la semejanza de los triángulos P*xv*, PEC) $\frac{PE}{PC} = \frac{AC}{PC}$. Y también $\frac{L \times Pv}{Gv \times Pv} = \frac{L}{Gv}$; y (por la naturaleza de las cónicas), el rectángulo $\frac{Gv \times vP}{Qv^2} = \frac{PC^2}{CD^2}$; y (por el corolario 2 del lema VII) cuando P y Q coinciden $Qx^2 = Qv^2$; y $\frac{Qx^2}{QT^2}$, o $\frac{Qv^2}{QT^2} = \frac{EP^2}{PF^2} = \frac{CA^2}{PF^2}$ o (por el lema XII)= $\frac{CD^2}{CB^2}$ y multiplicando (ordenadamente) todas estas razones resultara que $\frac{L \times QR}{QT^2} = \frac{AC \times L \times PC^2 \times CD^2}{PC \times Gv \times CD^2 \times CB^2} = \frac{2CB^2 \times PC^2 \times CD^2}{PC \times Gv \times CD^2 \times CB^2} = \frac{2PC}{Gv}$. Pero al coincidir P y Q resulta que 2PC=G*v*. Luego también se igualan L x QR y QT^2 proporcionales a los anteriores. Multiplicando estas igualdades por $\frac{SP^2}{QR}$ tendremos que L x SP^2= $\frac{SP^2 \times QT^2}{QR}$. Luego (por los Corolarios 1 y 5 de la Proposición VI) la fuerza centrípeta es inversamente como L x SP^2, es decir, inversamente como el cuadrado de la distancia SP. Q. E. I.



Lo mismo de otra manera

Hállese la fuerza que parte desde el centro C de la hipérbola. Será proporcional a la distancia CP. Por ello la fuerza tendente al foco S será (por el Corolario 3 de la Proposición VIII) como $\frac{PE^3}{SP^2}$ esto es, por estar dada PE, inversamente como SP^2. Q. E. I.

Del mismo modo se demuestra que un cuerpo cuya fuerza centrípeta se cambia en centrífuga se moverá en una hipérbola opuesta.

Lema XIII

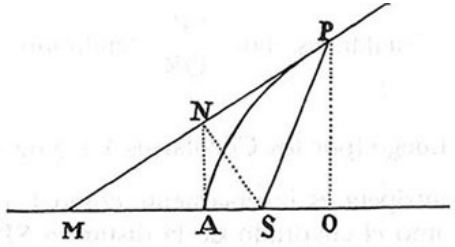
El «latus rectum» correspondiente a un vértice cualquiera de una parábola es cuatro veces la distancia de dicho vértice al foco de la figura.

Es evidente por las cónicas.

Lema XIV

La perpendicular que se traza desde el foco de una parábola a su tangente es media proporcional entre las distancias desde el foco al punto de contacto y al vértice principal de la figura.

Sea AP la parábola, S su foco, A el vértice principal, P el punto de contacto, PO la ordenada al diámetro principal, PM la tangente al diámetro principal concurrente en M, y SN la perpendicular desde el foco a la tangente. Únanse A, N y por la igualdad de MS y SP, MN y NP, MA y AO; las rectas AN y OP serán paralelas; y por tanto el triángulo SAN será rectángulo en A y semejante a los triángulos iguales SNM, SNP; luego PS es a SN como SN a SA. Q. E. D.



Corolario 1. PS^2 es a SN^2 como PS a SA.

Corolario 2. Y por estar dada SA, SN^2 es como PS.

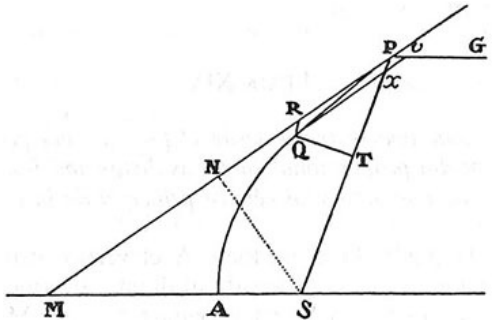
Corolario 3. Y el cruce de una tangente cualquiera PM con la recta SN que le es perpendicular desde el foco, incide sobre la recta AN que es tangente a la parábola en el vértice principal.

Proposición XIII. Problema VIII

Si un cuerpo se mueve en el perímetro de una parábola, hállese la ley de la fuerza centrípeta tendente al foco de dicha figura.

Sea la misma figura construida para el Lema, y sea P el cuerpo en el perímetro de la parábola, y desde el punto Q, hacia el que el cuerpo se mueve, tracemos QR paralela a SP así como la perpendicular QT y la recta Q*v* paralela a la tangente y que corta tanto al diámetro PG en *v* como a SP en *x*. Ahora bien por la semejanza de los triángulos P*xv*, SPM, y por la igualdad de los lados SM, SP de uno, son iguales los lados P*x* o QR y Pu del otro. Pero (por las cónicas) el cuadrado de la ordenada Q*v* es igual al rectángulo bajo el «latus rectum» y el segmento P*v* del diámetro; esto es (por el Lema XIII) al rectángulo 4PS x P*v* o 4PS x QR; y al coincidir P y Q la razón Q*v* a Q*x* (por el Corolario 2 del Lema VII) es una razón de igualdad. Luego Qx^2, en tal

caso, es igual al rectángulo 4PS x QR. Por tanto (por la semejanza de los triángulos Q*x*T, SPN) $\frac{Qx^2}{QT^2} = \frac{PS^2}{SN^2}$ (esto es, por el Corolario 1 del Lema XIV) $\frac{PS}{SA} = \frac{4PS \text{ x } QR}{4SA \text{ x } QR}$, y de aquí (por la Proposición IX del Libro V de los *Elementos*), QT^2 = 4SA x QR. Multiplicando estas igualdades por $\frac{SP^2}{QR}$ obtendremos que $\frac{SP^2 \text{ x } QT^2}{QR} = SP^2$ x 4SA; y por esto (por los Corolarios 1 y 5 de la Proposición VI) la fuerza centrípeta es inversamente como SP^2 x 4SA, esto es, por estar dado 4SA, inversamente como el cuadrado de la distancia SP. Q. E. I.



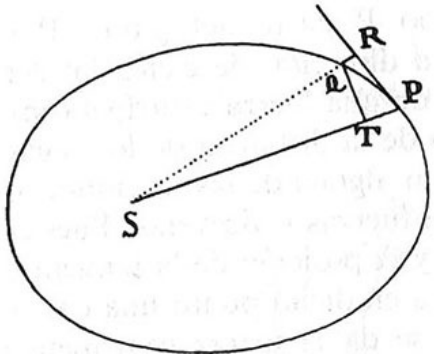
COROLARIO 1. Se sigue de las tres últimas proposiciones que si un cierto cuerpo P parte del punto P con una velocidad cualquiera según la dirección de cierta línea recta PR y bajo la acción simultánea de una fuerza centrípeta inversamente proporcional al cuadrado de la distancia de los lugares al centro, dicho cuerpo se moverá en alguna de las secciones cónicas que tenga su foco en el centro de fuerzas y viceversa. Pues conocidos el foco, el punto de contacto y la posición de la tangente, se puede describir la cónica que tendrá en dicho punto una curvatura dada. Pero se da la curvatura si se da la fuerza centrípeta y la velocidad del cuerpo; y dos órbitas tangentes entre sí con la misma fuerza centrípeta y la misma velocidad no pueden describirse.

COROLARIO 2. Si la velocidad con que un cuerpo parte de su lugar P es tal que el segmento PR puede ser recorrido en una parte infinitésima de tiempo y la fuerza centrípeta es tal que en el mismo tiempo puede mover al cuerpo a lo largo del espacio QR, dicho cuerpo se moverá en una sección cónica cuyo «latus rectum» principal es la susodicha cantidad $\frac{QT^2}{QR}$ en el último instante cuando las líneas PR, QR, disminuyen *in infinitum*. En estos Corolarios considero al círculo como una elipse; y exceptúo el caso en que el cuerpo desciende hacia el centro según una línea recta.

PROPOSICIÓN XIV. TEOREMA VI

Si varios cuerpos giran en torno a un centro común y la fuerza centrípeta es inversamente como el cuadrado de la distancia de los lugares al centro, digo que los «latera recta» principales de las órbitas son como los cuadrados de las áreas descritas en tiempos iguales por los radios trazados al centro.

Pues (por el Corolario 2 de la Proposición XIII) el «latus rectum» L es igual a la magnitud $\frac{QT^2}{QR}$ en el último instante cuando P y Q tienden a coincidir. Pero la pequeña línea QR en un tiempo dado es como la fuerza centrípeta generatriz, esto es (por hipótesis) inversamente como SP^2. Luego $\frac{QT^2}{QR}$ es como QT^2 x SP^2, esto es, el «latus rectum» L por el cuadrado del área QT x SP. Q. E. D.



COROLARIO. De aquí que el área total de la elipse así como la del rectángulo, proporcional a ella, comprendido entre los ejes es como el producto de la raíz cuadrada del «latus rectum» y el tiempo periódico. Pues el área total es como el área QT x SP, descrita en un tiempo dado, multiplicada por el tiempo periódico.

PROPOSICIÓN XV. TEOREMA VII

Supuesto esto, digo que los tiempos periódicos en las elipses son como los ejes mayores elevados a la potencia $^3/_2$.

Puesto que el eje menor es media proporcional entre el eje mayor y el «latus rectum» y por tanto el rectángulo comprendido entre los ejes es como el producto de

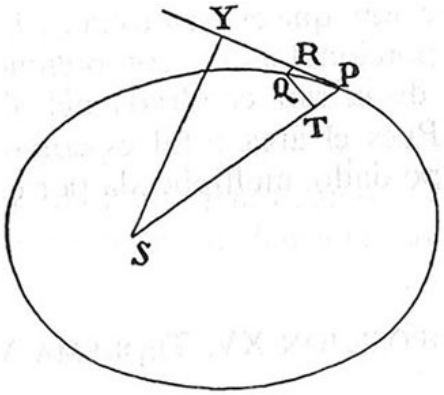
la raíz cuadrada del «latus rectum» y del eje mayor elevado a $^3/_2$. Pero dicho rectángulo (por el Corolario de la Proposición XIV) es como el producto de la raíz cuadrada del «latus rectum» y del tiempo periódico. Dividamos uno y otro por la raíz cuadrada del «latus rectum» y tendremos que el tiempo periódico sigue en la misma razón que la potencia $^3/_2$ del eje mayor. Q. E. D.

COROLARIO. Los tiempos periódicos en las elipses son iguales a los de los círculos, cuyos diámetros son iguales a los ejes mayores de las elipses.

PROPOSICIÓN XVI. TEOREMA VIII

Supuesto lo dicho y trazadas a los cuerpos líneas rectas tangentes también en ese punto a la órbita, y trazadas perpendiculares desde el foco común a dichas tangentes: digo que las velocidades de los cuerpos están en razón compuesta, inversamente, de las perpendiculares y, directamente, de la raíz cuadrada de los «latera recta» principales.

Tracemos desdé el foco S a la tangente PR la perpendicular SY y la velocidad del cuerpo P será inversamente como la raíz de la cantidad $\frac{SY^2}{L}$. Pues dicha velocidad es como el arco infinitamente pequeño PQ descrito en una partícula instantánea de tiempo, esto es (por el Lema VII) como la tangente PR, lo que es lo mismo, por la proporcionalidad de PR a QT y de SP a SY, como $\frac{SP \times QT}{SY}$, o también, inversamente como SY y directamente como SP x QT; y SP x QT es como el área descrita en un tiempo dado, esto es (por la Proposición XIV) como la raíz cuadrada del «latus rectum». Q. E. D.



COROLARIO 1. Los «latera recta» principales están en razón compuesta de los cuadrados de las perpendiculares y del cuadrado de las velocidades.

COROLARIO 2. Las velocidades de los cuerpos en las distancias máximas y mínimas al foco común están en razón compuesta inversamente de las distancias y directamente de la raíz cuadrada de los «latera recta» principales. Pues las perpendiculares son ahora las distancias mismas.

COROLARIO 3. Y, por tanto, la velocidad en una sección cónica, en la distancia máxima o mínima al foco, es a la velocidad en un círculo, a la misma distancia del centro, como la raíz cuadrada del «latus rectum» principal al doble de dicha distancia.

COROLARIO 4. Las velocidades de los cuerpos que giran en elipses a distancias medias del foco común son iguales a las de los cuerpos que giran en círculos a las mismas distancias; esto es (por el Corolario 6 de la Proposición IV) inversamente como la raíz cuadrada de las distancias. Pues ahora las perpendiculares son semiejes menores y éstos son como las medias proporcionales entre las distancias y los «latera recta». Multiplíquense la inversa de esta razón por la raíz cuadrada de la razón directa de los «latera recta» y obtendremos la raíz cuadrada de la razón inversa de las distancias.

COROLARIO 5. En la misma figura, o incluso en figuras distintas, cuyos «latera recta» principales son iguales, la velocidad del cuerpo es inversamente como la perpendicular trazada desde el foco a la tangente.

COROLARIO 6. En una parábola la velocidad es inversamente como la raíz cuadrada de la distancia del cuerpo al foco de la figura; en una elipse varía más y en una hipérbola menos que la razón susodicha. Pues (por el Corolario 2 del Lema XIV) la perpendicular trazada desde el foco a la tangente de la parábola es como la raíz cuadrada de la distancia. En la hipérbola la perpendicular varía menos y en la elipse varía más.

COROLARIO 7. En una parábola la velocidad de un cuerpo a cualquier distancia del foco es a la velocidad de un cuerpo que gira en un círculo a la misma distancia del centro, como la raíz cuadrada de la razón de 2 a 1; en la elipse es menor que esta razón y en la hipérbola mayor. Pues, por el Corolario segundo de esta Proposición, la velocidad en el vértice de la parábola está en esta razón, y, por el Corolario 6 de esta Proposición y por la cuarta, se conserva la misma proporción en todas las distancias. Y por tanto también, en una parábola, la velocidad es en todas partes igual a la velocidad de un cuerpo que gire en un círculo a la mitad de distancia, y es menor en una elipse y mayor en una hipérbola.

COROLARIO 8. La velocidad de un cuerpo que gira en una sección cónica cualquiera es a la velocidad de un cuerpo que gira en un círculo a la distancia del semi «latus rectum» principal de la sección como dicha distancia es a la perpendicular trazada desde el foco a la tangente de la sección. Es evidente por el Corolario quinto.

Corolario 9. De donde, dado que (por el Corolario 6 de la Proposición iv) la velocidad de un cuerpo que gira en este círculo es a la velocidad de otro cuerpo que gira en cualquier otro círculo inversamente como la raíz cuadrada de las distancias, por lo mismo se seguirá que la velocidad de un cuerpo que gira en una sección cónica será a la velocidad de un cuerpo que gira en un círculo a la misma distancia como es la media proporcional entre dicha distancia común y la mitad del «latus rectum» principal de la sección, a la perpendicular trazada desde el foco común a la tangente de la sección.

Proposición xvii. Problema ix

Supuesto que la fuerza centrípeta es inversamente proporcional al cuadrado de la distancia de los lugares al centro y que la cantidad absoluta de dicha fuerza es conocida, hállese la línea descrita por el cuerpo desde un punto dado, con una velocidad dada y partiendo en la dirección de una recta dada.

Sea la fuerza centrípeta tendente a un punto S la fuerza con la que gira un cuerpo *p* en una órbita dada *pq* y supongamos conocida su velocidad en el punto *p*. Parta el cuerpo P desde el punto P según la línea PR con una velocidad dada, e inmediatamente, obligado por la fuerza centrípeta, desvíese por la sección cónica PQ. La recta PR será tangente a la sección en P. Del mismo modo, la recta *pr* sea tangente a la órbita *pq* en *p* y si se supone que se trazan perpendiculares a dichas tangentes desde S, ocurrirá (por el Corolario 1 de la Proposición xvi) que el «latus rectum» principal de la sección cónica estará, respecto al «latus rectum» principal de la órbita, en razón compuesta de la razón cuadrada de las perpendiculares y de la razón cuadrada de las velocidades y, por ello, está dada. Sea L el «latus rectum» de la sección. Se da también el foco S de la misma sección cónica. Sea el ángulo RPH el suplementario de RPS y tendremos dada en posición la línea PH en la que está el otro foco H. Una vez trazada la perpendicular SK sobre PH, supongamos que se traza el semieje conjugado BC y tendremos que:

$$SP^2 - 2PH \times PK + PH^2 = SH^2 = 4CH^2 = 4(BH^2 - BC^2) =$$
$$(SP + PH)^2 - L(SP + PH) =$$
$$SP^2 + 2PS \times PH + PH^2 - L(SP + PH)$$

añádase a ambos miembros

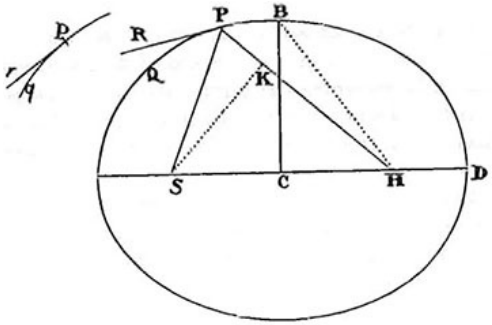
$$2PK \times PH - SP^2 - PH^2 + L(SP + PH)$$

y resultará

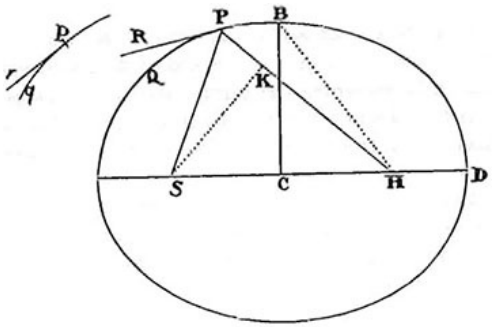
$$L(SP + PH) = 2PS \times PH + 2PK \times PH,$$

o también

$$\frac{(SP + PH)}{PH} = \frac{2(SP + KP)}{L}$$



De donde tenemos dada tanto en longitud, como en posición, PH. Esto es, si la velocidad del cuerpo en P es tal que el «latus rectum» L fuera menor que 2SP + 2KP entonces PH caerá, con la línea PS, del mismo lado de la tangente PR y, por tanto, la figura será una elipse, y vendrá dada, por estar dados los focos S, H y el eje principal SP + PH. Pero si la velocidad del cuerpo es tal que el «latus rectum» L fuese igual a 2SP + 2KP la longitud de PH será infinita; y por tanto la figura será una parábola que tiene por eje a SH paralelo a la línea PK, por tanto, también estará dada. Y si el cuerpo parte de P con una velocidad aún mayor, la longitud PH estará al otro lado de la tangente y, por tanto, la tangente pasa por entre ambos focos, siendo entonces la figura una hipérbola que tiene como eje principal la diferencia de las líneas SP y PH; y por eso estará dada. Pues si el cuerpo en todos estos casos gira en una sección cónica así hallada, está demostrado en las Proposiciones xi, xii, xiii que la fuerza centrípeta será inversamente como el cuadrado de la distancia del cuerpo al centro de fuerzas S; y por tanto la línea PQ representa correctamente la trayectoria descrita por el cuerpo con semejante fuerza, partiendo de un punto dado P, con una velocidad dada y según una recta PR dada en posición. Q. E. F.



COROLARIO 1. De aquí que en toda sección cónica, dados el vértice principal D, el «latus rectum» L y el foco S, se obtiene el otro foco H, tomando DH a DS como el «latus rectum» a la diferencia entre el «latus rectum» y 4DS. Pues la proporción $\frac{SP + PH}{PH} = \frac{2SP + 2KP}{L}$ se convierte para este Corolario en $\frac{DS + DH}{DH} = \frac{4DS}{L}$ y $\frac{DS}{DH} = \frac{4DS - L}{L}$.

COROLARIO 2. De donde, dada la velocidad de un cuerpo en el vértice principal D, se hallará directamente la órbita, a saber, tomando su «latus rectum» al doble de la distancia DS, en razón cuadrada de esta velocidad dada a la velocidad del cuerpo que gira en el círculo a la distancia DS (por el Corolario 3 de la Proposición XVI), y tomando después DH a DS como el «latus rectum» a la diferencia entre el «latus rectum» y 4DS.

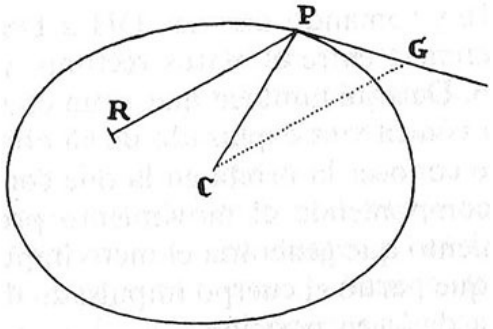
COROLARIO 3. De aquí también que, si un cuerpo se mueve en cualquier sección cónica y es expulsado de su órbita por cualquier impulso, se puede conocer la órbita en la que continuará después su curso. Pues, componiendo el movimiento propio del cuerpo con aquel movimiento que generaría el mero impulso, se tendrá el movimiento con que partió el cuerpo impulsado de un lugar dado y según una recta dada en posición.

COROLARIO 4. Y si dicho cuerpo es perturbado continuamente por una fuerza impresa desde fuera, podemos conocer muy aproximadamente su curso calculando los cambios que en algunos puntos produce dicha fuerza y estimando a partir de la analogía de la serie los cambios continuos en los lugares intermedios[23].

ESCOLIO

Si un cuerpo P, con una fuerza centrípeta tendente a un punto dado cualquiera R, se

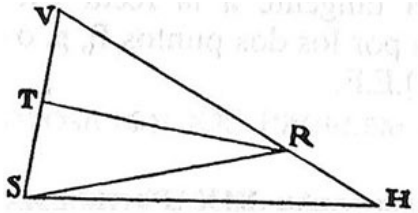
mueve en el perímetro de una sección cualquiera dada cuyo centro sea C, y se desea hallar la ley de la fuerza centrípeta, trácese GC paralela al radio RP y cortando a la tangente PG en G; y entonces la fuerza mencionada será (por el Corolario 1 y el Escolio de la Proposición X, y por el Corolario 3 de la Proposición VII) como $\frac{CG^3}{RP^2}$ [24].



Sección IV
DE COMO HALLAR ÓRBITAS ELÍPTICAS, PARABÓLICAS E HIPERBÓLICAS A PARTIR DE UN FOCO DADO

Lema XV

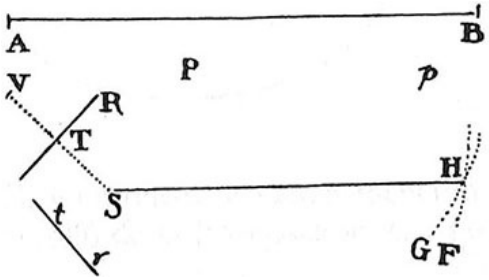
Si de los dos focos S, H, de una elipse o de una hipérbola trazamos hasta un tercer punto V las rectas SV, HV, una de las cuales HV es igual al eje principal de la figura, esto es, al eje en que están situados los focos y la otra SV es bisecada en T por la perpendicular TR trazada sobre ella, dicha perpendicular TR tocará a la sección cónica en algún punto; y al contrario, si la toca, HV será igual al eje principal de la figura.



Corte la perpendicular TR a la recta HV, prolongada si es preciso, en R, y únanse S y R. Por ser iguales TS y TV serán también iguales las rectas SR, VR, y los ángulos TRS y TRV. De donde el punto R estará en la sección cónica y la perpendicular TR será tangente a ella; «et contra». Q. E. D.

Proposición XVIII. Problema X

Dados el foco y los ejes principales trazar trayectorias elípticas e hiperbólicas que pasen por puntos dados y toquen a rectas dadas en posición.

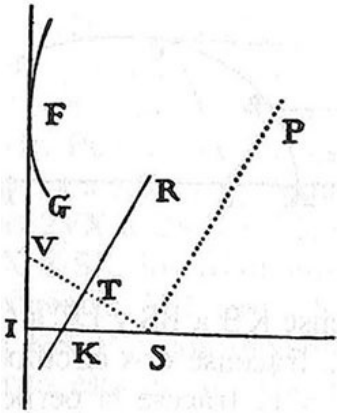


Sea S el foco común de las figuras; sea AB la longitud del eje principal de una trayectoria cualquiera; P el punto por el que debe pasar la trayectoria y sea TR la recta que debe tocar. Con centro en P y radio AB - SP, si la órbita es una elipse, AB + SP si fuere una hipérbola, trácese el círculo HG. Trácese la perpendicular ST sobre la tangente TR y prolónguese aquella hasta V, de modo que TV sea igual a ST y con centro en V y radio AB trácese el círculo FH. De esta manera, ya se den dos puntos P, *p*, ya dos tangentes, TR, *tr*, ya un punto P y la tangente TR, hay que trazar dos círculos. Sea H su intersección, y desde los focos S, H, con el eje dado trácese la trayectoria. Digo que está dada. Pues la trayectoria descrita (dado que PH + SP en una elipse y PH - SP en una hipérbola, es igual al eje) pasará por el punto P y (por el Lema anterior) será tangente a la recta TR. Y por el mismo argumento o pasará por los dos puntos P, *p*, o será tangente a las dos rectas TR, *tr*. Q. E. F.

Proposición XIX. Problema XI

Trazar en torno a un foco dado una trayectoria parabólica que pase por unos puntos dados y sea tangente a rectas dadas en posición.

Sea S el foco, P el punto y TR la tangente a la trayectoria que se ha de trazar. Trácese el círculo FG, con centro en P y radio PS.

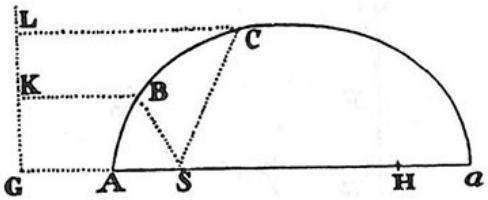


Trácese desde el foco a la tangente la perpendicular ST y prolónguese hasta V de tal modo que TV sea igual a ST. Del mismo modo habría que trazar el otro círculo *fg* si se diera otro punto *p*, o también hallar otro punto *v* si se diera otra tangente *tr*; después trácese la recta IF, tangente a los dos círculos FG, *fg* si se dan dos puntos P, *p*, o que pase por los puntos V, *v*, si se dan dos tangentes TR, *tr*, o que sea tangente al círculo FG y pase por el punto V si se da el punto P y la tangente TR. Trácese sobre FI la perpendicular SI, bisecándola en K; y con eje SK y vértice principal en K, trácese la parábola. Digo que está resuelto. Pues la parábola, por ser iguales SK, IK entre sí y SP, FP, también, pasará por el punto P; y (por el Lema XIV, Corolario 3) al ser iguales ST y TV y ser recto el ángulo STR, será tangente a la recta TR. Q. E. F.

Proposición XX. Problema XII

Trazar en torno a un foco dado una trayectoria dada de cualquier género que pase por puntos dados y sea tangente a rectas dadas en posición.

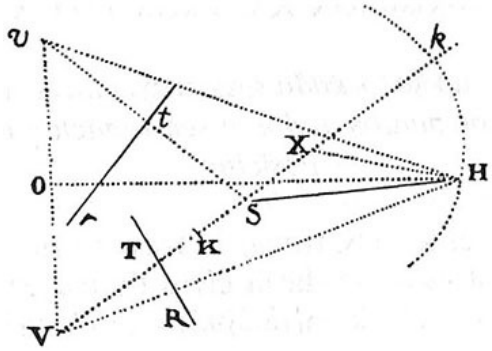
Caso 1. Dado el foco S, trazar la trayectoria ABC por los dos puntos B, C. Puesto que se da la clase de trayectoria también se dará la razón entre el eje principal y la distancia entre focos.



Según esa razón tómense KB a BS y LC a CS. Con centros en B y C, y radios BK y

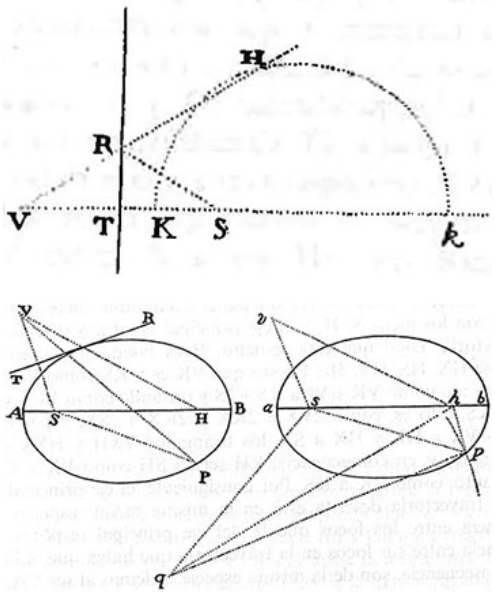
CL, trácense dos círculos y, sobre la recta KL tangente a ellos en K y L, trácese la perpendicular SG y córtese ésta en A y a de tal modo que GA sea a AS y G*a* a *a*S como KB es a BS y con eje A*a* y vértices en A, *a*, trácese la trayectoria. Digo que está resuelto. Pues sea H el otro foco de la figura trazada y al ser GA a AS como G*a* a *a*S, restando también estará G*a* - GA (o sea A*a*) a *a*S - AS (o sea SH) en la misma razón y, por tanto, en la misma razón en que está el eje principal de la figura que debe ser trazada con la distancia entre sus focos; y por tanto la figura trazada es del mismo tipo que la que había que trazar. Y, dado que KB respecto a BS y LC respecto a CS están en la misma razón, esta figura pasará por B, C, como se sabe por las cónicas.

Caso 2. Dado el foco S trazar la trayectoria que sea tangente a las rectas TR, *tr*, en cualquier punto. Trácense desde el foco a las tangentes las perpendiculares ST, *st*, y prolónguense hasta V, *v*, de tal modo que TV, *tv*, sean iguales respectivamente a TS y *ts*. Bisecando V*v* en O trácese desde ésta una perpendicular infinita OH y prolónguese al infinito la recta VS cortándola en los puntos K, *k* de tal modo que VK sea a KS y V*k* a *k*S como el eje principal de la figura que ha de hallarse es a la distancia entre focos. Sobre el diámetro K*k* trácese el círculo que corte a OH en H; y con los focos S, H, y el eje principal igual a VH trácese la trayectoria. Digo que está resuelto. Pues biséquese K*k* en *x* y únanse HX, HS, HV, H*v*. Puesto que VK es a KS como V*k* a *k*S; y sumando como VK + V*fc* a KS + *k*S y restando como V*k* - VK a *k*S - KS, esto es, como 2VX a 2KX y 2KX a 2SX, y por tanto como VX a HX y HX a SX, los triángulos VXH y HXS serán semejantes y, en consecuencia, VH será a SH como VX a XH y, por tanto, como VK a KS. Por consiguiente, el eje principal VH de la trayectoria descrita está en la misma razón respecto a la distancia entre los focos que la del eje principal respecto a la distancia entre sus focos en la trayectoria que había que hallar y, en consecuencia, son de la misma especie. Además al ser VH, *u*H, iguales al eje principal y estar VS, *v*S, bisecadas perpendicularmente por las rectas TR, *tr*, es evidente (por el Lema XV) que dichas rectas son tangentes a la trayectoria descrita. Q. E. F.



Caso 3. Dado el foco S trazar una trayectoria que sea tangente en el punto dado R

a la recta TR. Trácese la perpendicular ST a la recta TR y prolónguese aquélla hasta V, de tal modo que TV sea igual a ST. Únase VR, y con la recta VS prolongada infinitamente y cortada en K y *k*, de tal modo que VK sea a SK y *vk* sea a S*k* como el eje principal de la elipse que hay que describir es a la distancia entre sus focos, y trazado un círculo sobre el diámetro K*k*, corte aquél en H a la prolongación de la recta VR, y con los focos S, H, y el eje principal igual a la recta VH trácese la trayectoria. Digo que está resuelto. Pues que VH es a SH como VK es a SK y, por tanto, como el eje principal de la trayectoria que hay que trazar es a la distancia entre sus focos, es evidente por las demostraciones del caso segundo y, por tanto, que la trayectoria descrita es de la misma especie que la que había que describir, mientras que la recta TR, bisectriz del ángulo VRS, es tangente a la trayectoria en el punto R, es evidente por las cónicas. Q. E. F.



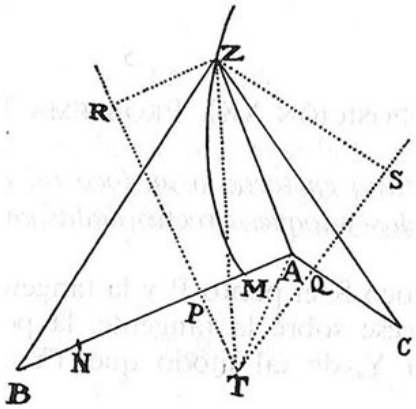
Caso 4. Trácese ahora en torno al foco S la trayectoria APB tangente a la recta TR y que pase por un punto cualquiera P dado fuera de la tangente. Y que sea semejante a la figura *apb* cuyo eje principal es *ab* y trazada con los focos *s, h*. Trácese sobre la tangente TR, la perpendicular ST y prolónguese hasta V de modo que sea TV igual a ST. Constrúyanse los ángulos *hsq*, *shq*, iguales a VSP, SVP; con centro en *q* y con radio tal que sea a *ab* como SP a VS trácese un círculo que corte la figura *apb* en *p*. Únase *sp* y hágase que SH sea a *sh* como SP a *sp* y éstas construyan así el ángulo PSH igual al ángulo *psh* y el ángulo VSH igual al *psq*. Por fin con los focos S, H, el eje principal AB, igual a la distancia VH, trácese la sección cónica. Digo que está resuelto. Pues si se hace que *sv* sea a *sp* como *sh* es a *sq*, y que el

ángulo *vsp* sea igual al ángulo *hsq* y el ángulo *vsh* igual al ángulo *psq*, entonces los triángulos *svh* y *spq* serán semejantes y, por tanto, *vh* será a *pq* como *sh* es a *sq*, esto es (por la semejanza de los triángulos VSP, *hsq*), como VS es a SP o como *ab* es a *pq*. Son pues, iguales *vh* y *ab*. Pero, por la semejanza de los triángulos VSH, *vsh*, ocurre que VH es a SH como *vh* es a *sh*, esto es, el eje de la sección cónica descrita es a la distancia entre sus focos, como el eje *ab* a la distancia *sh* entre sus focos; y por tanto, la figura descrita es semejante a la figura *apb*. Pasará, pues, dicha figura por el punto P, dado que el triángulo PSH es semejante al triángulo *psh*; y dado que VH es igual a su eje y que VS es bisecada perpendicularmente por la recta TR, también será tangente a dicha recta TR. Q. E. F.

Lema xvi

Trazar desde tres puntos dados a otro cuarto punto no dado tres rectas cuyas diferencias o se dan o son nulas.

Caso 1. Sean los puntos dados A, B, C, y el cuarto punto Z, que es necesario hallar; por la diferencia conocida de las líneas AZ, BZ, el punto Z se situará en una hipérbola cuyos focos son A y B, y el eje principal la mencionada diferencia dada. Sea tal eje MN. Tómese PM a MA como MN a AB y trácese PR perpendicular a AB y también tírese ZR perpendicular a PR; por la naturaleza de esta hipérbola ocurrirá que ZR es a AZ como MN es a AB. Por un razonamiento similar el punto Z estará ubicado en otra hipérbola cuyos focos son AC y cuyo eje principal es la diferencia entre AZ y CZ, y puede ahora trazarse QS perpendicular a AC y si después se traza ZS, normal a QS, desde el punto Z de la hipérbola, dicha ZS será a AZ como la diferencia entre AZ y CZ es a AC. Se tienen dadas, pues, las razones entre ZR y ZS con AZ y, por tanto, también las de ZR y ZS entre sí; por lo cual si las rectas RP, SQ, se encuentran en T, y se trazan TZ y TA la figura TRZS estará dada en especie y la recta TZ en la que estará contenido en algún lugar el punto Z estará dada en posición. También estará dada la recta TA lo mismo que el ángulo ATZ; y puesto que están dadas las razones de AZ y TZ a ZS, estarán dadas sus razones entre sí; y de aquí se tendrá dado el triángulo ATZ cuyo vértice es el punto Z. Q. E. I.



CASO 2. Si dos de las tres líneas, por ejemplo, AZ y BZ se igualan, trácese la recta TZ, de tal modo que biseque la recta AB; después hállese el triángulo ATZ «ut supra».

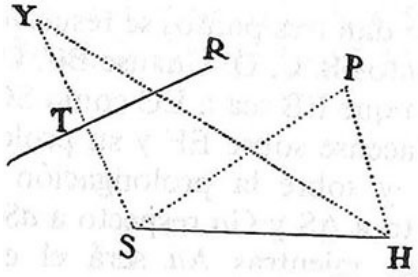
CASO 3. Si las tres son iguales, el punto Z está ubicado en el centro del círculo que pase por los puntos A, B, C. Q. E. I.

Este Lema problemático se resuelve también por el «Liber tactionum» de *Apolonio*, restituido por *Vieta*[25].

PROPOSICIÓN XXI. PROBLEMA XIII

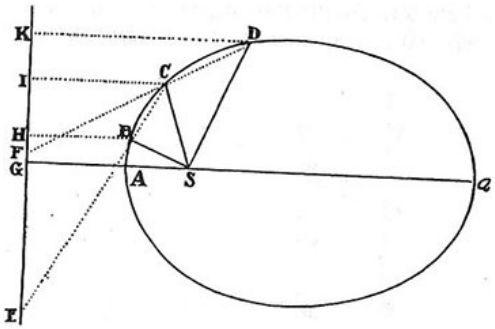
Trazar una trayectoria en torno a un foco tal que pase por unos puntos dados y toque a rectas dadas en posición.

Supóngase el foco S, el punto P y la tangente TR y hállese el otro foco H. Trácese sobre la tangente, la perpendicular ST y prolónguese hasta Y, de tal modo que TY sea igual a ST y ocurrirá que YH será igual al eje principal. Únase SP y PH, y SP será la diferencia entre HP y el eje principal. De este modo si se dieran varias tangentes TR, o varios puntos P, se llegará siempre a líneas tales como YH o PH trazadas desde dichos puntos Y o P al foco H, las cuales o son iguales a los ejes o se diferencian de ellos según las longitudes dadas SP y que, por tanto, o son iguales entre sí o tienen entre sí diferencias dadas; en consecuencia, por el Lema anterior, el otro foco H está dado. Hallados los focos y la longitud del eje (el cual o es YH, o, si la trayectoria es una elipse, PH + SP, o, si es una hipérbola, PH - SP) se tendrá la trayectoria. Q. E. I.



ESCOLIO

Cuando la trayectoria es una hipérbola no incluyo bajo la definición de ésta a su hipérbola opuesta. Pues un cuerpo, permaneciendo en su movimiento, no puede pasar a la hipérbola opuesta.



El caso en que se dan tres puntos se resuelve más rápidamente así. Sean los tres puntos B, C, D. Únanse BC, CD, y prolónguense hasta E, F, de modo que EB sea a EC como SC a SD y FC a FD como SC a SD. Trácense sobre EF y su prolongación las rectas normales SG, BH, y sobre la prolongación al infinito de GS tómense GA respecto a AS y G*a* respecto a *a*S como HB es a BS y A será el vértice, mientras A*a* será el eje principal de la trayectoria; la cual, según que GA sea mayor, igual o menor que AS, será elipse, parábola o hipérbola; en el primer caso el punto *a* caerá al mismo lado de la línea GF que el punto A, en el segundo caso se alejará al infinito y en el tercero caerá al lado opuesto que la línea GF. Pues si se trazan sobre GF las perpendiculares CI, DK, entonces IC será a HB como EC a EB, esto es, como SC a SB; y al revés, IC a SC como HB a SB o como GA a S*a*. Y por un argumento semejante se probaría que KD está en la misma razón con SD. Están situados, pues, los puntos B, C, D, en la sección cónica descrita en torno al foco S, de tal modo que todas las rectas trazadas desde el foco S a cada punto de la sección estén en la susodicha razón con respecto a las perpendiculares trazadas desde dichos puntos a la

recta GF.

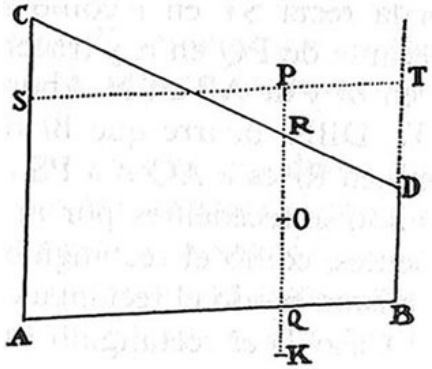
El ilustre geómetra *de la Hire* resuelve este problema por un método no muy distinto en su libro VIII de las cónicas, Proposición XXV.

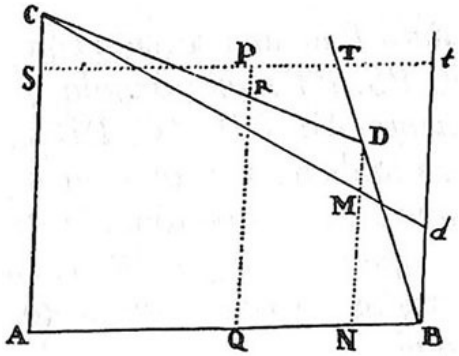
Sección V
LA OBTENCIÓN DE ÓRBITAS CUANDO NO SE DA NINGUNO DE LOS FOCOS

Lema XVII

Si de un punto cualquiera P de una sección cónica dada se trazan las líneas rectas PQ, PR, PSy PT según ángulos dados sobre los lados prolongados infinitamente AB, CD, AC, DB de un trapecio ABDC que se halla inscrito en la dicha sección cónica y trazadas las líneas una a cada uno; ocurrirá que el rectángulo de las rectas trazadas sobre los dos lados opuestos PQ x PR estará con respecto al rectángulo de las trazadas sobre los otros dos lados opuestos PS x PT en una razón dada.

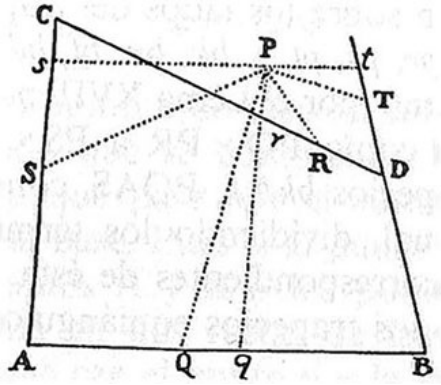
Caso 1. Supongamos, en primer lugar, que las líneas trazadas sobre los lados opuestos son paralelas a otro de los restantes lados, por ejemplo PQ y PR al lado AC, y PS y PT al lado AB. Además, sean dos de los lados opuestos, por ejemplo AC y BD, paralelos entre sí. Entonces también la recta que biseca dichos lados paralelos será uno de los diámetros de la sección cónica y bisecará también a RQ. Sea O el punto en que RQ es bisecada, y PO será ordenada aplicada al diámetro. Prolónguese PO hasta K de tal modo que OK sea igual a PO y OK será ordenada aplicada al diámetro en sentido opuesto. Dado que los puntos A, B, P y K están en la sección cónica y que PK corta a AB en un ángulo dado, tendremos (por las prop. 17, 19, 21 y 23 del Lib. III de las cónicas de *Apolonio*) que el rectángulo PQK está con el rectángulo AQB en una razón dada. Pero QK y PR son iguales, por cuanto son diferencias iguales de OK, OP y OQ, OR, y por tanto también los rectángulos PQK y PQ x PR son iguales; y en consecuencia el rectángulo PQ x PR está con el rectángulo AQB, esto es, con el rectángulo PS x PT en una razón dada. Q. E. D.





CASO 2. Supongamos ahora que los lados opuestos del trapecio AC y BD no sean paralelos. Tracemos B*d* paralela a AC y que corte tanto a la recta ST en *t* como a la sección cónica en *d*. Unamos C*d*, secante de PQ en *r*, y tracemos DM paralela a PQ y secante de C*d* en M y de AB en N. Ahora, por la semejanza de los triángulos BT*t*, DBN, ocurre que B*t* o PQ es a T*t* como DN a NB. Y así también R*r* es a AQ o a PS como DM es a AN.

Luego multiplicando antecedentes por antecedentes y consecuentes por consecuentes, como el rectángulo PQ x R*r* es al rectángulo PS x T*t* del mismo modo el rectángulo NDM es al rectángulo ANB, y (por el Caso 1) el rectángulo PQ x P*r* es al rectángulo PS x P*t* y, dividiendo, también es el rectángulo PQ x PR al rectángulo PS x PT. Q. E. D.

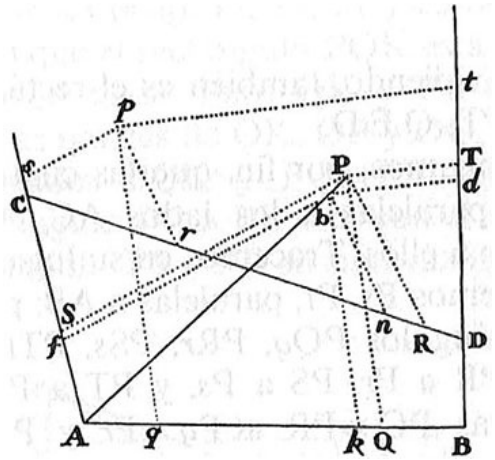


CASO 3. Supongamos, por fin, que las cuatro líneas PQ, PR, PS, PT, no son paralelas a los lados AC, AB, sino que son inclinadas respecto a ellos. Tracemos en su lugar P*q*, P*r*, paralelas al lado AC; y tracemos P*s*, P*t*, paralelas a AB; por estar dados los ángulos de los triángulos PQ*q*, PR*r*, PSS, PT*t*, se darán las razones PQ a P*q*, PR a P*r*, PS a P*s*, y PT a P*t*; y por tanto las razones compuestas PQ x PR a P*q* x P*r* y PS x PT a P*s* x P*t*. Pero según lo demostrado más arriba, la razón P*q* x P*r* a P*s* x P*t* está dada: luego también lo estará la razón PQ x PR a PS x PT. Q. E. D.

LEMA XVIII

Con los mismos supuestos, si el rectángulo de las dos líneas trazadas a dos lados opuestos del trapecio, PQ x PR, está al rectángulo PS x PT de las trazadas sobre los otros dos lados en una razón dada, entonces el punto P desde el que se trazan dichas líneas es tangente a la sección cónica descrita en torno al trapecio.

Supongamos una sección cónica descrita por los puntos A, B, C, D, e infinitos puntos P como por ejemplo *p*: digo que el punto P siempre será tangente a esta sección. Si lo negaras, únase AP secante a la sección cónica en un punto distinto de P, si ello es posible, como por ejemplo en *b*. Por consiguiente, si de estos puntos *p* y *b* se trazan sobre los lados del trapecio según ángulos dados las rectas *pq, pr, ps, pt*, y *bk, bn, bf, bd*; tendremos que *bk* x *bn* será a *bf* x *bd* como (por el Lema XVII) *pq* x *pr* a *ps* x *pt* y por tanto (por hipótesis) como PQ x PR a PS x PT. Y debido a la semejanza de los trapecios *bk*A*f*, PQAS, como es *bk* a *bf* así es PQ a PS. Por lo cual, dividiendo los términos de la primera proporción por los correspondientes de ésta, *bn* será a *bd* como PR a PT. Por tanto los trapecios equiángulos D*nbd*, DRPT son semejantes y sus diagonales D*b*, DP coinciden. Y de este modo *b* cae en la intersección de las líneas AP, DP y por tanto coincide con P. Por ello sea cual sea el punto donde se suponga P caerá en la sección cónica supuesta. Q. E. D.

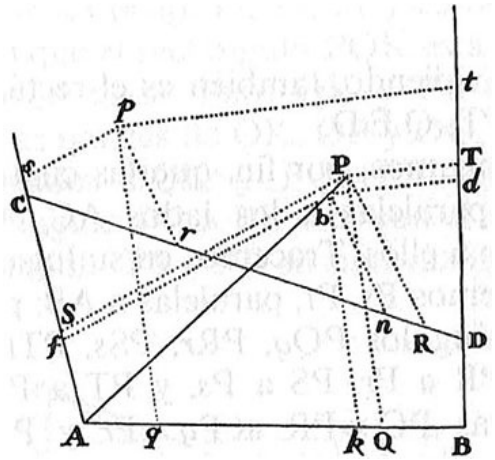


COROLARIO. De aquí que, si desde un punto común P se trazan tres rectas PQ, PR, PS sobre otras tres dadas en posición, una a cada una y según ángulos dados y el rectángulo formado por dos de las trazadas PQ x PR es al cuadrado de la tercera PS según una razón dada, el punto P desde el que fueron trazadas estará ubicado en la sección cónica tangente a las líneas AB, CD en A y C; y viceversa. Pues aproxímese hasta juntarse la línea BD a la línea AC manteniendo su posición las tres AB, CD, AC; después hagamos lo mismo con PT hacia PS; y entonces el rectángulo PS x PT devendrá PS^2 y las rectas AB, CD que cortaban la curva en los puntos A y B, C y D ya no pueden cortar más la curva en aquellos puntos coincidentes, sino sólo ser tangentes en ellos.

ESCOLIO

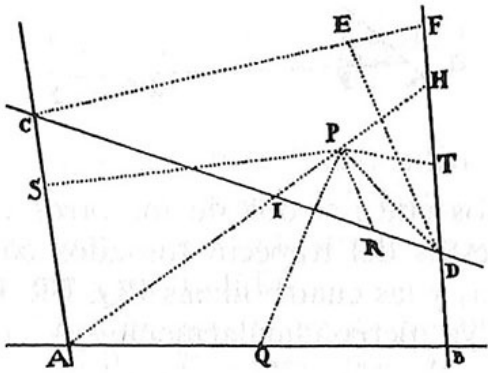
En este Lema el nombre de sección cónica se toma en sentido amplio, de modo que se comprenda bajo el mismo tanto la sección rectilínea que pase por el vértice del cono como la circular paralela a la base. Pues si el punto *p* cae en la recta con que se unen los puntos A y B o los puntos C y D la sección cónica se transforma en dos rectas de las cuales una coincide con la recta en la que cae el punto *p* y la otra es la recta con la que quedan unidos entre sí dos de los otros cuatro puntos. Si dos ángulos opuestos del trapecio tomados conjuntamente son iguales a dos rectos y las cuatro líneas PQ, PR, PS, PT, se trazan sobre sus lados ya perpendicularmente ya con un grado de inclinación dado y el rectángulo comprendido entre dos de ellas PQ x PR es igual a PS x PT comprendido entre las otras dos, entonces la sección cónica resultará un círculo. Y lo mismo ocurre si se trazan cuatro líneas con ángulos cualesquiera y el rectángulo PQ x PR comprendido entre dos de ellas es al rectángulo PS x PT comprendido entre las otras dos como el rectángulo comprendido entre los senos de

los ángulos S, T, en los que inciden las dos últimas PS, PT es al rectángulo comprendido entre los senos de los ángulos Q, R, en que inciden las dos primeras PQ, PR. En todos los demás casos el lugar del punto P será una de las tres figuras que comúnmente se denominan secciones cónicas. Pero en lugar del trapecio ABCD puede colocarse un cuadrilátero cuyos dos lados opuestos se crucen mutuamente como diagonales. Y uno o dos de los cuatro puntos A, B, C, D, pueden llevarse al infinito con lo que los lados de la figura que convergen en dichos puntos resultarán paralelos; y en tal caso la sección cónica pasará por los demás puntos mientras que en la dirección de las paralelas irá hacia el infinito.



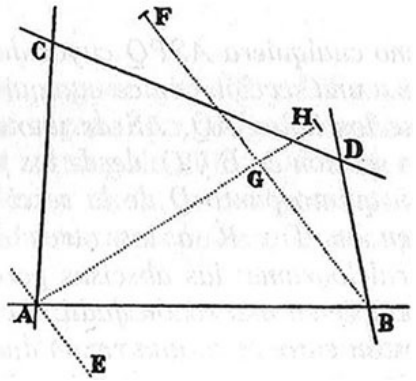
LEMA XIX

Hallar el punto P a partir del cual, si se trazan las cuatro rectas PQ, PR, PS, PT sobre otras cuatro en posición dada AB, CD, AC, BD, una a una y con ángulos dados, el rectángulo PQ x PR comprendido entre dos de aquéllas sea al rectángulo PS x PT comprendido entre las otras dos según una razón dada.



Hagamos que las líneas AB, CD, sobre las que se trazan las rectas PQ, PR, que contienen uno de los rectángulos, se crucen con las otras dos líneas de posición dada en los puntos A, B, C, D. Desde cualquiera de ellos A tracemos la recta AH en la que queremos hallar el punto P. Corte esta línea a las opuestas BD, CD, a saber, a BD en H y a CD en I y por estar dados todos los ángulos de la figura estarán dadas las razones de PQ a PA y de PA a PS y por tanto la de PQ a PS. Restando ésta de la razón dada de PQ x PR a PS x PT tendremos la razón de PR a PT, y sumando las razones dadas PI a PR y PT a PH tendremos la razón PI a PH, y por ende el punto P. Q. E. I.

COROLARIO 1. De aquí que también se pueda trazar una tangente a un punto cualquiera D del lugar de los infinitos puntos P. Pues la cuerda PD, cuando coinciden los puntos P y D, esto es, cuando discurre AH por el punto D, resulta una tangente. En tal caso la razón última de las evanescentes IP y PH, se halla como se ha visto más arriba. Tracemos CF, paralela a la propia AD, tal que corte a BD en F y cortada en E según la dicha razón última, y también de DE será tangente dado que CF y la evanescente IH son paralelas e igualmente cortadas en E y en P.



COROLARIO 2. De aquí también puede definirse el lugar de todos los puntos P. Tracemos por uno cualquiera de los puntos A, B, C, D, por ejemplo A, la tangente AE al lugar de todos los puntos y por otro punto cualquiera B tracemos una paralela BF a la tangente y que corte al lugar en F. Por el Lema XIX se hallará el punto F. Bisecando a BF en G y siendo indefinida la línea AG, ésta será la posición del diámetro al que vienen aplicadas ordenadamente BG y FG. Corte ahora AG al lugar en H y entonces AH será el diámetro o «latus transversum» al cual el «latus rectum» será como BG^2 a AG x GH. Si AG nunca cortase el lugar de los puntos, permaneciendo la línea AH infinita, entonces el lugar será una parábola y su «latus rectum» perteneciente al diámetro AG será $\frac{BG^2}{AG}$. Pero si corta al lugar en algún punto, entonces el lugar será una hipérbola cuando los puntos A y H estén situados al mismo lado del punto G; y una elipse cuando G esté entre A y H, salvo cuando por casualidad se dé que AGB sea recto y por tanto BG^2 sea igual al rectángulo AGH, caso en el que tendremos un círculo.

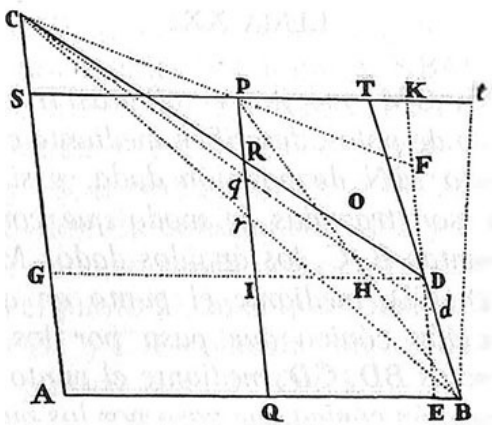
Y de este modo tenemos resuelto en el corolario, no por cálculo sino por composición geométrica como querían los antiguos, el problema por ellos planteado de las cuatro líneas, sugerido por *Euclides* y replanteado por *Apolonio*.

LEMA XX

Sea un paralelogramo cualquiera ASPQ cuyos dos ángulos opuestos A y P son tangentes a una sección cónica cualquiera en los puntos A y P; y prolónguense los lados AQ, AS de uno de ellos al infinito, hasta cortar a dicha sección en B y C; desde los puntos de cruce B y C trácense hasta un quinto punto D de la sección otras dos rectas BD, CD que corten en T y R a los otros dos lados PS, PQ prolongados del paralelogramo: las abscisas parciales PR y PT de los lados estarán siempre en una razón dada. Y al contrario, si tales abscisas parciales están entre sí en una razón dada, el punto D será tangente a la sección cónica que pase por los dichos cuatro puntos A, B, C, P.

CASO 1. Únanse BP, CP y trácense desde el punto D las dos rectas DG, DE, de las que la primera DG sea paralela a AB y corte a PB, PQ, CA, en H, I, G; y la otra DE sea paralela a AC y corte a PC, PS, AB, en F, K, E: (por el Lema XVII) el rectángulo DE x DF será al rectángulo DG x DH según una razón dada. Pero PQ es a DE (o IQ) como PB a HB, y por tanto como PT a DH; y miembro a miembro, PQ es a PT como DE a DH. Y además PR es a DF como RC es a DC, y por tanto como PS (o IG) a DG, y alternativamente PR es a PS como DF a DG; y conjugando las razones, el rectángulo PQ x PR es al rectángulo PS x PT como el rectángulo DE x DF al

rectángulo DG x DH, y por ende en razón dada. Pero PQ y PS están dados, luego también está dada la razón PR a PT. Q. E. D.



CASO 2. Pero si se supone que PR y PT están entre sí en una razón dada, entonces razonando de igual modo en sentido inverso se seguirá que el rectángulo DE x DF está con el rectángulo DG x DH en una razón dada, y por tanto el punto D (por el Lema XVIII) cae en la sección cónica que pasa por los puntos A, B, C, P. Q. E. D.

COROLARIO 1. De aquí que si se hace a BC secante de PQ en *r*, y se toma P*t* sobre PT en razón a P*r* precisamente en la misma que guarda PT a PR: entonces B*t* será tangente a la sección cónica en el punto B. Pues imaginemos al punto D acercándose al punto B de modo que, al tender a desaparecer la cuerda BD, BT resulte tangente; y CD lo mismo que BT coincidirían con CB y B*t*.

COROLARIO 2. Y viceversa, si B*t* es tangente y BD, CD coinciden en un punto D cualquiera de una sección cónica, entonces PR será a PT como P*r* a P*t*. Y, por el contrario, si PR es a PT como P*r* a P*t* entonces BD, CD coinciden en un punto cualquiera D de una sección cónica.

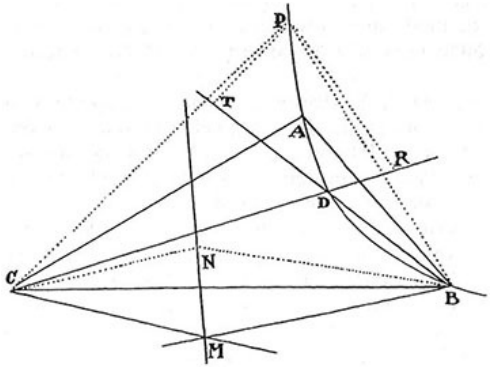
COROLARIO 3. Una sección cónica no corta a otra sección cónica más que en cuatro puntos. Pues, si posible fuera, pasen dos secciones cónicas por cinco puntos A, B, C, P, O, y córtelas la recta BD en los puntos D, *d*, mientras que la recta C*d* corta a PQ en *q*. En consecuencia, PR es a PT como P*q* a PT; de donde PR y P*q* son iguales entre sí, contra la hipótesis.

LEMA XXI

Si dos rectas BM, CM, móviles e infinitas trazadas por puntos dados B, C, a modo de polos, describen mediante el punto M en que se cruzan una recta MN de posición dada, y si otras dos rectas infinitas BD, CD, son trazadas de modo que constituyan

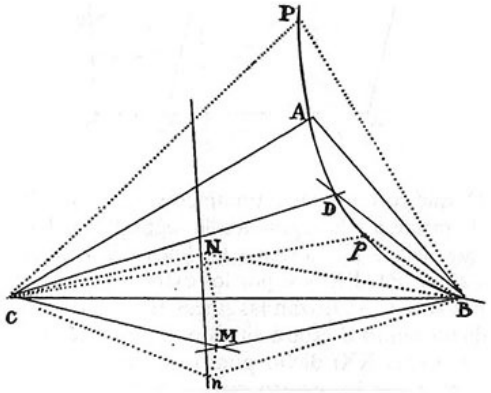
con las primeras en los puntos B, C, los ángulos dados MBD, MCD, digo estas dos BD, CD, mediante el punto en que se cruzan D, describen una sección cónica que pasa por los puntos B, C. Y viceversa, si las rectas BD, CD, mediante el punto en que se cruzan D describen una sección cónica que pasa por los puntos B, C, A, y el ángulo DBM es siempre igual al ángulo dado ABC y el ángulo DCM es siempre igual al ángulo dado ACB entonces el punto M cae en una recta de posición dada.

Pues trácese sobre la recta MN el punto N y cuando el punto móvil M coincide con el inmóvil N, coincida también el punto móvil D con el inmóvil P. Únanse CN, BN, CP, BP, y desde el punto P trácense las rectas PT, PR, que cortan a BD, CD, en T y R y que forman el ángulo BPT igual al ángulo dado BNM y el ángulo CPR igual también al ángulo dado CNM. Dado (por hipótesis) que los ángulos MBD, NBP son iguales así como los ángulos MCD, NCP, si se restan los ángulos comunes NBD y NCD, permanecerán iguales los ángulos NBM y PBT, NCM y PCR: Y por tanto los triángulos NBM, PBT son semejantes lo mismo que los triángulos NCM, PCR. Por lo cual PT es a NM como PB a NB, y PR a NM como PC a NC. Pero los puntos B, C, N, P son inmóviles. En consecuencia PT y PR están en razón dada a NM y por ende en razón dada entre sí; consecuentemente (por el Lema XX) el punto D, en el que constantemente concurren las rectas móviles BT y CR, coincide con los de la sección cónica que pasa por los puntos B, C, P. Q. E. D.



Y por el contrario, si el punto móvil D coincide con la sección cónica que pasa por los puntos B, C, A, y el ángulo DBM es siempre igual al ángulo dado ABC y el ángulo DCM también es siempre igual al ángulo dado ACB, y cuando el punto D coincide sucesivamente en dos puntos cualesquiera inmóviles *p*, P de la sección, el punto móvil M coincide también sucesivamente en dos puntos inmóviles n, N: trácese por dichos puntos *n*, N la recta *n*, N y ella será el lugar constante de dicho punto móvil M. Pues, si fuera posible, hágase correr al punto M por una curva. El punto D será tangente a una sección cónica que pase por cinco puntos B, C, A, *p*, P, mientras el punto M es tangente constantemente a una línea curva. Pero, según lo ya

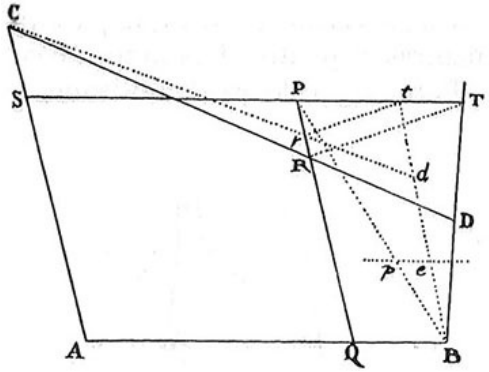
demostrado, el punto D toca también la sección cónica que pasa por los mismos cinco puntos B, C, A, *p*, P, mientras el punto M toca constantemente una línea recta. Luego las dos secciones cónicas pasan por los mismos cinco puntos, contra el Corolario 3 del Lema XX. Por tanto hacer al punto M recorrer una línea curva es absurdo. Q. E. D.



Proposición XXII. Problema XIV

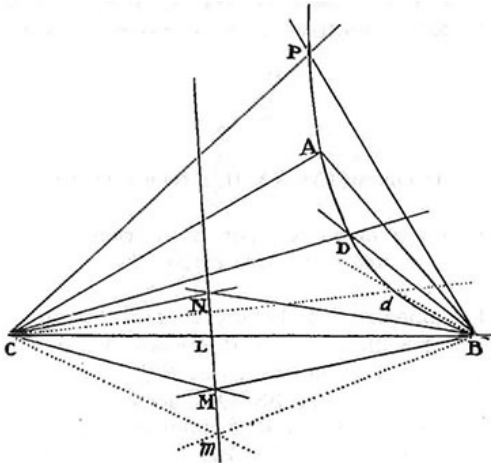
Trazar una trayectoria por cinco puntos dados.

Sean los cinco puntos A, B, C, P, D, dados. Desde uno de ellos cualquiera A hasta otros dos cualesquiera B, C, considerados como polos, tracemos las rectas AB, AC, así como las paralelas a estas TPS, PRQ, por el cuarto punto P. Tracemos después desde los dos polos B, C, por el quinto punto D, las dos rectas infinitas BDT, CRD, que corten a las últimamente trazadas TPS, PRQ (la primera a la primera y la segunda a la segunda) en T y R. Por fin, trazada *tr* paralela a TR, córtense las líneas P*t*, P*r*, tales que sean proporcionales a PT, PR; y si por los extremos *t, r*, de tales líneas y por los polos B, C, se trazan las líneas B*f*, C*r* concurrentes en *d*, entonces dicho punto *d* estará situado en la trayectoria buscada. Pues (por el Lema XX) dicho punto está situado en la sección cónica que pasa por los cuatro puntos A, B, C, P; y al tender a desaparecer las líneas R*r*, T*t*, tiende a coincidir el punto *d* con el punto D. Luego la sección cónica pasa por los cinco puntos A, B, C, P, D. Q. E. D.



Lo mismo de otra manera

Únanse tres de entre los puntos dados, A, B, C; y en torno a dos tomados como polos, B, C, háganse girar los ángulos ABC, ACB, de magnitud dada y aplíquense los lados BA, CA primero al punto D y después al punto P y señálense los puntos M, N en los que se cruzan entre sí los otros dos lados BL, CL. Trácese la recta infinita MN y háganse girar los mencionados ángulos móviles en torno a sus polos B, C, con la condición de que la intersección de los lados BL, CL o de BM, CM, supongámosla m, caiga siempre sobre la recta infinita MN; y la intersección de los lados BA, CA o BD, CD, supongámosla d, definirá la trayectoria buscada PAD*d*B. Pues (por el Lema XXI) el punto *d* coincidirá con la sección cónica que pase por los puntos B, C: y cuando el punto *m* llega a los puntos L, M, N, el punto *d* (por construcción) llegará a los puntos A, D, P. Describirá pues una sección cónica que pase por los cinco puntos A, B, C, P, D. Q. E. F.



Corolario 1. De aquí que, con facilidad, se pueda trazar una recta que toque a la trayectoria en un punto dado B cualquiera. Llévese el punto *d* hasta el punto B y la recta B*d* será la tangente buscada.

Corolario 2. De donde también pueden hallarse los centros de las trayectorias, los diámetros y los «latera recta», como en el Corolario 2 del Lema xix.

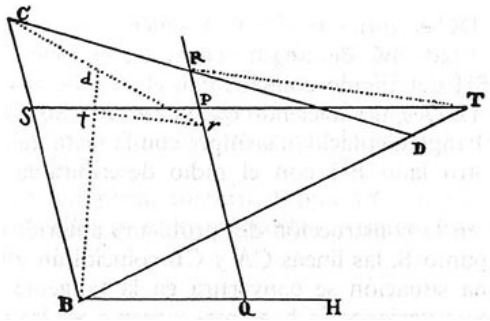
Escolio

La construcción primera resulta un poco más simple uniendo BP y tomando sobre ella, prolongada si es menester, B*p* a BP como PR es a PT; y trazando por *p* la recta infinita *pe* paralela a SPT, tomando además sobre ella a *pe* siempre igual a P*r* y haciendo que las rectas B*e*, C*r*, concurran en *d*. Pues dado que serán iguales las razones de P*r* a P*t*, PR a PT, *p*B a PB, *pe* a P*t*, también lo serán *pe* y P*r*. Con este método los puntos de la trayectoria se hallan muy fácilmente, salvo que se prefiera describir la curva mecánicamente, como en la segunda construcción.

Proposición xxiii. Problema xv

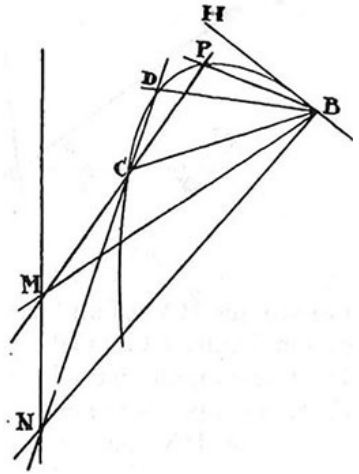
Trazar una trayectoria que pase por cuatro puntos y sea tangente a una recta dada en posición.

Caso 1. Supóngase dada la tangente HB, el punto de contacto B, y los otros tres puntos C, D, P. Únanse BC, procurando que PS sea paralela a la recta BH y que PQ lo sea a la recta BC; complétese el paralelogramo BSPQ. Trácese BD secante de SP en T, y CD secante de PQ en R. Finalmente, trazada la línea *tr* paralela a TR, tomemos sobre PQ, PS los segmentos P*r*, P*t*, respectivamente proporcionales a PR, PT; y trazadas las líneas C*r*, B*t*, el punto *d* en que se cortan (por el Lema xx) caerá siempre en la trayectoria que hay que describir.



De otro modo

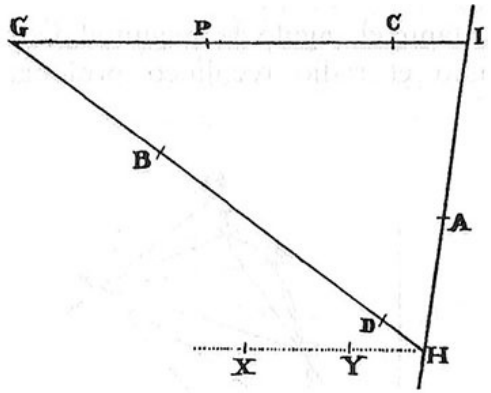
Hágase girar tanto el ángulo de magnitud dada CBH en torno al polo B, como el radio rectilíneo prolongado en ambas direcciones DC en torno al polo C. Señálense los puntos M, N en los cuales el lado BC del ángulo corta a dicho radio cuando el otro lado BH del ángulo coincide con el susodicho radio en los puntos P y D. Después haciendo que el radio dicho CP o CD y el lado BC del ángulo coincidan siempre con la recta infinita MN, el cruce del otro lado BH con el radio describirá la trayectoria buscada.



Pues si en la construcción del problema anterior el punto A alcanza el punto B, las líneas CA y CB coincidirán y la línea AB en su última situación se convertirá en la tangente BH; y por tanto las construcciones hechas antes vienen a ser las mismas que las hechas ahora. En consecuencia, la concurrencia del lado BH con el radio describirá la sección cónica por los puntos C, D, P y la recta BH será tangente en el punto B. Q. E. F.

Caso 2. Sean los cuatro puntos dados B, C, D, P situados fuera de la tangente HI. Únanse dos a dos con las líneas BD, CP concurrentes en G y que cortan a la tangente en H e I. Córtese la tangente en A de tal modo que HA sea a IA como es el rectángulo bajo la media proporcional entre CG y GP y la media proporcional entre BH y HD al rectángulo bajo la media proporcional entre DG y GB y la media proporcional entre PI e IC; y será A el punto de contacto. Pues si HX, paralela a PI, cortase a la trayectoria en cualesquiera puntos X e Y, habrá que colocar (por las cónicas) al punto A de tal modo que HA^2 sería a AI^2 en razón compuesta de la razón del rectángulo XHY al rectángulo BHD, o del rectángulo CGP al rectángulo DGB, y de la razón del rectángulo BHD al rectángulo PIC. Así pues, una vez hallado el punto de contacto A,

se describirá la trayectoria como en el caso primero. Q. E. F.



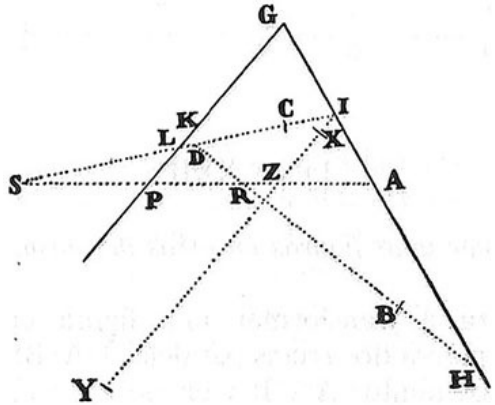
Sin embargo, puede tomarse el punto A entre los puntos H e Y o fuera de ellos; y, en consecuencia, describir por duplicado la trayectoria.

Proposición xxiv. Problema xvi

Trazar una trayectoria que pase por tres puntos dados y sea tangente a dos rectas de posición dada.

Sean las tangentes dadas HI, KL y los puntos B, C, D. Trácese por dos cualesquiera puntos B, D la recta infinita BD que cruza a las tangentes en los puntos H, K. Después por otros dos puntos cualesquiera C, D, trácese la recta infinita CD que cruce a las tangentes en los puntos I, L. Córtese a las rectas así trazadas en R y S de tal modo que sea HR a KR como es la media proporcional entre BH y HD a la media proporcional entre BK y KD, e IS a LS como la media proporcional entre CI e ID a la media proporcional entre CL y LD. Pero córtese por un punto arbitrario, ya entre les puntos K y H, I y L, ya fuera de ellos; trácese RS secante de las tangentes en A y P y serán A y P los puntos de contacto. Pues si se supone que A y P fuesen puntos de contacto situados en otros pintos de las tangentes y por uno de los puntos H, I, K, L, por ejemplo, I, situado en otra tangente HI, se traza la recta IY paralela a la otra tangente KL y que corte a la curva en X e Y, en la cual recta se toma además IZ como media proporcional entre IX e IY el rectángulo XIY, o IZ^2 (por las cónicas), será a LP^2 como el rectángulo CID al rectángulo CLD, esto es (por constricción) como SI^2 a SL^2 y por tanto IZ a LP como SI a SL. Los puntos S, P, Z, pertenecen, por tanto, a una misma recta. Ahora bien, al concurrir las tangentes en G, ocurrirá por las cónicas que el rectángulo XIY o IZ^2 será a IA^2 como GP^2 a GA^2 y por Unto IZ a IA como GP a GA. Pertenecen pues los puntos P, Z, y A a la misma recta y por tanto los puntos S,

P y A están en la misma recta. Y con un argumento semejante se probará que los puntos R, P y A están en la misma recta. Pertenecen, pues, los puntos de contacto A y P a la recta RS. Y una vez hallados éstos, la trayectoria se trazará como en el caso primero del problema anterior Q. E. F.



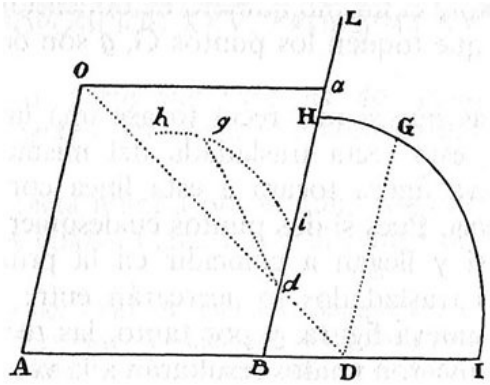
En esta proposición, y en el caso segundo de la anterior, las constricciones son las mismas, tanto si la recta XY corta a la trayectoria en X e Y como si no la corta; tales construcciones no dependan de dicha sección. Sino que, una vez demostradas las constricciones cuando la recta corta a la trayectoria, quedan también patentes las construcciones cuando no la corta; y por amor a la brevedad no me detengo en demostrarlas más profusamente.

Lema xxii

Transformar unas figuras en otras del mismo género.

Sea la figura a transformar una figura cualquiera HGI. Trácense según plazca dos rectas paralelas OA, BL secantes a una tercera AB en los puntos A y B y en posición dada, y desde un punto cualquiera G de la figura trácese una línea GD, paralela a la misma OA. Después desde un punto O, dado en la línea OA, trácese hasta el punto D la recta OD que corte a la recta BL en *d*, y desde el punto de cruce elévese la recta *dg* de modo que contenga con la recta BL un ángulo dado, y además tal que tenga respecto a O*d* la misma razón que DG tiene respecto a OD; será *g* entonces en la nueva figura *hgi* el punto correspondiente a G. Del mismo modo cada punto de la primera figura vendrá a dar cada punto de la nueva figura. Pues imagínese que el punto G recorre con un movimiento continuo todos los puntos de la primera figura, y que el punto *g* también con movimiento continuo recorre todos los puntos de la nueva figura describiéndola de paso. Para distinguirlas llamemos DG a la primera ordenada,

dg a la ordenada nueva; llamemos AD a la abscisa primera y *ad* a la abscisa nueva; llamemos O al polo y OD al radio abscindente, OA al radio ordenado primero y O*a* (con el que se cierra el paralelogramo OAB*a*) al radio ordenado nuevo.



Digo ahora que si el punto G toca a la línea recta en posición dada, el punto *g* tocará también a la línea recta de posición dada. Si el punto G toca la sección cónica, el punto *g* tocará también la sección cónica. Incluyo aquí el círculo entre las secciones cónicas. Pues si el punto G toca una línea de tercer orden analítico, el punto *g* tocará una línea del mismo orden; y lo mismo para curvas de orden superior. Dos líneas que sean tocadas por los puntos G, *g* serán siempre del mismo orden analítico. Puesto que *ad* es a OA como son O*d* a OD, *dg* a DG y AB a AD; por tanto, AD es $\frac{\text{OA x AB}}{ad}$, y DG es igual a $\frac{\text{OA x } dg}{ad}$. Ahora si el punto G toca una línea recta y por tanto en la ecuación que contenga la relación entre la abscisa AD y la ordenada DG, dichas líneas indeterminadas AD y DG no alcanzan más que una única dimensión, escribiendo en esta ecuación $\frac{\text{OA x AB}}{ad}$ en lugar de AD, y $\frac{\text{OA x } dg}{ad}$ en lugar de DG, se obtendrá una nueva ecuación en la que la abscisa nueva *ad* y la nueva ordenada *dg* no alcanzarán más que una sola dimensión, y precisamente la que da por resultado una línea recta. Pero si AD y DG, o una de ellas, alcanzase en la primera ecuación una segunda dimensión, también *ad* y *dg* alcanzarán segundas dimensiones en la segunda ecuación. Y lo mismo para tres o más dimensiones. Las indeterminadas *ad*, *dg* en la segunda ecuación y AD, DG en la primera alcanzarán siempre el mismo número de dimensiones, y por tanto las líneas a las que toquen los puntos G, *g* son del mismo orden analítico.

Digo además que si una recta tocase una línea curva en la primera figura; esta

recta trasladada del mismo modo con la curva a la nueva figura tocará a esta línea curva en la nueva figura; y viceversa. Pues si dos puntos cualesquiera de la curva se acercan entre sí y llegan a coincidir en la primera figura, los mismos puntos trasladados se acercarán entre sí y llegarán a coincidir en la nueva figura; y por tanto, las rectas con las que estos puntos estuvieron unidos resultarán a la vez tangentes de las curvas en ambas figuras.

Pueden construirse demostraciones de estas afirmaciones de modo más geométrico. Pero prefiero la brevedad.

Por tanto, si se ha de transformar una figura rectilínea en otra bastará trasladar las intersecciones de las rectas que constituyen la primera figura y por ellas trazar las rectas de la nueva. Si se ha de tratar, en cambio, de transformar una figura curvilínea, hay que trasladar los puntos, las tangentes y demás rectas en cuya virtud se define la curva. Sirve así este lema para resolver los problemas más difíciles, trasformando las figuras propuestas en otras más sencillas. Por tanto, cualesquiera rectas convergentes pueden trasformarse en paralelas si se toma como primer radio ordenado una línea recta cualquiera que pasa por el punto de convergencia; y esto porque dicho punto de convergencia cae de este modo en el infinito; y líneas paralelas son aquellas que nunca se encuentran. Pero, después que se resuelve el problema en la nueva figura, si realizando operaciones inversas, esta figura se trasforma en la primitiva, se tendrá la solución buscada.

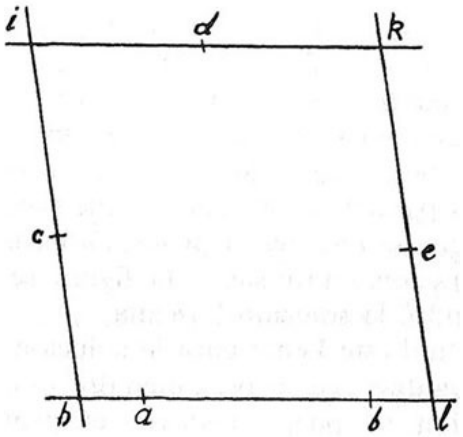
También es útil este Lema para la solución de problemas de sólidos. Pues cuantas veces se presenten dos secciones cónicas con cuya intersección se pueda resolver el problema, se puede transformar cualquiera de ellas, si es una hipérbola o una parábola en una elipse: después la elipse se transforma fácilmente en un círculo. También una recta y una sección cónica, en la construcción de problemas de planos, se transforman en una recta y un círculo[26].

Proposición xxv. Problema xvii

Describir una trayectoria que pase por dos puntos dados y sea tangente a tres rectas de posición dada.

Trácese una recta infinita por el punto de intersección mutua de dos tangentes cualesquiera y por el punto de intersección de una tercera tangente con la recta que pase por los dos puntos dados; tomando como primer radio ordenado la recta así trazada, transfórmese la figura, por el lema anterior, en una nueva figura. En esta nueva figura las dos tangentes aquéllas resultarán entre sí paralelas, y la tangente tercera vendrá a ser paralela a la recta que pase por los dos puntos dados. Sean *hi*, *kl* dichas tangentes paralelas, *ik* la tercera tangente y *hl* la recta paralela a esta tercera tangente y que pasa por los puntos *ab*, por los que en esta nueva figura debe pasar la

sección cónica, y que completa el paralelogramo *hikl*. Córtense las rectas *hi, ik, kl* en *c, d, e*, de modo que *he* sea al lado cuadrado del rectángulo *ahb*, que *ic* sea a *id*, y *ke* a *kd* como la suma de las rectas *hi* y *kl* es a la suma de las tres líneas de las cuales la primera es la recta *ik*, y las otras dos son los lados cuadrados de los rectángulos *ahb y alb*; y serán *c, d, e* los puntos de contacto. Pues, por las cónicas, hc^2 es al rectángulo *ahb* como ic^2 es a id^2 y como ke^2 es a kd^2 y como el^2 es al rectángulo *alb*; y por tanto están en dicha razón subduplicada *he* al lado cuadrado del rectángulo *ahb, ic* a *id, ke* a *kd* y *el* al lado cuadrado del rectángulo *alb*, y sumando estarán en la razón dada todos los antecedentes *hi* y *kl* respecto a todos los consecuentes que son el lado cuadrado del rectángulo *alb*[27]. Se tienen, pues, a partir de aquella razón dada los puntos de contacto *c, d, e*, en la nueva figura. Por las operaciones inversas del último Lema trasládense estos puntos sobre la primera figura, y en ella (por el Prob. XIV) se trazará la trayectoria. Q. E. F. Por lo demás, como los puntos *a, b*, bien caerán entre los puntos *h, l*, bien fuera, los puntos *c, d, e*, deberán caer o entre los puntos *h, i, k, l*, o fuera. Si uno de los puntos *a, b*, cae entre los puntos *h, l*, y el otro cae fuera, el problema es imposible.

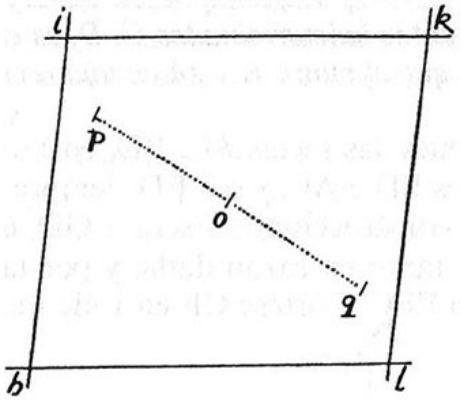


Proposición XXVI. Problema XVIII

Trazar una trayectoria que pase por un punto dado y sea tangente a cuatro rectas de posición dada.

Trácese desde la intersección de dos tangentes cualesquiera una recta infinita hasta la intersección de las otras dos tangentes; y por tanto también están en dicha razón subduplicada tomando dicha recta por primer radio ordenado, trasfórmese la figura (por el Lema XXII) en una nueva figura, y los dos pares de tangentes, que

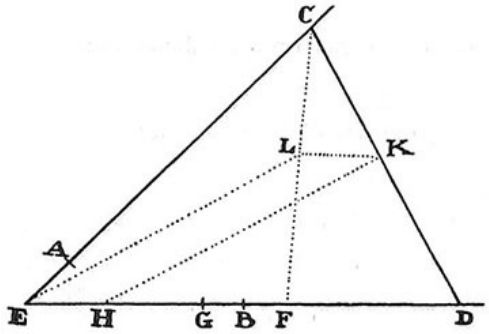
concurrían en el primer radio ordenado, resultarán paralelas. Sean éstas *hi* y *kl, ik* y *hl* que completan el paralelogramo *hikl*. Sea *p* el punto correspondiente en esta nueva figura al punto dado en la primera. Trácese por el centro O de la figura la línea *pq*, y siendo O*q* igual a O*p* será *q* el otro punto por el que pasará la sección cónica en esta nueva figura. Por la operación inversa del Lema XXII trasládese este punto a la figura primera y se tendrán en ella dos puntos por los que debe trazarse la trayectoria. Por ellos, en efecto, por el Problema XVII, puede trazarse dicha trayectoria. Q. E. F.



Lema XXIII

Si dos rectas de posición dada AC, BD terminan en dos puntos dados A, B, y están entre sí en una razón dada y la recta CD con la que se unen los puntos indeterminados C, D, es cortada en K en una razón dada: digo que el punto K está en una recta de posición dada.

Concurran, pues, las rectas AC, BD, en E, y sobre BE tómese BG a AE como es BD a AC, y sea FD siempre igual a la distancia dada EG; y por construcción EC será a GD, esto es, a EF como AC a BD, y por tanto en razón dada, y por tanto estará dado el tipo de triángulo EFC. Córtese CF en L de modo que CL esté en razón a CF como CK a CD; y por la razón dada anterior, se dará también el tipo de triángulo EFL; y por tanto, el punto L estará situado en la recta EL de posición dada. Únase LK, y los triángulos CLK, CFD serán semejantes; y por estar dadas FD y la razón de LK a FD, estará dada también LK. Igual a ésta tómese EH y ELKH será siempre un paralelogramo. Se halla pues el punto K en el lado de posición dada HK de dicho paralelogramo. Q. E. D.

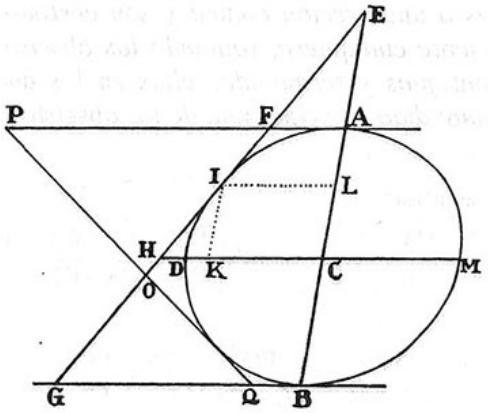


COROLARIO. Por estar la figura EFLC dada en especie, las tres rectas EF, EL y EC, esto es, GD, HK y EC están entre sí en una razón dada.

LEMA XXIV

Si tres rectas fuesen tangentes a cualquier sección cónica de las cuales dos fuesen paralelas y de posición dada, digo que el semidiámetro de la sección paralelo a estas dos será media proporcional entre los segmentos de las mismas comprendidos entre los puntos de contacto y la tercera tangente.

Sean AF, GB dos paralelas tangentes a la sección cónica ADB en los puntos A y B; sea EF una tercera recta tangente a la sección cónica en I, y que se corte con las dos tangentes anteriores en F y G; y sea CD el semidiámetro de la figura paralelo a las tangentes: digo que AF, CD, BG, están en proporción continua.



Pues si los diámetros conjugados AB, DM cortasen a la tangente FG en E y H y

se cortasen mutuamente en C y se completase el paralelogramo IKCL; ocurrirá por la naturaleza de las cónicas que EC será a CA como CA a CL, y también como las diferencias EC - CA a CA - CL, esto es como EA a AL y sumando, EA a EA + AL, o sea EL como EC a EC + CA, o sea EB; y por tanto, por la semejanza de los triángulos EAF, ELI, ECH, EBG, AF es a LI como CH es a BG. Al igual que, por la naturaleza de las cónicas, LI o CK es a CD como CD es a CH; y por tanto permutadas adecuadamente AF es a CD como CD es a BG. Q. E. D[28].

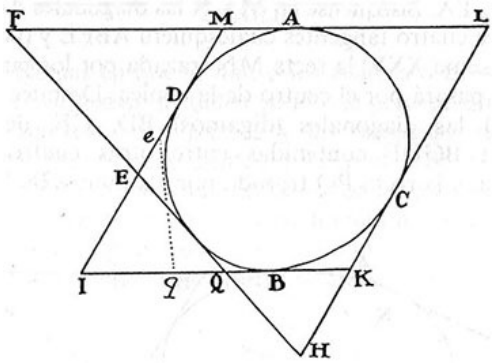
COROLARIO 1. De aquí que si dos tangentes FG, PQ cortasen a las tangentes paralelas AF, BG en F y G, P y Q y se cortasen entre sí en O; permutadas adecuadamente ocurriría que AF es a BQ como AP es a BG, y por separación como FP a GQ, y por tanto como FO a OG.

COROLARIO 2. De donde también dos rectas PG, FQ trazadas por los puntos P y G, F y Q concurren en la recta ACB que pase por el centro de la figura y por los puntos de contacto A, B.

LEMA XXV

Si los cuatro lados de un paralelogramo, prolongados infinitamente, fuesen tangentes a una sección cónica y son cortados a su vez por una quinta tangente cualquiera; tomando las abscisas de dos lados cualesquiera contiguos y terminadas ellas en los ángulos opuestos del paralelogramo: digo que cada una de las abscisas será al lado del que es abscisa como la parte del otro lado contiguo comprendida entre el punto de contacto y el tercer lado es a la otra parte de las abscisas.

Sean los cuatro lados ML, IK, KL, MI del paralelogramo MLIK tangentes a la sección cónica en A, B, C, D, y corte la quinta tangente FQ a dichos lados en F, Q, H y E; tómese sobre los lados MI, KI, las abscisas ME, KQ, o sobre los lados KL, ML, las abscisas KH, MF: digo que ME será a MI como BK a KQ; y KH a KL como AM a MF. Pues, por el corolario primero del Lema anterior ME es a El como AM o BK a BQ, y sumando ME a MI como BK a KQ. Q. E. D. Y también KH a HL como BK o AM a AF, y dividiendo KH a KL como AM a MF. Q. E. D.



COROLARIO 1. De aquí que si se tiene dado el paralelogramo IKLM, descrito en torno a una sección cónica dada, se tendrá dado también el rectángulo KQ x ME, al igual que el rectángulo igual a este KH x MF. Pues dichos rectángulos son iguales por la semejanza de los triángulos KQH, MFE.

COROLARIO 2. Y si se traza una sexta tangente *eq* que incida sobre las tangentes KI, MI en *q* y *e*; el rectángulo KQ x ME será igual al rectángulo K*q* x M*e*; y KQ será a Me como K*q* a ME y sustrayendo como Q*q* a E*e*.

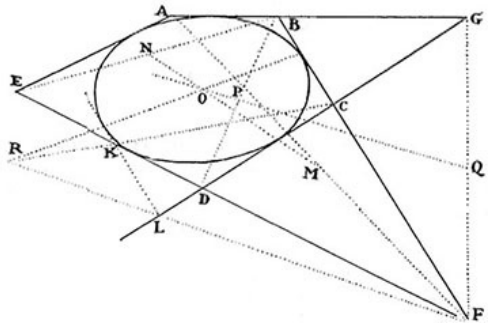
COROLARIO 3. De donde también, si se unen y se bisecan E*q* y *e*Q y se traza una recta por los puntos de las bisecciones, pasará esta recta por el centro de la sección cónica. Puesto que al ser Q*q* a E*e* como KQ a M*e* dicha recta pasará por la mitad de todas las rectas E*q*, *e*Q, M*k* (por el Lema XXIII) y el medio de la recta MK es el centro de la sección.

PROPOSICIÓN XXVII. PROBLEMA XIX

Describir una trayectoria que sea tangente a cinco rectas de posición dada.

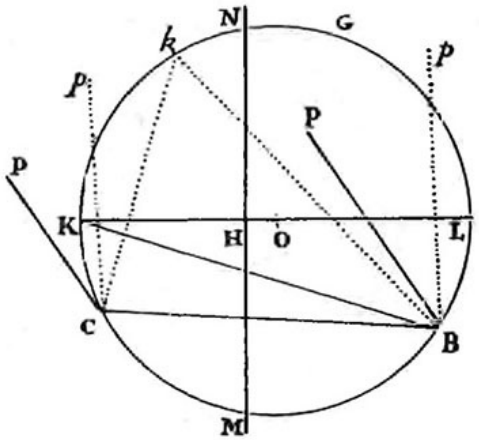
Supónganse dadas en posición las tangentes ABG, BCF, GCD, FDE, EA. Biséquense en M y N las diagonales AF, BE, contenidas entre cuatro tangentes cualesquiera ABFE y (por el Corolario 3 del Lema XXV) la recta MN trazada por los puntos de las bisecciones pasará por el centro de la cónica. Después biséquense en P y Q las diagonales (digamos) BD, GF, de la figura cuadrilátera BGDF contenidas entre otras cuatro tangentes cualesquiera, y la recta PQ trazada por los puntos de la bisección pasará por el centro de la trayectoria cónica. Se tendrá pues el centro en el punto de concurso de las bisecantes. Sea O dicho punto. Trácese a cualquiera tangente BC una paralela KL, tal que esté situada a la distancia precisa para que el centro O quede en el punto medio entre las paralelas, y trazada KL será tangente a la trayectoria a describir. Por los puntos de contacto C y K, F y L de estas tangentes no

paralelas CL, FK, con las paralelas CF, KL, trácense CK, FL, concurrentes en R, y la recta OR trazada y prolongada cortará a las tangentes paralelas CF, KL, en los puntos de contacto. Esto es evidente por el Corolario 2 del Lema XXIV. Del mismo modo es posible hallar los demás puntos de contacto y finalmente, por la construcción del prob. XIV, describir la trayectoria. Q. E. F.



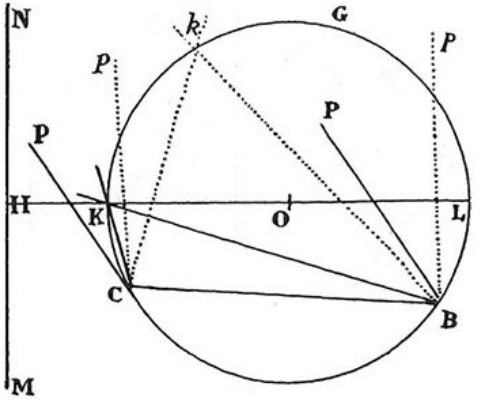
ESCOLIO

Los problemas en que se dan o los centros o las asíntotas de las trayectorias quedan incluidos en lo que precede. Pues dados a la vez los centros y los puntos y las tangentes, están dados también todos los demás puntos y tangentes de la otra parte equidistantes del centro. Una asíntota debe ser considerada como si fuera una tangente y su término infinitamente distante (por decirlo así) como el punto de contacto. Imagínese que el punto de contacto de una tangente cualquiera se alejase hasta el infinito, en tal caso la tangente se transforma en asíntota y las construcciones de los problemas precedentes se transforman en construcciones en las que se da una asíntota.



Una vez descrita la trayectoria es posible hallar los ejes y los focos de la misma por este método. En la construcción y figura del Lema XXI hágase que los lados BP, CP, de los ángulos móviles PBN, PCN, con cuya intersección se describía la trayectoria, sean entre sí paralelos y, conservando su misma posición, giren sobre sus polos B, C, en dicha figura. Entre tanto, los otros lados CN, BN, de dichos ángulos describirán con su intersección en K o *k* el círculo BGKC. Sea O el centro de este círculo. Desde este centro trácese sobre la regla MN, en la que concurrían los lados CN, BN, mientras se trazaba la trayectoria cónica, la normal OH que corta el círculo en K y L. Y cuando los otros lados CK, BK, concurren en el punto susodicho K que está más cerca de la regla, los lados primeros CP, BP, serán paralelos al eje mayor, y perpendiculares al menor; y lo contrario ocurrirá si los mismos lados concurren en el punto más distante L. De donde si se da el centro de la trayectoria se darán también los ejes. Y dados éstos, los focos están a un paso.

Pero los cuadrados de los ejes están entre sí como KH a LH, y de aquí es fácil describir una cónica de una especie dada por cuatro puntos dados. Pues si dos de los puntos dados se toman como los polos C, B, el tercero dará los ángulos móviles PCK, PBK; dados éstos se puede trazar el círculo BGKC. Entonces, por estar dada la especie de trayectoria cónica, se tendrá la razón OH a OK y por tanto la propia OH. Con centro en O y con distancia OH trácese otro círculo y la recta tangente a este círculo que pasa por el cruce de los lados CK, BK, cuando los lados primeros CP, BP, concurren en el cuarto punto dado, será la regla aquella MN con cuya ayuda se traza la trayectoria. De donde también a veces un trapecio de especie dada (si se exceptúan algunos casos imposibles) puede inscribirse en cualquiera sección cónica dada.

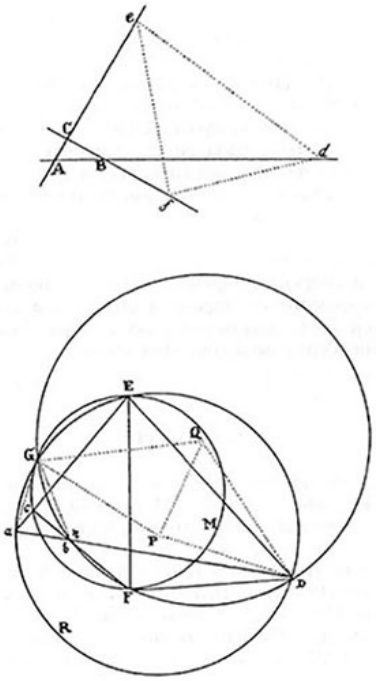


Hay otros Lemas en cuya virtud pueden trazarse trayectorias de una especie dada, si se dan los puntos y las tangentes. De este género es el de que si se traza una línea recta por un punto de posición dada que corte a una sección cónica dada en dos puntos y se biseca el intervalo comprendido entre las intersecciones, el punto de la bisección es tangente a otra sección cónica de la misma especie que la anterior y cuyos ejes son paralelos a los ejes de la anterior. Pero paso a cosas más útiles.

LEMA XXVI

Situar uno a uno los tres ángulos de un triángulo de especie y magnitud dadas sobre tres rectas dadas también en posición y que no sean entre sí paralelas ninguna.

Se tienen en posición tres rectas infinitas AB, AC, BC y es preciso colocar el triángulo DEF de tal manera que su ángulo D toque la línea AB, el ángulo E toque la línea AC y el ángulo F la línea BC. Sobre DE, DF y EF trácense tres segmentos circulares DRE, DGF, EMF, que comprendan ángulos respectivamente iguales a los ángulos BAC, ABC, ACB. Pero se han de trazar estos segmentos sobre aquellas partes de las líneas DE, DF, EF, de tal modo que al girar las letras DRFD lo hagan en el mismo orden que las letras BACB y que las letras DGFE giren en el mismo orden que las letras ACBA; después complétense estos segmentos circulares en sus círculos completos. Sea G el punto en que se corten los dos primeros círculos y sean sus centros P y Q. Unidos GP, PQ, tómese G*a* a AB como GP a PQ y, con centro en G y con distancia G*a* trácese un círculo que corte al círculo primero DGE en *a*. Únase por una parte *a*D secante del círculo segundo DFG en *b* y por otra *a*E secante del círculo tercero EMF en *c*. Y ya es factible la figura ABC*def* semejante e igual a la figura *abc*DEF. Con lo cual se completa el problema.



Trácese, pues, F*c* que toque a *a*D en *n*, y únanse *a*G, *b*G, QG, QD, PD. Por construcción, el ángulo E*a*D es igual al ángulo CAB, y el ángulo *ac*F igual al ángulo ACB, y por tanto el triángulo *anc* equiángulo con el triángulo ABC. Luego el ángulo *anc* o F*n*D, es igual al ángulo ABC y por tanto al ángulo F*b*D, y por eso el punto *n* coincide con el punto *b*. Además, el ángulo GPQ, que es la mitad del ángulo central GPD, es igual al ángulo ni la circunferencia G*a*D, y el ángulo GQP, que es la mitad del ángulo central GQD, es igual al complemento para dos rectos del ángulo en la circunferencia G*b*D y por tanto igual al ángulo G*ba*; v así los ángulos GPQ, G*ab* son semejantes; y G*a* es respecto a *ab* como GP es a PQ; esto es (por construcción) como G*a* a AB. Pero son iguales *ab* y AB y por tanto los triángulos *abc* y ABC, que acabamos de probar que son semejantes, son también iguales. En consecuencia, puesto que los ángulos D, E, F, del triángulo DEF caen respectivamente sobre los lados *ab*, *ac*, *bc*, del triángulo *abc*, la figura ABC*def* puede ser completada semejante e igual a la figura *abc*DEF, y construyéndola quedará resuelto el problema. Q. E. F.

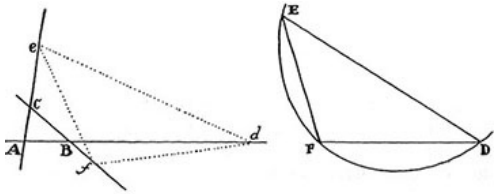
COROLARIO. Ahora puede trazarse una recta cuyas partes de longitud dada queden interceptadas entre tres rectas de posición duda. Supongamos el triángulo DEF, cuyo punto D se acerca al ludo EF, y cuyos lados DE, DF, están colocados sobre una recta, convirtiéndose en una línea recta cuya parte dada DE estará interceptada por las rectas de posición dada AB, AC y la otra parte dada DF estará interceptada por las rectas de posición dada AB, BC; y aplicando la construcción precedente a este caso

quedará el problema resuelto.

PROPOSICIÓN XXVIII. PROBLEMA XX

Describir una trayectoria dada en cuanto a forma y magnitud cuyas partes dadas queden interceptadas por tres rectas de posición dada.

La trayectoria que se ha de describir habrá de ser semejante e igual a la curva DEF y habrá de estar cortada por tres rectas AB, AC, BC, de posición dada en partes semejantes e iguales a las partes dadas de esta DE y EF.



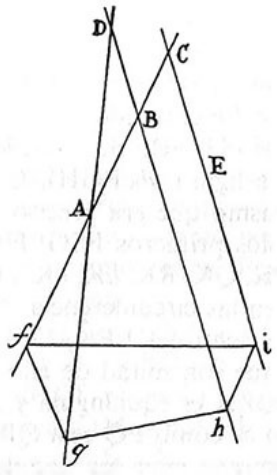
Tracemos las rectas DE, EF, DF y hagamos que los ángulo D, E, F, de este triángulo DEF caigan sobre las rectas de posición dada (por el Lema XXVI) y después tracemos la trayectoria en torno al triángulo semejante e igual a la curva DEF. Q. E. F.

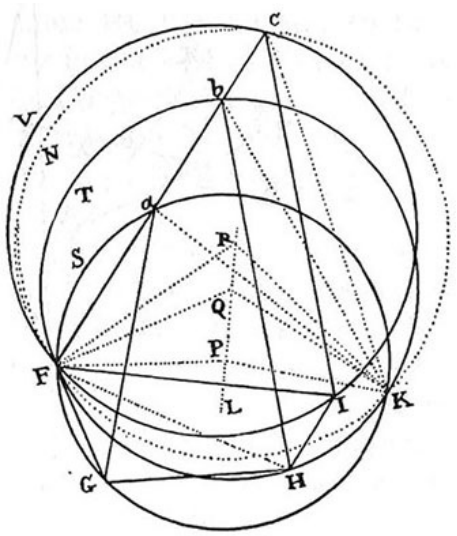
LEMA XXVII

Describir un trapecio de forma dada cuyos ángulos toquen cada uno a cada una de cuatro rectas dadas en posición tales que ni toda sean paralelas ni converjan en un punto común.

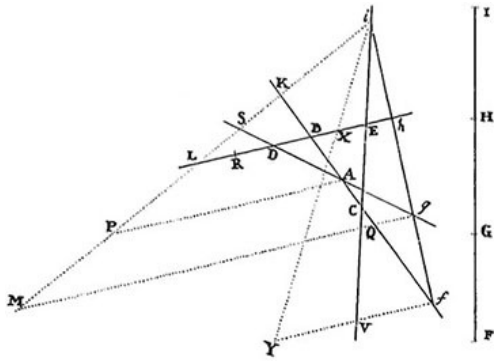
Sean las cuatro rectas dadas en posición ABC, AD, BD, CE de las cuales, la primera corta a la segunda en A, a la tercera en I y a la cuarta en C; y hay que describir un trapecio *fghi*, que sea semejante al trapecio FGHI; y cuyo ángulo *f* igual al ángulo dado F, toque a la recta ABC y el resto de los ángulos *g*, *h*, *i*, iguales a los ángulos también dados G, H, I, toquen a las otra líneas AD, BD, CE, respectivamente. Únase FH y sobre FG, FH FI, trácense otros tantos segmentos circulares FSG, FTH, FVI, de los cuales el primero FSG comprenda un ángulo igual al ángulo BAD, el segundo FTH comprenda un ángulo igual al ángulo CBD y el tercero FVI comprenda un ángulo igual al ángulo ACE Pero deben trazarse los segmentos hacia aquellas partes de la: líneas FG, FH, FI, de modo tal que el orden de las letras FSGI sea el mismo orden circular que el de las letras BADB y que e orden

de las letras FTHF sea el mismo que el de las letras CBDC y que el de las letras FVIF sea el mismo en su giro que el de las letras ACEA. Complétense los segmentos en círculos completos y sea P el centro del primer círculo FSG, y Q el centro del segundo ITH. Únanse y prolónguese PQ y tómese sobre ella QR en la misma proporción a PQ que tiene BC a AB. Pero tómese QR hacia el lado del punto Q tal que el orden de las letras P, Q, R, sea el mismo que el de las letras A, B, C; y con centro en R y con distancia RF trácese un cuarto círculo FN*c* que corte al círculo tercero FVI en *c*. Únase F*e* cortando al círculo primero en *a* y al segundo en *b*. Trácense *a*G, *b*H, *c*I, y puede obtenerse la figura ABC*fghi* semejante a la figura *abc*FGHI. Una vez hecho esto el trapecio *fghi* será el mismo que era preciso construir.





Pues, los dos círculos primeros FSG, FTH, córtense mutuamente en K. Únanse PK, QK, RK, *a*K, *b*K, *c*K, y prolónguese QP hasta L. Los ángulos en las circunferencias F*a*K, F*b*K, F*d*C, son mitades de los ángulos centrales FPK, FQK, FRK, y por tanto iguales a los ángulos que son mitad de ellos, LPK, LQK, LRK. Por tanto, la figura PQRK es equiángula y semejante a la figura *abc*K y por ello *ab* es a *bc* como PQ es a QR, esto es, como AB a BC. Y además por construcción los ángulos *f*A*g*, *f*B*h*, *f*C*i* son iguales a los ángulos F*a*G, F*b*H, F*el*. Luego puede construirse la figura ABC*fghi* semejante a la figura *abc*FGHl. Hecho esto se tendrá el trapecio *fghi*, semejante al trapecio FGHI, y con sus ángulos *f*, *g*, *h*, *i*, tocará a las rectas ABC, AD, BD, CE. Q. E. F.



COROLARIO. Ahora se puede trazar una recta cuyas partes, interceptadas por cuatro rectas dadas en posición según un cierto orden, tendrán entre sí una proporción

dada. Auméntense los ángulos FGH, GHI, hasta el punto que las rectas FG, GH, HI, se hallen sobre una misma recta y para construir el problema en el presente caso se trazará la recta *fghi* cuyas partes *fg*, *gh*, *hi*, interceptadas por las cuatro rectas de posición dada AB y AD, AD y BD, BD y CE, serán entre sí como las líneas FG, GH, HI, y mantendrán entre sí el mismo orden. Pero esto mismo se halla más directamente como sigue.

Prolónguese AB hasta K y BD hasta L, de modo que BK sea u AB como HI a GH, y DL a BD como GI a FG, y únanse KL con la prolongación de CE pasando por *i*. Prolónguese *i*L hasta M de modo que LM sea a *i*M como GH es a HI y trácese ahora MQ paralela a LB y que corte a la recta AD en *g* así como *gi* que corte a AB, BD en *f*, *h*. Y digo que está hecho.

Pues corte M*g* a la recta AB en Q, y AD a la recta KL en S, y trácese AP que sea paralela a BD y toque a *i*L en P, y estarán en idéntica razón *g*M a L*h* (*gi* a *hi*, M*i* a L*i*, GI a HI, AK a BK) y AP a BL. Córtese DL en R de modo que DL sea a RL en la misma dicha razón; y por ser proporcionales *g*S a *g*M, AS a AP y DS a DL, por lo mismo, como *g*S es a L*h* será AS a BL y DS a RL; y conjuntamente BL - RL a L*h* - BL como AS - DS a *g*S - AS. Esto es, BR a B*h* como AD a A*g*, y por tanto, como BD a *g*Q. Y alternativamente BR a BD como *Bh* a *g*Q, o sea *fh* a *fg*. Pero por construcción la línea BL fue cortada en razón idéntica en D y en R a como lo fue la línea FI en G y en H: por ello BR es a BD como FH es a FG. Luego *fh* es a *fg* como FH es a FG. Pero como también sea *gi* a *hi* como M*i* a L*i*, esto es, como GI a HI, es evidente que las líneas FI, *fi* están cortadas de modo igual en G, H y en *g*, *h*. Q. E. F.

En la construcción de este Corolario, después de trazar LK secante de CE en *i*, se puede prolongar *i*E hasta V de modo que EV sea a E*i* como FH es a HI, y trazar V*f* paralela a BD. Lo mismo ocurre si con centro en i y distancia IH se describe un círculo que corte a BD en X y se prolonga *i*X hasta Y de modo que *i*Y sea igual a IF, y se traza Y*f* paralela a BD.

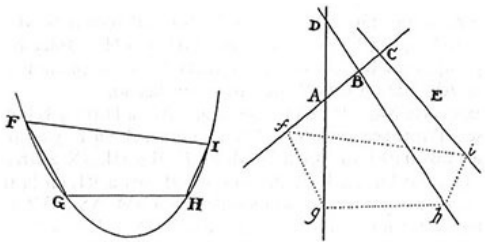
De este problema ya dieron otras soluciones *Wren* y *Wallis* hace tiempo.

Proposición xxix. Problema xxi

Describir una trayectoria de forma dada que esté cortada en partes dadas en orden, forma y proporción por cuatro rectas dadas en posición.

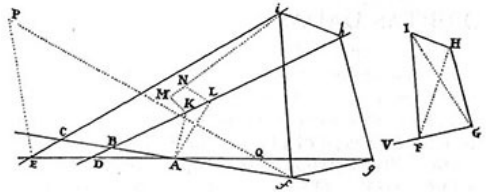
Ha de trazarse una trayectoria que sea semejante a la línea curva FGHI y cuyas partes, semejantes y proporcionales a las partes de aquella FG, GH, HI, estén interceptadas entre las rectas AB y AD, AD y BD, BD y CE dadas en posición, la primera entre las primeras, la segunda entre las segundas, la tercera entre las terceras. Trazadas las rectas FG, GH, HI, FI, trácese (por el Lema xxvii) el trapecio *fghi*, semejante al trapecio FGHI, y tal que sus ángulos *f*, *g*, *h*, *i*, toquen las rectas de

posición dada AB, AD, BD, CE, cada uno a cada una en el orden dicho. Después trácese la trayectoria en torno a este trapecio y será semejante a la línea curva FGHI.



Escolio

También se puede construir este problema como sigue. Unidos FG, GH, HI, FI, prolónguese GF hasta V, y únanse FH, IG, y háganse los ángulos CAK, DAL iguales a los ángulos FGH, VFH. Sean concurrentes AK y AL con la recta BD en K y en L, y desde aquí trácense KM y LN tales que, KM forme el ángulo AKM igual al ángulo GHI y sea como HI es a GH, y LN forme el ángulo ALN igual al ángulo FHI y sea a AL como HI es a FH. Pero han de trazarse AK, KM, AL, LN hacia aquel lado de las líneas AD, AK, AL que permita que las letras CAKMC, ALKA, DALND giren en el mismo orden que las letras FGHIF; y trazada MN vaya a encontrar a CE en *i*. Hágase el ángulo *i*EP igual al ángulo IGF, y sea PE a E*i* como FG a GI; y trácese por P la recta PQ*f* que contenga con la recta ADE al ángulo PQE igual al ángulo FIG, y corte a la recta AB en *f*, y únase *fi*. Pero trácense PE y PQ hacia los lados de las líneas CE, PE de modo que la disposición circular de las letras PE*i*P y PEQP sea la misma que la de las letras FGHIF, y si sobre la línea *fi*, también con el mismo orden de las letras, se forma el trapecio *fghi*, semejante al trapecio FGHI, y se circunscribe la trayectoria de forma dada, se habrá resuelto el problema.



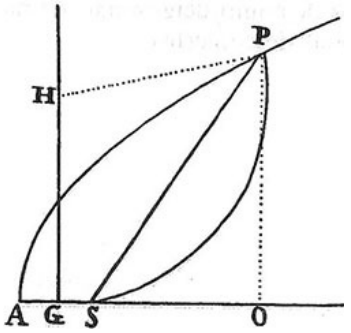
Hasta aquí hemos tratado de cómo hallar las órbitas. Falta ahora que tratemos de cómo determinar los movimientos de los cuerpos en las órbitas descubiertas.

Sección VI
SOBRE COMO HALLAR LOS MOVIMIENTOS EN ORBITAS DADAS

Proposición xxx. Problema xxii

Hallar el lugar, en un momento dado, de un cuerpo que se mueve en una trayectoria parabólica dada.

Sea S el foco y A el vértice principal de la parábola, y 4AS x M sea igual al área de la sección parabólica APS que ha sido descrita por el radio SP, bien sea después del paso del cuerpo por el vértice, bien se haya de describir hasta su llegada al dicho vértice. Se conoce la magnitud del área de dicha sección por su proporcionalidad al tiempo. Elévese la perpendicular GH tal que corte en dos partes iguales a AS en G y sea igual a 3M, y el círculo trazado con centro en H y radio HS cortará a la parábola en el lugar buscado P. Pues, trazada PO perpendicular al eje y trazada PH, ocurre que $AG^2 + GH^2[= HP^2 = (AO - AG)^2 + (PO - GH)^2] = A0^2 + P0^2 - 2GAO - 2GH \times PO + AG^2 + GH^2$. De donde $2GH \times PO[= AO^2 + PO^2 - 2GAO] = AO^2 + \frac{3}{4}P0^2$. En lugar de AO^2 póngase $AO \times \frac{PO^2}{4AS}$, dividiendo todos los términos por 3PO y multiplicándolos por 2AS resultará $\frac{4}{3}GH \times AS [= \frac{1}{6}AO \times PO + \frac{1}{2}AS \times PO = \frac{AO + 3AS}{6} \times PO = \frac{4AO - 350}{6} \times PO$ = área de (APO - SPO)] = al área de APS.



Pero GH era 3M y por tanto $\frac{4}{3}GH \times AS$ es $4AS \times M$. Luego el área comprendida en la sección APS es igual a la de la sección que corresponde a $4AS \times M$. Q. E. D.

Corolario 1. De aquí que GH es a AS como el tiempo en el que el cuerpo describe el arco AP al tiempo en que el cuerpo describe el arco comprendido entre el vértice A y la perpendicular al eje elevada por el foco S.

Corolario 2. Y, en el círculo ASP que pasa perpetuamente por el cuerpo en movimiento P, la velocidad en el punto H es a la velocidad que tenía el cuerpo en el vértice A como 3 a 8; y por tanto la línea GH está en esa misma razón respecto a la recta que podría describir en el tiempo de su movimiento desde A hasta POG, con la velocidad que tenía en el vértice A.

Corolario 3. Y viceversa, también puede hallarse, a partir de aquí, el tiempo en que un cuerpo describe un arco cualquiera determinado AP. Únase AP y por su punto medio elévese una perpendicular que toque a la recta GH en H.

Lema xxviii[29]

No existe figura oval alguna cuya área, seccionada por rectas cualesquiera, pueda hallarse de modo general mediante ecuaciones finitas en cuanto al número de términos y dimensiones.

Dese dentro del círculo oval un punto cualquiera, en torno al cual como polo gire perpetuamente una línea recta, con movimiento uniforme, y mientras tanto, sobre dicha línea, salga desde el polo un punto móvil que camine siempre con una velocidad que sea como el cuadrado de la recta dentro del óvalo. Con este movimiento dicho punto describirá una espiral de infinitos giros. Ahora bien, si pudiese hallarse mediante una ecuación finita la parte de área oval cortada por la tal recta, también se hallaría mediante esa misma ecuación la distancia desde el polo al punto ya que es proporcional a dicha área, y en consecuencia, podrían hallarse todos los puntos de la espiral mediante una ecuación imita; y por tanto también podría hallarse mediante una ecuación imita la intersección de cualquier recta dada en

posición con la espiral. Pero toda recta prolongada infinitamente corta a la espiral en infinitos puntos, y la ecuación por la que se halla la intersección de dos líneas contiene todas las intersecciones entre ellas mediante otras tantas raíces y alcanza por tanto tantas dimensiones como intersecciones haya. Puesto que dos círculos se cortan mutuamente en dos puntos, una intersección no se halla más que mediante una ecuación de dos dimensiones, mediante la cual también puede hallarse la otra. Puesto que las intersecciones de dos secciones cónicas pueden ser cuatro, no podrá hallarse de modo general una de ellas más que mediante una ecuación de cuatro dimensiones, la cual permitirá hallarlas todas a la vez. Pues si se buscan separadamente cada una de dichas intersecciones, dado que la ley y condición de cada una es la misma, el cálculo será el mismo en todos los casos y por ello la conclusión será siempre la misma, la cual deberá además por ello contener a la vez todas las intersecciones y expresarlas indistintamente. De donde también se sigue que las intersecciones de las secciones cónicas con las curvas de tercer grado, dado que pueden ser seis, se resolverán conjuntamente mediante ecuaciones de seis dimensiones, y las intersecciones de dos curvas de tercer grado, puesto que pueden ser nueve, se resolverán a la vez mediante ecuaciones de nueve dimensiones. Si esto no fuese así de modo necesario, sería posible reducir todos los problemas de sólidos a problemas de planos y los de orden superior a los sólidos a problemas de sólidos. Hablo aquí de curvas de grado irreductible. Pues si la ecuación por la que se define la curva se puede reducir a un grado inferior, no sería una sola curva sino compuesta de dos o más cuyas intersecciones pueden hallarse por separado mediante cálculos distintos. Del mismo modo las dos intersecciones de las rectas y las secciones cónicas se resuelven siempre mediante ecuaciones de dos dimensiones; las tres intersecciones de las rectas con las curvas irreductibles de tercer grado mediante ecuaciones de tres dimensiones, las cuatro intersecciones de rectas con curvas irreductibles de cuarto grado mediante ecuaciones de cuatro dimensiones y así hasta el infinito. Por tanto, las infinitas intersecciones entre una recta y una espiral, dado que esta curva es simple y no se puede reducir a otras varias curvas, requieren ecuaciones de infinito número de dimensiones y raíces mediante las cuales puedan resolverse todas las intersecciones a la vez. Ello porque la ley y el cálculo de todas son idénticos. Pues si se traza una perpendicular desde el polo sobre dicha recta secante y se hace girar dicha perpendicular en torno al polo junto con la secante, las intersecciones de la espiral se sucederán mutuamente una a otra y la que estaba la primera o más cercana será la segunda al cabo de una revolución, la tercera después de dos, y así sucesivamente; pero mientras tanto no cambiará la ecuación salvo en la medida en que cambien las magnitudes de las cantidades por cuyo medio se determina la posición de la secante. Por lo cual, dado que dichas cantidades vuelven a ser, después de cada revolución, las mismas que eran al principio, la ecuación también vuelve a ser de la misma forma y por tanto, una y la misma ecuación ofrecerá todas las intersecciones y por ello tendrá el infinito número de raíces por medio de las cuales sería posible ofrecer todas

aquellas. En consecuencia, no se puede hallar de modo general la intersección de una recta y una espiral mediante una ecuación finita, y por eso mismo no existe figura oval cuyo área cortada por rectas cualesquiera pueda ser determinada de modo general por medio de ecuación alguna de este tipo.

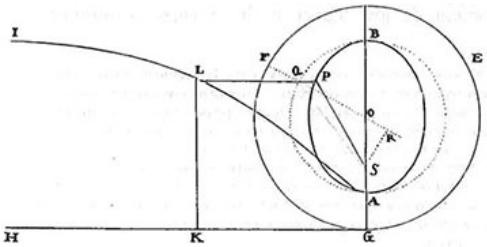
Mediante el mismo argumento puede probarse que, si el intervalo entre el polo y el punto con el que se describe la espiral se tomase proporcionalmente al perímetro de la figura oval abscindida, tampoco es posible hallar la longitud del perímetro de un modo general mediante una ecuación finita. Pero aquí hablo de óvalos a los que no son tangentes figuras conjugadas que evaden al infinito.

COROLARIO. De aquí que el área de una elipse descrita por el radio trazado desde un foco hasta un cuerpo en movimiento, no puede hallarse a partir del tiempo dado por medio de una ecuación finita; y por eso no se podrá determinar mediante la descripción de curvas racionales geométricamente. Denomino racionales geométricamente a las curvas cuyos puntos todos pueden ser establecidos por medio de longitudes definidas mediante ecuaciones, esto es, por razones complejas de longitudes; a las demás (espirales, cuadráticas, trocoides) las denomino irracionales. Pues las longitudes que son o que no son como número a número (según el libro décimo de los Elementos) son racionales o irracionales aritméticamente. Así pues, de una elipse segrego el área proporcional al tiempo mediante una curva geométricamente irracional del modo siguiente.

PROPOSICIÓN XXXI. PROBLEMA XXIII

Hallar el lugar en un momento dado de un cuerpo que se mueve en una trayectoria elíptica dada.

Sea A el vértice principal, S el foco y O el centro de la elipse APB, y sea P el lugar a hallar del cuerpo. Prolónguese OA hasta G de modo que OG sea a OA como OA a OS. Elévese la perpendicular GH, y con centro en O y distancia OG trácese el círculo GEF, y sobre la regla GH a modo de base hágase avanzar la rueda GEF girando sobre su eje y mientras tanto va describiendo con su punto A la trocoide ALI. Hecho esto, tómese GK en razón al perímetro de la rueda GEFG como el tiempo en que el cuerpo partiendo de A describió el arco AP al tiempo de una revolución en la elipse. Elévese la perpendicular KL que se encuentra con la trocoide en L, y trazada LP paralela a la propia KG, encontrará a la elipse en el lugar buscado, P, del cuerpo.

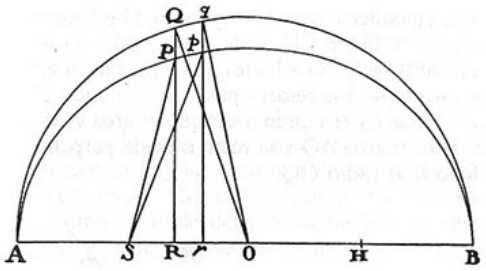


Pues con centro en O y distancia OA trácese el semicírculo AQB y LP, prolongada si es preciso, toque en Q al arco AQ, y únanse SQ y OQ. Toque OQ al arco EFG en F y sobre la propia OQ trácese la perpendicular SR. El área APS es como el área AQS, esto es, como la diferencia entre el sector PQA y el triángulo OQS o también como la diferencia entre los rectángulos ½OQ x AQ y ½OQ x SR, esto es, por estar dado ½OQ, como la diferencia entre el arco AQ y la recta SR, y por tanto (al ser iguales las razones dadas SR al seno del arco AQ, OS a OA, OA a OG, AQ a GF, y dividiendo, AQ - SR a GF - seno del arco AQ) como GK, diferencia entre el arco GF y el seno del arco AQ. Q. E. D.

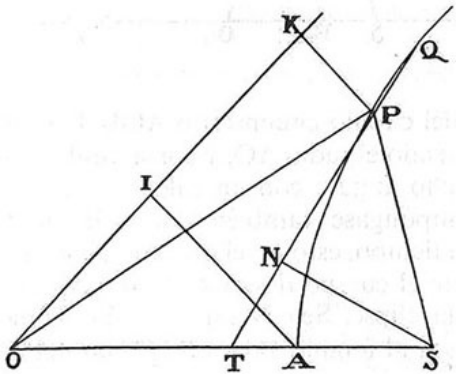
Escolio[30]

Por lo demás, como la descripción de esta curva sea dificultosa, es mejor dar una solución aproximada. Hállese entonces un cierto ángulo B que sea al ángulo de 57,29578 grados, que es el que subtiende un arco igual al radio, como la distancia SH entre focos al diámetro AB de la elipse; a continuación hállese una determinada longitud L que esté inversamente en la misma razón al radio. Una vez halladas ambas cosas, puede resolverse el problema mediante el análisis siguiente. Supóngase conocido el lugar P del cuerpo, próximo a su verdadero lugar *p*, a través de cualquiera construcción, o incluso mediante una conjetura. Trazada sobre el eje de la elipse la ordenada PR, por la proporción de los diámetros de la elipse, tendremos dada la ordenada RQ del círculo circunscrito AQB, la cual es el seno del ángulo AOQ, siendo el radio AO, y corta también a la elipse en P. Hasta hallar dicho ángulo con un cálculo rudimentario en valor aproximado. Supóngase también conocido el ángulo que es proporcional al tiempo, esto es, el que sea a cuatro rectos como el tiempo en el que el cuerpo describe el arco A*p* al tiempo de una revolución en la elipse. Sea N este ángulo. Tómense ahora un ángulo D que sea al ángulo B como el seno del ángulo AOQ al radio, y un ángulo E que sea al ángulo N - AOQ + D como la longitud L a la propia longitud L, disminuida o aumentada en el enseno del ángulo AOQ según este ángulo sea menor o mayor que un recto. Después tómense un ángulo F que sea al ángulo B mino el seno del ángulo AOQ + E al radio y un ángulo G que sea al ángulo N - AOQ - E + F como la longitud L es a la propia longitud L disminuida o

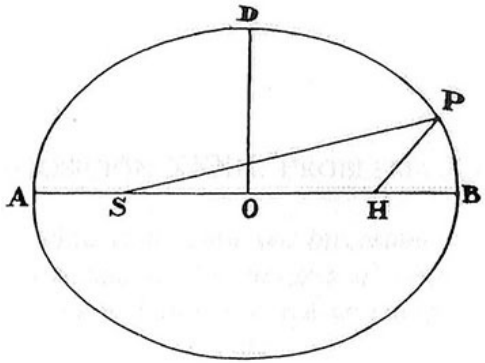
aumentada en el coseno del ángulo AOQ + E según este ángulo sea menor o mayor que un recto. Por tercera vez tómense un ángulo H que sea al ángulo B como el seno del ángulo AOQ + E + G es al radio y un ángulo I que sea al ángulo N - AOQ - E - G + H como la longitud L es a la propia longitud L disminuida o aumentada en el coseno del ángulo AOQ + E + G según este ángulo sea menor o mayor que un recto. V así es posible seguir hasta el infinito. Tómese por fin el ángulo AO*q* igual al ángulo AOQ + E + G + I + etc. Y a partir de su coseno O*r* y de su ordenada *pr*, que es a su seno *qr* como es el eje menor de la elipse al eje mayor, se tendrá el lugar correcto *p* del cuerpo. Cuando el ángulo N - AOQ + D resulta negativo, el signo I de E debe cambiarse siempre en - y el signo - en +. Lo mismo ha de entenderse para los signos de G e I, cuando resultan negativos N - AOQ - E + F y N - AOQ - E - G + H. Pero la serie infinita AOQ + E + G + I, etc., converge tan rápidamente que muy escasamente será necesario pasar del segundo término E. Y el cálculo se basa en el teorema de que el área APS es como la diferencia entre el arco AQ y la recta trazada perpendicularmente desde el foco S al radio OQ.



Con un cálculo similar se resuelve el problema en la hipérbola. Sea O su centro, A su vértice, S el foco y OK la asíntota. Sea conocida la cantidad del área a segregar proporcional al tiempo. Sea ésta A, y hágase la conjetura sobre la posición de la recta SP que segrega muy aproximadamente el área verdadera de APS. Únase OP y trácense desde A y P hasta la asíntota AI, PK paralelas a la otra asíntota, y se tendrá, por la tabla de logaritmos el área AIKP y el área igual a ésta de OPA, la cual restada del área del triángulo OPS da como resto el área segregada APS. Aplicando el doble de la diferencia del área a segregar A y de la segregada APS, 2APS - 2A o 2A - 2APS, a la línea SN, que es perpendicular desde el foco S a la tangente TP, resultará la longitud de la cuerda PQ. Pero esta cuerda PQ ha de inscribirse entre A y P si el área segregada APS es mayor que el área a segregar A, de no ser así ha de inscribirse hacia el otro lado de P: y el punto Q será un lugar más exacto del cuerpo. Y repitiendo el cálculo se hallará siempre con mayor exactitud.



Así pues, con estos cálculos el problema alcanza una solución analítica general. Pero para usos astronómicos es más cómodo el cálculo particular siguiente. Sean AO, OB, OD los semiejes de la elipse, L el «latus rectum» de la misma, D la diferencia entre el semieje menor OD y la mitad del «latus rectum», ½L; hállese entonces el ángulo Y cuyo seno sea al radio como el rectángulo bajo dicha diferencia D y la semisuma de los ejes AO + OD es al cuadrado del eje mayor AB; después hállese el ángulo Z cuyo seno sea al radio como el doble del rectángulo bajo la distancia SH entre los focos y dicha diferencia D al triple del cuadrado del semieje mayor AO. Una vez hallados éstos, el lugar del cuerpo se determinará a continuación del modo siguiente. Tómese el ángulo T proporcional al tiempo en que se ha descrito el arco BP, o igual al movimiento medio (como se dice); y un ángulo V, primera ecuación del movimiento medio, al ángulo Y, primera ecuación máxima, como es el seno del doble del ángulo T al radio; y un ángulo X, segunda ecuación, al ángulo Z, segunda ecuación máxima, como el cubo del seno del ángulo T es al cubo del radio. De los ángulos T, V, X, tómese el ángulo BHP, ecuación del movimiento medio, tal que sea igual o a la suma de T + X + V si el ángulo T es menor que un recto, o a la diferencia de T + X - V si fuese mayor que un recto y menor que dos; y si HP toca a la elipse en P, trazada SP, segregará el área BSP muy aproximadamente proporcional al tiempo. Esta fórmula parece bastante fácil, habida cuenta de que basta hallar las dos o tres primeras figuras de los ángulos muy pequeños V y X establecidos, si se quiere, en minutos segundos. Y además es suficientemente exacta para la teoría de los planetas. Pues en la órbita de Marte, cuya ecuación máxima del centro es de diez grados, el error apenas será mayor que un minuto segundo. Pero una vez hallado el ángulo BHP, ecuación del movimiento medio, ya se tiene inmediatamente por un método bien conocido tanto el ángulo del movimiento real BSP como la distancia SP[31].

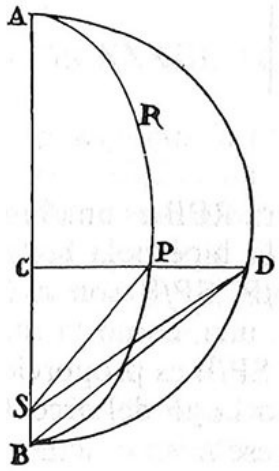


Hasta aquí sobre el movimiento de los cuerpos en líneas curvas. Pero puede ocurrir que un móvil ascienda o descienda por una línea recta, y paso a exponer ahora lo que se refiere a estos movimientos.

Sección VII
DEL ASCENSO Y DESCENSO RECTILÍNEO DE LOS CUERPOS

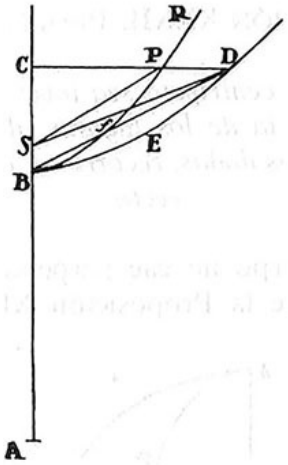
Proposición XXXII. Problema XXIV

Supuesto que la fuerza centrípeta sea inversamente proporcional al cuadrado de la distancia de los lugares al centro, determinar los espacios que, en tiempos dados, recorre un cuerpo cayendo en línea recta.

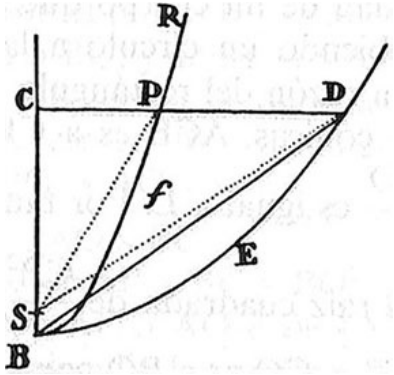


Caso 1. Si un cuerpo no cae perpendicular mente describirá (por el Corolario 1 de la Proposición XIII) una cierta sección cónica cuyo foco coincidirá con el centro de fuerzas. Sea ARPB dicha sección cónica y S su foco. Primero en el caso de que dicha figura sea una elipse; sobre su eje mayor AB trácese el semicírculo ADB, y pase por el cuerpo que cae la recta DPC perpendicular al eje; y trazadas DS, PS, el área ASD será proporcional al área ASP, así como al tiempo. Manteniendo el eje AB, disminuya la latitud de la elipse continuamente y el área ASD seguirá siendo continuamente proporcional al tiempo. Disminúyase infinitamente aquella latitud y, al coincidir la órbita APB con el eje AB y el foco S con el extremo B del eje, también resultará el

área ABD proporcional al tiempo. Y así se tendrá el espacio AC que será el descrito por el cuerpo que cae perpendicularmente desde A en un tiempo dado, siempre que se tome el área ABD proporcional al tiempo y se trace desde el punto D hasta la recta AB la perpendicular DC. Q. E. I.



Caso 2. Si dicha figura RPB es una hipérbola, trácese sobre su diámetro principal AB la hipérbola rectangular BED; y puesto que las áreas CSP, CBfP, SPfB son a las áreas CSD, CBED, SDEB, cada una a cada una, como la razón dada de las alturas CP, CD; y como el área SPfB es proporcional al tiempo en el que el cuerpo se moverá a lo largo del arco PfB, el área SDEB será también proporcional a ese mismo tiempo. Disminúyase el «latus rectum» de la hipérbola RPB hasta el infinito manteniendo el lado transverso, y el arco PB coincidirá con la recta CB y el foco S con el vértice B y la recta SD con la recta BD. Por tanto, el área BDEB será proporcional al tiempo en el que el cuerpo C en caída recta describe la línea CB. Q. E. I.



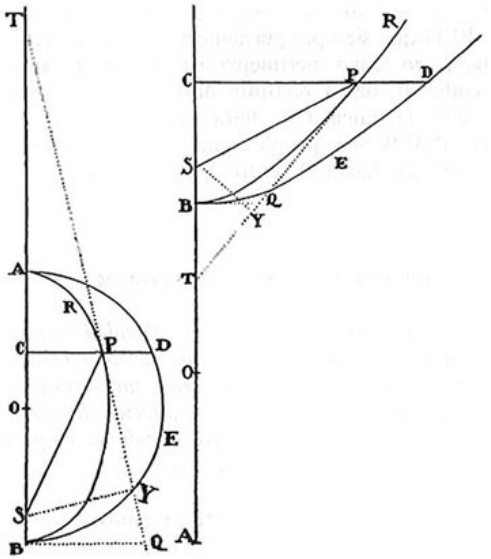
Caso 3. Y con similar argumento si la figura RPB es una parábola, y por el mismo vértice principal B se describe otra parábola BED que siempre permanezca dada, mientras la primera parábola, en cuyo perímetro se mueve el cuerpo P, va disminuyendo su «latus rectum» hasta que se anula, llegará a coincidir con la línea CB; hará entonces que el segmento parabólico BDEB sea proporcional al tiempo en que dicho cuerpo P o C cae hasta el centro S o B. Q. E. I.

Proposición xxxiii. Teorema ix

Supuesto lo ya descubierto, digo que la velocidad de un cuerpo que cae en un lugar C es a la velocidad de un cuerpo que gira en círculo con centro B y distancia BC, como la raíz cuadrada de la razón de AC, distancia del cuerpo al vértice A ulterior del círculo o de la hipérbola rectangular, respecto a ½AB, semidiámetro principal de la figura.

Hágase en O la bisección de AB, diámetro común de ambas figuras RPB, DEB; y trácese la recta PT, tangente a la figura RPB en P, y que corte también a dicho diámetro común AB (prolongado si es necesario) en T; y sea SY perpendicular a dicha recta, lo mismo que BQ al susodicho diámetro, y supóngase que L sea el «latus rectum» de la figura RPB. Se sabe por el Corolario 9 de la Proposición xvi que la velocidad de un cuerpo que se mueve sobre la línea RPB en torno al centro S en un lugar cualquiera P será a la velocidad de un cuerpo que se mueve en torno al mismo centro describiendo un círculo a la distancia SP, como la raíz cuadrada de la razón del rectángulo ½L x SP a SY al cuadrado. Pues, por las cónicas, ACB es a CP^2 como 2AO a L, y por tanto, $\frac{2CP^2 \text{ x } AO}{ACB}$ es igual a L. Por tanto dichas velocidades son entre sí como la raíz cuadrada de $\frac{CP^2 \text{ x } AO \text{ x } SP}{AC \text{ x } CB}$ es a SY^2. Además, por las cónicas, CO es a BO como BO es a TO y, componiendo o dividiendo, como CB a BT. De donde, dividiendo o componiendo, será BO ± CO a BO como CT a BT, esto es, AC a AO como CP a BQ; y de aquí que $\frac{CP^2 \text{ x } AO \text{ x } SP}{ACB}$ sea igual a $\frac{BQ^2 \text{ x } AC \text{ x } SP}{AO \text{ x } BC}$. Disminúyase ahora la anchura CP de la figura RPB hasta el infinito de modo que el punto P coincida con el punto C y el punto S con el punto B y la línea BQ; y ahora la velocidad del cuerpo que cae por la línea recta CB será a la velocidad de un cuerpo que describe un círculo de centro B y distancia BC como la raíz cuadrada de la razón

$\frac{BQ^2 \text{ x } AC \text{ x } SP}{AO \text{ x } BC}$ a SY^2, esto es (despreciando las razones de la igualdad de SP a BC y BQ^2 a SY^2), como la raiz cuadrada de la razón de AC a AO o de ½AB. Q. E. D.

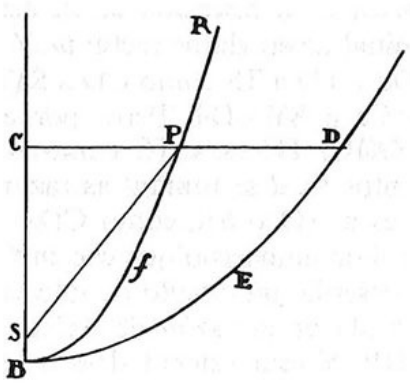


Corolario 1. Al coincidir los puntos B y S, TC será a TS como AC a AO.

Corolario 2. Si un cuerpo gira en un círculo de determinada distancia al centro, al desviar su movimiento hacia arriba subirá hasta duplicar su distancia al centro.

Proposición xxxiv. Teorema x

Si la figura BED es una parábola, digo que la velocidad de un cuerpo que cae en cualquier punto C es igual a la velocidad con la que un cuerpo puede describir uniformemente un círculo de centro B y con la mitad de su intervalo BC.



Pues la velocidad de un cuerpo que describe una parábola de centro S siendo la parábola RPB será en un punto P cualquiera (por el Corolario 7 de la Proposición XVI) igual a la velocidad de un cuerpo que describe un círculo uniformemente en torno al mismo centro S a la mitad del intervalo SP. Disminúyase infinitamente la anchura CP de la parábola hasta que el arco parabólico P*f*B coincida con la recta CB, el centro S con el vértice B y el intervalo SP con el intervalo BC, y la proposición será evidente. Q. E. D.

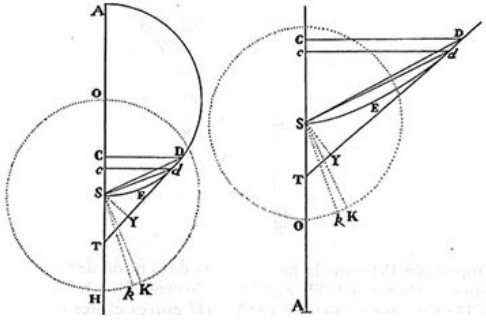
Proposición XXXV. Teorema XI

Con los mismos supuestos, digo que el área de la figura DES, descrita con radio indefinido SD, será igual al área que puede describir en el mismo tiempo un cuerpo girando uniformemente en torno al centro S y con radio igual a la mitad del «latus rectum» de la figura DES.

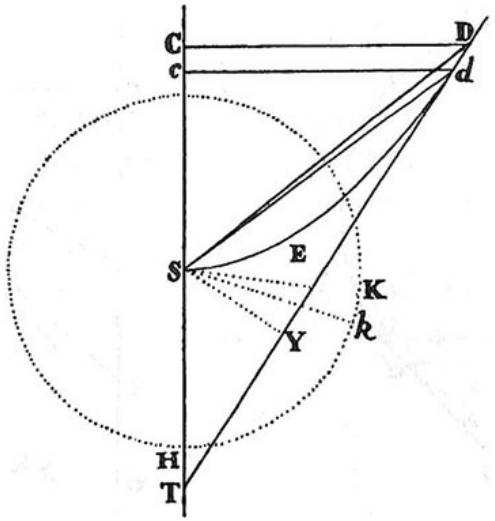
Pues imagínese que un cuerpo C describe al caer en la mínima partícula de tiempo el segmento C*c*, mientras otro cuerpo K, girando uniformemente en el círculo OK*k* en torno al centro S, describe el arco K*k*. Elévense las perpendiculares CD, *cd* que se encuentran con la figura DES en D y en *d*. Únanse SD, S*d*, SK, S*k* y trácese D*d* que encuentra al eje AS en T y sobre ella hágase descender la perpendicular SY.

Caso 1. Si ahora la figura DES es un círculo o una hipérbola rectangular trácese en O la bisección de su diámetro transverso AS, y SO será la mitad de su «latus rectum». Y puesto que TC es a TD como C*c* a D*d* y TD a TS como CD a SY, también será TC a TS como CD x C*c* a SY x D*d*. Pero (por el Corolario 1 de la Proposición XXXIII) TC es a TS como AC es a AO si al confundirse los puntos D, *d* se toman las razones últimas de las líneas. Luego AC es a AO o SK como CD x C*c* es a SY x D*d*. Además la velocidad de un cuerpo que cae en C es a la velocidad de un cuerpo que describe un círculo de intervalo SC y centro S como la raíz cuadrada de la razón de

AC a AO o SK (por la Proposición XXXIII). Y esta velocidad es a la velocidad de un cuerpo que describa el círculo OK*k* como la raíz cuadrada de la razón de SK a SC (por el Corolario 6 de la Proposición IV) y por lanío la velocidad primera es a la última, esto es el segmento C*c* al arco K*k*, como la raíz cuadrada de la razón de AC a SC, esto es, en razón de AC a CD. Por lo cual CD x C*c* es igual a AC x K*k*, y por tanto AC es a SK como AC x K*k* a SY x D*d*, y de aquí que SK x K*k* sea igual a SY x D*d*, y ½SK x K*k* igual a ½SY x D*d*, esto es, el área KS*k* igual al área SD*d*. Por lo tanto en cada partícula de tiempo se generan partículas de dos áreas KS*k* y SD*d* que si disminuyen en magnitud y aumentan en número infinitamente, llegan a ser iguales, y por consiguiente (por el Corolario del Lema IV), todas las áreas generadas a la vez son siempre iguales. Q. E. D.



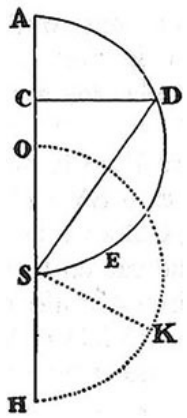
Caso 2. Pero si la figura DES fuese una parábola, hallaremos como antes que CD x C*c* es a SY x D*d* como TC a TS, esto es, como 2 a 1, y por tanto ¼CD x C*c* es igual a ½SY x D*d*. Pero la velocidad de un cuerpo que cae en C es igual a la velocidad con que puede describir un círculo uniformemente con intervalo de ½SC (por la Proposición XXXIV). Y esta velocidad es a la velocidad con la que puede describir un círculo de radio SK, esto es, el segmento C*c* al arco K*k* (por el Corolario 6 de la Proposición IV) como la raíz cuadrada de la razón de SK a ½SC, esto es, en la razón de SK a ½CD. Por lo cual ½SK x K*k* es igual a ¼CD x C*c*, y por tanto igual a ½SY x D*d*, esto es, el área KS*k* igual al área SD*d*, como antes. Q. E. D.



Proposición xxxvi. Problema xxv

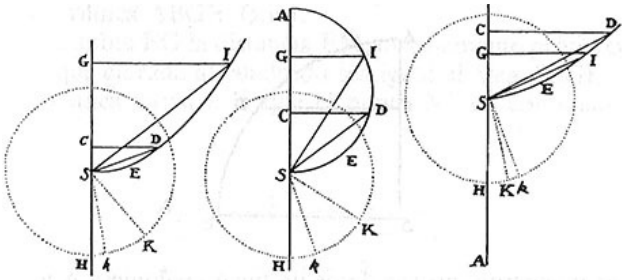
Determinar los tiempos de descenso de un cuerpo que cae desde un punto dado A.

Sobre el diámetro AS, distancia inicial del cuerpo al centro, trácese el semicírculo ADS, y también el semicírculo OKH igual al anterior y con centro en S. Desde C, lugar cualquiera del cuerpo, elévese la ordenada CD. Únase SD, y hágase el sector OSK igual al área ASD. Por la Proposición xxxv es evidente que un cuerpo al caer recorre el espacio AC en el mismo tiempo en que otro cuerpo, girando uniformemente en torno al centro S, puede describir el arco OK. Q. E. F.



Proposición xxxvii. Problema xxvi

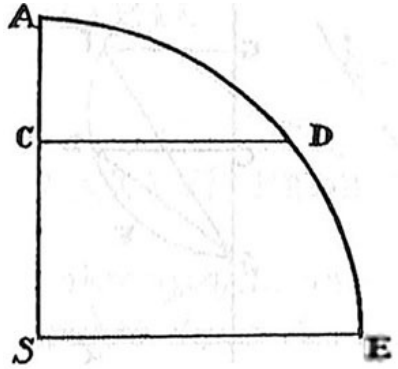
Determinar los tiempos de ascenso o caída de un cuerpo lanzado hacia arriba o hacia abajo desde un punto dado.



Sea G el punto de partida del cuerpo siguiendo la línea GS y con una velocidad cualquiera. Tómese GA a ½AS como el cuadrado de la razón de dicha velocidad a la velocidad uniforme con que un cuerpo puede girar en un círculo de intervalo dado G y centro S. Si dicha razón es de 2 a 1, el punto A dista infinitamente, en cuyo caso ha de trazarse una parábola con vértice S, eje SG, y «latus rectum» cualquiera. Esto es evidente por la Proposición xxxiv. Pero si dicha razón es menor o mayor que 2 a 1, en el primer caso, debe describirse un círculo sobre el diámetro SA y en el segundo una hipérbola rectangular. Es evidente por la Proposición xxxiii. Ahora con centro S e intervalo igual a la mitad del «latus rectum» trácese el círculo H*k*K, y desde el punto G del cuerpo que asciende o cae y desde otro punto cualquiera C elévense las perpendiculares GI, CD que tocan a la sección cónica o al círculo en I y en D. Después, unidas SI, SD, iguálense los segmentos SEIS, SEDS a los sectores HSK, HS*k*, y por la Proposición xxxv, el cuerpo G recorrerá el espacio GC en el mismo tiempo en que el cuerpo K puede describir el arco K*k*. Q. E. F.

Proposición xxxviii. Teorema xii

Supuesto que la fuerza centrípeta sea proporcional a la altura o distancia de los lugares al centro, digo que los tiempos de caída, las velocidades y los espacios recorridos son respectivamente proporcionales a los arcos, a los senos rectos y a los senos versos de los arcos.



Sea un cuerpo que cae desde un lugar cualquiera A y según la línea AS; y desde el centro de fuerzas S, y con intervalo AS, trácese el cuadrante circular AE, y sea CD el seno recto de un arco cualquiera AD; y el cuerpo A, en el tiempo AD, recorrerá al caer el espacio AC, y en el punto C alcanzará la velocidad CD.

Se demuestra esto partiendo de la Proposición X del mismo modo que partiendo de la Proposición XI se demostró la Proposición XXXII.

COROLARIO 1. De aquí que sean iguales los tiempos en que un cuerpo que cae desde el punto A llega hasta S y otro cuerpo girando recorre el arco cuadrantal ADE.

COROLARIO 2. Por lo tanto, los tiempos todos en que los cuerpos que caen desde un lugar cualquiera hasta el centro son iguales. Pues todos los tiempos periódicos de cuerpos en revolución son iguales (por el Corolario 3 de la Proposición IV).

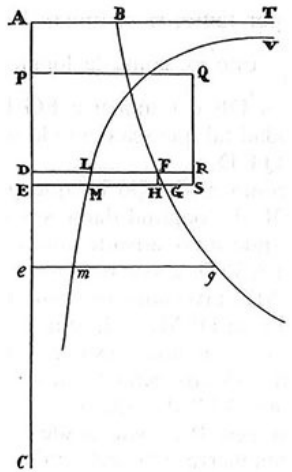
PROPOSICIÓN XXXIX. PROBLEMA XXVII

Supuesta una fuerza centrípeta de cualquier clase, y concedidas las cuadraturas de las figuras curvilíneas, se pide la velocidad de un cuerpo que asciende o cae según una recta en cada punto de la misma, y también el tiempo en que llegará a cada uno de los puntos; y viceversa.

Sea el cuerpo E que cae desde el punto cualquiera A según la recta ADEC, y desde su lugar E elévese la perpendicular EG, proporcional siempre a la fuerza centrípeta tendente en ese lugar hacia el centro C: sea BFG una línea curva continuamente tangente del punto G. Al comienzo del movimiento coincida EG con la perpendicular AB, y la velocidad del cuerpo en un punto cualquiera E será como la recta que elevada al cuadrado sea igual al área curvilínea ABGE. Q. E. I.

Tómese sobre EG la distancia EM, inversamente proporcional a la recta que elevada al cuadrado sea igual al área ABGE, y sea VLM una línea curva a la que el punto M sea continuamente tangente, y cuya asíntota sea la prolongación de la recta

AB; y el tiempo en que un cuerpo al caer recorre el espacio AE será como el área curvilínea ABTVME. Q. E. I.

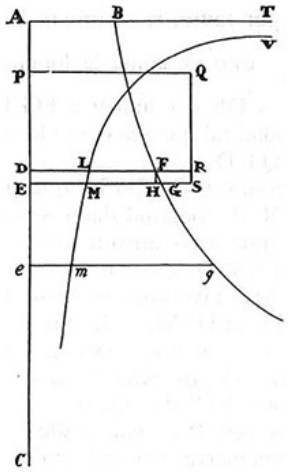


Pues sobre la recta AE tómese el segmento mínimo DE de una longitud dada, y sea DLF el lugar de la línea EMG, cuando el cuerpo pasaba por D; y si la fuerza centrípeta fuese tal que la recta, cuyo cuadrado es igual al área ABGE, sea como la velocidad del cuerpo que cae, el área misma será como el cuadrado de la velocidad, esto es, si se representan las velocidades en D y E por V y V + I, el área ABFD será como VV, y el área ABGE como VV + 2VI + II y, restando, el área DFGE como 2VI + II, por lo cual $\frac{DFGE}{DE}$ como $\frac{2VI + II}{DE}$, esto es, si se toman las razones primeras de estas cantidades en su momento inicial, la longitud DF es como $\frac{2VI}{DE}$, y en consecuencia también como la mitad de esta cantidad $\frac{I \times V}{DE}$. Pero el tiempo en el que el cuerpo caer recorre el segmento aquel DE es directamente como dicho segmento e inversamente como la velocidad V, y la fuerza es directamente como el incremento 1 de la velocidad e inversamente como el tiempo, y por tanto, si se toman las razones primeras iniciales, como, $\frac{I \times V}{DE}$, esto es, como la longitud DF. Luego una fuerza proporcional a DF o también a EG hará que el cuerpo caiga con una velocidad tal que sea como la recta cuyo cuadrado sea el área ABGE. Q. E. D.

Por lo demás, como el tiempo en que puede recorrerse un segmento mínimo DE de longitud dada, sea inversamente como la velocidad y por ende inversamente como

la línea recta cuyo cuadrado sea el área ABFD; y, por otra parte, sea DL, y también el área naciente DLME, inversamente como dicha línea recta, el tiempo será como el área DLME, y la suma de todos los tiempos como la suma de todas las áreas, esto es (por el Corolario del Lema IV), la totalidad de los tiempos en que se recorre la línea AE como el área total ATVME. Q. E. D.

COROLARIO 1. Si fuese P el lugar desde el que debiese caer un cuerpo urgido por una fuerza centrípeta uniforme conocida (como se supone generalmente la gravedad) de modo que en el punto D adquiera una velocidad igual a la de otro cuerpo en ese punto D que la ha adquirido al caer merced a otra fuerza cualquiera, y sobre la perpendicular DF se tomara DR tal que sea a DF como dicha fuerza uniforme es a la otra fuerza en D, y se completase el rectángulo PDRQ, y se segregase el área ABFD igual al mismo; entonces A será el lugar desde el cual cayó el otro cuerpo. Pues, completando el rectángulo DRSE, como el área ABFD sea al área DFGE como VV a 2VI, y por lo mismo como ½V a I, esto es, como la mitad de la velocidad total al incremento de la velocidad del cuerpo que cae empujado por la fuerza no uniforme; y similarmente el área PQRD es al área DRSE como la mitad de toda la velocidad al incremento de la velocidad del cuerpo que cae con fuerza uniforme, y como dichos incrementos (por la igualdad de los tiempos nacientes) son como las fuerzas generatrices, esto es, como las ordenadas DF, DR, y por tanto, como las áreas nacientes DFGE, DRSE; en consecuencia, las áreas totales ABFD, PQED serán entre sí como las mitades de las velocidades totales, y por lo tanto, serán iguales por ser las velocidades iguales.



COROLARIO 2. De aquí que si un cuerpo es lanzado desde cualquier punto D hacia arriba o hacia abajo con una velocidad dada, y se supone la ley de la fuerza centrípeta, su velocidad en otro punto *e* se hallará elevando la ordenada *eg*, y

tomando dicha velocidad a la velocidad en el punto D como la recta, cuyo cuadrado equivalga al rectángulo PQRD aumentado en el área curvilínea DF*ge* si el punto *e* está por debajo del punto D y disminuido si está por arriba, es a la recta cuyo cuadrado equivalga sólo al rectángulo PQRD.

COROLARIO 3. También se halla el tiempo elevando la ordenada *em* inversamente proporcional al lado cuadrado de PQRD ± DF*ge*, y tomando el tiempo en que el cuerpo hubiese descrito la línea D*e* al tiempo en que otro cuerpo cayendo con fuerza uniforme desde P llega hasta D, como el área curvilínea DL*me* al rectángulo 2PD x DL. Puesto que el tiempo, en el que cayendo con fuerza uniforme un cuerpo describe la línea PD es al tiempo en el que el mismo cuerpo describe la línea PE, como la raíz cuadrada de la razón de PD a PE, esto es (considerando el segmento DE en el instante de su nacimiento), como la razón de PD a PD + ½DE o como 2PD + DE, y por partes, al tiempo en que dicho cuerpo recorrió el segmento DE como 2PD a DE, y por tanto como el rectángulo 2PD x DL al área DLME; y el tiempo en que un cuerpo describió el segmento DE es al tiempo en que el otro describió con movimiento variable la línea D*e* como el área DLME al área DL*me*, y por consiguiente el primer tiempo es al último como el rectángulo 2PD x DL al área DL*me*.

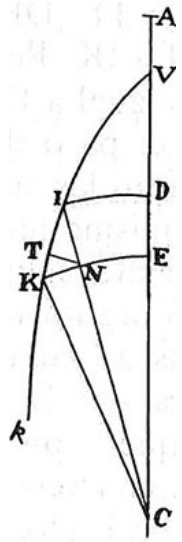
Sección VIII
DE CÓMO HALLAR ÓRBITAS EN LAS CUALES GIRASEN CUERPOS SUJETOS A FUERZAS CENTRÍPETAS CUALESQUIERA

Proposición XL. Teorema XIII

Si un cuerpo se mueve de cualquier forma sujeto a la acción de una fuerza centrípeta, y otro cuerpo sube o baja por una línea recta, y en algún momento de alturas iguales fuesen iguales sus velocidades, serán iguales sus velocidades en todas las alturas iguales.

Caiga un cuerpo desde A a través de D, E hacia un centro C, y muévase otro cuerpo desde V según la línea curva VIK*k*. Con centro en C y distancias cualesquiera trácense los círculos concéntricos DI, EK que tocan a la recta AC en D y E y a la curva VIK en I y K. Únase IC que cortará a KE en N; y sobre IK trácese la perpendicular NT; sea también muy pequeño el intervalo entre circunferencias DE o IN, y sean iguales las velocidades de los cuerpos en D y I. Puesto que las distancias CD, CI son iguales, las fuerzas centrípetas en D y en I serán iguales. Representen a estas fuerzas los segmentos iguales DE, IN; si una de las fuerzas IN (por el Corolario II de las Leyes) se descompone en dos NT y IT, la fuerza NT, actuando según la línea NT, perpendicular a la trayectoria ITK del cuerpo, en nada alterará la velocidad de dicho cuerpo en su curso, sino que solamente lo desviará de su curso rectilíneo y hará que dicho cuerpo, apartándose continuamente de la tangente de la órbita, camine según la trayectoria curva ITK*k*. Toda dicha fuerza se consumirá en la producción de este efecto: mientras que la otra fuerza IT, actuando según la marcha del cuerpo, toda ella lo acelerará, y en un tiempo mínimo dado generará una aceleración proporcional a ella misma. Por tanto las aceleraciones de los cuerpos en D e I alcanzadas en tiempos iguales (si se consideran las razones primeras de los segmentos nacientes DE, IN, IK, IT, NT) son como las líneas DE, IT: mientras que en tiempos desiguales son como el producto de dichas líneas por los tiempos. Por lo demás, los tiempos en que se describirán DE e IK, por la igualdad de velocidades, son como las trayectorias descritas DE e IK, y por tanto las aceleraciones en el paso de los cuerpos por DE e IK, son como el producto de DE e IT, DE e IK, esto es como DE cuadrado y el

rectángulo IT x IK. Pero el rectángulo IT x IK es igual a IN cuadrado, esto es igual a DE cuadrado y por tanto las aceleraciones generadas en el paso de los cuerpos desde Del hasta E y K son iguales. Luego las velocidades de los cuerpos en E y K son iguales: y por el mismo argumento siempre resultarán iguales a distancias subsiguientes iguales. Q. E. D.



Y además por el mismo argumento cuerpos con velocidades iguales y a iguales distancias del centro se desacelerarán igual al ascender a distancias iguales. Q. E. D.

Corolario 1. De aquí que, si un cuerpo oscila colgado de un hilo o se ve obligado a girar en círculo mediante un impedimento completamente pulido y lubrificado, y otro cuerpo sube o baja por una recta, y a una altura cualquiera igual son sus velocidades iguales, también serán iguales sus velocidades en cualesquiera otras alturas iguales. Pues con el hilo del cuerpo pendular o con la pared completamente lubrificada del recipiente se consigue exactamente lo mismo que con la fuerza transversal NT. Por ella el cuerpo no se acelera ni se desacelera, sino que solamente es obligado a apartarse de la trayectoria recta.

Corolario 2. De aquí también que, si la cantidad P fuese la máxima distancia al centro que puede alcanzar un cuerpo, tanto oscilando como girando en cualquier trayectoria, y la que alcanzaría por tanto lanzado hacia arriba desde un punto cualquiera de la trayectoria con la velocidad que tiene en ese punto; y fuese la cantidad A la distancia del cuerpo al centro en otro punto cualquiera de la órbita, y siendo siempre la fuerza centrípeta como una potencia de A tal como A^{n-1}, cuyo índice $n-1$ es cualquier número n disminuido en una unidad; a cualquier altura A la velocidad del cuerpo será como $\sqrt{P^n - A^n}$, y por tanto estará dada. Pues la velocidad

recta de ascenso o descenso (por la Proposición XXXIX) está en esta misma razón.

Proposición XLI. Problema XXVIII

Dada una fuerza centrípeta de cualquier clase y supuestas las cuadraturas de las figuras curvas, hállense tanto las trayectorias en que se mueven los cuerpos como los tiempos de los movimientos en las trayectorias halladas.

Tienda una fuerza cualquiera hacia el centro C y hállese la trayectoria VIK*k*. Sea dado el círculo VR con centro en C y trazado con una distancia cualquiera CV, y con el mismo centro trácense otros círculos ID, KE que corten a la trayectoria en I y K y a la recta CV en D y E. Trácese tanto la recta CNIX que corta a los círculos KE, VR en N y X, como la recta CKY que toca al círculo VR en Y. Y sean los puntos I y K muy próximos entre sí y que el cuerpo se mueva desde V por I y K hasta *k*; sea, así mismo A el punto desde el cual otro cuerpo debe caer para que adquiera en el punto D una velocidad igual a la del cuerpo anterior en I. Y, dado lo dicho en la Proposición XXXIX, el segmento IK, descrito en el menor tiempo dado, será como la velocidad y, por tanto, como la recta cuyo cuadrado sea igual al área ABFD, y sí tendrá dado el triángulo ICK proporcional al tiempo, y por lo mismo KN será inversamente como la altura IC, esto es, si se da una cantidad Q cualquiera y a la altura IC la denominamos A, como $\frac{Q}{A}$. A esta cantidad $\frac{Q}{A}$ llamémosla Z y supongamos que la magnitud Q es tal que en algún caso $\sqrt{ABFD}$ es a Z como IK a KN, entonces en todos los casos $\sqrt{ABFD}$ será a Z como IK a KN, y ABFD a ZZ como IK^2 a KN^2 y restando ABFD - ZZ: $ZZ = IN^2 : KN^2$,

por tanto $\sqrt{ABFD - ZZ} : Z$ o $\frac{Q}{A} = IN : KN$

y por ello $A \times KN = \frac{Q \times IN}{\sqrt{ABFD - ZZ}} = IN : KN$

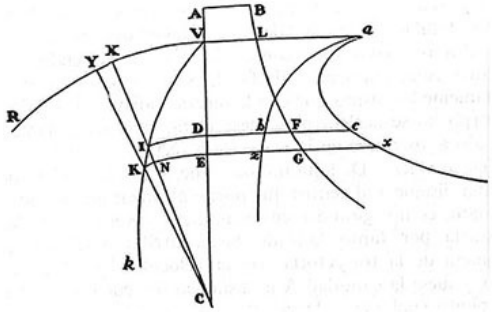
de donde, como $YX \times XC : A \times KN = CX^2 : AA$

será el rectángulo $XY \times XC = \frac{Q \times YN \times CX^2}{AA\sqrt{ABFD - ZZ}}$

Por tanto, si en la perpendicular DF se toman siempre D*b*, D*c* iguales respectivamente a

$\frac{Q}{2\sqrt{ABFD - ZZ}}$ y a $\frac{Q \times CX^2}{2AA\sqrt{ABFD - ZZ}}$

y se trazan las curvas *ab*, *ac*, a las cuales son siempre tangentes los puntos *b*, *c*; y se eleva hasta la línea AC desde el punto V la perpendicular V*a* que corta las áreas curvilíneas VD*ba*, VD*ca* y se elevan también las ordenadas E*z*, E*x*: dado que el rectángulo D*b* x IN o D*bz*E es igual a la mitad del rectángulo A x KN o al triángulo ICK; y el rectángulo De x IN o D*cx*E es igual a la mitad del rectángulo YX x XC, o sea, al triángulo XCY; esto es, puesto que las partículas nacientes D*bz*E, ICK de las áreas VD*ba*, VIC son siempre iguales, y las partículas nacientes D*cx*E, XCY de las áreas VD*ca*, VCX son siempre iguales, el área generada VD*ba* siempre será igual al área generada VIC y por tanto proporcional al tiempo, y el área generada VD*ca* igual al sector generado VCX. Así pues, dado el tiempo de partida de un cuerpo desde V, se tendrá dada también el área VD*ba* proporcional a él, y de aquí se tendrá también la altura del cuerpo CD o CI, así como el área VD*ca* y su igual el sector VCX a la vez que su ángulo VCI. Una vez dados el ángulo VCI y la altura CI se tiene el punto I en el que al completar ese tiempo se hallará el cuerpo. Q. E. I.

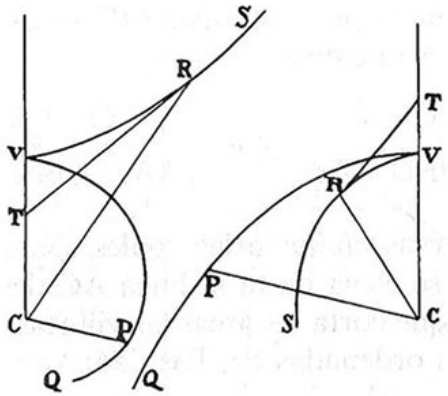


Corolario 1. De aquí que las alturas máximas y mínimas de los cuerpos, esto es, los ápsides de las trayectorias se puedan hallar fácilmente. Puesto que los ápsides son los puntos en los que la recta IC trazada por el centro cae perpendicularmente sobre la trayectoria VIK: cosa que ocurre cuando las rectas IK y NK se igualan, y por tanto cuando el área ABFD es igual a ZZ.

Corolario 2. Y también, partiendo de la altura dada del cuerpo IC, se halla fácilmente el ángulo KIN con el cual corta la trayectoria en cualquier punto a dicha línea IC; a saber, tomando el seno del ángulo al radio como KN a IK, esto es, como Z al lado cuadrado del área ABFD.

Corolario 3. Si con centro en C y vértice principal en V se describe una sección cónica cualquiera VRS, y desde un punto malquiera R dé la misma se traza una tangente que corte al eje principal CV prolongado infinitamente en un punto T;

después, uniendo CR, se traza la recta CP igual a la abscisa CT, y el ángulo VCP proporcional al sector VCR; y si además una fuerza centrípeta inversamente proporcional al cubo de las distancias de los lugares al centro tiende hacia el centro C, y un cuerpo sale con una velocidad precisa desde V según una línea perpendicular a la recta CV: dicho cuerpo circulará por la trayectoria VPQ siempre tangente al punto P; y por lo tanto, si la sección cónica VRS fuese una hipérbola, descenderá hacia el centro: pero si fuese una elipse, ascenderá continuamente y marchará hacia el infinito. Y por el contrario, si un cuerpo con velocidad cualquiera parte de V y desde allí empieza bien a descender oblicuamente hacia el centro o bien a ascender oblicuamente desde él, según sea la figura VRS una hipérbola o una elipse, se puede hallar la trayectoria aumentando o disminuyendo el ángulo VCP según una razón dada. Por lo demás, al convertirse la fuerza centrípeta en centrífuga, el cuerpo ascenderá oblicuamente por la trayectoria VPC, la cual se halla tomando el ángulo VCP proporcional al sector elíptico VRC y la longitud CP igual a CT como antes. Todas estas cosas se siguen de la proposición anterior, mediante la cuadratura de determinada curva cuyo hallazgo, por ser fácil, en aras de la brevedad, doy por hecho[32].

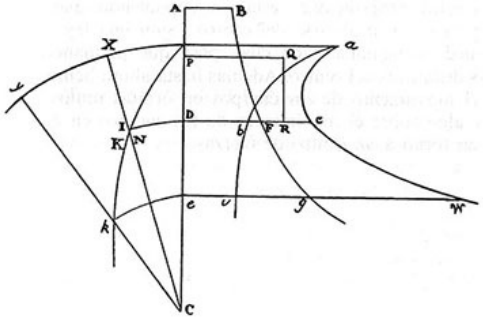


Proposición xlii. Problema xxix

Dada la ley de la fuerza centrípeta, hallar los movimientos de un cuerpo que parte de un lugar dado, con una velocidad dada y según una recta dada.

Supuesto lo dicho en las tres proposiciones precedentes; parta un cuerpo de I según el segmento IK con la velocidad que otro cuerpo puede adquirir en D cayendo desde P con fuerza centrípeta uniforme: y sea esta fuerza uniforme a la fuerza que actúa en I sobre el cuerpo anterior, como DR a DF. Diríjase pues el cuerpo hacia *k*; y

con centro en C y distancia C*k* trácese el círculo *ke* que corta a la recta PD en e y elévense *eg, ev, ew* ordenadas a las curvas BF*g*, *abv, acw*. A partir del rectángulo dado PDRQ y de la ley dada de la fuerza centrípeta con la que se mueve el primer cuerpo, se tiene dada la curva BF*g*, por construcción del Problema xxvii y su Corolario 1. De aquí, por estar dado el ángulo CIK, se tiene la proporción de las líneas nacientes IK, KN, y después, por la construcción del Problema xxviii se tiene la cantidad Q junto con las líneas curvas *abv, acw*: y por lo tanto, al cumplirse cualquier tiempo D*bve*, se tiene, tanto la altura del cuerpo C*e* o C*k*, como el área D*cwe* y el sector igual a ella XZ*y*, así como el ángulo ICK y el punto *k* en el cual se hallará el cuerpo en ese momento. Q. E. I.



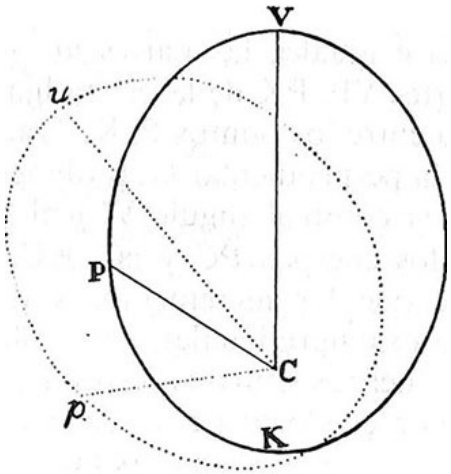
En estas proposiciones estamos suponiendo que la fuerza centrípeta varía al alejarse del centro según una ley cualquiera, que puede imaginarse cada cual, pero que permanece igual a iguales distancias del centro. Además hasta ahora hemos considerado el movimiento de los cuerpos en órbitas inmóviles. Falta añadir algo sobre el movimiento de los cuerpos en órbitas que giren en torno a un centro de fuerzas.

Sección IX
DEL MOVIMIENTO DE CUERPOS EN ÓRBITAS MÓVILES Y DEL MOVIMIENTO DE APSIDES

Proposición XLIII. Problema XXX

Hágase que un cuerpo, que se halla en una trayectoria cualquiera la cual gira en torno a un centro de fuerzas, se pueda mover igual que otro cuerpo en esa misma trayectoria en reposo.

Gire el cuerpo P en la órbita de posición dada VPK partiendo de V hacia K. Desde el centro C trácese C*p*, que sea siempre igual a CP, y el ángulo VC*p* resulte proporcional al ángulo VCP; y el área descrita por la línea C*p* será al área VCP que resulta simultáneamente descrita por la línea CP como la velocidad de la línea descriptora C*p* a la velocidad de la línea descriptora CP; esto es, como el ángulo VC*p* al ángulo VCP, y por tanto en una razón dada, y por tanto proporcional al tiempo. Al ser proporcional al tiempo el área descrita por C*p* en un plano inmóvil, es evidente que un cuerpo bajo la acción de una adecuada fuerza centrípeta puede girar junto con el punto *p* en la línea curva que por la razón expuesta describe el punto *p* en el plano inmóvil. Háganse iguales el ángulo VC*u* y el ángulo PC*p*, la línea *Cu* y la línea CV y la figura *u*C*p* con la figura VCP, y el cuerpo siempre presente en *p* se moverá en el perímetro de la figura en revolución *u*C*p*, además describirá el arco de la misma *up* en el mismo tiempo en que otro cuerpo P podría describir el arco VP igual al anterior sobre la figura en reposo VPK. Hállese por consiguiente, por el Corolario 5 de la Proposición VI, la fuerza centrípeta con la que un cuerpo puede girar en la curva que describe el punto *p* en el plano inmóvil, y quedará resuelto el problema. Q. E. F.

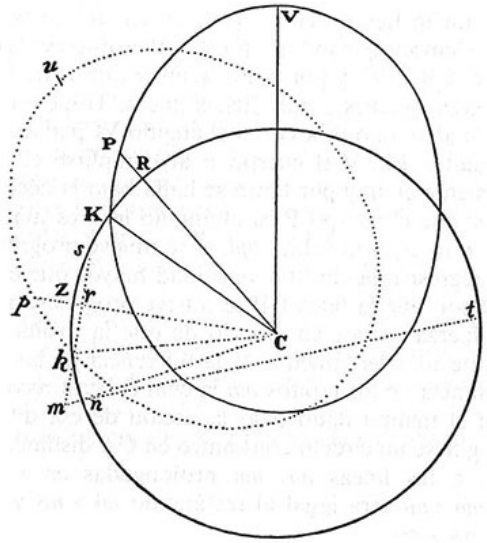


Proposición XLIV. Teorema XIV

La diferencia entre las fuerzas por las que dos cuerpos pueden moverse igualmente, el uno en una órbita fija y el otro en una órbita igual en revolución, está en razón inversa del cubo de las alturas comunes.

Sean semejantes e iguales las partes *up*, *pk* de la órbita en revolución a las partes VP, PK de la órbita fija; supóngase que es mínima la distancia entre los puntos P, K. Trácese desde el punto *k* sobre la recta *p*C la perpendicular *kr*, prolongándola hasta *m*, de modo que *mr* sea a *kr* como el ángulo VC*p* al ángulo VCP. Dado que las alturas de los cuerpos PC y *p*C, KC y *k*C siempre son iguales, es evidente que los incrementos y decrementos de las líneas PC y *p*C serán siempre iguales, y por ello si el movimiento de cada uno de los cuerpos situados respectivamente en P y en *p* se divide en dos (por el Corolario II de las Leyes) uno de los cuales tiende al centro, o sea, según las líneas PC, pC, y el otro, transversal al anterior, tiene una dirección perpendicular a dichas lineas PC, *p*C; el movimiento hacia el centro será igual en ambos, mientras el movimiento transversal del cuerpo *p* será al movimiento transversal del cuerpo P como el movimiento angular de la línea *p*C al movimiento angular de la línea PC, esto es, como el ángulo VC*p* al ángulo VCP. Por consiguiente, en el mismo tiempo en el que el cuerpo P con sus dos movimientos alcanza el punto K, el cuerpo *p* con movimiento igual hacia el centro se moverá igualmente desde *p* hacia C, y por tanto, al cumplirse dicho tiempo, se hallará en algún lugar sobre la línea *mkr* que es perpendicular a *p*C pasando por *k*; y con el movimiento transversal habrá adquirido una distancia de *p*C que será a la distancia adquirida por el otro cuerpo P respecto a

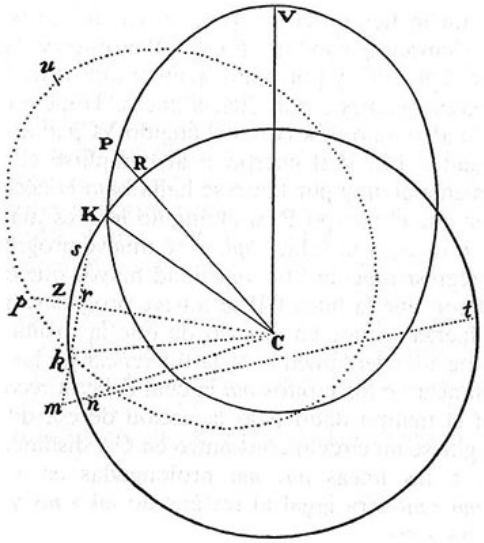
PC como el movimiento transversal del cuerpo *p* al movimiento transversal del cuerpo p. Por lo cual, al ser *kr* igual a la distancia adquirida por el cuerpo P respecto a la línea PC, y al ser *mr* a *kr* como el ángulo VC*p* al ángulo VCP, esto es, como el movimiento transversal del cuerpo *p* al movimiento transversal del cuerpo P, es evidente que al cumplirse dicho tiempo el cuerpo *p* se hallará en el punto *m*. Esto será así siempre que los cuerpos *p* y P se muevan igualmente según las líneas *p*C, PC, y por tanto siempre que se hallen bajo la acción de fuerzas iguales según dichas líneas. Tómese en cambio el ángulo *p*C*n* al ángulo *p*C*k* como el ángulo VC*p* al ángulo VCP y sea *n*C igual a *k*C, y el cuerpo *p* al cumplirse el tiempo se hallará realmente en *n*; y por tanto se halla bajo la acción de una fuerza mayor que el cuerpo P, si el ángulo *n*C*p* es mayor que el ángulo *k*C*p*, esto es, si la órbita *upk*, o se mueve progresivamente o se mueve regresivamente, con velocidad mayor que el doble de la velocidad con que la línea CP se mueve progresivamente; y es urgido con fuerza menor en el caso de que la órbita se mueva regresivamente más lentamente. Y la diferencia de las fuerzas es como la distancia de los puntos *mn* la cual debería recorrer dicho cuerpo *p* en el tiempo dado bajo la acción de esa diferencia de fuerzas. Imagínese un círculo con centro en C y distancia C*n* o C*k* que cortase a las líneas *mr*, *mn* prolongadas en *s* y *f*, y el rectángulo *mn* x *mt* será igual al rectángulo *mk* x *ms* y por tanto *mn* igual a $\frac{mk \text{ x } ms}{mt}$. Como, por otra parte, los triángulos *p*C*k*, *p*C*n* para un tiempo dado alcancen magnitudes dadas, *kr* y *mr*, y su diferencia *mk* y su suma *ms* son inversamente como la altura *p*C, por tanto, el rectángulo *mk* x *ms* es inversamente como el cuadrado de la altura *p*C. Y además, *mt* es directamente como ½*mt*, esto es, como la altura *p*C. Estas son las razones primeras de las lineas nacientes; y por ello ocurre que $\frac{mk \text{ x } ms}{mt}$, esto es, que el segmento naciente *mn*, y la diferencia de fuerzas proporcional a ella sea como el inverso del cubo de la altura *p*C. Q. E. D.



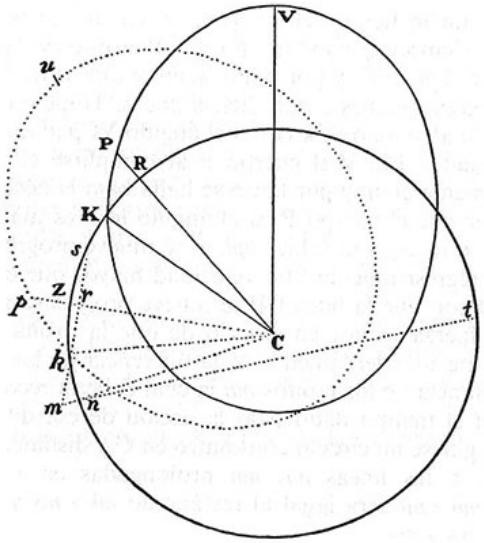
COROLARIO 1. De aquí que la diferencia de fuerzas en los puntos P y *p*, o K y *k*, es a la fuerza con que un cuerpo podría girar con movimiento circular desde R a K en el mismo tiempo en que el cuerpo P describe en la órbita fija el arco PK, como el segmento naciente *mn* al seno verso del arco naciente RK, esto es, como $\frac{mk \text{ x } ms}{mt}$ a $\frac{rk^2}{2k\text{C}}$, o como *mk* x *ms* a *rk* cuadrado; esto es, si se toman cantidades dadas, F, G, mutuamente en la misma proporción que tienen entre sí el ángulo VCP a VC*p*, como GG - FF a FF. Y por tanto, si con centro en C y distancia cualquiera CP o C*p* se traza un sector circular igual al área de todo VPC, el cual ha sido descrito por el cuerpo P girando en la órbita inmóvil en mi tiempo cualquiera con el radio trazado al centro, la diferencia de fuerzas, con las que el cuerpo P en la órbita inmóvil y el cuerpo *p* en la órbita en revolución están girando, será a la fuerza centrípeta con la que otro cuerpo con radio trazado al centro puede describir uniformemente ese sector en el mismo tiempo en que se describe el área VPC, como GG - FF a FF. Pues dicho sector y el área *p*C*k* son entre sí como los tiempos en que son descritos.

COROLARIO 2. Si la órbita VPK fuese una elipse con su foco en C y su ápside superior en V; y se supone la elipse *upk* semejante e igual a ella de modo que *p*C sea siempre igual a PC y el ángulo VC*p* respecto a VCP en una razón dada G a F, y por la altura PC o *p*C se escribe A, y por el «latus rectum» de la elipse 2R: la fuerza, con la que un

cuerpo puede girar en la elipse móvil, sera como $\frac{FF}{AA} + \frac{RGG - RFF}{A^3}$ y viceversa. Represéntese la fuerza con la cual un cuerpo puede girar en una elipse fija por la cantidad $\frac{FF}{AA}$ y la fuerza en V será $\frac{FF}{CV^2}$. Pero la fuerza con la que un cuerpo puede girar en un círculo a la distancia CV con la velocidad que tiene un cuerpo en V girando en una elipse, es a la fuerza con la que un cuerpo girando en una elipse se ve urgido en el ápside V, como la mitad del «latus rectum» de la elipse al semidiámetro CV del círculo, siendo por tanto $\frac{RFF}{CV^3}$; y la fuerza que sea a ésta como GG - FF a FF vale $\frac{RGG - RFF}{CV^3}$; y esta fuerza (por el Corolario 1 de esta Proposición) es la diferencia de fuerzas en V con las cuales giran el cuerpo P en la elipse inmóvil VPK y el cuerpo *p* en la elipse móvil *upk*. De donde, puesto que (por esta Proposición) dicha diferencia en otra altura cualquiera A es a sí misma en la altura CV como a $\frac{1}{A^3}$ a $\frac{1}{CV^3}$, dicha diferencia para toda altura A valdrá $\frac{RGG - RFF}{A^3}$. Por lo tanto añádase a la fuerza $\frac{FF}{AA}$ con la cual un cuerpo puede girar en una elipse fija VPK el exceso $\frac{RGG - RFF}{A^3}$ y se tendrá la fuerza total $\frac{FF}{AA} + \frac{RGG - RFF}{A^3}$ con la cual puede un cuerpo girar en la elipse móvil *upk* en tiempos iguales.

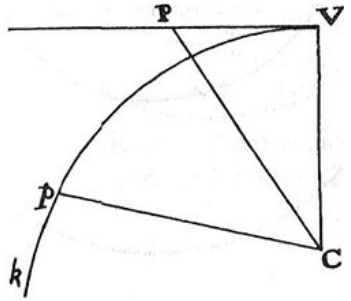


COROLARIO 3. De igual modo se hallará que si la órbita inmóvil VPK fuese una elipse cuyo centro está en el centro de fuerzas C; y si se supone la elipse móvil *upk* semejante, igual y concéntrica a ella; y fuese 2R el «latus rectum» principal de tal elipse, y 2T el «latus transversum» o eje mayor, y el ángulo VC*p* es constantemente a VCP como G a F; las fuerzas, con las cuales los cuerpos pueden girar en la elipse fija y en la móvil en tiempos iguales, serán respectivamente como $\frac{FFA}{T^3}$ y $\frac{FFA}{T^3} + \frac{RGG - RFF}{A^3}$.



COROLARIO 4. Y, en general, si llamamos T a la altura máxima del cuerpo CV, y llamamos R al radio de la curvatura que tiene en V la órbita VPK, esto es, al radio de un círculo de curvatura igual, y se denomina $\frac{FFA}{TT}$ a la fuerza centrípeta con que un cuerpo puede girar en cualquier trayectoria inmóvil VPK ni el punto V, mientras en otros puntos P se denomina indeterminadamente X, y se denomina A a la altura CP, y se loma G respecto a F en la razón dada del ángulo VC*p* al ángulo VCP: la fuerza centrípeta con la cual el mismo cuerpo podría o alizar los mismos movimientos en los mismos tiempos sobre la misma trayectoria movida circularmente *upk* será como la suma de las fuerzas $X + \frac{VRGG - VRFF}{A^3}$.

COROLARIO 5. Por lo tanto, dado el movimiento de un cuerpo en una órbita fija cualquiera, su movimiento angular en torno al centro de fuerzas puede aumentarse o disminuirse en una razón dada, de donde se pueden encontrar órbitas fijas nuevas en las cuales giren los cuerpos con nuevas fuerzas centrípetas.



COROLARIO 6. Así pues, si sobre la recta CV de posición dada se eleva la perpendicular VP de longitud indeterminada y se une CP y se traza C*p* igual a la anterior formando el ángulo VC*p* tal que sea al ángulo VCP según una razón dada; la fuerza con que un cuerpo puede girar en la curva V*pk*, que describe continuamente el punto *p*, será inversamente como el cubo de la altura C*p*. Pues el cuerpo P, por la fuerza de inercia, no actuando ninguna otra fuerza, continuará por la recta VP. Añádase la fuerza tendente hacia el centro C inversamente proporcional al cubo de la altura CP o C*p* y (por lo que se acaba de demostrar) el movimiento aquel rectilíneo se desviará según la línea curva V*pk*. Pero esta curva V*pk* es la misma que aquella curva VPK hallada en el Corolario 3 de la Proposición XLI por la cual, dijimos allí, ascienden oblicuamente los cuerpos atraídos por fuerzas como éstas.

PROPOSICIÓN XLV. PROBLEMA XXXI[33]

Hallar el movimiento de ápsides en órbitas muy próximas a círculos.

Se resuelve aritméticamente este problema haciendo que la órbita descrita en un plano fijo por un cuerpo que gira en una elipse móvil (como en los Corolarios 2 y 3 de la Proposición anterior) se convierta en la figura de la órbita cuyos ápsides se buscan, y calculando después los ápsides de la órbita descrita por bicho cuerpo en el plano fijo. Pero las órbitas devienen de formas iguales si las fuerzas centrípetas con las que son descritas, comparadas entre sí, resultan proporcionales a alturas iguales. Sea el punto V el ápside superior, y escríbase T por la altura máxima CV, A por la altura cualquiera CP o C*p*, y X por la diferencia de alturas CV - CP; entonces, la fuerza con la que un cuerpo se mueve en la elipse que gira en torno a su foco C (como en el Corolario 2) y que en el Corolario 2 era como $\frac{FF}{AA} + \frac{RGG - RFF}{A^3}$, esto es, como $\frac{FFA + RGG - RFF}{A^3}$, sustituyendo T - X por A, sera como

$\frac{RGG - RFF + TFF - FFX}{A^3}$. De modo semejante ha de reducirse otra fuerza centrípeta a una fracción cuyo denominador sea A^3 y los numeradores habrán de ser análogos mediante la correlación de términos homólogos entre sí. Resultará más claro con ejemplos.

EJEMPLO I. Supongamos que la fuerza centrípeta es uniforme y, por consiguiente, como $\frac{A^3}{A^3}$, o (escribiendo T - X por A en el numerador) como $\frac{T^3 - 3TTX + 3TXX - X^3}{A^3}$; y al poner en correspondencia los términos de los numeradores, a saber, los valores dados con los dados y los no dados con los no dados, resultará RGG - RFF + TFF a T^3 como -FFX a $-3TTX + 3TXX - X^3$ o como -FF a $- 3TT + 3TX - XX$. Al suponer ahora que la órbita se aproxima mucho a un círculo, hágase coincidir con él, y al hacerse iguales R y T, y disminuir X a la vez hasta el infinito, las razones últimas serán RGG a T^3 como -FF a -3TT, o también, GG a TT como FF a 3TT y, viceversa, GG a FF como TT a 3TT, esto es, como 1 a 3; y por tanto G a F, esto es, el ángulo VC*p* al ángulo VCP como 1 a √3. Por tanto, dado que un cuerpo en una elipse inmóvil completa al caer desde el ápside superior al inferior un ángulo (por así decirlo) de 180 grados; el otro cuerpo en una elipse móvil, pero también en la órbita inmóvil de que tratamos, al caer desde el ápside superior al inferior completará un ángulo de $\frac{180}{\sqrt{3}}$ grados: esto también gracias a la semejanza de esta órbita, descrita por un cuerpo bajo la acción de una fuerza centrípeta uniforme, con la órbita descrita por un cuerpo que gira en una elipse en revolución sobre un plano fijo. Por la correlación de términos anterior devienen dichas órbitas semejantes, no universalmente sino sólo cuando se aproximan muy sensiblemente a la forma circular. Por tanto, cuando un cuerpo bajo la acción de una fuerza centrípeta gira en una órbita casi circular completará siempre entre los ápsides superior e inferior un ángulo de $\frac{180}{\sqrt{3}}$ grados, o sea, de 103.º, 55′, 23″ en el centro; cada vez que pasa del ápside superior al inferior completa dicho ángulo, y al volver desde aquí hasta el ápside superior de nuevo completa el mismo ángulo; y así siempre hasta el infinito.

EJEMPLO II. Supongamos que la fuerza centrípeta fuese como una potencia cualquiera de la altura A tal como A^{n-3}, o $\frac{A^n}{A^3}$, donde *n*-3 y *n* significan exponentes enteros o fraccionarios cualesquiera, racionales o irracionales, positivos o negativos. Dicho

numerador A^{n-3} o $(T - X)^n$ convertido en serie indeterminada por nuestro método de series convergentes devendrá en $T^n - nXT^{n-1} + \frac{nn - n}{2} XXT^n$, etc. Y comparados sus términos con los del otro numerador RGG - RFF + TFF - FFX resulta que RGG - RFF + TFF es a T^n como -FF a $-nT^{n-1} + \frac{nn - n}{2} XT^{n-2}$, etc. Y tomando las razones últimas en el momento en que las órbitas se aproximan a la forma de círculos RGG será a T^n como -FF a $-nT^{n-1}$ o también GG a T^{n-1} como FF a nT^{n-1} y alternativamente, GG a FF como T^{n-1} a nT^{n-1}, esto es, como 1 a *n*; y por tanto G es a F, esto es, el ángulo VC*p* es al ángulo VCP, como 1 a √*n*. Por lo cual, al ser de 180 grados el ángulo descrito por el cuerpo en su descenso desde el ápside superior al inferior de una elipse, el ángulo VC*p*, descrito en el descenso del cuerpo desde el ápside superior al inferior de una órbita cuasicircular la cual es descrita por el cuerpo que gire con una fuerza centrípeta proporcional a la potencia A^{n-3}, será igual a un ángulo de $\frac{180}{\sqrt{n}}$ grados, y al repetirse este ángulo el cuerpo regresará al ápside superior y así hasta el infinito. De modo que si la fuerza centrípeta fuese como la distancia del cuerpo al centro, esto es, como A o como $\frac{A^4}{A^3}$ entonces *n* será igual a 4, y √*n* igual a 2; y por lo tanto, el ángulo entre ápsides superior e inferior igual a $\frac{180}{2}$ grados, o sea, de 90 grados. Por consiguiente, al completar la cuarta parte de una revolución el cuerpo llegará al ápside inferior, y al completar otra cuarta parte llegará al ápside superior, y así cada vez hasta el infinito. Lo cual es también evidente por la Proposición X. Pues un cuerpo bajo la acción de dicha fuerza centrípeta gira en una elipse inmóvil cuyo centro está en el centro de fuerzas. Pero si la fuerza centrípeta es inversamente como la distancia, esto es, como $\frac{1}{A}$ o como $\frac{A^2}{A^3}$, *n* será igual a 2, y por tanto el ángulo entre ápsides superior e inferior sera de $\frac{180}{\sqrt{2}}$ o sea de 127.º, 16′, 45″ y por lo mismo, el cuerpo que gire con semejante fuerza se moverá desde el ápside superior al inferior y de éste al superior repitiendo siempre este ángulo. Pero si la fuerza fuese como el inverso de la raíz cuarta de la onceava potencia de la altura, esto es, inversamente como $A^{11/4}$, y por tanto, directamente como $\frac{1}{A^{11/4}}$, o como $\frac{A^{1/4}}{A^3}$, entonces *n* sera igual a ¼ y $\frac{180}{\sqrt{n}}$ grados sera igual a 360 grados y por tanto el cuerpo

que descienda desde el ápside superior en su descenso continuo desde él alcanzará el ápside inferior cuando complete una revolución completa, y después, al cumplir desde allí en su ascenso continuo otra revolución completa, regresará al ápside superior; y así sucesivamente para siempre.

EJEMPLO III. Si se toman *m* y *n* por cualesquiera índices de potencias de la altura y *b*, *c*, por números dados cualesquiera, supongamos que la fuerza centrípeta fuese como $\frac{bA^m + cA^n}{A^3}$, esto es, como

$$\frac{b(T - X)^m + c(T - X)^n}{A^3}$$

o (por nuestro método de series convergentes) como

$$\frac{bT^m + cT^n - mbXT^{m-1} - ncXT^{n-1} + \frac{mm-m}{2}bXXT^{m-2} + \frac{nn-n}{2}cXXT^{n-2}\ldots}{A^3},$$

y al comparar los términos de los numeradores ocurrirá que RGG-RFF + TFF será a $bT^m + cT^n$ como -FF a $-mbT^{m-1} - ncT^{n-1} + \frac{mm-m}{2}bXT^{m-2} + \frac{nn-n}{2}cXT^{n-2}$, etc. Y tomando las razones últimas resultantes cuando las órbitas se aproximan a la forma circular resultará que GG es a $bT^{m-1} + cT^{n-1}$ como FF a $mbT^{m-1} + ncT^{n-1}$ y, viceversa, GG a FF como $bT^{m-1} + cT^{n-1}$ a $mbT^{m-1} + ncT^{n-1}$. Expresando la altura máxima CY o T aritméticamente por la unidad, esta proporción se transforma en GG a FF como $b + c$ a $mb + nc$ y, por tanto, como 1 a $\frac{mb+nc}{b+c}$. De donde G es a F, esto es, el ángulo VC*p* al ángulo CVP, como 1 a $\sqrt{\frac{mb+nc}{b+c}}$. Y por tanto, al ser de 180 grados el ángulo VCP entre los ápsides superior e inferior en la elipse fija, el ángulo VC*p* entre los mismos ábsides en la órbita descrita por un cuerpo con fuerza centrípeta proporcional a $\frac{bA^m + cA^n}{A^3}$, será igual a un ángulo de 180.º $\sqrt{\frac{b+c}{mb+nc}}$. Y por el mismo argumento si la fuerza centrípeta fuese como $\frac{bA^m - cA^n}{A^3}$, el ángulo entre ábsides resultará de 180.º $\sqrt{\frac{b-c}{mb-nc}}$. Y el problema se resuelve de igual modo

en casos más difíciles. La cantidad, a la cual es proporcional la fuerza centrípeta, debe descomponerse siempre en series convergentes cuyo denominador sea A^3. Después hay que suponer en la misma razón la parte dada del numerador proveniente de dicha operación respecto a la parte no dada del mismo, la parte dada de este numerador RGG - RFF + TFF - FFX a su otra parte no dada: eliminando las cantidades superfluas y tomando como unidad a T, se obtiene la proporción G a F.

COROLARIO 1. De aquí que, si la fuerza centrípeta fuese como cualquier potencia de la altura, podría hallarse dicha potencia a partir del movimiento de los ápsides y viceversa. Efectivamente, si la totalidad del movimiento angular con el cual el cuerpo retorna al mismo ápside es al movimiento angular de una sola revolución o a 360 grados como un número *m* a otro número *n* y se denomina A a la altura: entonces la fuerza será como aquella potencia de la altura $A^{nn/mm-3}$ cuyo índice es $\frac{nn}{mm} - 3$. Lo cual está claro por el segundo ejemplo. De donde es evidente que dicha fuerza al alejarse del centro no puede decrecer en razón mayor que el cubo de la altura: un cuerpo girando con tal fuerza y partiendo del ápside si empieza a descender, nunca llegará al apside inferior o altura mínima, sino que caerá hasta el centro, describiendo aquella línea curva que hemos visto en el Corolario 3 de la Proposición XLI. Pero si al partir del ápside, éste fuese el inferior y comenzase a ascender mínimamente, ascenderá hasta el infinito y jamás llegará al ápside superior. Pues describirá aquella curva de que se trató en el mismo Corolario y en el Corolario 6 de la Proposición XLIV. Así cuando la fuerza, al alejarse del centro, decrece en razón mayor que el cubo de la altura, el cuerpo que parte de un ápside, en cuanto comience a descender o a ascender, o bien desciende hasta el centro o bien asciende hasta el infinito. Pero si la fuerza, al alejarse del centro, o decrece en razón menor que el cubo de la altura, o crece en una razón cualquiera de la altura, el cuerpo nunca descenderá hasta el centro, sino que en algún momento alcanzará el ápside inferior; y por el contrario, si un cuerpo asciende y desciende alternativamente desde un ápside a otro y jamás alcanza el centro, la fuerza o bien crece con la lejanía al centro, o bien decrece en razón menor que el cubo de la altura; y cuanto más rápidamente retorne un cuerpo de un ápside a otro, más se aleja la razón de las fuerzas de la razón cúbica. Si el cuerpo hubiese de retornar desde y hasta el ápside superior mediante ascensos y descensos alternos en 8 revoluciones, o en 4, o en 2, o en 1½, esto es, si *m* fuese a *n* como 8 ó 4 ó 2 ó 1½ son a 1, y $\frac{nn}{mm} - 3$ valiese entonces $\frac{1}{64} - 3$, o $\frac{1}{16} - 3$, o $\frac{1}{4} - 3$, o $\frac{4}{9} - 3$, la fuerza será como $A^{1/64-3}$, o $A^{1/16-3}$, o $A^{1/4-3}$, o $A^{4/9-3}$, esto es, inversamente como $A^{3-1/64}$, o $A^{3-1/16}$, o $A^{3-1/4}$, o $A^{3-4/9}$. Si después de cada revolución el cuerpo regresa al mismo ápside y éste permanece inmóvil, entonces *m* será a *n* como 1 a 1 y por ello

$A^{nn/mm - 3}$ será igual a A^{-2}, o también $\frac{1}{AA}$; de donde el decremento de las fuerzas estará en razón cuadrada de la altura, como antes se ha demostrado. Si el cuerpo alcanzase el mismo ápside en tres cuartos, o en dos tercios, o en un tercio, o en un cuarto de revolución completa, entonces *m* sera a *n* como $\frac{3}{4}$, o $\frac{2}{3}$, o $\frac{1}{3}$, o $\frac{1}{4}$, a 1 respectivamente, y por lo tanto $A^{nn/mm - 3}$ será igual a $A^{16/9 - 3}$, o $A^{9/4 - 3}$, o A^{9-3}, o A^{16-3}, de tal modo que la fuerza o bien es inversamente como $A^{11/9}$, o $A^{3/4}$, o directamente como A^6 o A^{13}. Por fin, si el cuerpo en su recorrido desde el ápside superior hasta alcanzar de nuevo el mismo ápside hiciese una revolución completa y tres grados más y, por tanto, en cada revolución del cuerpo dicho ápside avanza tres grados, entonces *m* será a 363 grados a 360 grados, o como 121 a 120, y por tanto $A^{nn/mm - 3}$ será igual a $A^{-29523/14641}$; y en consecuencia la fuerza centrípeta será inversamente como $A^{29523/14641}$ o inversamente como $A^{2\frac{4}{243}}$ aproximadamente. La fuerza centrípeta decrece, por tanto, en razón algo mayor que la del cuadrado, pero esta razón es 59¾ veces más próxima a la razón cuadrada que a la razón cúbica.

COROLARIO 2. De aquí también que, si un cuerpo bajo la acción de una fuerza centrípeta que sea como el inverso del cuadrado de la altura girase en una elipse cuyo foco se hallase en el centro de fuerzas, y a dicha fuerza centrípeta se añadiese o se quitase otra fuerza externa cualquiera, podría conocerse (por el ejemplo tercero) el movimiento de los ápsides originado por dicha tuerza externa: y viceversa. Pues si la fuerza con la cual gira el cuerpo en la elipse fuese como $\frac{1}{AA}$, y la fuerza externa aplicada fuese como *c*A, y por tanto la fuerza restante fuese como $\frac{A - cA^4}{A^3}$; (en el ejemplo tercero) *b* será igual a 1, *m* igual a 1 y *n* igual a 4, y por lo tanto el ángulo de revolución entre ápsides será igual a un angulo de $180\sqrt{\frac{1 - c}{1 - 4c}}$ grados. Supongamos que la fuerza externa fuese 357,45 veces menor que la otra fuerza con que el cuerpo se mueve en la elipse, esto es, $\frac{100}{35745}$, siendo A o T igual a 1, y $180\sqrt{\frac{1 - c}{1 - 4c}}$ resultará valer $180\sqrt{\frac{35645}{35345}}$ o sea 180, 7623, esto es, 180.º, 45′, 44″. Por lo tanto, el cuerpo que parte del ápside superior con un movimiento angular de 180.º, 45′, 44″, llegará al ápside inferior, y repitiendo este movimiento regresará al superior: y en consecuencia el ápside superior vendrá a progresar en cada revolución 1.º, 31′, 28″.

Aproximadamente dos veces más veloz es el ápside de la Luna.

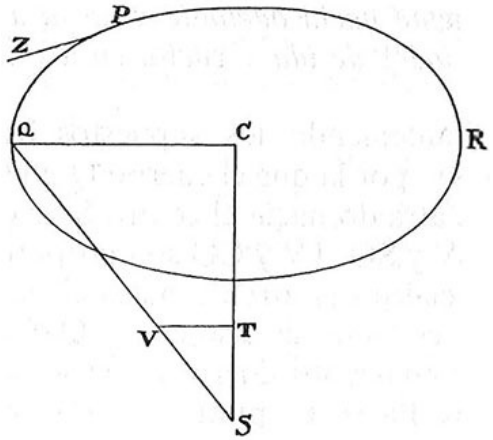
Hasta aquí sobre el movimiento de cuerpos en órbitas cuyos planos pasan por el centro de fuerzas. Falta determinar también los movimientos en planos excéntricos. Pues los autores que tratan el movimiento de los graves suelen considerar los ascensos y descensos de los pesos tanto perpendicular como oblicuamente en cualesquiera planos dados: y con igual razón se ha de considerar aquí el movimiento de cuerpos tendentes a un centro bajo cualesquiera fuerzas y moviéndose en planos excéntricos. Suponemos que dichos planos son perfectamente lisos y lubricados de modo que no retarden a los cuerpos. Además, en estas demostraciones, en lugar de planos sobre los que se deslizan los cuerpos y a los que tocan al deslizarse, utilizaremos planos paralelos a ellos, en los cuales se mueven los centros de los cuerpos y describen órbitas al moverse. Y por la misma ley determinamos después los movimientos de los cuerpos realizados en superficies curvas.

Sección X
DEL MOVIMIENTO DE CUERPOS EN SUPERFICIES DADAS, Y DEL MOVIMIENTO DE VAIVÉN DE LOS PÉNDULOS

Proposición xlvi. Problema xxxii

Supuesta una fuerza centrípeta de cualquier tipo, y dado tanto el centro de fuerzas como un plano cualquiera en el que gira el cuerpo, y concedidas las cuadraturas de las figuras curvas: hállese el movimiento de un cuerpo que parte de un lugar dado, con velocidad dada y según una recta dada en dicho plano.

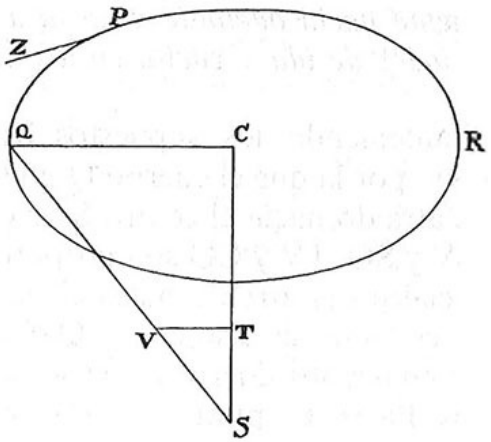
Sea S el centro de fuerzas, SC la distancia mínima desde ese centro al plano dado, P un cuerpo que parte del punto P según la recta PZ, Q el mismo cuerpo girando en su trayectoria, y PQR dicha trayectoria descrita en dicho plano y que ha de hallarse. Únanse CQ, QS, y si sobre QS se toma SV proporcional a la fuerza centrípeta con la cual el cuerpo es atraído hacia el centro S, y se traza VT paralela a CQ cortando a SC en T: la fuerza SV (por el Corolario II de las Leyes) se descompone en las fuerzas ST, TV, de las cuales ST atrayendo al cuerpo según la línea perpendicular al plano en nada cambia su movimiento en este plano. Pero la otra fuerza TV, actuando según la posición del plano, atrae al cuerpo directamente hacia el punto C dado en el plano y, por tanto, hace que por ello dicho cuerpo se mueva en ese plano como si la fuerza ST desapareciese y el cuerpo girase únicamente con la fuerza TV en torno al centro C en un espacio libre. Mas, estando dada la fuerza centrípeta TV con la que el cuerpo Q gira en un espacio libre en torno al centro dado C, está dada (por la Proposición XLII) también la trayectoria PQR, descrita por el cuerpo, así como el punto Q donde se hallará el cuerpo en un tiempo cualquiera dado y, finalmente, la velocidad del cuerpo en dicho punto Q. Y viceversa. Q. E. I.



Proposición xlvii. Teorema xv

Supuesto que la fuerza centrípeta sea proporcional a la distancia desde el cuerpo al centro, todos los cuerpos que giren en planos cualesquiera describirán elipses y completarán sus revoluciones en tiempos iguales; mientras que aquellos que se muevan según líneas rectas, alternativamente hacia adelante y hacia atrás, completarán sus diferentes períodos de ida y vuelta en los mismos tiempos.

Puesto que, manteniendo los supuestos de la Proposición anterior, la fuerza SV, por la que el cuerpo Q que gira en un plano cualquiera PQR es atraído hacia el centro S, es como la distancia SQ; y puesto que SV y SQ, TV y CQ son proporcionales, la fuerza TV, con la cual el cuerpo es atraído hacia el punto dado C en el plano de la órbita, es como la distancia CQ. Consecuentemente, las fuerzas, por las cuales los cuerpos que se hallan en el plano PQR son atraídos hacia el punto C, son proporcionales, a distancias iguales, a las fuerzas por las que los mismos cuerpos son atraídos en dondequiera hacia el centro S; y por tanto, los cuerpos se moverán en iguales tiempos y con iguales figuras, en cualquier plano PQR que rodee al punto C, tal y como lo harían en espacios libres en torno al centro S; y en consecuencia (por el Corolario 2 de la Proposición X y el Corolario 2 de la Proposición XXXVIII) siempre en tiempos iguales, o bien describirán elipses en dicho plano en torno al centro C, o bien completarán períodos moviéndose adelante y atrás según líneas rectas trazadas en dicho plano por el centro C. Q. E. D.



ESCOLIO

Afines a éstos son los ascensos y descensos de los cuerpos en superficies curvas. Imagínense líneas curvas descritas en un plano que giran luego en torno a ciertos ejes dados que pasan por el centro de fuerzas y, mediante esa revolución, describen superficies curvas; y también que los cuerpos se mueven de tal forma que sus centros se encuentren siempre en dichas superficies. Si tales cuerpos ascendiendo y descendiendo oblicuamente marchan hacia adelante y hacia atrás, realizarán sus movimientos en planos que pasan por el eje y, por tanto, en líneas curvas por cuya revolución se generaron aquellas superficies. En estos casos, por consiguiente, será suficiente considerar el movimiento en dichas líneas curvas.

PROPOSICIÓN XLVIII. TEOREMA XVI

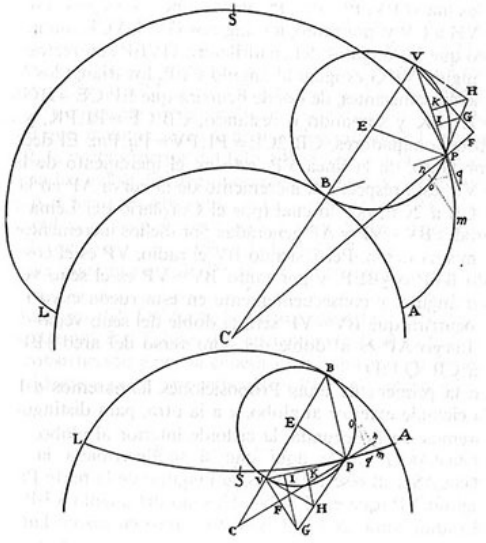
Si una rueda se apoya en el exterior de un globo formando ángulo recto y, girando como una rueda, describiese un círculo máximo, la longitud del camino curvilíneo descrito por un punto cualquiera dado en el perímetro de la rueda desde que toca al globo (que se puede llamar cicloide o epicicloide) será al doble del seno verso de la mitad del arco que ha tocado al globo al ir recorriéndolo desde aquel momento, como la suma de los diámetros del globo y la rueda al semidiámetro del globo.

PROPOSICIÓN XLIX. TEOREMA XVII

Si una rueda se apoya en el interior de un globo cóncavo formando ángulo recto y

avanza rodando por un círculo máximo, la longitud del camino curvilíneo descrito por un punto cualquiera dado en el perímetro de la rueda desde que toca al globo será al doble del seno verso de la mitad del arco que ha tocado al globo al ir recorriéndolo durante todo ese tiempo, como la diferencia de los diámetros del globo y la rueda al semidiámetro del globo.

Sea ABL el globo, C su centro, BPV la rueda que se apoya en él, E el centro de la rueda, B el punto de contacto y P el punto dado en el perímetro de la rueda. Imaginemos a esta rueda moviéndose sobre el círculo máximo ABL desde A por B hacia L, y girando de tal modo al desplazarse que los arcos AB, PB sean siempre iguales entre sí y mientras tanto el punto P dado en el perímetro de la rueda va describiendo el camino curvilíneo AP. Sea AP el total del camino curvilíneo descrito desde que la rueda tocó al globo en A, y la longitud AP de este camino será al doble del seno verso del arco ½PB como 2CE a CB. Pues toque la recta CE (prolongada si es preciso) a la rueda en V, y únanse CP, BP, EP, VP y descienda VF normal sobre la prolongación de CP. Sean PH, VH tangentes al círculo en P y V y concurrentes en H, y corte PH a VF en G, y sobre VP caigan las perpendiculares GI, HK. De nuevo con centro en C e intervalo cualquiera descríbase el círculo *nom* que corte a la recta CP en *n*, al perímetro BP de la rueda en *o*, y al camino curvilíneo AP en m; y con centro en V e intervalo V*o* descríbase el círculo que corta en *q* a VP prolongada.



Puesto que la rueda al moverse gira siempre en torno al punto de contacto B, es evidente que la recta BP es perpendicular a aquella línea curva AP descrita por el punto P de la rueda y, por lanío, que la recta VP es tangente a dicha curva en el punto P. El radio del círculo *nom* aumentado o disminuido gradualmente iguale siempre a la

distancia CP; y por la semejanza de la figura evanescente P*nomq* y de la figura PFGVI, la razón última de los segmentos evanescentes P*m*, P*n*, P*o*, P*q*, esto es, la razón de las variaciones instantáneas de la curva AP, de la recta CP, del arco circular BP, y de la recta VP será respectivamente la misma que la de las líneas PY, PF, PG, PI. Pero como VF es perpendicular a CF y VH a CV y, por tanto, los ángulos HYG, VCF son iguales; y (puesto que los ángulos del cuadrilátero HVEP son rectos en Y y P) el ángulo VHG es igual al ángulo CEP, los triángulos VHG y CEP serán semejantes; de donde ocurrirá que EP: CE = HG: HV, o HP = KI: PK y sumando o restando, CB: CE = PI: PK, y duplicando denominadores, CB: 2CE = PI: PV = P*q*: Pm. El decremento, por tanto, de la línea VP, esto es, el incremento de la línea BV - VP, está respecto al incremento de la curva AP en la razón dada CB a 2CE, por lo cual (por el Corolario del Lema IV) las longitudes BV - VP y AP generadas por dichos incrementos están en la misma razón. Pero, siendo BV el radio, VP es el coseno del ángulo BVP o ½BEP, y por tanto BV - VP es el seno verso del mismo ángulo; y consecuentemente en esta rueda cuyo radio es ½BV, ocurrirá que BV - VP será el doble del seno verso del arco ½BP. Luego AP es al doble del seno verso del arco ½BP como 2CE a CB. Q. E. D.

En la primera de estas Proposiciones llamaremos a la línea AP la cicloide exterior al globo, y a la otra, para distinguirlas, la llamaremos, en la segunda, la cicloide interior al globo.

COROLARIO 1. De aquí que si se describiera la cicloide completa ASL, al bisecarla en S, la longitud de la parte PS será a la longitud VP (que es el doble del seno del ángulo VBP, siendo EB el radio) como 2CE a CB, y por tanto en razón dada.

COROLARIO 2. Y la longitud del semiperímetro de la cicloide AS será igual a la recta que sea al diámetro BV de la rueda como 2CE a CB.

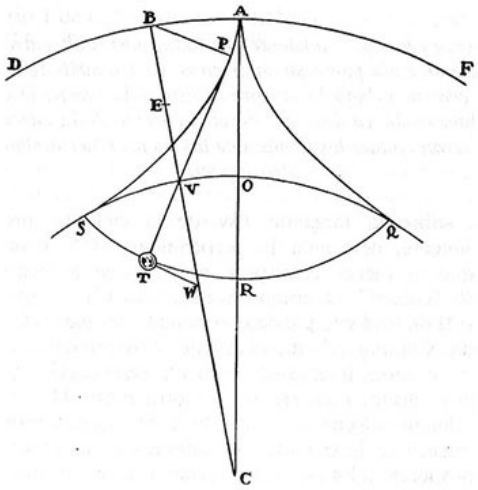
PROPOSICIÓN L. PROBLEMA XXXIII

Hacer oscilar a un cuerpo pendular en una cicloide dada.

Dentro del globo QVS, descrito con centro en C, sea dada la cicloide QRS bisecada en R y que toca a ambos lados con sus puntos extremos Q y S a la superficie del globo. Trácese CR bisecante del arco QS en O y prolónguese hasta A, de modo que CA sea a CO como CO a CR. Y con centro en C e intervalo CA descríbase el globo exterior DAF, y dentro de este globo descríbanse por medio de una rueda cuyo diámetro sea AO las tíos semicicloides AQ y AS tangentes al globo interior en Q y S y que toquen al globo exterior en A. Desde dicho punto A, mediante un hilo APT igual en longitud a AR, suspéndase el cuerpo T y hágasele oscilar entre las semicicloides AQ, AS de modo tal que, cuantas veces se aparta el péndulo de la perpendicular AR, la parte superior del hilo AP se aplica a la semicicloide APS hacia

la cual se dirige el movimiento, plegándose en torno a ella como si fuera un obstáculo, y la parte restante PT a la cual no obstaculiza la cicloide permanezca extendida en línea recta; y el peso T oscilará en la cicloide dada QRS. Q. E. F.

Hágase, pues, que el hilo PT se encuentre con la cicloide QRS en T, y con el círculo QOS en V, y trácese CV; y elévense las perpendiculares BP, TW hasta la parte recta del hilo PT desde los puntos extremos P y T, tales que encuentren a la recta CV en B y W. Por construcción y por la génesis de las figuras semejantes AS, SR, es evidente que las perpendiculares PB, TW cortan sobre CV las longitudes VB, VW iguales a los diámetros OA, OR de las ruedas. Por lo tanto TP es a VP (doble del seno del ángulo VBP, siendo V el radio) como BW a BV, o AO 4 + OR a AO, esto es (siendo proporcionales CA a CO, CO a CR y por división AO a OR) como CA + CO a CA, o, bisecando a BY en E, como 2CE a CB. Por consiguiente (por el Corolario 1 de la Proposición XLIX) la longitud de la parte recta del hilo PT es siempre igual al arco PS de la cicloide, y el hilo completo APT es siempre igual a la mitad del arco cicloide APS, esto es (por el Corolario 2 de la Proposición XLIX) a la longitud AR. Y por lo mismo a la inversa, si el hilo permanece siempre igual a la longitud AR el punto T se moverá en la cicloide dada QRS. Q. E. D.



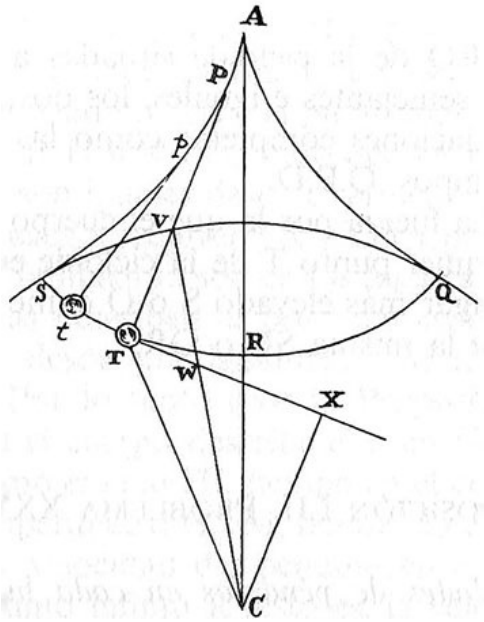
COROLARIO. El hilo AR es igual a la semicicloide AS y, por lo tanto, está en la misma razón respecto al semidiámetro AC del globo exterior, que la semicicloide SR semejante a aquélla tiene con respecto al semidiámetro CO del globo interior.

PROPOSICIÓN LI. TEOREMA XVIII

Si una fuerza centrípeta tendente por todas partes al centro C de un globo fuese en

cada punto singular como la distancia de cada lugar desde el centro, y bajo la acción de esta sola fuerza el cuerpo T oscilase (del modo ya descrito) en el perímetro de la cicloide QRS: digo que serán iguales los tiempos de las oscilaciones desiguales bajo otro aspecto.

Pues, sobre la tangente TW de la cicloide, prolongada indefinidamente, descienda la perpendicular CX y únase CT. Puesto que la fuerza centrípeta por la que el cuerpo T es impulsado hacia C es como la distancia CT y ésta (por el Corolario II de las Leyes) se descompone en las partes CX, TX, de las cuales CX empujando directamente al cuerpo desde P estira al hilo PT y se agota íntegramente en esa resistencia sin producir ningún otro efecto; mientras que la otra parte TX, al empujar transversalmente al cuerpo, o sea, hacia X, acelera directamente su movimiento en la cicloide; es evidente que la aceleración del cuerpo, proporcional a esta fuerza aceleratriz, es en cada momento como la longitud TX, esto es, por estar dadas CV, WV y las proporcionales a ellas TX, TW, como la longitud TW, esto es (por el Corolario 1 de la Proposición XLIX) como la longitud del arco cicloide TR. Por tanto, si se desvían desigualmente de la perpendicular AR dos péndulos APT, A*pt* y se dejan caer a la vez, sus aceleraciones serán siempre como los arcos a describir TR, *t*R. Pero las partes descritas en el inicio son como las aceleraciones, esto es, como el total a describir desde el principio y, por lo tanto, las partes que restan por describir y las aceleraciones subsiguientes, proporcionales a estas partes, son también como el total; y así sucesivamente. Por tanto las aceleraciones, así como las velocidades generadas y las partes descritas con estas velocidades y las partes a describir son siempre como el total; y en consecuencia, las partes a describir, al conservar entre sí una razón dada, evanescen a la vez, esto es, los dos cuerpos oscilantes llegarán a la vez a la perpendicular AR. Y puesto que el respectivo ascenso de los péndulos desde el punto inferior R, realizado a través de los mismos arcos cicloides con movimiento retrógrado, es retardado en cada lugar por las mismas fuerzas por las que eran acelerados los descensos, es evidente que las velocidades de los ascensos y descensos realizados por los mismos arcos son iguales y que, por tanto, se cumplen en tiempos iguales; y por lo mismo, como las dos partes RS y RQ de la cicloide situadas a cada lado de la perpendicular son semejantes e iguales, los dos péndulos realizarán, tanto las oscilaciones completas como las mitades, siempre en los mismos tiempos. Q. E. D.



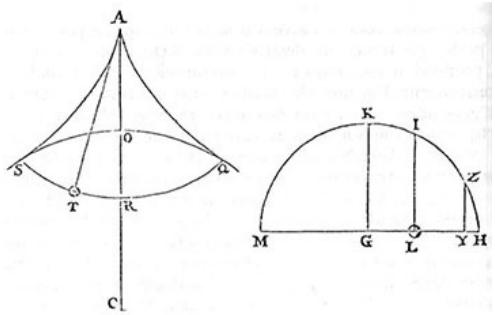
COROLARIO. La fuerza por la que el cuerpo T es acelerado o retardado en cualquier punto T de la cicloide es al peso total de ese cuerpo en el lugar más elevado S o Q como el arco TR de la cicloide al arco de la misma SR o QR.

PROPOSICIÓN LII. PROBLEMA XXXIV

Definir las velocidades de péndulos en cada lugar así como los tiempos en que se completan tanto las oscilaciones como cada parte de ellas.

Con centro cualquiera G, con radio GH igual al arco de la cicloide RS, trácese el semicírculo HLM bisecado por el semidiámetro GK. Y si una fuerza centrípeta proporcional a las distancias de los lugares al centro tendiera hacia el centro G y fuera en el perímetro HIK igual a la fuerza centrípeta tendente Inicia el mismo centro en el perímetro del globo QOS, y a la vez que se deja caer el péndulo desde el punto más alto S se deja caer otro cuerpo L desde H hasta G, dado que las fuerzas que actúan sobre los cuerpos son iguales desde el principio y proporcionales siempre a los espacios a recorrer TR y LG y por lo mismo, si TR y LG son iguales, también lo son en los lugares T y L, es evidente que dichos cuerpos describen al principio espacios iguales ST y HL y continúan desde allí igualmente urgidos y recorriendo espacios iguales. Por lo tanto (por la Proposición XXXVIII) el tiempo en el cual el cuerpo describe el arco ST es al tiempo de una oscilación como el arco HI, tiempo en el cual

el cuerpo H va hasta L, es al semiperímetro HKM, tiempo en el cual el cuerpo H va hasta M. Y la velocidad del péndulo en el punto T es a su velocidad en el punto ínfimo R (esto es, la velocidad del cuerpo H en el punto L a su velocidad en el punto G, o también el incremento momentáneo de la línea HL al incremento momentáneo de la línea HG, siendo constante la velocidad de crecimiento de los arcos HI y HK) como la ordenada LI al radio GK, o también como $\sqrt{(SR^2 - TR^2)}$ a SR. De donde, como para oscilaciones desiguales los arcos descritos en tiempos iguales son proporcionales a los arcos enteros de las oscilaciones, si se tienen dados los tiempos se tendrán tanto las velocidades como los arcos descritos en cualquiera oscilación. Que era lo primero que había que hallar[34].



Oscilen ahora los cuerpos pendulares en cicloides distintas descritas en el interior de globos diferentes cuyas fuerzas absolutas son también distintas; y, si se denomina V a la fuerza absoluta de un globo cualquiera QOS, la fuerza aceleratriz que actúa sobre el péndulo en la circunferencia de dicho globo cuando empieza a moverse directamente hacia su centro será como la distancia del cuerpo pendular a dicho centro conjuntamente con la fuerza absoluta del globo, esto es, como CO x V. Y, por tanto, el segmento HY, que es como la fuerza aceleratriz CO x V, será descrito en un tiempo dado: y si se eleva la perpendicular YZ que toque a la circunferencia en Z, el arco naciente HZ denotará dicho tiempo dado. Pero este arco naciente HZ es como la raíz cuadrada del rectángulo GHY y, por tanto, como $\sqrt{GH \times CO \times V}$. De aquí que el tiempo de una oscilación completa en la cicloide QRS (siendo directamente como el semiperímetro HKM, que denota esa oscilación completa; e inversamente como el arco HZ, que análogamente denota el tiempo dado) será directamente como GH e inversamente como $\sqrt{GH \times CO \times V}$, esto es, puesto que GH y SR son iguales, como $\sqrt{\frac{SR}{CO \times V}}$ o (por el Corolario de la Proposición L) como $\sqrt{\frac{AR}{AC \times V}}$. Por consiguiente, las oscilaciones en todos los globos y cicloides, efectuadas con cualesquiera fuerzas absolutas, están en razón compuesta directamente de la raíz cuadrada de la longitud de la cuerda e inversamente de la raíz cuadrada de la

distancia entre el punto de suspensión y el centro del globo y también inversamente de la raíz cuadrada de la fuerza absoluta del globo. Q. E. I.

COROLARIO 1. De aquí también que puedan compararse entre sí los tiempos de los cuerpos en oscilación, en caída o en giro. Pues si el diámetro de la rueda con la cual se describe la cicloide dentro del globo se supone igual al semidiámetro del globo, la cicloide resultará una línea recta que pasa por el centro del globo, y la oscilación será ahora un descenso y posterior ascenso en dicha recta. De donde están dados tanto el tiempo del descenso desde cualquier lugar hasta el centro como el tiempo igual a éste en el cual el cuerpo, girando uniformemente en torno al centro del globo y a cualquier distancia, describe el arco cuadrante. Pues dicho tiempo (por el Caso 2) es al tiempo de una semioscilación en toda cicloide QRS como 1 a $\sqrt{\frac{AR}{AC}}$.

COROLARIO 2. De aquí se siguen también los descubrimientos de *Wren* y de *Huygens* sobre la cicloide común. Pues si se aumenta infinitamente el diámetro de un globo, la superficie esférica del mismo se convertirá en un plano y la fuerza centrípeta actuará uniformemente en la dirección de las líneas perpendiculares a dicho plano, con lo que nuestra cicloide viene a ser una cicloide común. En este caso la longitud del arco de la cicloide entre ese plano y el punto descriptor resultará ser cuatro veces el seno verso del semiarco de la rueda entre dicho plano y el punto descriptor, como lo mostró *Wren*: y el péndulo entre dos cicloides semejantes oscilará con tiempos iguales en una cicloide semejante e igual, como demostró *Huygens*. Y también el descenso de los graves en el tiempo de una oscilación será el indicado por *Huygens*.

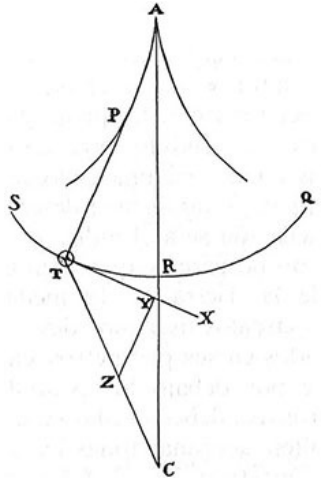
Se adaptan las proposiciones que hemos demostrado a la constitución real de la Tierra en la medida en que ruedas circulantes sobre sus círculos máximos describirán, con el movimiento de clavos fijados en sus perímetros, cicloides exteriores al globo; mientras que por debajo los péndulos suspendidos en minas y cavernas terrestres deben oscilar en cicloides interiores al globo para que resulten isócronas todas las oscilaciones. Pues la gravedad (como se mostrará en el Libro tercero) decrece al alejarse de la superficie terrestre; hacia arriba como el cuadrado de la distancia al centro y hacia abajo en razón simple.

PROPOSICIÓN LIII. PROBLEMA XXXV

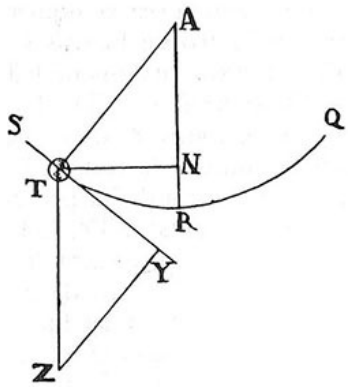
Concedidas las cuadraturas de figuras curvas, hallar las fuerzas con las cuales cuerpos que se muevan en curvas dadas realicen siempre sus oscilaciones en tiempos iguales.

Oscile el cuerpo T en una curva cualquiera STRQ, cuyo eje sea AR que pasa por

el centro de fuerzas C. Trácese TX, que loque a dicha curva en un lugar cualquiera T del cuerpo, y en esa tangente TX tómese TY igual al arco TR. La longitud de dicho arco es conocida por los métodos usuales para la cuadratura de figuras. Trácese desde el punto Y la recta YZ perpendicular a la tangente. Trácese CT que corte a dicha perpendicular en Z y la Fuerza centrípeta será proporcional a TZ. Q. E. I.



Ya que si la fuerza con la cual es atraído el cuerpo desde T hacia C se expresa por la recta TZ tomada proporcionalmente a ella, dicha fuerza se descompondrá en las fuerzas TY e YZ, de las cuales YZ, atrayendo el cuerpo según la longitud del hilo PT, en nada cambia su movimiento, mientras la otra fuerza TY acelera o retarda directamente su movimiento en la curva STRQ. Por tanto, al ser ésta como el espacio a recorrer TR, las aceleraciones o retardos del cuerpo al recorrer partes (mayor y menor) proporcionales de dos oscilaciones, siempre serán como esas partes, y por ende harán que esas partes siempre sean descritas simultáneamente. Pero los cuerpos que siempre describen a la vez partes proporcionales al todo, describirán los todos a la vez. Q. E. D.



COROLARIO 1. De aquí que si el cuerpo T pende del hilo rectilíneo AT desde el centro A y describe el arco circular STRQ, y entre tanto es urgido hacia abajo según líneas paralelas por una fuerza que sea a la fuerza uniforme de la gravedad como el arco TR a su seno TN, los tiempos de cada oscilación serán iguales. Puesto que TZ y AR son paralelas, los triángulos ATN y ZTY serán semejantes y, por tanto, TZ será a AT como TY a TN; esto es, si la fuerza uniforme de la gravedad se expresa por la longitud dada AT, la fuerza TZ, por la que las oscilaciones resultan isócronas, será a la fuerza AT de la gravedad como el arco TR, igual a TY, es a TN, seno del mismo.

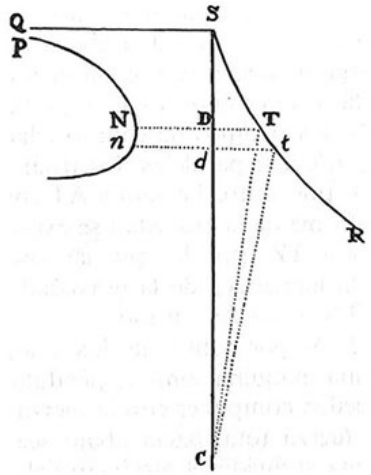
COROLARIO 2. Y por tanto en los relojes, si las fuerzas impresas por alguna máquina sobre el péndulo para conservar el movimiento se pueden componer con la fuerza de la gravedad de tal modo que la fuerza total hacia abajo sea siempre como la línea que se obtenga al dividir el producto del arco TR y el radio AR por el seno TN, todas las oscilaciones serán isócronas.

PROPOSICIÓN LIV. PROBLEMA XXXVI

Concedidas las cuadraturas de figuras curvas, hallar los tiempos en que con cualquier fuerza centrípeta los cuerpos descenderán o ascenderán por cualesquiera líneas curvas descritas sobre un plano que atraviesa el centro de fuerza.

Descienda un cuerpo desde un lugar cualquiera S a través de una cierta línea curva ST*t*R dada en un plano que pasa por el centro de fuerzas C. Únase CS y divídase en innumerables partes iguales, y sea una de ellas D*d*. Con centro en C y distancias CD y C*d* trácense los círculos DT *dt* que encuentran a la curva ST*r*R en T y *t*. Y estando dadas tanto la ley de la fuerza centrípeta como la altura CS desde la que cayó el cuerpo, estará dada la velocidad del cuerpo en cualquiera otra altura CT (por la Proposición XXXIX). Pero el tiempo en que el cuerpo describe el segmento T*t* es como la longitud de este segmento, esto es, directamente como la secante del

ángulo *t*TC; e inversamente como la velocidad. Sea proporcional a este tiempo la ordenada DN perpendicular a CS en el punto D, y, por estar dada D*d*, el rectángulo D*d* x DN, esto es el área DN*nd*, será proporcional a dicho tiempo. Por lo tanto, si PN*n* fuese aquella línea curva a la que siempre es tangente el punto N y su asíntota fuese la recta SQ que cae perpendicularmente sobre la recta CS, el área SQPND será proporcional al tiempo en que el cuerpo al caer ha descrito la línea ST; por tanto, hallada esta área estará dado el tiempo. Q. E. I.

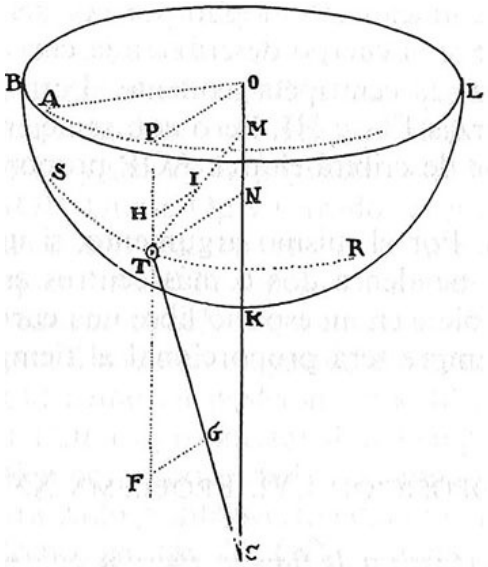


Proposición lv. Teorema xix

Si un cuerpo se mueve en cualquier superficie curva cuyo eje pasa por el centro de fuerzas y se traza una perpendicular desde el cuerpo al eje; y desde cualquier punto dado del eje se traza otra línea igual y paralela a la anterior, digo que dicha paralela describirá un área proporcional al tiempo.

Sea BKL la superficie curva, T el cuerpo que gira en ella, STR la trayectoria descrita en ella por el cuerpo, S el comienzo de la trayectoria, OMK el eje de la superficie curva, TN la perpendicular del cuerpo al eje, OP la paralela e igual a ésta trazada desde el punto dado O en el eje; sea AP la proyección de la trayectoria descrita por el punto P de la línea giratoria OP sobre el plano AOP; sea A el comienzo del trazo correspondiente al punto S, TC la recta trazada desde el cuerpo al centro, TG una parte de la misma proporcional a la fuerza centrípeta con que el cuerpo es ut raído hacia el centro C, TM una recta perpendicular a la superficie curva, TI una parte de la misma que es proporcional a la presión con la que el cuerpo actúa sobre la superficie y a la vez es repelido por ella hacia M; PTF una paralela al eje y que pasa

(por el Corolario ii de las Leyes) se descompone en las fuerzas TF, desde los puntos G e I sobre dicha paralela PHTF. Digo ahora que el área AOP, descrita por el radio OP desde el inicio del movimiento, es proporcional al tiempo. Puesto que la fuerza TG (por el Corolario 2 de las Leyes) se descompone en las fuerzas TF, FG; y la fuerza TI en las fuerzas TH, HI: ahora bien, las fuerzas TF, TH al actuar según la línea PF perpendicular al plano AOP sólo cambian el movimiento del cuerpo respecto a la perpendicular a dicho plano. Y por tanto su movimiento, en tanto que realizado según la posición del plano, esto es, el movimiento del punto P, con el que se describe el trazo de la trayectoria AP en dicho plano, es el mismo que si se suprimiesen las fuerzas TF, TH y el cuerpo fuese únicamente urgido por las fuerzas FG, HI; esto es, el mismo que si el cuerpo describiese la curva AP en el plano AOP con una fuerza centrípeta tendente al centro O e igual a la suma de las fuerzas FG y HI. Pero con semejante fuerza (por la Proposición i) se describirá el área AOP proporcional al tiempo. Q. E. D.



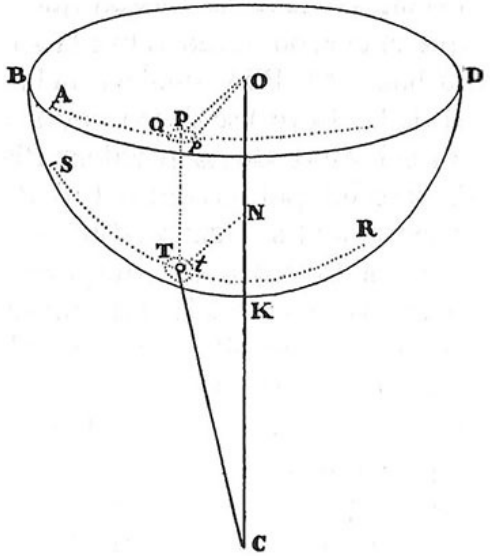
Corolario. Por el mismo argumento, si un cuerpo, urgido por fuerzas que tienden a dos o más centros en la misma recta dada CO, describiese en un espacio libre una curva cualquiera ST, el área AOP siempre será proporcional al tiempo.

Proposición lvi. Problema xxxvii

Concedida la cuadratura de figuras curvilíneas y supuesto que tanto la ley de la

fuerza centrípeta tendente a un centro dado como la superficie curva cuyo eje pasa por dicho centro están dadas, hallar la trayectoria que en esa superficie describirá un cuerpo que parte de un lugar dado, con una velocidad dada y hacia una dirección dada en esa superficie.

Manteniendo la construcción de la Proposición anterior, salga el cuerpo T desde el lugar dado S según una recta de posición dada hacia la trayectoria a hallar STR, cuya proyección en el plano BDO sea AP. Y al estar dada la velocidad del cuerpo en la altura SC estará dada su velocidad en cualquier otra altura TC. Con dicha velocidad describa en un tiempo dado mínimo la parte 17 de su trayectoria, y sea P*p* la proyección de la misma descrita sobre el plano AOP. Únase O*p* y trazado un pequeño círculo en la superficie curva con centro en T e intervalo T*t* sea su proyección en el plano AOP la elipse *p*Q. Y puesto que está dado en magnitud el pequeño círculo T*t*, y está dada su distancia TN o PO del eje CO, estará dada en magnitud dicha elipse *p*Q tanto en especie y magnitud como en posición respecto a la recta PO. Y dudo que el área PO*p* es proporcional al tiempo y está dada al estar dado el tiempo, estará dado el ángulo PO*p*. Y, por consiguiente, estará dado *p*, intersección común de la elipse y de la recta O*p*, así como el ángulo OP*p*, ángulo con el que la proyección AP*p* de la trayectoria corta a la línea OP. Y de aquí (aplicando la Proposición XLI y su Corolario 2) se ve fácilmente el modo de determinar AP*p*. Entonces, a partir de cada punto P de la proyección, elevando sobre el plano AOP las perpendiculares PT que corten a la superficie curva en T, quedará dado cada punto T de la trayectoria. Q. E. I.



Sección XI
DEL MOVIMIENTO DE CUERPOS QUE TIENDEN UNOS A OTROS CON FUERZAS CENTRÍPETAS

He expuesto hasta aquí los movimientos de cuerpos atraídos hacia un centro inmóvil, aunque puede que tal cosa no exista en la naturaleza de las cosas. Pues las atracciones suelen darse hacia los cuerpos, y las acciones de los cuerpos atrayentes y atraídos son siempre mutuas e iguales, por la Ley Tercera: hasta el punto de que, si fuesen dos cuerpos, ni el atrayente ni el atraído podrían estar en reposo, sino que ambos (por el Corolario IV de las Leyes) girarán en torno a un centro común de gravedad, como con atracción mutua; y, si fuesen varios los cuerpos, bien sean atraídos por uno y éste por los otros, bien se atraigan todos mutuamente, habrán de moverse entre sí de tal modo que o bien esté en reposo el centro común de gravedad o bien se mueva uniformemente en línea recta. Por lo cual paso ahora a exponer el movimiento de cuerpos que se atraen mutuamente, considerando a las fuerzas centrípetas como atracciones, aunque quizá, si hablásemos en términos físicos, se denominarían más propiamente impulsos. Pero ahora nos movemos en matemáticas y, por tanto, dejando de lado disputas físicas, hacemos uso de un lenguaje común en el cual podemos ser comprendidos más fácilmente por lectores matemáticos[35].

Proposición LVII. Teorema XX

Dos cuerpos que se atraen mutuamente describen figuras semejantes tanto en torno a su centro común de gravedad como uno en torno al otro mutuamente.

Pues las distancias de los cuerpos a su centro común de gravedad son inversamente proporcionales a los cuerpos, y se hallan, por tanto, en una razón dada entre sí, y componiendo razones, en una razón dada respecto a la distancia total entre cuerpos. Tales distancias se desplazan en torno a su extremo común con igual movimiento angular, puesto que al estar sobre una recta jamás cambian su inclinación mutua. Mas las líneas rectas que guardan entre sí una razón dada y giran en torno a sus extremos con igual movimiento angular en planos que, o bien reposan a la vez

que dichos extremos, o bien se mueven con cualquier movimiento no angular, describen en torno a tales extremos figuras totalmente semejantes. Por consiguiente, las figuras descritas con los giros de estas distancias, son semejantes. Q. E. D.

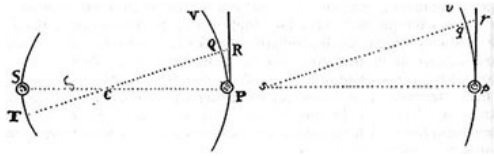
Proposición lviii. Teorema xxi

Si dos cuerpos se atraen mutuamente con fuerzas cualesquiera, y giran también en torno al centro común de gravedad, digo que con las mismas fuerzas y en torno a uno cualquiera de los dos cuerpos en reposo se puede describir una figura semejante e igual a las figuras que describen los cuerpos uno en torno a otro con los movimientos mencionados.

Giren los cuerpos S, P en torno al centro común de gravedad C, progresando de S a T y de P a Q: desde el punto dado *s* trácense *sp*, *sq* siempre iguales y paralelas a SP, TQ; y la curva *pqv*, descrita por el punto *p* al girar en torno al punto inmóvil *s*, será semejante e igual a las curvas descritas por los cuerpos S, P al girar mutuamente uno en torno a otro: por tanto (por el Teorema xx) será semejante a las curvas ST y PQV descritas por los mismos cuerpos en torno al centro común de gravedad C: y esto porque las proporciones de las líneas SC, CP, y SP o sp entre sí están dadas.

Caso 1. Dicho centro común de gravedad C, por el Corolario iv de las Leyes, o está en reposo o se mueve uniformemente en línea recta. Primero supongamos que reposa y que en *s* y *p* se sitúan dos cuerpos, uno inmóvil en *s* y otro móvil en *p*, semejantes e iguales a los cuerpos S y P. Hágase después que las rectas PR y *pr* sean tangentes a las curvas PQ y *pq* en P y *p* y prolónguense CQ y *sq* hasta R y *r*. Y por la semejanza de las figuras CPRQ, RQ será a *rg* como CP a *sp* y, por tanto, en una razón dada. En consecuencia, si la fuerza con la cual es atraído el cuerpo P hacia el cuerpo S y, por tanto, hacia el centro intermedio C, se hallase respecto a la fuerza con la cual el cuerpo *p* es atraído hacia el centro *s* en esa misma razón dada, esas fuerzas en tiempos iguales atraerán siempre a los cuerpos desde las tangentes PR y *pr* hacia los arcos PQ y *pq* a través de los intervalos RQ y *rq* proporcionales a dichas fuerzas; por ello, la segunda fuerza hará que el cuerpo *p* gire en la curva *pqv*, semejante a la curva PQV en la cual la primera fuerza hará girar al cuerpo P; y completarán revoluciones en los mismos tiempos. Pero como dichas fuerzas no son entre sí como CP a *sp* sino que (por la semejanza e igualdad de los cuerpos S y *s*, P y *p*, y la igualdad de las distancias SP, *sp*) son iguales entre sí, los cuerpos en tiempos iguales son atraídos igualmente de las tangentes; y por tanto, para que el cuerpo segundo *p* sea atraído a través de un intervalo mayor *rq*, se requiere un tiempo mayor y esto como la raíz cuadrada de los intervalos, por cuanto que (por el Lema x) los espacios recorridos en el mismo comienzo del movimiento son como el cuadrado de los tiempos. Supóngase, pues, que la velocidad del cuerpo *p* es a la velocidad del cuerpo P como

la raíz cuadrada de la distancia *sp* a la distancia CP, de tal modo que en tiempos que estén en esta misma razón se describan los arcos *pq*, PQ que están en razón simple: los cuerpos P, *p* atraídos siempre por fuerzas iguales describirán alrededor de los centros en reposo C y *s* las figuras semejantes PQV, *pqv*, de las cuales la segunda *pqv* es semejante e igual a la descrita por el cuerpo P al girar en tomo al cuerpo móvil S. Q. E. D.



Caso 2. Ahora supongamos que el centro común de gravedad junto con el espacio en el cual los cuerpos se mueven entre ellos, avanza uniformemente en línea recta; y (por el Corolario vi de las Leyes) todos los movimientos en tal espacio se producirán igual que antes y, por tanto, los cuerpos describirán entre sí las mismas figuras que antes, y consecuentemente iguales y semejantes a la figura *pqv*. Q. E. D.

Corolario 1. De aquí que dos cuerpos que se atraigan entre sí con fuerzas proporcionales a su distancia (por la Proposición x) describen alrededor de su centro común de gravedad y uno alrededor del otro elipses concéntricas; y viceversa, si las figuras descritas son tales, entonces las fuerzas son proporcionales a la distancia.

Corolario 2. Y dos cuerpos, con fuerzas inversamente proporcionales al cuadrado de la distancia entre ellos, describirán (por las Proposiciones xi, xii, xiii) tanto alrededor de su centro común de gravedad como el uno en torno al otro secciones cónicas que tienen su foco en el centro en torno al cual se describen las figuras. Y viceversa, si las figuras descritas son tales, entonces las fuerzas centrípetas son inversamente proporcionales al cuadrado de la distancia.

Corolario 3. Dos cuerpos cualesquiera girando en torno al centro común de gravedad describen áreas proporcionales a los tiempos con los radios trazados a dicho centro y a sí mismos.

Proposición lix. Teorema xxii

El tiempo periódico de dos cuerpos S y P en giro alrededor de su centro común de gravedad C es al tiempo periódico de uno de los cuerpos P en giro alrededor del otro S en reposo y describiendo una figura semejante e igual a las figuras que resultan al girar uno en torno a otro como la raíz cuadrada de un cuerpo S a la raíz cuadrada de la suma de los cuerpos S + P.

Pues, por la demostración de la Proposición anterior, los tiempos en que son

descritos cualesquiera arcos semejantes PQ y *pq* son como las raíces cuadradas de CP y SP o *sp*, esto es, como la raíz cuadrada del cuerpo S a la de la suma de los cuerpos S + P. Y componiendo, las sumas de los tiempos en que son descritos todos los arcos semejantes PQ y *pq*, esto es, los tiempos totales en que son descritas todas las figuras semejantes, se hallan en la susodicha razón subduplicada. Q. E. D.

Proposición LX. Teorema XXIII

Si dos cuerpos S y P que se atraen mutuamente con fuerzas inversamente proporcionales al cuadrado de sus distancias giran en torno a su centro común de gravedad: digo que el eje principal de la elipse que uno de los cuerpos, P, describe con su movimiento en torno al otro, S, será al eje principal de la elipse que dicho cuerpo P pudiera describir en el mismo tiempo periódico en torno al otro cuerpo S en reposo como la suma de los dos cuerpos S + P a la primera de las dos medias proporcionales entre esta suma y el otro cuerpo S.

Pues si las elipses descritas fuesen iguales entre sí, los tiempos periódicos (por el Teorema anterior) serán como la raíz cuadrada de la razón del cuerpo S a la suma de los cuerpos S + P. Disminúyase el tiempo periódico en la última elipse según la dicha razón y los tiempos periódicos se harán iguales; pero el eje principal de la elipse (por la Proposición XV) disminuirá en una razón, de la cual ésta es sesquiplicada, esto es, en una razón respecto a la cual la razón S a S + P es cúbica, por lo cual será al eje principal de la otra elipse como la primera de dos medias proporcionales entre S + P y S a S + P. Y, ala inversa, el eje principal de la elipse descrita en torno al cuerpo móvil será al eje principal de la descrita en torno al inmóvil como S + P a la primera de dos medias proporcionales entre S + P y S. Q. E. D.

Proposición LXI. Teorema XXIV

Si dos cuerpos se atraen mutuamente con cualquier fuerza y no son empujados ni obstruidos externamente y se mueven de cualquier modo; sus movimientos vienen a ser los mismos que si no se atrajesen entre ellos, sino que ambos fuesen atraídos por un tercero situado en el centro común de gravedad con las mismas fuerzas; y la, ley de tales fuerzas de atracción será la misma tanto respecto a la distancia a dicho centro común como respecto a la distancia total entre cuerpos.

Pues las fuerzas con las cuales los cuerpos se atraen mutuamente, al tender hacia los cuerpos, tienden al centro intermedio común de gravedad; y por eso son las mismas que si procediesen del cuerpo intermedio. Q. E. D.

Y como está dada la razón de la distancia de cada cuerpo respecto a dicho centro común a la distancia entre cuerpos, está dada la razón entre cualquier potencia de una distancia y la misma potencia de la otra distancia; y también la razón de cualquier cantidad derivada de cualquier forma de una distancia y de cantidades dadas, respecto a otra cantidad derivada similarmente de la otra distancia y de cuantas cantidades dadas que posean respecto a la primera la dicha razón dada de distancias. Por lo cual, si la fuerza por la que un cuerpo es atraído por otro fuese directa o inversamente como la distancia mutua entre cuerpos, o como cualquier potencia de esta distancia, o, finalmente, como cualquier cantidad derivada de cualquier forma de esta distancia y cantidades dadas: la fuerza por la cual el dicho cuerpo es atraído hacia el centro común de gravedad será también directa o inversamente como la distancia del cuerpo atraído al centro común de gravedad, o como la dicha potencia de esta distancia o, por fin, como la cantidad derivada de esta distancia y de análogas cantidades dadas. Esto es, la ley de la fuerza de atracción será la misma para una y otra distancia. Q. E. D.

Proposición LXII. Problema XXXVIII

Determinar los movimientos de dos cuerpos que atrayéndose mutuamente con fuerzas inversamente proporcionales a los cuadrados de la distancia entre ellos, se dejan caer desde lugares dados.

Los cuerpos (por el último Teorema) se moverán de modo semejante a como lo harían siendo atraídos por un tercero situado en el centro común de gravedad; dicho centro, por hipótesis, estará en reposo en el comienzo mismo del movimiento, y por lo tanto (por el Corolario IV de las Leyes) siempre estará en reposo. Por lo tanto, los movimientos de los cuerpos (por el Problema XXV) han de ser determinados como si fuesen impelidos por fuerzas tendentes a ese centro; y se tendrán los movimientos de los cuerpos que se atraen mutuamente. Q. E. I.

Proposición LXIII. Problema XXXIX

Determinar los movimientos de dos cuerpos que se atraen mutuamente con fuerzas inversamente proporcionales a los cuadrados de su distancia y parten de lugares dados, según rectas dadas y con velocidades dadas.

Al estar dados los movimientos de los cuerpos en el inicio, también está dado el movimiento del centro común de gravedad, así como el movimiento del espacio que se mueve uniformemente en línea recta junto con dicho centro, lo mismo que los

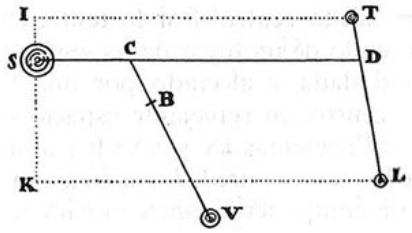
movimientos iniciales de los cuerpos respecto a dicho espacio. Pero los movimientos siguientes (por el Corolario V de las Leyes y el Teorema anterior) tendrán lugar en ese espacio como si éste reposase junto con el centro común de gravedad y los cuerpos no se atrajesen mutuamente, sino que fuesen atraídos por un tercer cuerpo situado en dicho centro. Por lo tanto, el movimiento de cada cuerpo, partiendo de un lugar dado, según una recta dada y con una velocidad dada y afectado por una fuerza centrípeta tendente a dicho centro, en semejante espacio móvil, ha de determinarse por los Problemas IX y XXVI: y a la vez se tendrá el movimiento del otro cuerpo alrededor del mismo centro. Con este movimiento ha de componerse aquel movimiento uniforme del sistema espacial y de los cuerpos que giran en él y se tendrá el movimiento absoluto de los cuerpos en el espacio inmóvil. Q. E. I[36].

Proposición LXIV. Problema XL

Si los cuerpos se atraen mutuamente con fuerzas que crecen en razón simple de sus distancias a los centros, hállense los movimientos de varios cuerpos entre sí.

Primero supónganse dos cuerpos T y L que tienen su centro común de gravedad en D. Describirán (por el Corolario I del Teorema XXI) elipses con centro en D, cuyas magnitudes son conocidas por el Problema V.

Ahora un tercer cuerpo S atraiga a los dos primeros T y L con las fuerzas aceleratrices ST, SL y sea a la vez atraído por ellos. La fuerza ST (por el Corolario II de las Leyes) se descompone en las fuerzas SD y DT; y la fuerza SL en las fuerzas SD y DL. Pero las fuerzas DT, DL, que son como su suma TL y, por tanto, como las fuerzas aceleratrices con las cuales los cuerpos T y L se atraen mutuamente, añadidas a las fuerzas de los cuerpos T y L, la primera a la primera y la segunda a la segunda, componen fuerzas proporcionales a las distancias DT y DL como al principio, pero ahora con fuerzas mayores que antes; y por tanto (por el Corolario I de la Proposición X y por los Corolarios 1 y 8 de la Proposición IV) hacen que describan elipses como antes, pero con movimiento más rápido. Las restantes fuerzas aceleratrices SD y SD, por las acciones motrices SD x T y SD x L, que son como los cuerpos, al atraer a dichos cuerpos igualmente y según las líneas TI, LK paralelas a la propia DS, en nada cambian sus posiciones relativas, sino que hacen que se acerquen a la línea IK, la cual ha de imaginarse trazada por el centro de S y perpendicular a DS.



Pero esta aproximación a IK se impedirá procurando que el sistema de cuerpos T y L por una parte, y el cuerpo S por otra, piren con velocidades precisas en torno al centro común de pravedad C. Con tal movimiento, dado que la suma de las fuerzas motrices SD x T y SD x L proporcional a la distancia CS tiende al centro C, el cuerpo S describe una elipse en torno al centro C; y el punto D, al ser proporcionales CS y CD, describirá en su zona tina elipse similar. Pero los cuerpos T y L, atraídos por las fuerzas motrices SD x T y SD x L, el primero por la primera y el segundo por la segunda, igualmente y según las paralelas TI y LK, como se dijo antes, continuarán describiendo (por los Corolarios V y VI de las Leyes) sus elipses en torno al centro móvil D, como anteriormente. Q. E. I.

Añádase ahora un cuarto cuerpo V y, con un razonamiento semejante se concluirá que dicho cuerpo junto con el punto C describirá elipses en torno al punto común de gravedad B; los movimientos de los cuerpos T, L y S en torno a los centros D y C, permanecerán, pero acelerados. Y con el mismo método podríamos añadir muchos cuerpos. Q. E. I.

Esto es así, aun cuando los cuerpos T y L se atrajesen mutuamente con fuerzas aceleratrices mayores o menores que aquellas con las cuales atraen a los demás cuerpos en razón de las distancias. Sean las atracciones aceleratrices de todos mutuamente como las distancias multiplicadas por los cuerpos atrayentes, y de lo dicho se desprende fácilmente que todos los cuerpos describirán elipses distintas en iguales tiempos periódicos en torno a su centro común de gravedad B, en el plano inmóvil. Q. E. I.

Proposición LXV. Teorema XXV

Varios cuerpos, cuyas fuerzas decrecen como el cuadrado de las distancias desde sus centros, pueden moverse entre sí en elipses, y describir mediante radios trazados a los focos áreas muy aproximadamente proporcionales a los tiempos.

Se ha demostrado en la Proposición anterior el caso en el cual los movimientos ocurren exactamente en elipses. Cuanto más se aparte la ley de las fuerzas de la ley allí propuesta, tanto más perturbarán los cuerpos sus mutuos movimientos; y tampoco es posible que los cuerpos, con atracción mutua según la ley aquí supuesta, se

muevan en elipses exactas, salvo que mantengan entre sí determinada proporción de distancias. En los casos siguientes, no obstante, no nos apartaremos mucho de las elipses.

Caso 1. Imagínese que varios cuerpos pequeños giran en torno de uno muy grande a distancias diferentes, y fuerzas absolutas, proporcionales a cada uno de ellos, tendiendo hacia cada uno. Y puesto que el centro común de gravedad de todos (por el Corolario iv de las Leyes) o está en reposo o se mueve uniformemente en línea recta, imaginemos que los cuerpos pequeños lo son tanto que el cuerpo mayor jamás dista sensiblemente de dicho centro: y dicho cuerpo mayor o está en reposo o se mueve uniformemente en línea recta sin desvío apreciable: los cuerpos pequeños girarán en elipses en torno del mayor y con radios trazados hasta éste describirán áreas proporcionales a los tiempos, salvo en la medida en que puedan ocurrir desvíos procedentes de la separación del cuerpo mayor del centro común de gravedad o de las interacciones mutuas de los cuerpos menores. Pero los cuerpos pequeños pueden disminuirse tanto que, tanto dicho desvío, como las interacciones mutuas, sean menores que unas dadas y, por tanto, hasta que las órbitas coincidan con elipses y las áreas respondan a los tiempos sin error mayor que uno dado. Q. E. O.

Caso 2. Supongamos ahora un sistema de cuerpos pequeños que giran en torno a uno muy grande del modo antes descrito; u otro sistema cualquiera de dos cuerpos, que giran uno en torno al otro, que se desplaza con movimiento uniforme y rectilíneo, y que a la vez es influido lateralmente por la fuerza de otro cuerpo mucho mayor y situado a gran distancia. Y, puesto que las fuerzas aceleratrices iguales que influyen en los cuerpos según líneas paralelas no cambian la situación de los cuerpos entre sí, sino que hacen que todo el sistema cambie de lugar preservándose los movimientos de las partes entre ellas, es evidente que en esos movimientos de los cuerpos atraídos no puede ocurrir ningún cambio procedente de la atracción del mayor, sino sólo de las desigualdades de las atracciones aceleratrices o de las inclinaciones de las líneas según las cuales ocurren las respectivas atracciones. Supóngase, pues, que todas las atracciones aceleratrices hacia el cuerpo mayor sean entre sí inversamente como el cuadrado de las distancias; aumentando, entonces, la distancia del cuerpo mayor hasta que las diferencias entre longitudes de las rectas trazadas desde él hasta los otros, y las inclinaciones respectivas de dichas líneas sean menores que unas dadas, los movimientos de las partes del sistema entre sí proseguirán sin errores que no sean menores que unos dados. Y como la distancia mutua entre dichas partes es muy pequeña, el sistema entero, como si fuera un solo cuerpo, es atraído y se moverá bajo dicha atracción como si de un solo cuerpo se tratara; esto es, con su centro de gravedad describirá alrededor del cuerpo mayor alguna sección cónica (i. e. una hipérbola o una parábola si la atracción es suave y una elipse si es más fuerte) y con un radio trazado hasta el cuerpo mayor describirá áreas proporcionales a los tiempos sin más errores que los originados de las distancias de las partes, desde luego pequeñas, y disminuíbles cuanto se quiera. Q. E. O.

Con argumentos similares pueden tratarse casos más complejos hasta el infinito.

Corolario 1. En el Caso segundo, cuanto más se acerca el cuerpo mayor al sistema de dos o más cuerpos, tanto más resultarán perturbados los movimientos de las partes del sistema cutre sí; puesto que mayores serán las inclinaciones de las líneas trazadas desde este cuerpo mayor hasta los otros y también será mayor la desigualdad de la proporción.

Corolario 2. Pero resultará máxima la perturbación suponiendo que las atracciones aceleratrices de las partes del sistema hacia el cuerpo mayor no son entre sí como el inverso del cuadrado de las distancias desde dicho cuerpo mayor; sobre todo si la desigualdad de dicha proporción es mayor que la desigualdad de la proporción de las distancias desde dicho cuerpo mayor. Porque si la fuerza aceleratriz, al actuar igualmente y según líneas paralelas, no produce perturbación alguna entre los movimientos del sistema, es necesario que surja alguna perturbación debida a la desigualdad de acción, y que sea mayor o menor según sea mayor o menor dicha desigualdad. El exceso de impulsos mayores actuando sobre algunos cuerpos y no sobre otros cambiará necesariamente su situación respectiva. Y esta perturbación, añadida a la que surge de la inclinación y desigualdad de las líneas, aumenta la perturbación total.

Corolario 3. Por tanto, si las partes de este sistema se mueven en elipses o en círculos sin perturbaciones notables, es evidente que o no se hallan impelidas hacia otros cuerpos por fuerzas aceleratrices más que de modo insensible o lo son de un modo uniforme y según líneas paralelas muy aproximadamente.

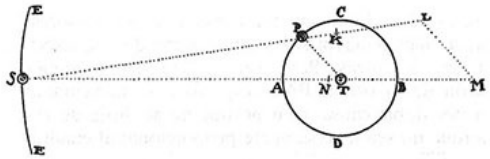
Proposición lxvi. Teorema xxvi

Si tres cuerpos, cuyas fuerzas decrecen como el cuadrado de las distancias, se atraen mutuamente y las atracciones aceleratrices de dos cualesquiera sobre el tercero son entre ellas inversamente como el cuadrado de las distancias, y los más pequeños giran en torno al mayor, digo: que el interior más próximo al central y mayor, con radios trazados hasta él describirá áreas más proporcionales a los tiempos, y una figura más cercana a la de una elipse con foco en el punto de intersección de los radios, tanto si el cuerpo mayor es perturbado por dichas atracciones, como si permanece en reposo no siendo atraído por los pequeños o si, siendo atraído unas veces mucho y otras poco, resultase perturbado unas veces mucho y otras poco.

Casi se sigue de la demostración del segundo Corolario de la Proposición anterior; pero se prueba mediante una demostración más clara y estricta del modo siguiente.

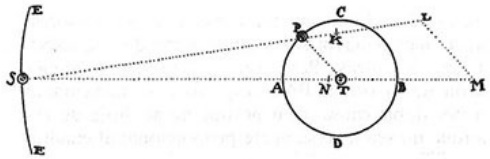
Caso 1. Supongamos que en torno al cuerpo mayor T giran los cuerpos menores P y S en el mismo plano, de los cuales P describe la órbita interior PAB, y S la

exterior ESE. Sea SK la distancia media entre los cuerpos P y S, y valga también como expresión de la atracción aceleratriz de P hacia S a esa distancia media. Tómese SL a SK como el cuadrado de la razón de SK a SP, y SL será la atracción aceleratriz de P hacia S a cualquier distancia SP. Únase PT, y paralela a ella trácese LM que toca a ST en M; y la atracción SL se descompondrá (por el Corolario II de las Leyes) en las atracciones SM, LM. El cuerpo P será urgido de este modo por una triple fuerza aceleratriz. Una fuerza tiende hacia T, y se origina por la atracción mutua entre los cuerpos T y P. Con esta fuerza sola el cuerpo P debería describir, en torno al cuerpo T y con el radio PT, áreas proporcionales a los tiempos así como una elipse cuyo foco se halla en el centro del cuerpo T, y esto tanto si T permanece en reposo como si oscila por causa de esta atracción, listo es evidente por la Proposición XI y por los Corolarios 2 y 3 del Teorema XXL La otra fuerza es la de la atracción LM, que porque tiende de P hacia T se superpone a la fuerza anterior y coincide con ella, y de este modo hará que las áreas sigan siendo proporcionales a los tiempos, por el Corolario 3 del Teorema XXL Pero, puesto que no es proporcional al inverso del cuadrado de la distancia PT, al ser añadida a la anterior compondrá con ella una fuerza que se aparta de dicha proporción; y esto tanto más cuanto mayor sea la proporción de esta fuerza respecto a la anterior, permaneciendo igual el resto. Por tanto, ya que (por la Proposición XI y por el Corolario 2 del Teorema XXI) la fuerza con la cual se describe la elipse en torno al foco T ha de tender hacia dicho foco y ser inversamente proporcional al cuadrado de dicha distancia PT, tal fuerza compuesta, al apartarse de dicha proporción, hará que la órbita PAB se aparte de la forma elíptica con foco en T, y esto tanto más cuanto mayor sea la diferencia con respecto a dicha proporción; y por lo tanto también cuanto mayor es la proporción de la segunda fuerza LM respecto a la primera, manteniendo el resto igual. Pero, ahora, la tercera fuerza, SM, al atraer al cuerpo P en una dirección paralela a ST, compone con las fuerzas anteriores una fuerza que ya no se dirige de P hacia T; y que diverge tanto más de esa dirección cuanto mayor es la proporción de esta tercera fuerza respecto a las anteriores, *caeteris paribus*; la cual, por tanto, hace que el cuerpo P, con radio TP, ya no describa más áreas proporcionales a los tiempos, y que la aberración con respecto a esta proporcionalidad sea mayor cuanto mayor sea la proporción de esta tercera fuerza respecto de las otras. Pero esta tercera fuerza aumentará la aberración de la órbita PAB respecto a la mencionada forma elíptica por doble causa, bien porque no se dirija de P hacia T, bien porque no sea inversamente proporcional al cuadrado de la distancia PT. Comprendido esto, es evidente que las áreas se tornan más proporcionales a los tiempos, permaneciendo las demás fuerzas, cuanto más pequeña sea la tercera fuerza; y que la órbita PAB se aproxima máximamente a la mencionada forma elíptica cuando tanto la segunda como la tercera fuerza, y sobre todo la tercera, se hacen mínimas, permaneciendo la primera.



Represéntese mediante la linea SN la atracción aceleratriz del cuerpo T hacia S; y si las atracciones aceleratrices SM, SN fuesen iguales, al atraer éstas a los cuerpos T, P, igualmente y según líneas paralelas, en nada cambiarán sus situaciones respectivas. Los movimientos de dichos cuerpos entre sí serían iguales (por el Corolario VI de las Leyes) que si desapareciesen dichas atracciones, Y por similar razón, si la atracción SN fuese menor que la SM, la propia SN restará una parte de la atracción SM, y quedará sólo la parte MN con la que resultarán perturbadas tanto la proporcionalidad de tiempos y áreas como la susodicha forma elíptica de la órbita. De manera semejante si la atracción SN fuese mayor que la atracción SM, surgirá de la sola diferencia MN la perturbación de la proporcionalidad y de la órbita. Así la atracción SN siempre reduce la atracción SM, tercera de antes, a la atracción MN, permaneciendo iguales la primera y la segunda; y por lo tanto las áreas y los tiempos se aproximan de modo máximo a la proporcionalidad, y la órbita PAB a la figura elíptica mencionada, cuando la atracción MN es nula o la mínima posible; esto es, cuando las atracciones aceleratrices de los cuerpos P y T realizadas hacia S se aproximan cuanto sea posible a la igualdad; o sea, cuando la atracción SN ni es nula ni es inferior a la más pequeña de todas las atracciones SM, sino como la media entre la máxima y la mínima de las atracciones SM, esto es, ni mucho mayor ni mucho menor que la atracción SK. Q. E. D.

CASO 2. Ahora supóngase que los cuerpos menores S, P giran en distintos planos en torno a otro mayor T; y la fuerza LM, al actuar según la línea PT situada sobre el plano de la órbita PAB, producirá el mismo efecto que antes, y no desviará al cuerpo P del plano de su órbita. Pero, la otra fuerza MN, al actuar según una línea paralela a ST (y por tanto, cuando el cuerpo S se mueve fuera de la línea nodal, inclinada hacia el plano de la órbita PAB) produce, además de la perturbación del movimiento ya mencionada de sentido longitudinal, otra de sentido latitudinal, al atraer al cuerpo P fuera de su plano orbital. Y esta perturbación, para cualquier situación mutua dada de los cuerpos P y T, será como la fuerza generadora MN; y resultará, por tanto, mínima cuando MN sea mínima, esto es (como ya dije) cuando la atracción SN no es ni mucho mayor ni mucho menor que la atracción SK. Q. E. D.

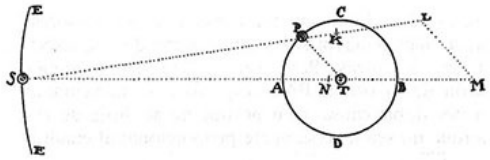


COROLARIO 1. De esto se sigue fácilmente que si varios cuerpos pequeños, P, S, R, etc., giran en torno a otro mayor T, el movimiento del cuerpo más interior P será perturbado mínimamente por las atracciones de los exteriores, toda vez que el cuerpo mayor T es atraído y perturbado en igual medida por los demás, ch razón de las fuerzas aceleratrices, que los otros entre sí.

COROLARIO 2. Pero en un sistema de tres cuerpos, T, P, S, si las atracciones aceleratrices de dos cualesquiera sobre un tercero son entre sí como el inverso del cuadrado de las distancias; el cuerpo P, con radio PT, describirá el área en torno al cuerpo T más rápidamente en las inmediaciones de la conjunción A y de la oposición B, que en las inmediaciones de las cuadraturas C, D. Pues toda fuerza que actúa sobre el cuerpo P y no actúa sobre el cuerpo T, puesto que no actúa según la línea PT, acelera o retarda la descripción del área, según actúe por delante o por detrás. Tal es la fuerza MN. Esta fuerza, al ir el cuerpo P desde C hacia A atrae hacia adelante y acelera el movimiento; después hasta llegar a D actúa por detrás y retarda el movimiento, después por delante hasta B, y finalmente por detrás al ir desde B hasta C.

COROLARIO 3. Y por el mismo razonamiento es claro que el cuerpo P, permaneciendo iguales las demás cosas, se mueve más rápidamente en la conjunción y en la oposición que en las cuadraturas.

COROLARIO 4. La órbita del cuerpo P, permaneciendo igual lo demás, es más curva en las cuadraturas que en la conjunción y en la oposición. Puesto que cuanto más rápidamente se mueve un cuerpo menos se desvía de la línea recta. Y además, la fuerza KL, o MN, en la conjunción y en la oposición es contraria a la fuerza con la cual el cuerpo T atrae al cuerpo P, y por lo tanto hace disminuir a la tal fuerza; y entonces el cuerpo P se desviará de la línea recta tanto menos cuanto menos es atraído hacia el cuerpo T.



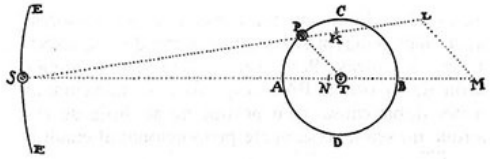
COROLARIO 5. De donde, el cuerpo P, permaneciendo el resto igual, se alejará más del cuerpo T en las cuadraturas que en la conjunción y la oposición. Esto es así

excluyendo el movimiento de excentricidad. Puesto que si la órbita del cuerpo P fuese excéntrica, su excentricidad (como se verá en el siguiente Corolario 9) resultará máxima cuando los ápsides se hallen en las sicigias; y por ello puede ocurrir que el cuerpo P, al alcanzar el ápside máximo, se aleje más del cuerpo T en las sicigias que en las cuadraturas.

COROLARIO 6. Dado que la fuerza centrípeta del cuerpo central T, mediante la cual el cuerpo P es retenido en su órbita, aumenta en las cuadraturas por la suma de la fuerza LM y disminuye en las sicigias por la resta de la fuerza KL, y, dada la magnitud de la fuerza KL, disminuye más que aumenta; y además, puesto que dicha fuerza centrípeta (por el Corolario 2 de la Proposición IV) está en razón compuesta directamente como el radio TP e inversamente como el cuadrado del tiempo periódico, es evidente que dicha razón compuesta es disminuida por la acción de la fuerza KL; y por lo mismo que el tiempo periódico aumenta, si el radio TP de la órbita permanece constante, y ello como la raíz cuadrada de la razón en que disminuye dicha fuerza centrípeta; y, por tanto, al aumentar o disminuir este radio, el tiempo periódico aumentará más o disminuirá menos que la potencia $^{3}/_{2}$ de dicho radio (por el Corolario 6 de la Proposición IV). Si la fuerza del cuerpo central disminuyera paulatinamente, el cuerpo P se alejaría siempre más del centro T al estar cada vez menos atraído; y, por el contrario, si dicha fuerza aumenta se acercará cada vez más. Por tanto, si la acción del cuerpo lejano S, mediante la cuál disminuye la susodicha fuerza, aumenta y disminuye alternativamente, aumentará y disminuirá alternativamente el radio TP; y también el tiempo periódico aumentará y disminuirá según una razón compuesta de la potencia $^{3}/_{2}$ del radio y de la raíz cuadrada de la cantidad en que disminuyó o aumentó la fuerza centrípeta del cuerpo central T por causa del aumento o disminución de la acción del cuerpo lejano S.

COROLARIO 7. De lo antedicho se desprende también que el eje de la elipse descrita por el cuerpo P, o línea de los ápsides, en cuanto a su movimiento angular, avanza y retrocede alternativamente, aunque los avances son mayores, y gracias a estos excesos, en conjunto se desplaza hacia adelante. Pues la fuerza con la cual el cuerpo P es atraído hacia el cuerpo T en las cuadraturas, momento en que la fuerza MN se desvanece, se compone de la fuerza LM y de la fuerza centrípeta con la cual el cuerpo T atrae al cuerpo P. Si se aumenta la distancia PT la primera fuerza LM aumenta casi en la misma proporción que la distancia, mientras la segunda disminuye como el cuadrado de dicha distancia; con lo cual la suma de ambas fuerzas decrece en una razón menor que la del cuadrado de la distancia PT, y por lo mismo (por el Corolario 1 de la Proposición XLV) el auge, o ápside superior, resultará retrasado. Pero en la conjunción y en la oposición la fuerza con la cual el cuerpo P es atraído hacia el cuerpo T es la diferencia entre la fuerza con la cual el cuerpo atrae al cuerpo P y la fuerza KL; y esta diferencia, por cuanto que la fuerza KL aumenta muy aproximadamente en razón de la distancia PT, disminuye en razón mayor que la del

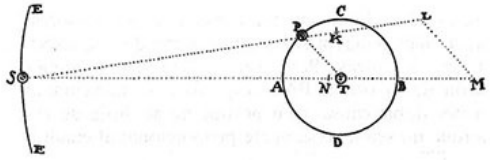
cuadrado de la distancia PT, y por tanto (por el Corolario 1 de la Proposición XLV) hará que el auge se adelante. En los puntos comprendidos entre sicigias y cuadraturas el movimiento del auge depende conjuntamente de ambas causas, de modo que avanza o retrocede según el exceso de una o de otra. Puesto que la fuerza KL en las sicigias es casi el doble que la fuerza LM en las cuadraturas, el exceso se hallará hacia el lado de la fuerza KL y empujará hacia adelante al auge. Se comprenderá más fácilmente la verdad de este Corolario y la del precedente imaginando el sistema de los dos cuerpos T y P rodeado por todas partes de muchos cuerpos S, S, S, etc. colocados sobre la órbita ESE. Puesto que por la acción de éstos disminuye por todas partes la acción del propio T, disminuirá ésta más que en razón del cuadrado de la distancia.



COROLARIO 8. Pero como el progreso o retroceso de los ápsides depende de la disminución de la fuerza centrípeta, según resulte aquélla en razón mayor o menor que el cuadrado de la distancia TP, al ir pasando el cuerpo desde el ápside inferior al superior; lo mismo que del incremento análogo al regresar al ápside inferior; y es por tanto máximo cuando la proporción de la fuerza en el ápside superior a la fuerza en el ápside inferior resulte más alejada de la razón del inverso del cuadrado de las distancias: es evidente que los ápsides en sus sicigias progresarán más rápidamente —gracias a la fuerza sustractiva KL o NM - LM— mientras en sus cuadraturas regresarán más lentamente gracias a la fuerza aditiva LM. Al ser tan largo el período de tiempo en el cual ocurren tanto la velocidad del progreso como el retardo del regreso, dicha desigualdad se torna máxima a la larga.

COROLARIO 9. Si un cuerpo gira en una elipse en torno a un centro desde el cual es atraído por una fuerza inversamente proporcional al cuadrado de la distancia desde dicho centro; y después, al caer desde el ápside superior o auge hacia el ápside inferior, dicha fuerza, por la presencia continua de otra fuerza nueva, aumentase hasta ser mayor que la razón del cuadrado de la distancia disminuida; es evidente que el cuerpo, empujado siempre hacia el centro por la presencia continua de la nueva fuerza, tenderá más hacia dicho centro que si solamente se viese afectado por la fuerza creciente según el cuadrado de la distancia disminuida; y por lo tanto describirá una órbita interior a la órbita elíptica, a la vez que el ápside inferior se acercará más al centro que antes. Pues la órbita, con la presencia de esta nueva fuerza, se tornará más excéntrica. Pero si la fuerza, al retornar el cuerpo desde el ápside inferior hasta el superior, decreciese en las mismas proporciones en que antes

crecía, el cuerpo retornará a su distancia anterior, y por consiguiente, si la fuerza decrece en una razón mayor, el cuerpo atraído ahora en menor grado ascenderá hasta una distancia mayor, con lo cual se aumentará más aún la excentricidad de la órbita. Por lo cual, si la razón del incremento o decremento de la fuerza centrípeta aumenta en cada revolución, la excentricidad será siempre creciente; y, por el contrario, será decreciente si dicha razón disminuye. Ahora bien, en el sistema de cuerpos T, P, S, cuando los ápsides de la órbita PAB están en las cuadraturas, dicha razón de incrementos y decrementos es mínima, y máxima cuando los ápsides están en las sicigias. Cuando los ápsides se hallan en las cuadraturas, la razón es menor junto a las cuadraturas y mayor junto a las sicigias que el cuadrado de la razón de las distancias; y de dicha razón mayor surge el movimiento directo del auge, como ya se ha dicho. Pero si se considera globalmente la razón de aumento y disminución en el recorrido entre ápsides, ésta resulta menor que el cuadrado de la razón de las distancias. La fuerza en el ápside inferior es a la fuerza en el ápside superior en razón menor que la razón entre los cuadrados de la distancia entre el ápside superior y el foco de la elipse y la distancia entre el ápside inferior y el mismo foco; y al contrario, cuando los ápsides se hallan en las sicigias la fuerza en el ápside inferior está en razón mayor, con respecto a la fuerza en el ápside superior, que la razón de los cuadrados de las distancias. Puesto que en las cuadraturas las fuerzas LM sumadas a las fuerzas del cuerpo T componen fuerzas en menor razón, y las fuerzas KL restadas en las sicigias de las del cuerpo T dejan fuerzas que están en razón mayor. Por lo tanto, la razón global de incremento y decremento en el recorrido entre ápsides es mínima en las cuadraturas y máxima en las sicigias: por lo mismo en el recorrido de los ápsides desde las cuadraturas hasta las sicigias aumenta continuamente, y aumenta de paso la excentricidad de la elipse; mientras que en el recorrido desde las sicigias hasta las cuadraturas disminuye continuamente a la vez que decrece la excentricidad.



COROLARIO 10. Para poder calcular los errores de latitud, supongamos que permanece inmóvil el plano de la órbita EST; y supuesta la causa de los errores ya explicada, es evidente que de las dos fuerzas NM y ML que son la causa total antedicha, la fuerza ML actuando siempre según el plano PAB de la órbita jamás perturba los movimientos en latitud; y que la fuerza NM, cuando los nodos están en las sicigias, al actuar también según el mismo plano de la órbita, tampoco perturba estos movimientos; pero cuando están en las cuadraturas, los perturba máximamente y, al atraer continuamente al cuerpo P del plano de su órbita, disminuye la inclinación

del plano durante el tránsito del cuerpo de las cuadraturas a las sicigias y la aumenta cuando pasa de las sicigias a las cuadraturas. De donde viene a ocurrir que cuando el cuerpo se halla en las sicigias la inclinación resulte mínima entre todas, mientras que vuelve a alcanzar la magnitud anterior aproximadamente, cuando el cuerpo llega al nodo siguiente. Pero si los nodos se sitúan en los octantes tras las cuadraturas, esto es, entre C y A, D y B, de lo dicho se entiende que al pasar el cuerpo P desde cualquier nodo hasta el grado nonagésimo siguiente la inclinación del plano disminuye continuamente; y a partir de aquí durante los siguientes cuarenta y cinco grados hasta llegar a la siguiente cuadratura, la inclinación aumenta, para disminuir de nuevo mientras transita por los siguientes cuarenta y cinco grados hasta el nodo siguiente. Y así la inclinación disminuye más que aumenta, por lo cual es siempre menor en el nodo siguiente que en el anterior. Y por un razonamiento semejante, la inclinación aumenta más que disminuye cuando los nodos están en los otros octantes, entre A y D, B y C. Y por ello la inclinación es máxima cuando los nodos están en las sicigias. Al ir pasando los nodos de las sicigias a las cuadraturas la inclinación va disminuyendo con cada aproximación del cuerpo a los nodos; y llega a ser mínima cuando los nodos están en las cuadraturas y el cuerpo en las sicigias; después crece de nuevo al mismo paso que antes decreció; y al acercarse los nodos a las sicigias siguientes, alcanza la anterior magnitud.

COROLARIO 11. Puesto que el cuerpo P es atraído continuamente del plano de su órbita cuando los nodos se hallan en las cuadraturas, y ello en dirección a S durante su tránsito del nodo C hasta el D pasando por la conjunción A; y en dirección contraria durante su tránsito del nodo D al C pasando por la oposición B: es evidente que en su movimiento desde el nodo C el cuerpo se separa continuamente del primitivo plano de su órbita CD hasta su llegada al nodo siguiente; y por esto, al distanciarse mucho en este nodo del plano primitivo CD, pasa por el plano de la órbita EST no por el otro nodo D correspondiente a dicho plano, sino por un punto situado hacia el lado de S, punto que se convierte, por tanto, en el nuevo lugar del nodo en retroceso. Y, por un razonamiento similar, los nodos continúan retrocediendo cu cada tránsito del cuerpo de este nodo al siguiente. Por consiguiente los nodos situados en las cuadraturas retroceden continuamente; en las sicigias, donde los movimientos no se perturban en absoluto en el sentido de la latitud, están en reposo; en los puntos intermedios, al participar de ambas condiciones, retroceden más lentamente: y por lo tanto, al ser siempre o retrógrados o estacionarios, resultarán adelantados con cada revolución.

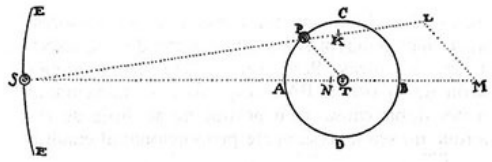
COROLARIO 12. Todos los errores descritos en estos Corolarios son algo mayores en las conjunciones de los cuerpos P, S que en sus oposiciones; y esto por ser mayores las fuerzas generadoras NM y ML.

COROLARIO 13. Pese a que las razones dadas en estos Corolarios no dependan de la magnitud del cuerpo S, todo lo que antecede se cumple cuando se atribuye al cuerpo S una magnitud tan grande como para que el sistema de los cuerpos T y P gire

en torno a él. Y al aumentar el tamaño del cuerpo S y con ello su fuerza centrípeta de la que surgen los errores del cuerpo P, resultarán mayores, a distancias iguales, los susodichos errores en este caso que en el otro, cuando el cuerpo S gira en torno al sistema de cuerpos P y T.

COROLARIO 14. Pero como las fuerzas NM y ML, cuando el cuerpo S se halla a gran distancia, son aproximadamente como la fuerza SK y la razón PT a ST conjuntamente, esto es, si se diesen tanto la distancia PT como la fuerza absoluta de S, como el inverso de ST^3; y son también las dichas fuerzas NM y ML las causas de todos los errores y efectos vistos en los anteriores Corolarios: es evidente que todos los efectos mencionados, permaneciendo el sistema de cuerpos P y T, y cambiando solamente la distancia ST y la fuerza absoluta del cuerpo S, estarán muy aproximadamente en razón compuesta de la razón directa de la fuerza absoluta del cuerpo S y de la razón inversa de la distancia ST al cubo. De donde se sigue que, si el sistema de cuerpos T y P gira en torno al cuerpo lejano S, las fuerzas NM y ML, así como sus efectos, serán (por los Corolarios 2 y 6 de la Proposición IV) inversamente como el cuadrado del tiempo periódico. Y además, si la magnitud del cuerpo S es proporcional a su fuerza absoluta, las fuerzas NM y ML y sus efectos serán en razón directa del cubo del diámetro aparente del cuerpo lejano S visto desde T, y viceversa. Ya que estas razones son las mismas que la anterior razón compuesta.

COROLARIO 15. Y aunque si las órbitas ESE y PAB, conservando sus formas, proporciones e inclinación mutua, alterasen su magnitud y si las fuerzas de los cuerpos S y T permaneciesen constantes o cambiasen en una razón dada, estas fuerzas (esto es, la fuerza del cuerpo T que hace al cuerpo P desviarse de su curso recto para seguir la órbita PAB, y la fuerza del cuerpo S que obliga al cuerpo P a desviarse de dicha órbita) siempre actuarán del mismo modo y en la misma proporción, necesariamente todos los efectos serán semejantes y proporcionales, como también serán proporcionales los tiempos de los efectos; esto es, todos los errores lineales serán como los diámetros de las órbitas, mientras los errores angulares serán como antes, y los tiempos de errores lineales semejantes, o de errores angulares iguales serán como los tiempos periódicos de las órbitas.



COROLARIO 16. De donde, si están dadas las formas de las órbitas y sus inclinaciones mutuas, y se alteran de cualquier modo las magnitudes, fuerzas y distancias de los cuerpos, de los errores dados y de los tiempos de los errores en un caso se pueden inferir los errores y los tiempos de los errores de otro caso con gran

aproximación; pero más brevemente aún con el método siguiente. Las fuerzas NM y ML, permaneciendo constante el resto, son como el radio TP; y sus efectos periódicos (por el Corolario 2 del Lema X) como las fuerzas y el cuadrado del tiempo periódico del cuerpo P conjuntamente. Estos son los errores lineales del cuerpo P; y de aquí que los errores angulares vistos desde el centro T (esto es, tanto el movimiento del auge y de los nodos como todos los errores aparentes en longitud y en latitud) son para cada revolución del cuerpo P aproximadamente como el cuadrado del tiempo de la revolución. Al componer estas razones con las razones del Corolario 14 resulta que en todo sistema de cuerpos T, P, S, en donde P gira en torno a T muy cercano a él, y T gira en torno a S muy distante, los errores angulares del cuerpo P visto desde T serán, en cada revolución del cuerpo P, directamente como el cuadrado del tiempo periódico del cuerpo P, e inversamente como el cuadrado del tiempo periódico del cuerpo T. Y por tanto el movimiento medio del auge estará en una razón dada con el movimiento medio de los nodos; y ambos movimientos serán directamente como el tiempo periódico de P, c inversamente como el cuadrado del tiempo periódico de T. Aumentando o disminuyendo la excentricidad e inclinación de la órbita PAB no se alteran los movimientos del auge y de los nodos de manera sensible, salvo que aquellos sean demasiado grandes.

COROLARIO 17. Dado que la línea LM resulta unas veces mayor y otras menor que el radio PT, represéntese el valor medio de la fuerza LM por dicho radio PT; y este valor será a la fuerza media SK o SN (que puede representarse también por ST) como la longitud PT a la longitud ST. Pero la fuerza media SN o ST, mediante la cual el cuerpo T es retenido en su órbita en torno a S, es a la fuerza, mediante la cual el cuerpo P es retenido en su órbita en torno a T, en razón compuesta de la razón del radio ST al radio PT y de la razón cuadrada del tiempo periódico del cuerpo P en torno a T al tiempo periódico del cuerpo T en torno a S. Y por lo mismo la fuerza media LM es a la fuerza que retiene al cuerpo P en su órbita en torno a T (o también por la que el susodicho cuerpo P podría girar en el mismo tiempo periódico en torno al punto móvil T a la distancia PT) en dicha razón cuadrada de los tiempos periódicos. Por tanto, dados los tiempos periódicos y la distancia PT, está dada la fuerza media LM; y dada ésta, también está dada muy aproximadamente la MN, por la analogía de las líneas PT, MN.

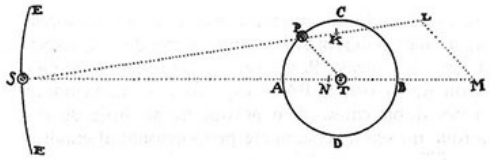
COROLARIO 18. Supongamos que muchos cuerpos fluidos se mueven en torno al mismo T a distancias iguales y bajo las mismas leyes con las que el cuerpo P gira en torno al cuerpo T; y que al llegar a ser contiguos acaban formando un anillo fluido, redondo y concéntrico con el cuerpo T; y cada parte del anillo, realizando todos sus movimientos según la ley del cuerpo P, se acercará más al cuerpo T, y se moverá más rápidamente en las conjunciones y oposiciones de la misma y del cuerpo S, que en las cuadraturas. Los nodos de dicho anillo, o sea, sus intersecciones con el plano de la órbita del cuerpo S o T, estarán en reposo en las sicigias; pero fuera de las sicigias, se moverán hacia atrás, y en las cuadraturas muy rápidamente, en otros puntos más

lentamente. La inclinación del anillo también variará y su eje oscilará en cada revolución, y al completarse ésta retornará a su estado anterior, salvo en la medida en que sea desplazada circularmente por la precesión de los nodos.

COROLARIO 19. Imaginemos ahora que la esfera del cuerpo T, constituida de materia no fluida, se expande y se extiende hasta alcanzar dicho anillo, y que hay agua en un foso excavado a lo largo de toda su circunferencia mientras gira uniformemente sobre su eje con el mismo movimiento periódico. Al ser este líquido acelerado y retardado alternativamente (como en el Corolario anterior) será más veloz en las sicigias y más lento en las cuadraturas que la superficie de la esfera, y así experimentará flujo y reflujo dentro del foso como el mar. Si se suprimiese la atracción del cuerpo S, el agua no adquiriría ningún movimiento de flujo o reflujo al girar en torno al centro en reposo de la esfera. La razón es la misma para una esfera que se desplaza uniformemente en línea recta mientras gira en torno a su centro (por el Corolario V de las Leyes) y para una esfera uniformemente desviada de su curso rectilíneo (por el Corolario VI de las mismas Leyes). Pero si se acerca el cuerpo S, por su desigual atracción, el agua será perturbada inmediatamente. Y será efectivamente mayor la atracción del agua más cercana y menor la de la más lejana. Pues la fuerza LM atraerá el agua en las cuadraturas hacia abajo, y la hará descender hasta las sicigias; y la fuerza KL la atraerá hacia arriba en las sicigias, detendrá su descenso y la hará ascender hasta las cuadraturas, salvo en la medida en que el flujo y reflujo del agua es obligado por el foso, y resulta algo retardado por el rozamiento.

COROLARIO 20. Si ahora el anillo se torna rígido y el globo disminuye, cesará el movimiento de flujo y reflujo; pero permanecerá el movimiento oscilatorio de la inclinación así como la precesión de los nodos. Supóngase que el globo tiene el mismo eje que el anillo y que completa sus revoluciones en el mismo tiempo, que toca por dentro al anillo con su superficie y esté adherido a él; al compartir el movimiento del anillo, el conjunto de los dos oscilará, y los nodos retrocederán. Pues el globo, como diremos ahora, es indiferente para recibir todas las impresiones. El ángulo máximo de inclinación para el anillo que rodea al globo ocurre cuando los nodos se hallan en las sicigias. Al pasar los nodos desde aquí hacia las cuadraturas, intenta reducir su inclinación y con este intento comunica un movimiento al globo entero. El globo retiene este movimiento impreso hasta que el anillo con un intento contrario suprime este movimiento e imprime otro en sentido contrario: de este modo el movimiento máximo de inclinación decreciente ocurre cuando los nodos se hallan en las cuadraturas, y el ángulo mínimo de inclinación en los octantes posteriores a las cuadraturas; y de nuevo el máximo movimiento de reclinación en las sicigias y el ángulo máximo en los octantes siguientes. Igual es el caso para el globo despojado de anillo cuando en las regiones ecuatoriales es algo más alto o conste de materia algo más densa que en los polos. Pues este exceso de materia en la parte ecuatorial hace las veces de anillo. Y aunque se suponga que, al aumentar la fuerza centrípeta de este globo en cualquier forma, todas sus partes tenderían hacia abajo, como las partes en

gravitación de la Tierra, apenas cambiarían por ello los fenómenos de este Corolario y del precedente; salvo en lo que variasen los lugares de máximas y mínimas alturas del agua. Pues el agua en su órbita ya no se sostiene y permanece por su fuerza centrífuga, sino por el foso donde fluye. Y además la fuerza LM atrae el agua hacia abajo en máximo grado en las cuadraturas, y la fuerza KL o NM - LM la atrae en el mayor grado hacia arriba en las sicigias. Y las dos fuerzas conjuntas dejan de atraer el agua hacia abajo y empiezan a atraerla hacia arriba en los octantes anteriores a las sicigias; y dejan de atraer el agua hacia arriba y empiezan a atraerla hacia abajo en los octantes posteriores a las sicigias. De aquí que la máxima altura del agua pueda ocurrir en los octantes posteriores a las sicigias, y la mínima aproximadamente en los octantes siguientes a las cuadraturas; salvo en la medida en la que el movimiento ascendente o descendente impreso por dichas fuerzas pueda continuar un poco por causa de la fuerza ínsita del agua, o detenerse un poco más pronto por causa de los obstáculos en el foso.



COROLARIO 21. Por la misma razón es por la que el exceso de materia del globo en el ecuador hace que los nodos sean retrógrados, y también que el aumento de dicho exceso aumente tal regreso, y que por su disminución disminuya y que con la supresión desaparezca. Si se suprimiese la materia redundante, esto es, si el globo en el ecuador fuese menos abultado o menos denso que en los polos, se originaría un movimiento de los nodos hacia adelante.

COROLARIO 22. Y de aquí, a su vez, que del movimiento de los nodos se desprenda la constitución del globo. Ya que si el globo conserva los mismos polos de modo constante y el movimiento de los nodos es retrógrado, en el ecuador hay un exceso de materia; pero si se mueven hacia adelante entonces hay falta de ella. Supóngase que un globo uniforme y perfectamente redondo se halle primero en reposo en un espacio libre; después es empujado por un impulso cualquiera ejercido oblicuamente sobre su superficie, y adquiere así un movimiento en parte circular y en parte rectilíneo. Puesto que semejante globo es completamente indiferente respecto a todos los ejes que pasan por su centro, ni es más propenso hacia un eje, o hacia una posición del eje, que hacia otra; es evidente que jamás cambiará ni de eje ni de posición del eje por su propia fuerza. Sea empujado ahora con un nuevo impulso oblicuamente ejercido sobre la misma parte de la superficie en que se ejerció antes; y como el antes o el después en nada cambia el efecto de un impulso, es evidente que estos dos impulsos ejercidos sucesivamente producen el mismo movimiento que si se

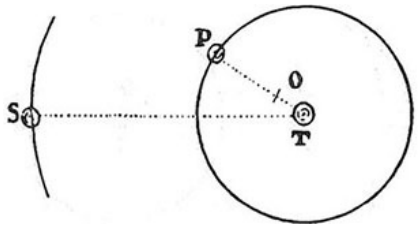
hubiesen ejercido a la vez, esto es, el mismo que tendría el globo si hubiese sido empujado por una sola fuerza compuesta de las dos (por el Corolario II de las Leyes), esto es, un movimiento simple en torno a un eje de inclinación dada. E igual es el caso si el segundo impulso se diese en cualquier lugar del ecuador del primero, lo mismo que ocurriría si el primer impulso se aplicase sobre un lugar cualquiera del ecuador del movimiento generado por el segundo impulso sin el concurso del primero; y por tanto, el de ambos impulsos ejercidos sobre cualesquiera lugares: dichos impulsos generarán el mismo movimiento circular que si se efectuasen a la vez y cada vez en el lugar de la intersección de los ecuadores de los movimientos generados por cada uno de ellos. Por tanto, un globo homogéneo y perfecto no retiene muchos movimientos diferenciados, sino que compone todos los movimientos impresos y los reduce a uno, y en tanto depende de él, gira siempre con un movimiento simple y uniforme en torno a un único eje dado y de inclinación invariable. Y ni siquiera la fuerza centrípeta podrá cambiar la inclinación del eje o la velocidad de rotación. Pues si suponemos el globo dividido en dos hemisferios por un plano que pase por su centro y por el centro al que se dirige la fuerza, dicha fuerza urgirá siempre por igual a ambos hemisferios y, por tanto, el globo no se inclinará hacia ningún lado en cuanto a su movimiento de rotación. Pero añádase en cualquier lugar entre el polo y el ecuador nueva materia acumulada en forma de monte, y ésta, con el intento continuo de alejarse de su centro de movimiento, perturbará el movimiento del globo y hará que sus polos vaguen por su superficie, describiendo círculos en torno de sí mismos y de sus puntos opuestos continuamente. Y tamaña desviación no se corregirá hasta que dicho monte no se sitúe, bien en cualquiera de los polos, en cuyo caso (por el Corolario 21) los nodos ecuatoriales se adelantarán; bien en el ecuador, y por esta causa (por el Corolario 20) retrocederán; o bien, finalmente, añadiendo materia nueva al otro lado del eje que venga a equilibrar el movimiento del monte, y de este modo los nodos se adelantarán o retrocederán según se hallen el monte y esta nueva materia más cerca del polo o del ecuador.

PROPOSICIÓN LXVII. TEOREMA XXVII

Supuestas las mismas leyes de las atracciones, digo que el cuerpo exterior S, con radios trazados al centro común de gravedad O de los cuerpos interiores P, T, describe en torno a dicho centro áreas más proporcionales a los tiempos y una órbita más cercana a la forma de una elipse con foco en dicho centro, que la que describiría en torno al cuerpo mayor y más interior T con radios trazados a él.

Puesto que las atracciones del cuerpo S hacia T y P componen su atracción absoluta, que se dirige más hacia O, centro común de gravedad de T y P, que hacia el cuerpo mayor T, y es más proporcional al inverso del cuadrado de la distancia SO que

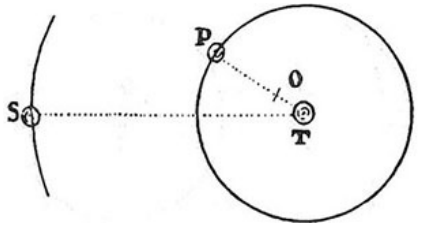
al del cuadrado de la distancia ST; como fácilmente verá quien piense en ello.



Proposición lxviii. Teorema xxviii

Supuestas las mismas leyes de las atracciones, digo que el cuerpo exterior S, con radios trazados al centro común de gravedad O de los cuerpos interiores P y T, describe áreas más proporcionales a los tiempos en torno a dicho centro y una órbita más semejante a la forma de una elipse con foco en dicho centro, si el cuerpo mayor y más interior fuese perturbado por tales atracciones igual que los demás, que si éste, o bien reposa sin atracción alguna, o bien es atraído mucho más o mucho menos o perturbado mucho más o mucho menos.

Se demuestra casi del mismo modo que la Proposición lxvi, pero con un argumento más prolijo, del cual prescindo por eso. Bastará considerarlo como sigue. De la última Proposición se sigue que el centro hacia el que se ve urgido el cuerpo S por las tuerzas conjuntas está próximo al centro común de gravedad de los otros dos. Si aquel centro coincide con este centro común y el centro común de gravedad de los tres estuviese en reposo, el cuerpo S por una parte, y el centro común de los otros dos por oirá, describirán elipses exactas en torno al centro común en reposo de los tres. Esto se sigue del Corolario 2 de la Proposición lviii, al compararlo con lo demostrado en las Proposiciones lxiv y lxv. Este movimiento elíptico es un tanto perturbado por la distancia entre el centro de los dos cuerpos y el centro hacia el cual es atraído el tercer cuerpo S. Añádase además un movimiento al centro común de los tres, y aumentará la perturbación. Por lo tanto, la perturbación será mínima cuando el centro común de los tres está en reposo; esto es, cuando el cuerpo mayor y más interior T es atraído con la misma ley de los otros: y se hace mayor cada vez cuando el centro común de los tres, al disminuir el movimiento del cuerpo T, empieza a moverse y se va agitando más y más.



Corolario. Y de aquí que, si varios cuerpos pequeños giran en torno a uno mayor, se pueda inferir que las órbitas descritas se aproximarán más a elipses, y la descripción de las áreas será más uniforme si todos los cuerpos se atraen y perturban mutuamente con fuerzas aceleratrices directamente proporcionales a sus fuerzas absolutas e inversamente proporcionales a los cuadrados de las distancias, y el foco de cada órbita se halla en el centro común de gravedad de todos los cuerpos interiores (a saber, si el foco de la órbita primera y más interior se halla en el centro de gravedad del cuerpo mayor y más interior; el foco de la segunda en el centro común de gravedad de los dos cuerpos más interiores; el de la tercera en el centro común de gravedad de los tres interiores, y así sucesivamente) que si el cuerpo interior estuviese en reposo y se convirtiera en el foco común de todas las órbitas.

Proposición lxix. Teorema xxix

En un sistema de varios cuerpos A, B, C, D, etc., si un cuerpo A atrae a todos los otros B, C, D, etc., con fuerzas aceleratrices que son inversamente como los cuadrados de las distancias al cuerpo atrayente; y otro cuerpo B atrae también a los otros cuerpos A, C, D, etc., con fuerzas que son inversamente como el cuadrado de las distancias al cuerpo atrayente: las fuerzas absolutas de los cuerpos atrayentes A, B, serán entre sí como los propios cuerpos A, B, a quienes corresponden dichas fuerzas.

Pues las atracciones aceleratrices de todos los cuerpos B, C, D, etc., hacia A son iguales entre sí a distancias iguales, por hipótesis; y de igual modo las atracciones aceleratrices de todos los cuerpos hacia B son iguales a distancias iguales. Pero la fuerza atractiva absoluta del cuerpo A es a la fuerza atractiva absoluta del cuerpo B como la atracción aceleratriz de todos los cuerpos hacia A es a la atracción aceleratriz de todos los cuerpos hacia B, a distancias iguales; y así es la atracción aceleratriz del cuerpo B hacia A a la atracción aceleratriz del cuerpo A hacia B. Pero la atracción aceleratriz del cuerpo B hacia A es a la atracción aceleratriz del cuerpo A hacia B, como la masa del cuerpo A a la masa del cuerpo B; puesto que las fuerzas motrices, que (por las Definiciones segunda, séptima y octava) son como las fuerzas aceleratrices y los cuerpos atraídos conjuntamente, son en este caso (por la Ley

tercera del movimiento) iguales entre sí. Por lo tanto, la fuerza atractiva absoluta del cuerpo A es a la fuerza atractiva absoluta del cuerpo B como la masa del cuerpo A es a la masa del cuerpo B. Q. E. D.

COROLARIO 1. Por tanto, si cada cuerpo del sistema, A, B, C, D, etc., considerado individualmente, atrae a todos los otros con fuerzas aceleratrices que son inversamente como los cuadrados de las distancias al atrayente; las fuerzas absolutas de todos estos cuerpos serán unas a otras como los propios cuerpos.

COROLARIO 2. Por la misma razón, si cada cuerpo del sistema, A, B, C, D, etc., considerado individualmente atrae a lodos los otros con fuerzas aceleratrices que están en razón directa o inversa de alguna potencia de las distancias al atrayente, o que vienen definidas por las distancias desde cada cuerpo atrayente según alguna ley común; está claro que las fuerzas absolutas de dichos cuerpos son como los cuerpos.

COROLARIO 3. En un sistema de cuerpos cuyas fuerzas decrecen en razón del cuadrado de las distancias, si los cuerpos menores giran en torno al mayor en elipses que tienen el foco común en el centro del cuerpo mayor y se aproximan lo más posible a dicha figura, a la vez que con radios trazados a dicho cuerpo mayor describen áreas lo más proporcionales posible a los tiempos: las fuerzas absolutas de dichos cuerpos serán entre sí en razón exacta o muy aproximada de los cuerpos; y viceversa. Es evidente por el Corolario de la Proposición XLVIII junto con el Corolario primero de esta Proposición.

ESCOLIO

Estas proposiciones nos conducen a la analogía entre fuerzas centrípetas y cuerpos centrales, hacia los cuales suelen dirigirse dichas fuerzas; pues parece razonable que las fuerzas que se dirigen hacia los cuerpos dependan de la naturaleza y magnitud de dichos cuerpos, como ocurre con los imanes. Y cuantas veces ocurren estos casos se tendrán que calcular las atracciones de los cuerpos atribuyendo a cada una de sus partículas las fuerzas adecuadas y obteniendo la suma de fuerzas. Tomo aquí la palabra atracción de modo genérico para cualquier conato de los cuerpos de acercarse mutuamente, tanto si tal conato acontece por la acción de los cuerpos, que se buscan unos a otros o se agitan mutuamente mediante emisión de espíritus, como si surge de la acción del éter o del aire o de cualquier otro medio corpóreo o incorpóreo que empuje de alguna forma a los cuerpos inmersos en él unos hacia otros. Y en el mismo sentido genérico utilizo el término impulso, ocupándome en este tratado no de las especies de fuerzas y cualidades físicas, sino de las cantidades y proporciones matemáticas, como expliqué en las definiciones. En matemáticas se han de investigar las magnitudes de las fuerzas y las razones que se siguen en cualesquiera condiciones supuestas: después, al descender a la física, hay que comparar estas razones con los fenómenos; para que aparezca cuáles condiciones de

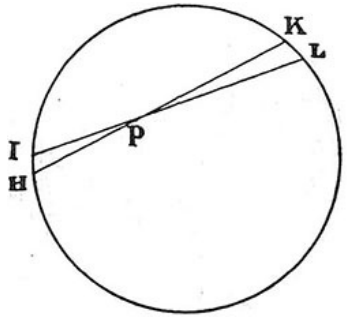
esas fuerzas corresponden a cada clase de cuerpos atractivos. Y sólo después será posible discutir con más seguridad sobre las clases de fuerzas, de las causas y razones físicas. Veamos, pues, con qué fuerzas deberán interaccionar entre sí los cuerpos esféricos constituidos de partículas atractivas de la manera ya dicha y qué movimientos van a seguirse de ello.

Sección XII
DE LAS FUERZAS ATRACTIVAS DE CUERPOS ESFÉRICOS

Proposición LXX. Teorema XXX

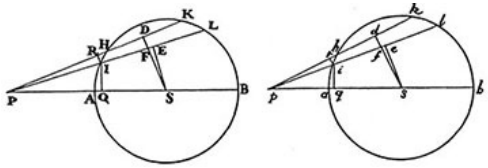
Si hacia cada punto de una superficie esférica se dirigiesen fuerzas centrípetas iguales y decrecientes en razón del cuadrado de la distancia desde dichos puntos: digo que un corpúsculo situado en el interior de dicha superficie no es atraído hacia ningún lado por tales fuerzas.

Sea HIKL la superficie esférica y P el corpúsculo ubicado en el interior. Trácense por P hasta esa superficie las dos líneas HK, IL, que intercepten arcos HI, KL, muy pequeños; y, por la semejanza de los triángulos HPI, LPK (por el Corolario 3 del Lema VII), dichos arcos serán proporcionales a las distancias HP, LP; y las partículas ubicadas en HI, KL, de la superficie esférica delimitada por rectas que pasen por P en cualquier dirección se hallarán bajo la susodicha razón del cuadrado. Luego las fuerzas de dichas partículas ejercidas hacia el cuerpo P son iguales entre sí. Puesto que son directamente como las partículas, e inversamente como el cuadrado de las distancias. Y esas dos razones componen una razón de igualdad. Pero las atracciones iguales hechas en sentido contrario se destruyen mutuamente. Por la misma razón todas las atracciones de toda la superficie esférica son destruidas por atracciones contrarias. Por consiguiente, el cuerpo P no será impelido hacia parte alguna por dichas atracciones. Q. E. D.



Proposición LXXI. Teorema XXXI

Con los mismos supuestos, digo que un corpúsculo situado fuera de una superficie esférica es atraído hacia el centro de la esfera con una fuerza inversamente proporcional al cuadrado de su distancia al centro de la esfera.



Sean AHKB, *ahkb*, dos superficies esféricas iguales descritas con centros en S, *s*, y diámetros AB, *ab*, y sean P, *p*, corpúsculos situados en el exterior sobre la prolongación de dichos diámetros. Desde los corpúsculos trácense las líneas PHK, PIL y *phk*, *pil*, seccionando de los círculos máximos AHB, *ahb*, los arcos iguales HK, *hk*, e IL, *il*: desciendan sobre ellas las perpendiculares SD, *sd*; SE, *se*; IR, *ir*; tales que SD, *sd*, corten a PL, *pl*, en F y *f*; caigan también sobre los diámetros las perpendiculares IQ, *iq*. Desvanézcanse los ángulos DPE, *dpe*; y por ser iguales las líneas DS y *ds*, ES y *es*, pueden considerarse iguales las líneas PE, PF, y *pe*, *pf*, así como los pequeños segmentos DF, *df*; puesto que su razón última es de igualdad al desvanecerse a la vez los ángulos DPE, *dpe*. Una vez hecho esto, PI será a PF como RI a DF, y *pf* a *pi* como *df* o DF a *ri*; y por tanto PI x *pf* a PF x *pi* como RI a *ri*, esto es (por el Corolario 3 del Lema VII) como el arco IH al arco *ih*. De nuevo PI a PS como IQ a SE, y *ps* a *pi* como *se* o SE a *iq*; y por tanto PI x *ps* a PS x *pi* como IQ a *iq*. Y el producto de las razones PI^2 x *pf* x *ps* es a pz^2 x PF x PS, como IH x IQ a *ih* x *ig*; esto es, como la superficie circular que describiría el arco IH al girar el semicírculo AKB sobre el diámetro AB es a la superficie circular que describiría el arco *ih* al girar el semicírculo *akb* sobre el diámetro *ab*. Y las fuerzas con las cuales estas superficies atraen según líneas tendentes hacia ellas a los corpúsculos P y *p* son (por hipótesis) directamente como dichas superficies e inversamente como el cuadrado de las distancias entre los cuerpos y las superficies, esto es, como *pf* x *ps* a PF x PS. Y estas fuerzas son a sus partes oblicuas que (una vez descompuestas las fuerzas según el Corolario II de las Leyes) tienden al centro según las líneas PS, ps, como PI a PQ, y *pi* a *pq*; esto es (por la semejanza de los triángulos PIQ y PSF, *piq* y *psf*) como PS a PF y *ps* a *pf*. De donde resulta que la atracción del corpúsculo P hacia S es a la atracción del corpúsculo *p* hacia *s* como $\frac{\text{PF x } pf \text{ x } ps}{\text{PS}}$ a $\frac{pf \text{ x PF x PS}}{ps}$, esto es, como ps^2 a PS^2. Y por igual razón las fuerzas con las cuales las superficies descritas por la revolución de los arcos KL y *kl* atraen a los corpúsculos serán como

ps^2 a PS^2. hallándose siempre en esta proporción las fuerzas de todas las superficies esféricas en que puedan dividirse ambas superficies esféricas, siempre que se tomen *sd* igual a SD y *se* igual a SE. Y, por composición, las fuerzas de todas las superficies esféricas ejercidas sobre los corpúsculos se hallarán en la misma proporción. Q. E. D.

Proposición lxxii. Teorema xxxii

Si hacia cada punto de una esfera tienden fuerzas centrípetas iguales y decrecientes como el cuadrado de las distancias desde dichos puntos; y si está dada tanto la densidad como la razón del diámetro de la esfera a la distancia desde su centro hasta el corpúsculo: digo que la fuerza con la cual el corpúsculo es atraído será proporcional al semidiámetro de la esfera.

Pues supongamos que dos corpúsculos son atraídos cada uno de ellos por dos esferas, uno por una y otro por otra, y que sus distancias a los centros de las esferas son respectivamente proporcionales a los diámetros de las esferas, a la vez que las esferas se hallan compuestas de partículas semejantes y semejantemente dispuestas respecto a los corpúsculos. En tal caso las atracciones de un corpúsculo respecto a cada partícula de una esfera serán a las atracciones del otro corpúsculo respecto a cada partícula de la otra esfera como la razón compuesta de la razón directa de las partículas y de la razón inversa del cuadrado de las distancias. Pero las partículas son como las esferas, esto es, como el cubo de los diámetros, y las distancias son como los diámetros; y la primera razón directamente junto con el cuadrado de la segunda inversamente es la razón de diámetro a diámetro. Q. E. D.

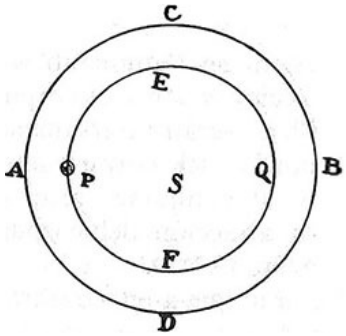
Corolario 1. De aquí que si giran unos corpúsculos en círculos en torno a esferas que constan de materia igualmente atractiva, y las distancias a los centros de esas esferas son proporcionales a los diámetros de las mismas; los tiempos periódicos serán iguales.

Corolario 2. Y viceversa, si los tiempos periódicos son iguales, las distancias serán proporcionales a los diámetros. Estos dos constan por el Corolario 3 de la Proposición iv.

Corolario 3. Si hacia cada punto de dos sólidos cualesquiera semejantes e igualmente densos tienden fuerzas centrípetas iguales y decrecientes según el cuadrado de las distancias a dichos puntos, las fuerzas, por las cuales los corpúsculos son atraídos hacia tales sólidos desde posiciones semejantes respecto a ellos, serán entre sí como los diámetros de los sólidos.

Proposición lxxiii. Teorema xxxiii

Si hacia cada punto de una esfera dada tienden fuerzas centrípetas iguales y decrecientes según el cuadrado de la distancia a dichos puntos, digo que un corpúsculo situado dentro de la esfera es atraído con una fuerza proporcional a su distancia al centro.



En la esfera ABCD, trazada con centro en S, póngase el corpúsculo P; y con centro en el mismo S y distancia SP supóngase trazada la esfera PEQF. Es evidente (por la Proposición lxx) que las superficies esféricas concéntricas de que se compone la diferencia entre esferas AEBF en nada actúa sobre el cuerpo P, al destruirse sus atracciones por las contrarias. Sólo queda la atracción de la esfera interior PEQF. Y (por la Proposición lxxii) ésta es como la distancia PS. Q. E. D.

Escolio

Las superficies de que se componen los cuerpos no son aquí puramente matemáticas, sino orbes tan tenues que su espesor es casi nulo; o sea, orbes evanescentes de los cuales consta la esfera al final cuando el número de dichos orbes aumenta y su espesor disminuye hasta el infinito. Igualmente hay que entender los puntos, de los que se dice que constituyen las líneas, las superficies y los sólidos, como partículas iguales de magnitud despreciable.

Proposición lxxiv. Teorema xxxiv

Con los mismos supuestos, digo que un corpúsculo situado fuera de una esfera es atraído con una fuerza inversamente proporcional al cuadrado de su distancia al centro de la misma.

Pues divídase la esfera en innumerables superficies esféricas concéntricas, y las atracciones sobre el corpúsculo resultantes de cada una de las superficies serán

inversamente proporcionales al cuadrado de la distancia del corpúsculo al centro (por la Proposición LXXI). Y al componer resultará la suma de las atracciones, esto es, la atracción del corpúsculo hacia toda la esfera, en la misma razón. Q. E. D.

COROLARIO 1. De aquí que a iguales distancias de los centros de esferas homogéneas las atracciones son como las esferas. Pues (por la Proposición LXXII) si las distancias son proporcionales a los diámetros de las esferas, las fuerzas serán como los diámetros. Disminúyase en esa razón la distancia mayor; y al ser ahora iguales las distancias, la atracción aumentará como el cuadrado de dicha razón; y, por lo tanto, será a la otra atracción como el cubo de dicha razón, esto es, como la razón de las esferas.

COROLARIO 2. A distancias cualesquiera las atracciones son como las esferas divididas por el cuadrado de las distancias.

COROLARIO 3. Si un corpúsculo se halla fuera de una esfera homogénea, es atraído con una fuerza inversamente proporcional al cuadrado de su distancia al centro de aquélla y la esfera está constituida de partículas atractivas; la fuerza de cualquier partícula decrecerá como el cuadrado de la distancia desde la partícula.

PROPOSICIÓN LXXV. TEOREMA XXXV

Si hacia cada punto de una esfera dada tienden fuerzas centrípetas iguales y decrecientes según el cuadrado de las distancias a cada punto; digo que otra esfera cualquiera semejante será atraída por ella con una fuerza inversamente proporcional al cuadrado de las distancias entre los centros.

Pues la atracción de cada partícula es inversamente como el cuadrado de su distancia al centro de la esfera atrayente (por la Proposición LXXIV) y, por tanto, resulta ser la misma que si toda la fuerza atractiva procediese de un único corpúsculo situado en el centro de la tal esfera. Pero dicha atracción es tanta cuanta, a su vez, sería la atracción del mismo corpúsculo si fuese atraído por cada partícula de la esfera atraída con una fuerza igual a aquélla con la que él mismo atrae. Pero dicha atracción del corpúsculo (por la Proposición LXXIV) sería inversamente proporcional al cuadrado de su distancia al centro de la esfera; y, por lo tanto, la atracción de una esfera igual a ella se halla también en la misma razón. Q. E. D.

COROLARIO 1. Las atracciones de esferas hacia otras esferas homogéneas son como las esferas atrayentes divididas por el cuadrado de las distancias entre sus centros y los de las atraídas.

COROLARIO 2. Lo mismo ocurre cuando la esfera atraída también atrae. Puesto que cada punto de ésta atrae a cada punió de la otra con la misma fuerza con que a su vez son atraídos; por lo tanto, dado que en toda atracción (por la Ley III) son urgidos

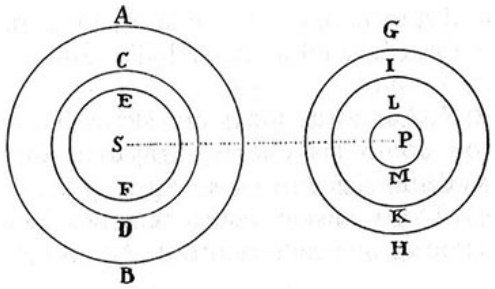
tanto el punto atrayente como el atraído, se igualarán las fuerzas de atracción mutua, manteniéndose las proporciones.

COROLARIO 3. Todo lo anteriormente demostrado sobre el movimiento de los cuerpos en torno al foco de las secciones cónicas vale cuando la esfera atrayente se sitúa en el foco y los cuerpos se mueven fuera de la misma.

COROLARIO 4. Y las cosas que se demostraron sobre el movimiento de los cuerpos en torno al centro de las secciones cónicas valen cuando el movimiento ocurra dentro de la esfera.

PROPOSICIÓN LXXVI. TEOREMA XXXVI

Si las esferas fueren desiguales de forma continua desde el centro hasta la circunferencia (en cuanto a densidad de materia y fuerza atractiva), mientras son iguales en todo lo demás, en todo su perímetro y para cada distancia dada al centro; y la fuerza atractiva de cada punto decrece según el cuadrado de la distancia del cuerpo atraído: digo que la fuerza total con la cual una esfera de este tipo atrae a otra será inversamente proporcional al cuadrado de la distancia entre centros.



Sean varias esferas concéntricas semejantes AB, CD, EF, etc., que al añadir las interiores a las exteriores compongan una materia más densa hacia el centro y al suprimirlas la dejen más tenue; éstas (por la Proposición LXXV) atraerán a cualesquiera otras esferas concéntricas semejantes, GH, IK, LM, etc., cada una a cada una, con fuerzas inversamente proporcionales al cuadrado de la distancia SP. Y por suma o resta, el total de todas esas fuerzas o el exceso de unas sobre otras, esto es, la fuerza con la cual toda la esfera AB, compuesta de todas las concéntricas o de todas las diferencias de las concéntricas, atrae a toda la esfera GH, compuesta de las concéntricas o de las diferencias de las concéntricas, estará en la misma razón. Auméntese hasta el infinito el número de esferas concéntricas de modo que la densidad de la materia a la vez que la fuerza atractiva crezca o decrezca según alguna ley desde la periferia hacia el centro; y añadiendo materia no atractiva complétese donde sea preciso la falta de densidad hasta que las esferas adquieran la forma

deseada; y la fuerza con la cual una atraerá a la otra también ahora estará, por el argumento anterior, en esa misma razón inversa del cuadrado. Q. E. D.

Corolario 1. De aquí que si muchas esferas de éstas, mutuamente semejantes en todo, se atraen entre sí, a distancias iguales cualesquiera entre centros, sus atracciones aceleratrices una a una serán como las esferas atrayentes.

Corolario 2. Y a distancias desiguales cualesquiera, serán como las esferas atrayentes divididas por los cuadrados de las distancias entre centros.

Corolario 3. Y las atracciones motrices o los pesos de una esfera hacia otra serán, a distancias iguales de los centros, como las esferas atrayentes y las atraídas conjuntamente, esto es, como el contenido de las esferas resultante de su multiplicación.

Corolario 4. Y a distancias desiguales, directamente como dichos contenidos e inversamente como los cuadrados de las distancias entre centros.

Corolario 5. Lo mismo vale cuando la atracción procede de la fuerza atractiva de ambas esferas atrayéndose mutuamente una a otra. Pues la atracción se duplica con ambas fuerzas, guardando la proporción.

Corolario 6. Si esferas de esta clase giran en torno a otras en reposo, una en torno a cada una, y las distancias entre los centros de las que giran y los de las que están en reposo fuesen proporcionales a los diámetros de las que están en reposo, los tiempos periódicos serán iguales.

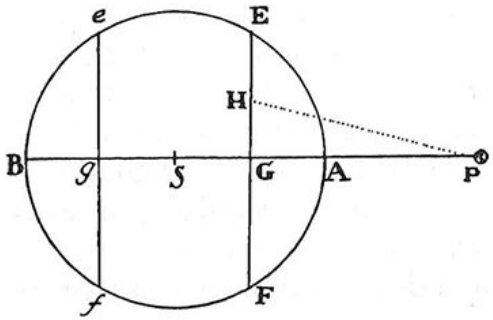
Corolario 7. Y viceversa, si los tiempos periódicos son iguales; las distancias serán proporcionales a los diámetros.

Corolario 8. A esto mismo se atienen todas las cosas demostradas más arriba sobre el movimiento de cuerpos en torno a focos de secciones cónicas, cuando en el foco se sitúa la esfera atractiva de cualquiera forma o condición de las descritas.

Corolario 9. Y de igual modo cuando los cuerpos que giran son también esferas atractivas de cualquier condición ya descrita.

Proposición lxxvii. Teorema xxxvii

Si hacia cada uno de los puntos de esferas tienden fuerzas centrípetas proporcionales a las distancias de los puntos a los cuerpos atraídos: digo que la fuerza compuesta con la cual se atraen mutuamente dos esferas es como la distancia entre los centros de las esferas.



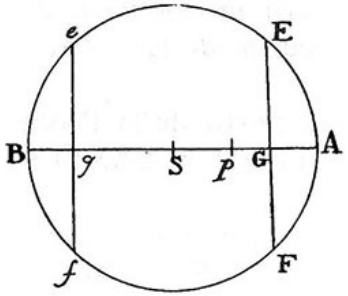
Caso 1. Sea la esfera AEBF, S su centro, P un corpúsculo atraído, PASB el eje de la esfera que pasa por el centro del corpúsculo, EF, ef, dos planos perpendiculares al eje y que cortan la esfera y equidistantes a un lado y otro del centro de la esfera; G, g, las intersecciones de los planos y el eje y H un punto cualquiera en el plano EF. La fuerza centrípeta del punto H sobre el corpúsculo P ejercida según la línea PH es como la distancia PH; y (por el Corolario ii de las Leyes) según la línea PG, o sea hacia el centro S, es como la longitud PG. Por lo tanto, la fuerza de todos los puntos del plano EF, esto es, la fuerza de todo el plano, con la cual el corpúsculo P es atraído hacia el centro S es como la distancia PG multiplicada por el número de puntos, esto es, como el sólido contenido bajo el plano EF y la distancia PG. E igualmente la fuerza del plano ef, por la cual el corpúsculo P es atraído hacia el centro S, es como dicho plano multiplicado por su distancia Pg, o como el plano EF igual al anterior multiplicado por esa distancia Pg; y la suma de las fuerzas de ambos planos es como el plano EF multiplicado por la suma de las distancias PG + Pg, esto es, como dicho plano multiplicado por el doble de la distancia PS del centro hasta el corpúsculo, o como el doble del plano EF multiplicado por la distancia PS, o como la suma de los planos iguales EF + ef multiplicada por esa misma distancia. Y por un argumento similar, las fuerzas de todos los planos en toda la esfera, equidistantes del centro de la misma, son como la suma de los planos multiplicada por la distancia PS, esto es, como toda la esfera y la distancia PS conjuntamente. Q. E. D.

Caso 2. El corpúsculo P atraiga ahora a la esfera AEBF. Por el mismo argumento se probará que la fuerza por la que dicha esfera es atraída será como la distancia PS. Q. E. D.

Caso 3. Sea ahora otra esfera compuesta de innumerables corpúsculos P; y puesto que la fuerza por la que es atraído cada corpúsculo es como la distancia del corpúsculo al centro de la primera esfera y la esfera conjuntamente, y por lo tanto es la misma que si toda ella procediera de un corpúsculo único situado en el centro de la esfera; la fuerza total con la cual todos los corpúsculos son atraídos hacia la segunda esfera, esto es, con la cual es atraída toda la dicha esfera, será la misma que si dicha esfera fuera atraída por una fuerza procedente de un único corpúsculo en el centro de la primera esfera, y por tanto proporcional a la distancia entre los centros de las

esferas. Q. E. D.

CASO 4. Si las esferas se atraen mutuamente se duplicarán las fuerzas manteniéndose la proporción. Q. E. D.



CASO 5. Sitúese ahora el corpúsculo *p* dentro de la esfera AEBF; y puesto que la fuerza del plano *ef* sobre el corpúsculo es como el sólido contenido bajo dicho plano y la distancia *pg*, y la fuerza contraria del plano EF es como el sólido contenido bajo dicho plano y la distancia *p*G, la fuerza compuesta de ambas será como la diferencia de los sólidos, esto es, como la suma de los planos iguales multiplicada por la semidiferencia de las distancias, esto es, como dicha suma multiplicada por la distancia *p*S, distancia del corpúsculo al centro de la esfera. Y, por un argumento similar la atracción de todos los planos EF, *ef*, de toda la esfera, esto es, la atracción de toda la esfera es conjuntamente como la suma de todos los planos, o sea la esfera toda y, como *p*S, distancia del corpúsculo al centro de la esfera Q. E. D.

CASO 6. Y si se compone una nueva esfera con innumerables corpúsculos *p* situada dentro de la esfera anterior AEBF, se probará como antes que la atracción, tanto simple de una esfera hacia otra, como mutua entre ambas una hacia otra, será como la distancia entre centros *p*S. Q. E. D.

PROPOSICIÓN LXXVIII. TEOREMA XXXVIII

Si al pasar del centro hacia la circunferencia las esferas son constantemente desiguales y heterogéneas, pero en cada círculo a distancias dadas del centro son continuamente iguales; y la fuerza atractiva de cada punto fuese como la distancia del cuerpo atraído: digo que la fuerza total con la cual dos esferas de este tipo se atraerán mutuamente será proporcional a la distancia entre los centros de las esferas.

Esto se demuestra a partir de la Proposición precedente, del mismo modo que la Proposición LXXVI se demostró a partir de la Proposición LXXV.

COROLARIO. Lo que se demostró más arriba en las Proposiciones X y LXIV sobre el

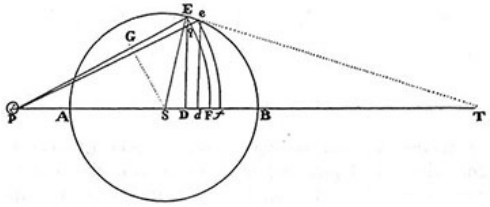
movimiento de los cuerpos en torno a los centros de secciones cónicas, tiene valor cuando todas las atracciones surgen de fuerzas de cuerpos esféricos del tipo recién descrito, y los cuerpos atraídos son esferas también del mismo tipo.

ESCOLIO

Acabo de exponer los dos casos más notables de atracciones; a saber, cuando las fuerzas centrípetas decrecen en razón del cuadrado de las distancias, o cuando crecen según la razón simple de las distancias, haciendo en ambos casos que los cuerpos giren en secciones cónicas, y componiendo con dicha ley las fuerzas centrípetas de los cuerpos esféricos al separarse del centro haciéndolas crecer o decrecer según el caso. Lo cual es digno de tenerse en cuenta. Resultaría demasiado prolijo contemplar uno a uno los demás casos que ofrecen soluciones menos elegantes. Prefiero incluirlos y determinarlos a la vez a todos juntos con el método general siguiente.

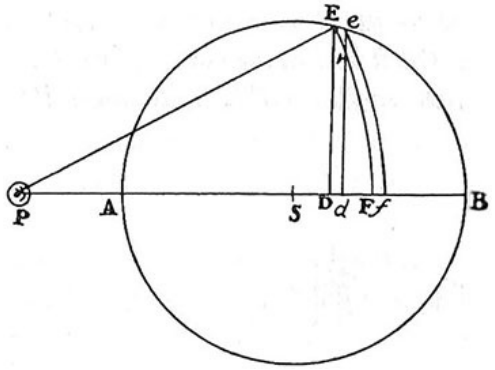
LEMA XXIX

Si en torno a un centro S se describe un círculo cualquiera AEB y se trazan con centro en P dos círculos EF, ef, que cortan al primero en E, e, y a la línea PS en F, f; y sobre PS descienden las perpendiculares ED, ed: digo que, si se supone que disminuyen infinitamente las distancias de los arcos EF, ef, la razón última de la línea evanescente Dd a la línea evanescente Ff será la misma que la de la línea PE a la línea PS.



Pues si la línea Pe corta al arco EF en *q*; y la recta E*e*, que coincide con el arco evanescente E*e*, prolongada, corta a la recta PS en T; y desde S se traza sobre PE la normal SG: por la semejanza de los triángulos DTE, *d*T*e*, DES; D*d* será a E*e* como DT a TE, o como DE a ES; y por la semejanza de los triángulos E*eq*, ESG (por el Lema VIII y el Corolario 3 del Lema VII), E*e* será a *eq* o F*f* como ES a SG; y, por consiguiente, D*d* es a F*f* como DE a SG; esto es (por la semejanza de los triángulos PDE, PGS) como PE a PS. Q. E. D.

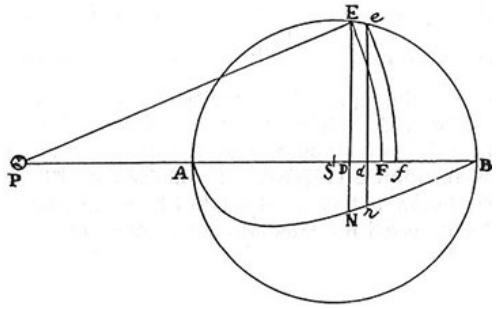
Si una superficie EF fe llega a ser evanescente al disminuir continuamente hasta el infinito su anchura y con su giro en torno al eje PS describe un sólido esférico cóncavo-convexo hacia cada una de cuyas partículas iguales tienden fuerzas centrípetas iguales: digo que la fuerza con la cual dicho sólido atrae a un corpúsculo situado en P está en razón compuesta de la razón del sólido DE^2 x Ff y la razón de la fuerza con la cual la partícula dada en el lugar Ff atraería al mismo corpúsculo.



Pues si consideramos primero la fuerza de la superficie esférica FE generada por el giro del arco FE y cortada en un punto cualquiera *r* por la línea *de*; la parte anular de la superficie, generada por el giro del arco *r*E, será como el segmento D*d*, manteniéndose el radio PE de la esfera (como demostró ARQUÍMEDES en el libro *De Sphera et Cylindro*). Y la fuerza de tal superficie, ejercida según las líneas PE o P*r* colocadas en la superficie cónica en cualquier punto, será como dicha superficie cónica anular; esto es, como el segmento D*d* o, lo que es lo mismo, como el rectángulo comprendido por el radio de la esfera PE y el segmento D*d*: pero si se ejerce según la línea PS tendente hacia el centro S será menor en la razón de PD a PE, y por tanto como PD x D*d*. Supóngase ahora que la línea DF se divide en innumerables partículas iguales, cada una de las cuales se denomina D*d*; entonces la superficie FE quedará dividida en otros tantos anillos iguales, cuyas fuerzas serán como la suma de todos los PD x D*d*, esto es, como $\frac{1}{2}$PF^2 - $\frac{1}{2}$PD^2 y, por lo tanto, como DE^2. Ahora multiplíquese la superficie FE por la altura F*f* y resultará la fuerza del sólido EF*fe* ejercida sobre el corpúsculo P como DE^2 x F*f*, esto cuando se dé la fuerza que una partícula dada F*f* ejerce sobre el corpúsculo P a la distancia PF. Pero si no se da dicha fuerza, la fuerza del sólido EF *fe* será como el sólido DE^2 x F*f* y la fuerza no dada conjuntamente. Q. E. D.

Si hacia cada partícula igual de una esfera ABE, descrita con centro en S, tienden fuerzas centrípetas iguales, y desde cada punto D sobre el eje de la esfera AB, en el que se halla situado un corpúsculo P, se elevan las perpendiculares DE que tocan a la esfera en E, y sobre ellas se toman las longitudes DN, que son conjuntamente como $\frac{DE^2 \text{ x } PS}{PE}$ *y la fuerza que la partícula de la esfera situada en el eje a la distancia PE ejerce sobre el corpúsculo P: digo que la fuerza total con la cual el corpúsculo P es atraído hacia la esfera es como el área ANB comprendida bajo el eje AB de la esfera y la línea curva ANB, de la cual el punto N es siempre tangente.*

Pues, manteniendo las construcciones del Lema y del Teorema anteriores, imagínese al eje de la esfera AB dividido en innumerables partículas iguales D*d*, y a toda la esfera dividida en otras tantas láminas esféricas cóncavo-convexas EFF*e*; y elévese la perpendicular *dn*. Por el Teorema anterior, la fuerza con la cual la lámina EF*fe* atrae al corpúsculo P es como DE^2 x F*f* y la fuerza de una sola partícula ejercida a la distancia PE o PF conjuntamente. Pero (por el Lema anterior) D*d* es a F*f* como PE a PS, y por tanto F*f* es igual a $\frac{\text{PS x D}d}{\text{PE}}$, y DE^2 x F*f* es igual a D*d* x $\frac{\text{DE}^2 \text{ x PS}}{\text{PE}}$, y por tanto la fuerza de la lámina EF*fe* es como D*d* x $\frac{\text{DE}^2 \text{ x PS}}{\text{PE}}$ y la fuerza de la partícula ejercida a la distancia PF conjuntamente, esto es (por hipótesis), como DN x D*d*, o el área evanescente DN*nd*. Por tanto, las fuerzas de todas las láminas ejercidas sobre el cuerpo P son como todas las áreas DN*nd*, esto es, la fuerza total de la esfera como toda el área ANB. Q. E. D.



COROLARIO 1. De aquí que si la fuerza centrípeta tendente hacia cada partícula

permanece siempre la misma a todas las distancias y DN es como $\frac{DE^2 \times PS}{PE}$, la fuerza total con la que el corpúsculo es atraído por la esfera será como el área ANB.

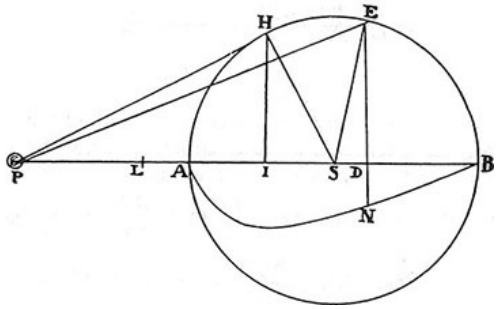
COROLARIO 2. Si la fuerza centrípeta de las partículas es inversamente como la distancia del corpúsculo atraído por ellas y DN como $\frac{DE^2 \times PS}{PE^2}$ la fuerza con la cual el corpúsculo P es atraído por toda la esfera será como el área ANB.

COROLARIO 3. Si la fuerza centrípeta de las partículas es inversamente como el cubo de la distancia del corpúsculo atraído por ellas y DN es como $\frac{DE^2 \times PS}{PE^4}$, la fuerza con la cual el corpúsculo es atraído por toda la esfera será como ANB.

COROLARIO 4. Y, en general, si se supone que la fuerza centrípeta tendente hacia cada partícula de la esfera es inversamente como la cantidad V y DN como $\frac{DE^2 \times PS}{PE \times V}$, la fuerza con la que el corpúsculo es atraído por toda la esfera será como el área ANB.

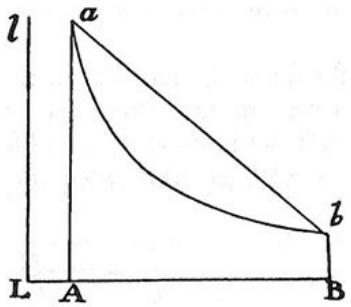
PROPOSICIÓN LXXXI. PROBLEMA XLI

Suponiendo lo anterior, mídase el área ANB.



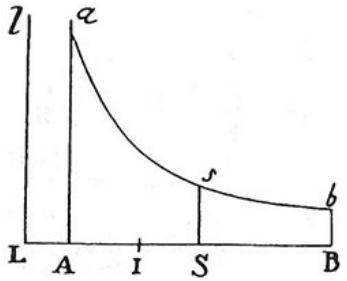
Desde el punto P trácese la recta PH tangente a la esfera en H, y sobre el eje PAB descienda la normal HI y biséquese PI en L; y (por la Proposición XII del libro II de los [*Elementos*]) $PE^2 = PS^2 + SE^2 + 2PSD$. Pero SE^2 o SH^2 (por la semejanza de los triángulos SPH, SHI) es igual al rectángulo PSI. Por tanto, PE^2 es igual al contenido bajo PS y PS + SI + 2SD, esto es, bajo PS y 2LS + 2SD, esto es, bajo PS y 2LD.

Además, DE^2 es igual a $SE^2 - SD^2$, o sea, $SE^2 - LS^2 + 2SLD - LD^2$, esto es, $2SLD - LD^2 - ALB$. Pues $LS^2 - SE^2$ o $LS^2 - SA^2$ (por la Proposición VI del libro II de los [*Elementos*]) es igual al rectángulo ALB. Si por DE^2 se escribe $2SLD - LD^2 - ALB$; y la cantidad $\frac{DE^2 \times PS}{PE \times V}$, que según el Corolario 4 de la Proposición anterior es como la longitud de la ordenada aplicada en DN, quedará descompuesta en tres partes $\frac{2SLD \times PS}{PE \times V} - \frac{LD^2 \times PS}{PE \times V} - \frac{ALB \times PS}{PE \times V}$: donde si en lugar de V escribimos la razón inversa de la fuerza centrípeta, y en lugar de PE la media proporcional entre PS y 2LD, dichas tres partes resultan ordenadas de las curvas correspondientes, cuyas áreas se obtienen por métodos ya comunes. Q. E. F.

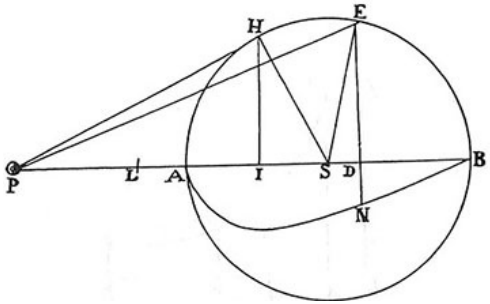


EJEMPLO 1. Si la fuerza centrípeta tendente hacia cada partícula de la esfera fuese inversamente como la distancia; escríbase por V la distancia PE y en lugar de PE^2 escríbase 2PS x LD, y DN será como $SL - \frac{1}{2}LD - \frac{ALB}{2LD}$. Supóngase a DN igual a su doble $2SL - LD - \frac{ALB}{LD}$; y transportando a lo largo de la longitud AB la parte dada de la ordenada 2SL describirá el área rectangular 2SL x AB; mientras la parte indefinida LD, transportada normalmente por la misma longitud, con un movimiento continuo tal que al moverse crezca o decrezca para permanecer siempre igual a la longitud LD, describirá el área $\frac{LB^2 - LA^2}{2}$, esto es, el área SL x AB; la cual restada del área anterior 2SL x AB deja el área SL x AB. Y la tercera parte $\frac{ALB}{LD}$, transportada también con un movimiento local normal sobre la misma longitud, describirá un área hiperbólica; la cual, restada del área SL x AB dejará el área buscada ANB. De donde surge la construcción siguiente del problema. Elévense por los puntos L, A, B, las

perpendiculares L*l*, A*a*, B*b*, haciendo que A*a* sea igual a LB y B*b* igual a LA. Siendo asíntotas L*l*, LB, trácese por los puntos *a*, *b*, la hipérbola *ab*. Y trazada la cuerda *ba* cerrará el área *aba* igual al área buscada ANB.



EJEMPLO 2. Si la fuerza centrípeta tendente hacia cada partícula de la esfera fuere inversamente como el cubo de la distancia, o (lo que es igual) como dicho cubo dividido por un plano dado; escríbase $\frac{PE^2}{2AS^2}$ en lugar de V, ya también 2PS x LD en lugar de PE^2, y DM será como $\frac{SL \times AS^2}{PS \times LD} - \frac{AS^2}{2PS} - \frac{ALB \times AS^2}{2PS \times LD^2}$, esto es (por la proporcionalidad continua de PS, AS, SI) como $\frac{LSI}{LD} - ½SI - \frac{ALB \times SI}{2LD^2}$. Si se llevan estas tres partes sobre la longitud AB, la primera de ellas $\frac{LSI}{LD}$ generará un área hiperbólica; la segunda ½SI el área ½AB x SI; la tercera $\frac{ALB \times SI}{2LD^2}$ el área $\frac{ALB \times SI}{2LA} - \frac{ALB \times SI}{2LB}$, esto es, ½AB x SI. Réstese de la primera suma de la segunda y la tercera y quedará el área buscada ANB. De donde la construcción del problema resulta como sigue. Por los puntos L, A, S, B, elévense las perpendiculares L*l*, A*a*, S*s*, B*b*, de las cuales S*s* sea igual a SI y por el punto *s* con las asíntotas L*l*, LB, trácese la hipérbola *asb* que encuentre a las perpendiculares A*a* y B*b* en *a* y *b*; y restando el rectángulo 2ASI del área hiperbólica A*asb*B quedará el área buscada ANB.



EJEMPLO 3. Si la fuerza centrípeta tendente hacia cada partícula de la esfera decrece como la cuarta potencia de la distancia de las partículas, escríbase $\frac{PE^4}{2AS^3}$ en lugar de V, después $\sqrt{(2PS \times LD)}$ en lugar de PE, y DN vendrá a ser como $\frac{SI^2 \times SL}{\sqrt{2SI}} \times \frac{1}{\sqrt{LD^3}} - \frac{SI^2}{2\sqrt{2SI}} \times \frac{1}{\sqrt{LD}} - \frac{SI^2 \times ALB}{2\sqrt{2SI}} \times \frac{1}{\sqrt{LD^5}}$; y al llevar estas tres partes sobre la longitud AB producen otras tantas áreas, a saber: $\frac{2SI^2 \times SL}{\sqrt{2SI}}$ en $\left(\frac{1}{\sqrt{LA}} - \frac{1}{\sqrt{LB}}\right)$; $\frac{SI^2}{\sqrt{2SI}}$ en $(\sqrt{LB} - \sqrt{LA})$; y $\frac{SI^2 \times ALB}{3\sqrt{2SI}}$ en $\left(\frac{1}{\sqrt{LA^3}} - \frac{1}{\sqrt{LB^3}}\right)$. Y tras la reducción debida dan $\frac{2SI^2 \times SL}{LI}$, SI^2, y $SI^2 + \frac{2SI^3}{3LI}$. De estas, a su vez, después de resta las últimas de la primera, resulta $\frac{4SI^3}{3LI}$. Por tanto la fuerza total con la cual el corpúsculo P es atraído hacia el centro de la esfera es como $\frac{SI^3}{PI}$ esto es, como el inverso de PS^3 x PI. Q. E. I.

Con el mismo método podemos determinar la atracción de un corpúsculo ubicado dentro de la esfera, pero es más directo por el Teorema siguiente.

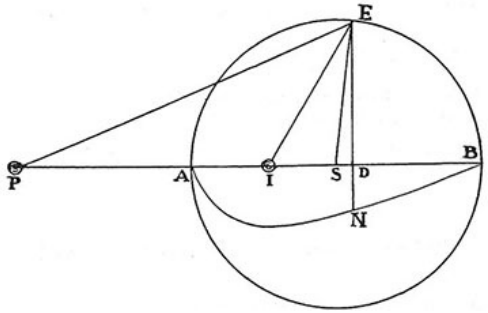
PROPOSICIÓN LXXXII. TEOREMA XLI

En una esfera descrita con centro en S y radio SA, si se toman SI, SA, SP, continuamente proporcionales: digo que la atracción de un corpúsculo dentro de la

esfera en un lugar cualquiera I es a la atracción del mismo juera de la esfera en el lugar P en razón compuesta de la raíz cuadrada de la razón de las distancias al centro IS, PS, y de la raíz cuadrada de la razón de las fuerzas centrípetas tendentes al centro en dichos lugares P, I.

Puesto que si las fuerzas centrípetas de las partículas de la esfera fuesen inversamente como la distancia al corpúsculo atraído por ellas, la fuerza, con la cual el corpúsculo situado en I es atraído por toda la esfera, será a la fuerza con la cual es atraído en P, en razón compuesta de la raíz cuadrada de la razón de las distancias SI a SP, y de la raíz cuadrada de la razón de la fuerza centrípeta en el lugar I de una partícula que la produce desde el centro, respecto a la fuerza centrípeta en P generada también por la misma partícula desde el centro, esto es, inversamente como la raíz cuadrada de las distancias respectivas SI, SP. Estas dos razones de ralees cuadradas componen razón de igualdad, y por tanto, las atracciones en I y P producidas por toda la esfera son iguales. Con un cálculo similar se encontrará que, si las fuerzas de las partículas de la esfera son inversamente como el cuadrado de la razón de las distancias, la atracción en I es a la atracción en P como la distancia SP al semidiámetro SA de la esfera; si dichas fuerzas son inversamente como la razón cúbica de las distancias, las atracciones en I y P serán entre sí como SP^2 a SA^2; si fuesen como la cuarta potencia, como SP^3 a SA^3. Por tanto, como la atracción en P, en este último caso, resulta ser inversamente como PS^3 x PI, la atracción en I será inversamente como SA^3 x PI, esto es (por estar dado SA^3) inversamente como PI. Y así sucesivamente hasta el infinito. El teorema, pues, se demuestra así:

Manteniendo las construcciones anteriores, y existiendo un corpúsculo en un lugar cualquiera P, la ordenada DN resultó ser como $\frac{DE^2 \text{ x } PS}{PE \text{ x } V}$. Luego si se traza IE, ordenada para otro lugar I del corpúsculo, «*mutatis mutandis*», resultará ser como $\frac{DE^2 \text{ x } IS}{IE \text{ x } V}$.

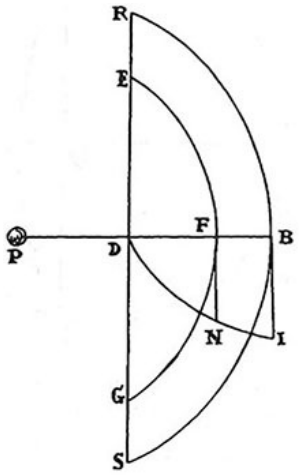


Supóngase que las fuerzas centrípetas procedentes de un punto E cualquiera de la esfera son entre sí a las distancias IE y PE como PE^n a IE^n (donde *n* representa la potencia de PE y de IE) y dichas ordenadas resultarán como $\frac{DE^2 \text{ x } PS}{PE \text{ x } PE^n}$ y $\frac{DE^2 \text{ x } IS}{IE \text{ x } IE^n}$, cuya razón mutua es como PS x IE x IE^n a IS x PE x PE^n. Por ser SI, SE, SP, continuamente proporcionales, los triángulos SPE, SEI, son semejantes, y por ello IE es a PE como IS a SE o SA; en lugar de la razón IE a PE escríbase la razón IS a SA; y la razón de las ordenadas resultará ser PS x IE^n a SA x PE^n. Pero la razón de PS a la raíz cuadrada de SA es la razón de las distancias PS, SI; y la razón de IE^n a la raíz cuadrada de PE^n (por ser IE a PE como IS a SA) es la razón de las fuerzas para las distancias PS, IS. Luego las ordenadas, y las áreas que describen las ordenadas, y las atracciones proporcionales a éstas, están en razón compuesta de las raíces cuadradas de aquellas razones. Q. E. D.

Proposición lxxxiii. Problema xlii

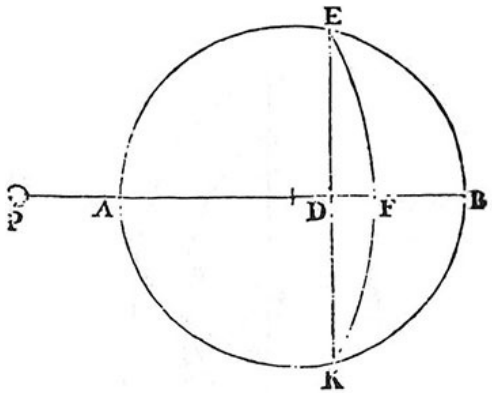
Hallar la fuerza con la cual un corpúsculo situado en el centro de una esfera es atraído hacia cualquier segmento de la misma.

Sea P un cuerpo en el centro de la esfera y RBSD el sector de la misma contenido entre el plano RDS y la superficie esférica RBS. Con la superficie esférica EFG descrita con centro en P córtese DB en F, y distínganse las dos partes del sector BREFGS y FEDG. Sea dicha superficie no algo puramente matemático, sino físico, que tiene una profundidad mínima. Llámese a esta profundidad O, y esta superficie será (por las demostraciones de Arquímedes) como PF x DF x O. Supongamos además que las fuerzas atractivas de las partículas de la esfera son inversamente como la potencia de las distancias de índice *n*; y la fuerza con la cual la superficie EFG atrae al cuerpo P será (por la Proposición lxxix) como $\frac{DE^2 \text{ x } O}{PF^n}$, esto es, como $\frac{2DF \text{ x } O}{PF^{n-1}} - \frac{DF^2 \text{ x } O}{PF^n}$. Sea la perpendicular FN multiplicada por O proporcional a esta cantidad; y el área curvilínea BDI, descrita por la ordenada FN al ser aplicada sobre la longitud DB con un movimiento continuo, será como la fuerza total con la cual el segmento completo RBSD atrae al cuerpo P. Q. E. I.



Proposición lxxxiv. Problema xliii

Hallar la fuerza con la cual un corpúsculo situado sobre el eje de cualquier segmento de una esfera y fuera del centro de la misma es atraído por dicho segmento.



Atraiga el segmento EBK al corpúsculo P situado sobre su eje ADB. Con centro en P y distancia PE descríbase la superficie esférica EFK mediante la cual se segregan las dos partes del segmento EBKFE y EFKDE. Hállese la fuerza de la parte primera por la Proposición lxxxi y la fuerza de la segunda por la Proposición lxxxiii; y la suma de las fuerzas será la fuerza del segmento completo EDKDE. Q. E. I.

Escolio

Explicadas las atracciones de los cuerpos esféricos, es ya hora de pasar a las leyes de las atracciones de otros cuerpos que constan también de partículas atractivas; pero tratarlas minuciosamente es menos necesario para mi propósito. Será suficiente, por su escaso uso en temas filosóficos, añadir algunas Proposiciones más generales sobre las fuerzas de tales cuerpos y de los movimientos resultantes.

Sección XIII
SOBRE LAS FUERZAS ATRACTIVAS DE CUERPOS NO ESFÉRICOS

Proposición LXXXV. Teorema XLII

Si la atracción padecida por un cuerpo es mucho mayor cuando está contiguo al atrayente que cuando están separados aunque sea por un intervalo mínimo; las fuerzas de las partículas del cuerpo atrayente decrecen al separarse del cuerpo atraído en razón mayor que el cuadrado de las distancias de las partículas.

Pues si las fuerzas decrecen en razón del cuadrado de las distancias de las partículas, la atracción hacia un cuerpo esférico, dado que (por la Proposición LXXIV) es inversamente como el cuadrado de la distancia al cuerpo atraído desde el centro de la esfera, no aumentará sensiblemente por el contacto; y todavía aumentará menos por el contacto si la atracción decrece en una proporción menor al retroceder el cuerpo atraído. La Proposición resulta, pues, evidente tratándose de esferas atractivas. Y lo mismo ocurre con orbes esféricos cóncavos que atraen a cuerpos externos. Y todavía es más claro para orbes que atraen a cuerpos ubicados en su interior, dado que las atracciones difundidas por las cavidades de los orbes en todas direcciones son (por la Proposición LXX) destruidas por las atracciones contrarias, y por ello son nulas hasta en el mismo contacto. Pero si de estas esferas y orbes esféricos se suprimen partes cualesquiera distantes del lugar de contacto y se añaden partes nuevas en otros lugares, pueden cambiarse a voluntad las figuras de los cuerpos atractivos, sin que las partes suprimidas o añadidas, al estar alejadas del lugar de contacto, aumenten notablemente el exceso de atracción que procede del contacto. Por tanto la Proposición se mantiene para cuerpos de cualesquiera figuras. Q. E. D.

Proposición LXXXVI. Teorema XLIII

Si las fuerzas de las partículas de que consta un cuerpo atractivo decrecen, al alejarse el cuerpo atraído, en razón cúbica o mayor de la distancia de las partículas, la atracción en el lugar de contacto será mucho mayor que cuando los cuerpos

atrayente y atraído se hallan separados por un intervalo aunque sea mínimo.

Pues, que la atracción aumenta hasta el infinito cuando el corpúsculo atraído se aproxima a una esfera atractiva de este tipo consta por la solución al Problema XLI, mostrada en los ejemplos segundo y tercero. Lo mismo se infiere fácilmente a partir de dichos ejemplos comparados con el Teorema XLI para el caso de las atracciones de cuerpos hacia orbes cóncavo-convexos, tanto si los cuerpos atraídos se sitúan fuera como dentro de dichas cavidades. Y añadiendo o quitando a estas esferas y orbes por doquier menos en el lugar de contacto cuanta materia atractiva se precise hasta que adquieran la figura requerida, la Proposición se mantendrá para todos los cuerpos. Q. E. D.

Proposición LXXXVII. Teorema XLIV

Si dos cuerpos entre sí semejantes que constan de materia igualmente atractiva atraen cada uno a un corpúsculo que le es proporcional y está ubicado respecto a ellos de modo semejante, las atracciones aceleratrices de los corpúsculos hacia los cuerpos enteros serán como las atracciones aceleratrices de los corpúsculos hacia partículas de aquellos cuerpos proporcionales a los todos y ubicados semejantemente en los todos.

Pues, si los cuerpos se dividen en partículas que sean proporcionales a los todos y ubicadas semejantemente en los todos, la atracción hacia una partícula cualquiera de un cuerpo será a la atracción hacia la correspondiente partícula del otro cuerpo como las atracciones hacia cada partícula del primer cuerpo son a las atracciones hacia cada partícula correspondiente del otro; y componiendo, también es así la atracción hacia todo el primer cuerpo respecto a la atracción hacia todo el segundo. Q. E. D.

Corolario 1. Luego si las fuerzas atractivas de las partículas, al aumentar las distancias de los corpúsculos atraídos, decreciesen en razón de alguna potencia de las distancias, las atracciones aceleratrices hacia los cuerpos enteros serán directamente como los cuerpos e inversamente como las dichas potencias de las distancias. Y si las fuerzas de las partículas decrecen en razón del cuadrado de las distancias respecto a los corpúsculos atraídos, y los cuerpos son como A^3 y B^3 y por ello tanto los lados cúbicos de los cuerpos como las distancias de los corpúsculos atraídos hasta los cuerpos son como A y B, las atracciones aceleratrices hacia los cuerpos serán como $\frac{A^3}{A^2}$ y $\frac{B^3}{B^3}$, esto es, como los lados cúbicos aquellos A y B de los cuerpos. Si las fuerzas de las partículas decrecen en razón cúbica de las distancias a los corpúsculos

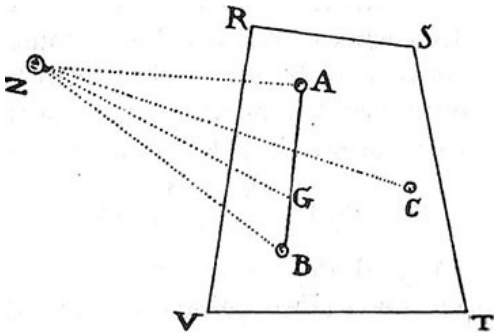
atraídos, las atracciones aceleratrices hacia los cuerpos completos serán como $\frac{A^3}{A^3}$ y $\frac{B^3}{B^3}$, esto es, iguales. Si las fuerzas decrecen en razón de la cuarta potencia, las atracciones hacia los cuerpos serán como $\frac{A^3}{A^4}$ y $\frac{B^3}{B^4}$, esto es, inversamente como los lados cúbicos A y B. Y así para los demás.

COROLARIO 2. De donde, a la inversa, de las fuerzas con las cuales cuerpos semejantes atraen a corpúsculos ubicados semejantemente respecto a ellos, se puede inferir la razón del decremento de las fuerzas atractivas de las partículas al alejarse el corpúsculo atraído, siempre que dicho decremento se halle en alguna razón directa o inversa respecto a las distancias.

PROPOSICIÓN LXXXVIII. TEOREMA XLV

Si las fuerzas atractivas de partículas iguales de un cuerpo cualquiera fuesen como las distancias de los lugares a las partículas, la fuerza del cuerpo entero tenderá hacia su centro de gravedad; y será la misma que la de un globo que constase de materia semejante e igual, y que tuviese su centro en el centro de gravedad.

Las partículas A, B, del cuerpo RSTV atraen a un corpúsculo cualquiera Z con fuerzas que, si las partículas son iguales entre sí, son como las distancias AZ, BZ; y si son desiguales, serán como dichas partículas y sus distancias AZ, BZ, conjuntamente, o también (por así decirlo) como dichas partículas multiplicadas respectivamente por sus distancias AZ, BZ. Expresemos estas fuerzas mediante dichos productos, A x AZ y B x BZ. Únase AB y córtesela en G de tal modo que AG sea a BG como la partícula B es a la partícula A; y G será el centro común de gravedad de las partículas A y B. La fuerza A x AZ (por el Corolario II de las Leyes) se descompondrá en las fuerzas A x GZ y A x AG, mientras la fuerza B x BZ se descompondrá en B x GZ y B x BG. Pero las fuerzas A x AG y B x BG, al ser proporcionales A respecto a B y BG respecto a AG, son iguales; y como tienen direcciones contrarias se anulan mutuamente. Quedan las fuerzas A x GZ y B x GZ. Estas tienden desde Z hacia el centro G y componen la fuerza (A + B) x GZ; esto es, la misma fuerza que si las partículas atractivas A, B, estuviesen ubicadas en su centro común de gravedad G, formando allí un globo.



Por el mismo argumento, si se añade una tercera partícula C, y se compone su fuerza con la fuerza (A + B) x GZ tendente hacia el centro G; la fuerza resultante de ello tenderá al centro común de gravedad de dicho globo en G y de la partícula C; esto es, hacia el común centro de gravedad de las tres partículas A, B, C; y será la misma que si el globo y la partícula C se hallasen en dicho centro común, formando allí un globo mayor. Y así se puede seguir hasta el infinito. Por lo mismo, la fuerza total de todas las partículas de un cuerpo cualquiera RSTV es la misma que si dicho cuerpo, manteniendo el centro de gravedad, tomase la forma de un globo. Q. E. D.

COROLARIO. De aquí que el movimiento del cuerpo atraído Z será el mismo que si el cuerpo atrayente RSTV fuese esférico; y por tanto lo mismo si el cuerpo atrayente reposa como si se mueve uniformemente en línea recta, el cuerpo atraído se moverá en una elipse que tiene su centro en el centro de gravedad del cuerpo atrayente.

PROPOSICIÓN LXXXIX. TEOREMA XLVI

Si se tienen varios cuerpos que constan de partículas iguales, cuyas fuerzas son como las distancias de los lugares hasta cada uno: la fuerza compuesta de todas las fuerzas con la cual es atraído un corpúsculo cualquiera tenderá al centro común de gravedad de los atrayentes; y será la misma que si dichos cuerpos atrayentes, manteniendo el mismo centro común de gravedad, se reuniesen y formasen un globo.

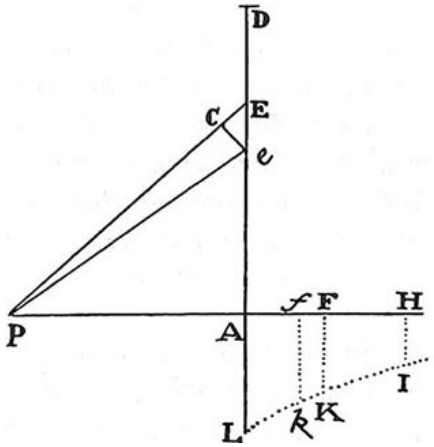
Se demuestra de igual modo que la Proposición anterior.

COROLARIO. Por tanto, el movimiento del cuerpo atraído será el mismo que si los cuerpos atrayentes, manteniendo el centro común de gravedad, se reuniesen y formasen un globo. Y, por lo mismo, si el centro común de gravedad de los cuerpos atrayentes reposa o progresa uniformemente en línea recta, el cuerpo atraído se moverá en una elipse cuyo centro está en el centro común de gravedad de los cuerpos atrayentes.

Proposición xc. Problema xliv

Si hacia cada punto de un círculo cualquiera tienden fuerzas centrípetas iguales, crecientes o decrecientes según alguna razón de las distancias, hállese la fuerza con la que es atraído un corpúsculo situado en un punto cualquiera de una recta levantada perpendicularmente desde el plano del círculo y en su punto central.

Imagínese un círculo descrito con centro en A y un radio cualquiera AD en un plano al que es perpendicular la recta AP; y hay que hallar la fuerza con la cual es atraído hacia él un corpúsculo cualquiera P. Desde un punto cualquiera del círculo, tal como E, trácese hasta el corpúsculo la recta PE. Sobre la recta PA tómese PF igual a PE y elévese la normal FK, que sea como la fuerza con la que el punto E atrae al corpúsculo P y sea IKL una línea curva a la que el punto K es continuamente tangente. Y toque ésta al plano del círculo en L. Sobre PA tómese PH igual a PD, y elévese la perpendicular HI que se encuentra con la curva antedicha en I; y la atracción del corpúsculo P hacia el círculo será como el área AHIL multiplicada por la altura AP. Q. E. I.



Efectivamente, tómese sobre AE el segmento mínimo E*e*. Únase P*e*, y sobre PE, PA, tómense PC, Pf, iguales a P*e*. Y puesto que la fuerza con la cual un punto cualquiera E del anillo descrito con centro en A y radio AE sobre el plano antedicho atrae hacia sí al cuerpo P supongamos que es como FK, y por tanto la fuerza con la que el dicho punto atrae al cuerpo P hacia A es como $\frac{AP \times FK}{PE}$, y la fuerza con la cual el anillo completo atrae al cuerpo P hacia A es como el anillo y $\frac{AP \times FK}{PE}$ conjuntamente; pero

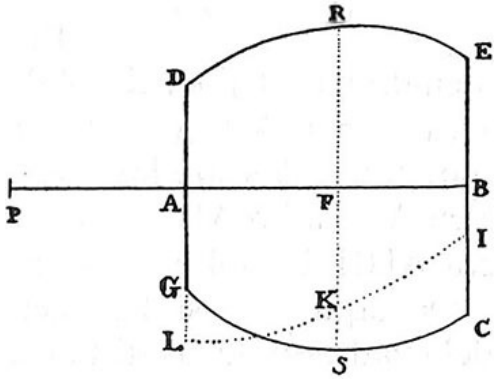
dicho anillo es como el rectángulo comprendido bajo el radio AE y la anchura E*e*, y este rectángulo (por la proporcionalidad de PE y AE, E*e* y CE) es igual al rectángulo PE x CE o PE x Ff; la fuerza con la cual este anillo atraerá al cuerpo P hacia A será como PE x Ff y $\frac{AP \times FK}{PE}$ conjuntamente, esto es, como el contenido bajo Ff x FK x AP, o también como el área FKkf multiplicada por AP. Y por tanto la suma de las fuerzas con las cuales todos los anillos sobre el círculo que se describe con centro en A y radio AD, atraen al cuerpo P hacia A es como toda el área AHIKL multiplicada por AP. Q. E. D.

Corolario 1. De aquí que si las fuerzas de los puntos decrecen en razón del cuadrado de las distancias, esto es, si FK se hace como $\frac{1}{PF^2}$ y por lo mismo el área AHIKL como $\frac{1}{PA} - \frac{1}{PH}$; la atracción del corpúsculo P hacia el círculo será como 1 - $\frac{PA}{PH}$, esto es, como $\frac{AH}{PH}$.

Corolario 2. Y, en general, si las fuerzas de los puntos a las distancias D fuesen inversamente como una potencia cualquiera D^n de las distancias; esto es, si FK fuese como $\frac{1}{D^n}$, y por tanto el área AHIKL como $\frac{1}{PA^{n-1}} - \frac{1}{PH^{n-1}}$, la atracción del corpúsculo P hacia el círculo sera como $\frac{1}{PA^{n-2}} - \frac{PA}{PH^{n-1}}$.

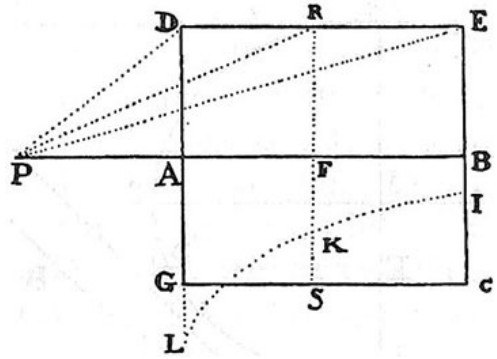
Corolario 3. Y si el diámetro del círculo se aumentase hasta el infinito y el número *n* es mayor que la unidad, la atracción del corpúsculo P hacia el plano entero infinito será inversamente como PA^{n-2}, ya que el otro termino $\frac{PA}{PH^{n-1}}$ se desvanece.

Proposición xci. Problema xlv



Hallar la atracción de un corpúsculo situado en el eje de un sólido redondo, hacia cada uno de cuyos puntos tienden fuerzas centrípetas iguales y decrecientes según una razón cualquiera de las distancias.

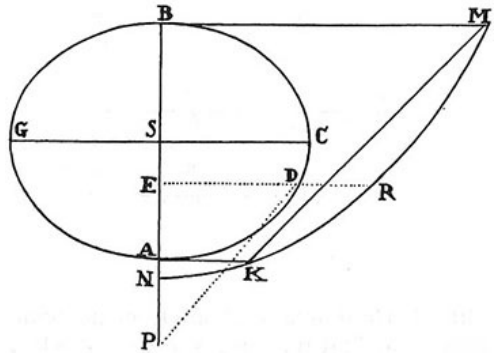
El corpúsculo P sea atraído hacia el sólido DECG y esté situado sobre su eje AB. Córtese el sólido con un círculo cualquiera RFS perpendicular a dicho eje, y en su semidiámetro FS, sobre un plano cualquiera PALKB que pasa por el eje, tómese (por la Proposición XC) la longitud FK proporcional a la fuerza con la que el corpúsculo P es atraído hacia dicho círculo. Sea el punto K tangente a la línea curva LKI que se encuentra con los planos más exteriores de los círculos AL y BI en L e I; y la atracción del corpúsculo P hacia el sólido será como el área LABI. Q. E. I.



COROLARIO 1. De donde, si el sólido es un cilindro descrito por el paralelogramo ADEB al girar sobre su eje AB, y las fuerzas centrípetas tendentes hacia cada uno de sus puntos fuesen inversamente como los cuadrados de las distancias hasta los puntos; la atracción del corpúsculo P hacia dicho cilindro será como AB - PE + PD.

Ya que la ordenada FK (por el Corolario 1 de la Proposición XC) será como $1 - \frac{PF}{PR}$. El miembro 1 de esta cantidad multiplicado por la longitud AB describe el área 1 x AB: y el otro miembro $\frac{PF}{PR}$ multiplicado por la longitud PB describe el área 1 x (PE - AD) (fácil de mostrar a partir de la cuadratura de la curva LKI); y de modo similar, aquel mismo miembro multiplicado por la longitud PA describe el área 1 x (PD - AD) y multiplicado por AB, que es la diferencia entre PB, PA, describe la diferencia de las áreas 1 x (PE - PD). Réstese del primer valor 1 x AB el del último 1 x (PE - PD) y el resto será el área LABI igual a 1 x (AB - PE + PD). Luego la fuerza proporcional a esta área es como AB - PE + PD.

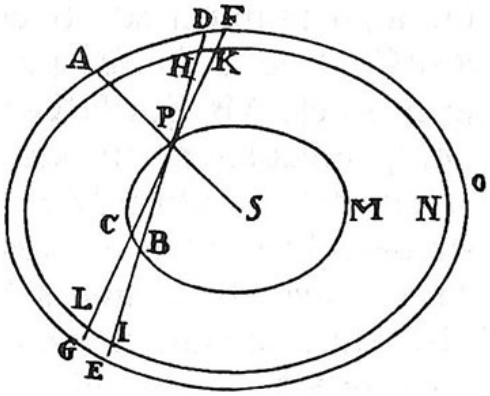
COROLARIO 2. De aquí también se obtiene la fuerza con la cual el esferoide AGBC atrae a un corpúsculo cualquiera P situado externamente en su eje AB. Sea NKRM la sección cónica cuya ordenada ER, perpendicular a PE, sea siempre igual a la longitud PD que se traza hasta el punto D en el cual la susodicha ordenada corta al esferoide. Desde los vértices del esferoide A, B, elévense, perpendiculares al eje AB, las líneas AK, BM, respectivamente iguales a AP, BP, que, por tanto, encontrarán a la sección cónica en K y M; y únanse KM, separando el segmento KMRK.



Sea S el centro del esferoide y SC el semidiámetro máximo; y la fuerza con la cual el esferoide atrae al cuerpo P será a la fuerza con la cual atrae al mismo cuerpo una esfera descrita con diámetro AB como $\frac{AS \times CS^2 - PS \times KMRK}{PS^2 + CS^2 - AS^2}$ a $\frac{AS^3}{3PS^2}$. Y con la misma base de cálculo es posible hallar las fuerzas de los segmentos del esferoide.

COROLARIO 3. Y si el corpúsculo se colocase sobre el eje dentro del esferoide, la atracción será como su distancia al centro. Cosa que se infiere fácilmente con el argumento siguiente, tanto si la partícula se halla en el eje como en otro diámetro dado cualquiera. Sea AGOF el esferoide atrayente, S su centro y P el cuerpo atraído.

Trácese por dicho cuerpo P tanto el semidiámetro SPA como las dos rectas cualesquiera DE, FG, que cortan al esferoide a uno y otro lado en D y E, y F y G; y sean PCM, HLN, las superficies de dos esferoides interiores, semejantes y concéntricas de la exterior, de las cuales la primera pasa por el cuerpo P y corta a las rectas DE y FG en B y C y la segunda corta a las mismas rectas en H, I y K, L. Tengan todos los esferoides el eje común y serán las partes de las rectas interceptadas a uno y otro lado iguales entre sí, DP y BE, FP y CG, DH e IE, FK y LG: por cuanto que las rectas DE, PB y HI son bisecadas en el mismo punto, al igual que las rectas FG, PC y KL. Supóngase ahora que DPF, EPG, designan conos opuestos, descritos por los ángulos verticales infinitamente pequeños DPF, EPG, y que las líneas DH, El, son también infinitamente pequeñas; y las partículas de los conos DHKF y GLIE, separadas por las superficies de los esferoides, serán entre sí, debido a la igualdad de las líneas DH, El, como los cuadrados de sus distancias al cuerpo P, y por ello atraerán igualmente a dicho corpúsculo. Y por la misma razón, si los espacios DPF, EGCB, se dividen en partículas mediante las superficies de innumerables esferoides semejantes, concéntricos y con eje común, todas ellas a ambos lados atraerán igualmente al cuerpo P en sentidos contrarios. Por tanto, las fuerzas del cono DPF y las del segmento cónico EGCB son iguales, y por ser contrarias se destruyen mutuamente. Y lo mismo ocurre con las fuerzas de toda materia exterior al esferoide más interior PCBM. En consecuencia, el cuerpo P es atraído únicamente por el esferoide interior PCBM, y por ello (por el Corolario 3 de la Proposición LXXII) su atracción es a la fuerza con la que el cuerpo A es atraído por el esferoide total AGOD, como la distancia PS a la distancia AS. Q. E. D.



Proposición XCII. Problema XLVI

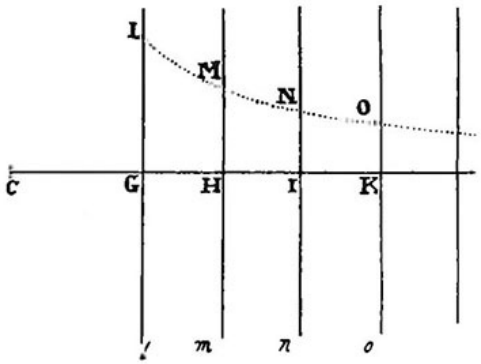
Dado un cuerpo atrayente, hallar la razón del decrecimiento de las fuerzas centrípetas tendentes hacia cada uno de sus puntos.

Del cuerpo dado hay que obtener una esfera o un cilindro u otra figura regular cualquiera cuya ley de atracción concuerde con alguna razón de decrecimiento calculable (por las Proposiciones LXXX, LXXXI y XCI). Después, mediante experimentos hay que hallar la fuerza de atracción a distintas distancias, y descubierta así la ley de la atracción hacia el todo, permitirá hallar la razón del decrecimiento de las fuerzas de cada parte, que era preciso hallar.

Proposición XCIII. Teorema XLVII

Si un sólido, plano por un lado e infinito por todos los demás, constase de partículas iguales e igualmente atrayentes, cuyas fuerzas al alejarse del sólido decrecen en razón de una potencia de las distancias mayor que el cuadrado de las mismas, y un corpúsculo situado a una u otra parte del plano es atraído por la fuerza del sólido entero; digo que la tal fuerza atractiva del sólido, al alejarse de su superficie plana, decrece en razón de una potencia cuya base es la distancia del corpúsculo al plano y su índice es 3 unidades menos que el índice de la potencia de las distancias.

Caso 1. Sea LG*l* el plano en que termina el sólido. Extiéndase éste hacia el lado del plano que mira hacia I, y descompóngase en innumerables planos *m*HM, *nl*N, *o*KO, etc., paralelos a GL. Y colóquese primero el cuerpo atraído C fuera del plano. Trácese CGHI perpendicular a esos innumerables planos, y decrezcan las fuerzas atractivas de los puntos del sólido en razón de una potencia de las distancias, cuyo índice *n* no sea menor que 3. Entonces (por el Corolario 3 de la Proposición XC) la fuerza con la que cualquier plano *m*HM atrae al punto C es inversamente como CH^{n-2}. Sobre el plano *m*HM tómese la longitud HM, inversamente proporcional a CH^{n-2}, y dicha fuerza será como HM. E igualmente para cada plano *l*GL, *n*IN, o KO, etc., tómense longitudes GL, IN, KO, etc., inversamente proporcionales a CG^{n-2}, CI^{n-2}, CK^{n-2}, etc.; y las fuerzas de tales planos serán como las longitudes tomadas, y por lo mismo la suma de las fuerzas como la suma de las longitudes, esto es, la fuerza del sólido entero como el área GLOK prolongada infinitamente hacia OK. Pero dicha área (por el conocido método de las cuadraturas) es inversamente como CG^{n-3}, y por tanto la fuerza del cuerpo entero es inversamente como CG^{n-3}. Q. E. D.



CASO 2. Ahora colóquese el corpúsculo C al otro lado del plano *l*GL dentro del sólido, y tómese la distancia CK igual a la distancia CG. Y la parte del sólido *l*GL*o*KO, delimitada por los planos paralelos *l*GL, o KO, no atraerá hacia ninguna parte al corpúsculo C, al estar situado en el medio, por cuanto que las acciones contrarias de los puntos opuestos se anulan en razón de su igualdad. Por lo cual el corpúsculo C será atraído por la sola fuerza del sólido que se halla más allá del plano OK. Pero esta fuerza (por el Caso 1) es inversamente como CK^{n-3}, esto es (por ser iguales CG y CK) inversamente como CG^{n-3}. Q. E. D.

COROLARIO 1. De aquí que si el sólido LGIN estuviese delimitado a cada lado por dos planos paralelos infinitos LG, IN, su fuerza atractiva resulta de restar de la fuerza atractiva de todo el sólido infinito LGKO la fuerza atractiva de la parte posterior NIKO prolongada infinitamente hacia KO.

COROLARIO 2. Si se despreciase la parte posterior de este sólido infinito, porque, al compararla con la atracción de la parte anterior, su atracción es insignificante, la atracción de dicha parte anterior decrecerá, al aumentar las distancias, en una razón muy próxima de la potencia CG^{n-3}.

COROLARIO 3. Y de aquí que, si un cuerpo finito y plano por un lado atrae a un corpúsculo desde la región media de dicho plano y la distancia entre el corpúsculo y el plano fuese muy pequeña comparada con las dimensiones del cuerpo atrayente, y éste constase de partículas homogéneas cuyas fuerzas atractivas decrecen en razón de alguna potencia de las distancias mayor que la cuarta; la fuerza atractiva del cuerpo entero decrecerá muy aproximadamente en razón de una potencia cuya base es aquella distancia muy pequeña y el índice 3 unidades menor que el índice de la potencia anterior. La afirmación no vale para un cuerpo compuesto de partículas cuyas fuerzas atractivas decrezcan en razón del cubo de las distancias; por cuanto que, en tal caso, la atracción de la parte posterior del cuerpo infinito del Corolario segundo es siempre infinitamente mayor que la atracción de la parte anterior.

ESCOLIO

Si un cuerpo cualquiera es atraído perpendicularmente hacia un plano dado y, a partir de una ley de atracción dada, se pide el movimiento del cuerpo, el problema se resuelve buscando (por la Proposición XXXIX) el movimiento del cuerpo descendiendo en línea recta hacia ese plano, y (por el Corolario II de las [*Leyes*]) componiendo dicho movimiento con un movimiento uniforme, realizado según líneas paralelas a dicho plano. Y viceversa, si se busca la ley de atracción hacia un plano según líneas paralelas y con la condición de que el cuerpo atraído se mueva en una línea curva cualquiera dada, el problema se resuelve operando al modo del Problema tercero.

Pero las operaciones pueden reducirse descomponiendo las ordenadas en series convergentes. Así, si a una base A se aplica con un ángulo dado la longitud ordenada B, que sea como una potencia cualquiera de la base $A^{m/n}$; y se busca la fuerza con la cual un cuerpo, tanto si es atraído como repelido respecto a la base según la posición de la ordenada, puede moverse en una línea curva a la que siempre toca la ordenada con su extremo superior: supongo que la base aumenta en una parte mínima O, y descompongo la ordenada $(A + O)^{m/n}$ en la serie infinita:

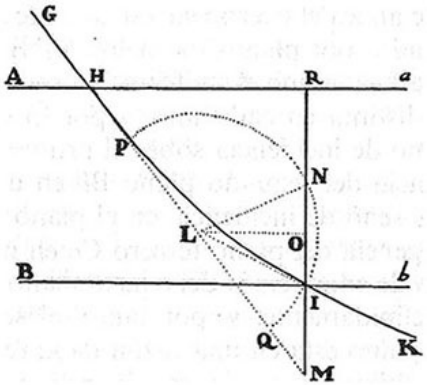
$$A^{m/n} + \frac{m}{n} OA^{m-n/n} + \frac{mm - mn}{2nn} OOA^{m-2n/n}, \text{ etc.,}$$

y supongo que la fuerza es proporcional al término en que O tiene dos dimensiones, esto es, al termino $\frac{mm - mn}{2nn} OOA \frac{m - 2n}{2n}$. La fuerza buscada es, pues, como $\frac{mm - mn}{nn} A^{m-2n/n}$ o, lo que es lo mismo, como $\frac{mm - mn}{nn} B^{m-2n/n}$. Y si la ordenada es tangente a una parábola, siendo $m = 2$ y $n = 1$ la fuerza será como $2B^{o}$ ya dado, y por tanto resultará dada. Por tanto, con una fuerza dada el cuerpo se moverá en una parábola como demostró Galileo. Y si la ordenada es tangente a una hipérbola, siendo $m = 0 - 1$ y $n = 1$, la fuerza será como $2A^{-3}$ o como $2B^{3}$; y por tanto con una fuerza que sea como el cubo de la ordenada el cuerpo se moverá en una hipérbola. Pero una vez dadas éstas, paso a otras Proposiciones sobre el movimiento que aún no he tocado[37].

Sección XIV
SOBRE EL MOVIMIENTO DE CUERPOS MÍNIMOS QUE SON SOMETIDOS A FUERZAS CENTRÍPETAS TENDENTES HACIA CADA PARTE DE OTRO CUERPO MUY GRANDE

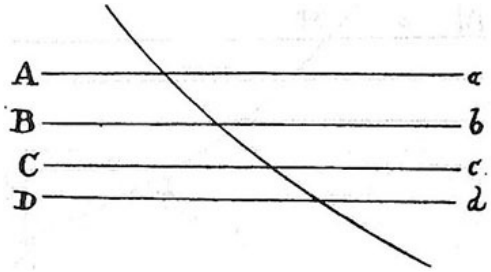
Proposición XCIV. Teorema XLVIII

Si dos medios semejantes se hallasen separados entre sí por un espacio delimitado a ambos lados por planos paralelos y un cuerpo al pasar por este espacio es atraído o impelido perpendicular mente hacia uno u otro medio sin ser ni agitado ni impedido por ninguna otra fuerza, y la atracción fuese la misma en todas partes a distancias iguales de uno y otro plano tomadas hacia la misma parte del mismo: digo que el seno de incidencia en uno u otro plano estará respecto al seno de emergencia del otro plano en una razón dada.



Caso 1. Sean A*a* y B*b* dos planos paralelos. Incida el cuerpo sobre el primer plano A*a* según la línea GH, y en todo su recorrido a través del espacio intermedio sea atraído o impelido hacia el medio de incidencia, y sea obligado por esa acción a describir la línea curva HI, y emerja según la línea IK. Elévese IM, perpendicular al plano de emergencia B*b*, y que se encuentre tanto con la prolongación de la línea de

incidencia en M como con el plano de incidencia A*a* en R; a su vez la línea de emergencia prolongada encuentre a HM en L. Con centro en L y radio LI trácese un círculo secante de HM en P y en Q, así como de la prolongación de MI en N; y, en primer lugar, si la atracción o el impulso se supone uniforme, la curva HI (por las demostraciones de Galileo) será una parábola que tiene la propiedad de que el rectángulo comprendido bajo un *latus rectum* dado y la línea IM es igual a HM^2; pero la línea HM será bisecada en L. De donde si se traza sobre MI la perpendicular LO, resultarán iguales MO y OR; y añadiéndoles las iguales ON, OI, los totales MN, IR, resultarán iguales. Por tanto, al estar dado IR, también está dado MN; y el rectángulo NMI es al rectángulo comprendido bajo el *latus rectum* y IM, esto es a HM^2, en una razón dada. Pero el rectángulo NMI es igual al rectángulo PMQ, esto es, a la diferencia de los cuadrados ML^2 y PL^2 o LI^2, y HM^2 guarda una razón dada con su cuarta parte ML^2: luego está dada la razón de ML^2 - LI^2 respecto a ML^2, y convirtiendo, la razón de LI^2 a ML^2 y la de la raíz cuadrada de LI a ML. Pero en todo triángulo LMI los senos de los ángulos son proporcionales a los lados opuestos. Luego se da la razón del seno del ángulo de incidencia LMR al seno del ángulo de emergencia LIR. Q. E. D.

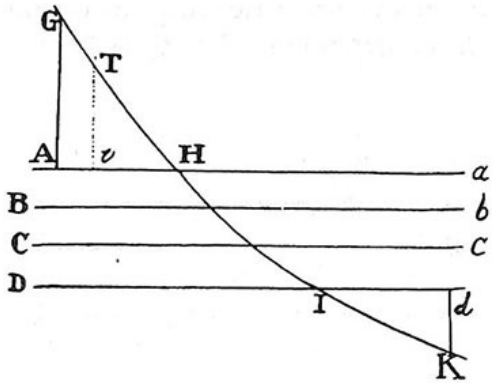


Caso 2. Pase ahora el cuerpo sucesivamente a través de varios espacios delimitados por planos paralelos A*ab*B, B*bc*C, etc., y sea afectado por una fuerza que es uniforme en cada uno de ellos por separado, pero distinta en cada uno; y por lo que acabamos de demostrar, el seno de incidencia sobre el primer plano A*a* será al seno de emergencia del segundo plano B*b* en una razón dada; y este seno, que es seno de incidencia en el plano segundo B*b*, será al seno de emergencia del plano tercero C*c* en una razón dada; y este seno al seno de emergencia del cuarto plano D*d* en una razón dada; y así indefinidamente: y, por tanto, el seno de incidencia sobre el primer plano está en una razón dada respecto al seno de emergencia del último plano. Disminúyanse ahora los intervalos entre los planos y auméntese su número hasta el infinito, para que la acción de atracción o impulso que obedezca a cualquier ley asignada se haga continua; y la razón del seno de incidencia sobre el primer plano respecto al seno de emergencia del último plano, al estar siempre dada, también ahora estará dada. Q. E. D.

Proposición xcv. Teorema xlix

Con los mismos supuestos digo que la velocidad del cuerpo antes de su incidencia es a su velocidad después de la emergencia como el seno de emergencia al seno de incidencia.

Tómense AH, Id, iguales y elévense las perpendiculares AG, *d*K, que corten a las líneas de incidencia y emergencia GH, IK en G y K. Sobre GH tómese TH igual a IK, y sobre el plano A*a* caiga la perpendicular T*v*. Y (por el Corolario ii de las [*Leyes*]) descompóngase el movimiento del cuerpo en dos, uno perpendicular a los planos A*a*, B*b*, C*c*, etc., y el otro paralelo a ellos. La fuerza de atracción o de impulso actuando según las líneas perpendiculares en nada cambia al movimiento según las paralelas, y por tanto con ese movimiento el cuerpo recorre en tiempos iguales los intervalos iguales según las paralelas que están entre la línea AG y el punto H, y entre el punto I y la línea *d*K; esto es, en tiempos iguales describe las líneas GH, IK. Por tanto, la velocidad antes de la incidencia es a la velocidad después de la emergencia como GH a IK o TH, esto es, como AH o Id a *p*H, esto es (respecto al radio TH OIK) como el seno de emergencia al seno de incidencia. Q. E. D.

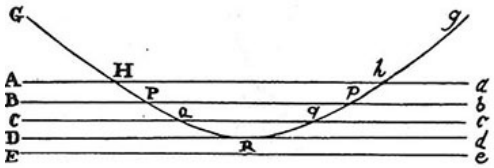


Proposición xcvi. Teorema l

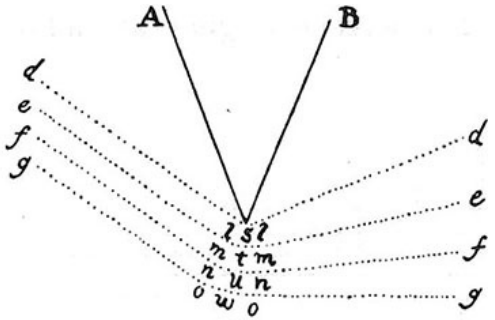
Supuesto lo anterior y que el movimiento antes de la incidencia es más veloz que después: digo que el cuerpo, inclinando la línea de incidencia, acabará por ser reflejado, y el ángulo de reflexión será igual al de incidencia.

Pues imagínese el cuerpo describiendo arcos parabólicos al pasar entre los planos paralelos A*a*, B*b*, C*c*, etc., como antes; y sean tales arcos HP, PQ, QR, etc. Y la oblicuidad de la línea de incidencia GH sea tal respecto al primer plano A*a* que haga

que el seno de incidencia sea al radio del círculo del que es seno en la misma razón en que se halla el propio seno de incidencia respecto al seno de emergencia del plano D*d* en el espacio D*de*E: al ser ahora el seno de emergencia igual al radio, el ángulo de emergencia será recto, y por tanto la línea de emergencia coincidirá con el plano D*d*. Alcance el cuerpo a este plano en el punto R; y puesto que la línea de emergencia coincide con dicho plano, es evidente que el cuerpo no puede pasar más allá hacia el plano E*e*. Pero tampoco puede proseguir por la línea de emergencia R*d*, ya que es atraído o impelido continuamente hacia el medio de incidencia. Regresará, por tanto, entre los planos C*c*, D*d*, describiendo el arco de parábola QR*q*, cuyo vértice principal (como demostró Galileo) está en R; cortará al plano C*c* con el mismo ángulo en *q* que antes en Q; después, continuando por los arcos parabólicos *qp*, *ph*, etc., semejantes e iguales a los primeros QP, PH, cortará a los demás planos con los mismos ángulos en *p*, *h*, etc., que antes en P, H, etc., y por fin emergerá en *h* con la misma oblicuidad con que incidió en H. Imagínese ahora que los intervalos de los planos A*a*, B*b*, C*c*, D*d*, E*e*, etc., se reducen y su número aumenta infinitamente, de modo que la atracción o impulso aplicados según una ley cualquiera determinada resulten continuos; y al permanecer siempre iguales el ángulo de emergencia y el ángulo de incidencia, también aquí lo seguirán siendo. Q. E. D.

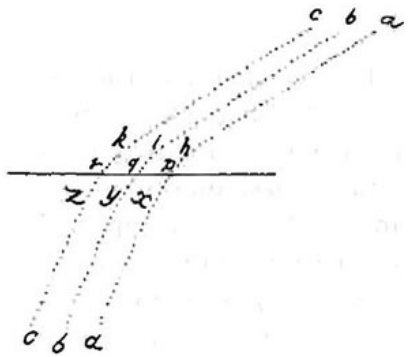


Escolio



Las reflexiones y refracciones de la luz no difieren mucho de estas atracciones, al efectuarse aquéllas según una razón dada de las secantes, como descubrió Snell, y, por consiguiente, según una razón dada de los senos, como explicó Descartes. Pues

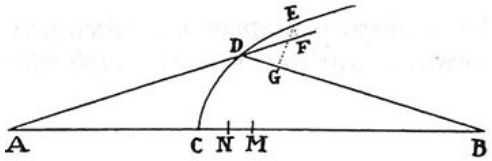
consta por los fenómenos de los satélites de Júpiter, confirmados por observaciones de diferentes astrónomos, que la luz se propaga sucesivamente y viene del Sol a la Tierra en un espacio de siete u ocho minutos. Y los rayos existentes en el aire (como recientemente descubrió Grimaldi al introducir luz en un cuarto oscuro a través de un orificio, y yo mismo he experimentado) al pasar junto a ángulos de cuerpos opacos o transparentes (tales como los bordes rectangulares de los círculos de monedas de oro, plata o bronce, y de los filos de cuchillos o de láminas de piedra o de vidrios rotos) se curvan hacia los cuerpos, como si fueran atraídos por ellos; y de esos rayos, los que pasan más cerca de los cuerpos, al pasar se curvan más, como si fueran más atraídos, como yo mismo he observado con todo cuidado. Y los que pasan a distancias mayores se curvan menos; y a distancias aún mayores, se curvan un tanto hacia el lado contrario, y forman tres haces de colores. En la figura sea *s* el corte de un cuchillo o cuña cualquiera AsB; y *gowog*, *fnunf*, *emtme*, *dlsld*, son rayos curvados hacia el cuchillo con arcos *owo*, *nun*, *mtm*, *lsl*, y esto más o menos según su distancia al cuchillo. Pero como dicha curvatura de los rayos ocurra en el aire fuera del cuchillo, deberán también los rayos incidentes en el cuchillo curvarse en el aire antes de alcanzar el cuchillo. Y lo mismo ocurre para los que inciden en un cristal. Por tanto, la refracción no ocurre en el punto de incidencia, sino paulatinamente por incurvación continua de los rayos, parte de la cual ocurre en el aire antes de alcanzar el cristal y parte en el cristal (si no me equivoco) después de haber penetrado en él: como aparece diseñado en los rayos *ckzc*, *biyb*, *ahxa*, incidiendo sobre *r*, *q*, *p*, etc., curvados entre *k* y *z*, *i* e *y*, *h* y *x*. Por consiguiente, dada la analogía existente entre la propagación de los rayos de luz y el movimiento de los cuerpos, parece oportuno añadir las proposiciones siguientes para usos ópticos, sin discutir cosa alguna sobre la naturaleza de los rayos (si son cuerpos o no) determinando tan sólo las trayectorias de cuerpos muy semejantes a las trayectorias de los rayos.



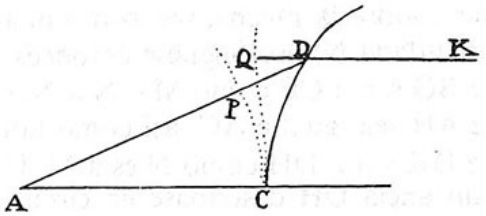
Proposición xcvii. Problema xlvii

Supuesto que el seno de incidencia sobre una superficie cualquiera esté respecto al seno de emergencia en una razón dada, y que la incurvación de la trayectoria de los cuerpos junto a dicha superficie ocurra en un espacio muy breve, que puede considerarse un punto; determinar la superficie que haga converger a todos los corpúsculos procedentes de un lugar dado sobre otro lugar dado.

Sea A el lugar desde el cual los corpúsculos divergen; B el lugar sobre el cual deben converger; CDE la línea curva que, al girar sobre el eje AB, describa la superficie buscada; D, E, dos puntos cualesquiera de esa curva; y EF, EG, perpendiculares que descienden sobre las trayectorias del cuerpo AD, DB. Aproxímese el punto D al punto E; y la razón última de la línea DF, con la que se incrementa AD, a la línea DG, con la que DB disminuye, será la misma que la del seno de incidencia al seno de emergencia. Luego está dada la razón del incremento de la línea AD a la disminución de la línea DB; y, por tanto, si sobre el eje AB se toma un punto cualquiera C, por el cual deba pasar la curva CDE, y se toma para AC el incremento CM, que esté respecto al decremento CN de la línea BC en la razón dada, y con centros en A, B, e intervalos AM, BN, se trazan dos círculos mutuamente secantes en D; dicho punto D será tangente a la curva buscada CDE y manteniéndose tangente a ella continuamente la determinará. Q. E. I.



Corolario 1. Haciendo que los puntos A o B ya se alejen al infinito, ya se acerquen hacia el punto C, se tendrán todas las figuras que Descartes expuso en la *Optica* y en la *Geometría* sobre las refracciones. Me pareció oportuno explicar en esta Proposición, ya que DESCARTES se lo reservó, la manera de hallarlas.



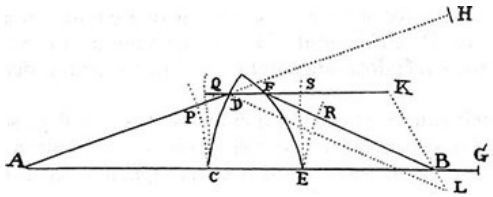
Corolario 2. Si un cuerpo incide sobre una superficie cualquiera CD, según la dirección AD, y la superficie obedece a una ley cualquiera, y el cuerpo emerge según otra recta cualquiera DK y por el punto C se imaginan trazadas dos curvas CP, CQ, siempre perpendiculares a AD y a DK: los incrementos de las líneas PD, QD y, por

tanto, las propias líneas PD, QD, resultantes de esos incrementos, serán entre sí como los senos de incidencia y emergencia respectivamente, y viceversa.

Proposición xcviii. Problema xlviii

Con los mismos supuestos, y trazando en torno al eje AB, una superficie atractiva CD cualquiera, sea regular o irregular, a cuyo través deben pasar los cuerpos procedentes de un lugar dado A: hallar una segunda superficie atractiva EF que haga converger a dichos cuerpos en otro lugar dado B.

Trazada la línea AB corte a la primera superficie en C y a la segunda en E, y fíjese D en un punto cualquiera. Supuesto que el seno de incidencia sobre la primera superficie respecto al seno de emergencia de la misma, y el seno de emergencia de la segunda al seno de incidencia sobre la misma, sea como una cantidad dada M a otra cantidad dada N: prolónguese entonces tanto AB hasta G de modo que BG sea a CE como M - N a N, como AD hasta H de modo que AH sea igual a AG, así como también DF hasta K, de modo que DK sea a DH como N es a M. Únase KB, y con centro en D y distancia DH descríbase un círculo que encuentre en L a KB prolongada, y trácese BF paralela a DL: y el punto F toca a la línea EF que al girar sobre el eje AB describe la superficie buscada. Q. E. F.



Pues imagínese que las lineas CP y CQ respecto a AD y a DF, y las líneas ER y ES respecto a FB y a FD son siempre perpendiculares, y por ello QS es siempre igual a CE; y (por el Corolario 2 de la Proposición xcvii) PD será a QD como M a N, y por tanto como DL a DK o FB a FK; y por partes, como DL - FB o PH - PD - FB a FD o FQ - QD; y conjuntamente, como PH - FB a FQ, esto es (por ser iguales PH y CG, QS y CE), como CE + BG - FR a CE - FS. Y (por ser proporcionales BG a CE y M - N a N) también ocurre que CE + BG es a CE como M a N; y parcialmente también FR es a FS como M a N; y por ello (por el Corolario 2 de la Proposición xcvii) la superficie EF obliga al cuerpo incidente sobre ella según la línea DF a proseguir según la línea FR hasta el lugar B. Q. E. D.

Escolio

Con el mismo método podríamos proseguir hasta tres o más superficies. Pero para usos ópticos es la esférica la figura más adecuada. Si los objetivos de los telescopios estuviesen construidos con dos cristales esféricos que encerrasen agua entre ellos, pudiera ser que los errores de refracción que ocurren en los extremos de los cristales, pudieran corregirse con bastante exactitud mediante las refracciones del agua. Semejantes objetivos deben preferirse a los cristales elípticos e hiperbólicos, no sólo por ser más fáciles de construir con precisión, sino también porque refractan más exactamente los haces de rayos situados fuera del eje del cristal. No obstante, la diversa refrangibilidad de los distintos rayos es el verdadero obstáculo con el que tropieza la óptica para perfeccionarse, tanto con formas esféricas como con otras. A no ser que se corrijan los errores que de ahí proceden, el esfuerzo por corregir los demás no tendrá éxito.

Libro segundo
DEL MOVIMIENTO DE LOS CUERPOS

Sección primera
DEL MOVIMIENTO DE LOS CUERPOS A LOS QUE SE RESISTE EN RAZÓN DE LA VELOCIDAD

Proposición i. Teorema i

La pérdida de movimiento, debida a la resistencia, de un cuerpo al que se resiste en razón de la velocidad es como el espacio recorrido en el movimiento.

Puesto que como el movimiento perdido en cada partícula igual de tiempo es como la velocidad, esto es, como la partícula de espacio recorrido: componiendo, el movimiento perdido en el tiempo total será como el espacio total recorrido. Q. E. D.

Corolario. Por lo tanto, si un cuerpo desprovisto de toda gravedad se mueve en los espacios libres con la sola fuerza ínsita en él; y se dan tanto el movimiento total inicial como el movimiento restante tras recorrer algún espacio, quedará dado también el espacio total que puede recorrer el cuerpo en un tiempo infinito. Pues dicho espacio será al espacio recorrido ya como el movimiento total inicial a la parte perdida de dicho movimiento.

Lema i

Las cantidades proporcionales a sus diferencias son continuamente proporcionales.

Sea A: A - B, como B: B - C, como C: C - D, etc.; y, al convertir, A será a B como B a C, como C a D, etc. Q. E. D.

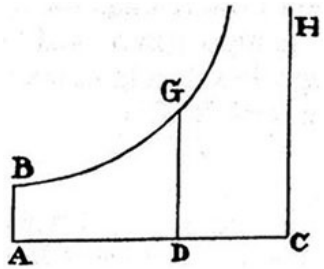
Proposición ii. Teorema ii

Si se resiste a un cuerpo en razón de la velocidad y se mueve en un medio homogéneo con su sola fuerza inercial y se toman tiempos iguales, las velocidades en los

momentos iniciales de cada tiempo están en progresión geométrica, y los espacios recorridos en cada tiempo son como las velocidades.

CASO 1. Divídase el tiempo en partículas iguales; y si en los comienzos de cada una de esas partículas actúa la resistencia con un único impulso que es como la velocidad: el decremento de la velocidad en cada partícula de tiempo será como la velocidad. Luego las velocidades son proporcionales a sus diferencias, y (por el Lema I del Libro II) por ello continuamente proporcionales. Por consiguiente, si con un número igual de partículas se componen cualesquiera tiempos iguales, las velocidades en los comienzos de esos tiempos serán como los términos de una progresión continua, tomados salteadamente, siempre que en cada salto se omitan igual número de términos intermedios. Pero las razones de tales términos se componen de las razones mutuas mismas de los términos intermedios repetidas igualmente, y por tanto también dichas razones compuestas son iguales entre sí. Por lo cual, las velocidades proporcionales a dichos términos están en progresión geométrica. Disminúyanse ahora aquellas partículas iguales de tiempo y auméntese su número hasta el infinito de modo que el impulso de la resistencia resulte continuo; y las velocidades al principio de tiempos iguales, siempre continuamente proporcionales, serán también en este caso continuamente proporcionales. Q. E. D.

CASO 2. Y, por partes, las diferencias de velocidades, esto es, las partes de las mismas perdidas en cada tiempo, son como los todos; ya que los espacios recorridos en cada tiempo son como las partes perdidas de las velocidades (por la Proposición I del Libro II) y por tanto también como los todos. Q. E. D.



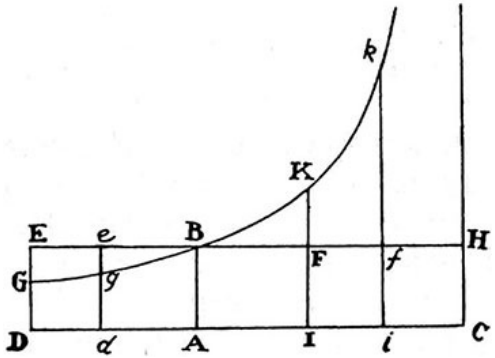
COROLARIO. De aquí que si se traza una hipérbola BG con las asíntotas rectangulares AC, CH y AB, DG son perpendiculares a la asíntota AC, y se representa tanto la velocidad del cuerpo como la resistencia del medio, al comienzo del movimiento, mediante una línea dada AC, y un tiempo transcurrido cualquiera mediante la línea indefinida DC: el tiempo puede representarse por el área ABGD y el espacio recorrido en dicho tiempo por la línea AD. Pues si dicha área se aumenta por el movimiento del punto D uniformemente con el tiempo, la recta DC decrecerá en razón geométrica al modo de la velocidad, mientras las partes de la recta AC,

descritas en tiempos iguales decrecerán en la misma razón.

PROPOSICIÓN III. PROBLEMA I

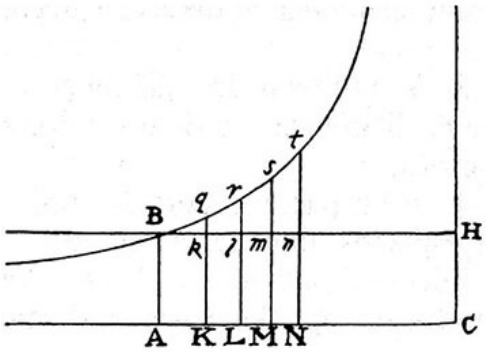
Definir el movimiento de un cuerpo al cual, mientras asciende o desciende en línea recta en un medio homogéneo, se resiste en razón de la velocidad y es urgido por una gravedad uniforme.

Para el cuerpo ascendente, represéntese la gravedad por un rectángulo cualquiera dado BACH, y la resistencia del medio al comienzo del ascenso por el rectángulo BADE tomado al lado contrario de la recta AB. Con las asíntotas rectangulares AC, CH, trácese por el punto B una hipérbola que corte a las perpendiculares DE, *de*, en G, *g*; y el cuerpo al ascender en el tiempo DG*gd* recorrerá el espacio EG*ge*, y en el tiempo DGBA el espacio de todo el ascenso EGB; en el tiempo ABKI el espacio de descenso BFK y en el tiempo IK*ki* el espacio de descenso KF*fk*; y las velocidades del cuerpo (proporcionales a la resistencia del medio) en esos períodos de tiempo serán ABED, AB*ed*, nula, ABFI, AB*fi*, respectivamente; y la velocidad máxima que puede adquirir el cuerpo al descender será BACH.



Pues, descompóngase el rectángulo BACH en innumerables rectángulos, A*k*, K*l*, L*m*, M*n*, etc., que sean como los incrementos de las velocidades producidos en los correspondientes tiempos iguales; y entonces, cero, A*k*, A*l*, A*m*, A*n*, etc., serán como las velocidades totales y, por tanto (por hipótesis) como las resistencias del medio al principio de cada tiempo igual. Hágase AC a AK o ABHC a AB*k*K como la fuerza de la gravedad a la resistencia al comienzo del segundo tiempo, y réstese de la fuerza de gravedad las de las resistencias, y ABHC, K*k*HC, L*l*HC, M*m*HC, etc., seguirán siendo como las fuerzas absolutas con las cuales el cuerpo es urgido al comienzo de cada uno de los tiempos, y por tanto (por la Ley II del movimiento) como los incrementos de las velocidades, esto es, como los rectángulos A*k*, K*l*, L*m*, M*n*, etc., y

por ello (por el Lema I del Libro II) en progresión geométrica. Por lo cual, si las rectas K*k*, L*l*, M*m*, N*n*, etc., al prolongarse cortan a la hipérbola en *q*, *r*, *s*, *t*, etc., las áreas AB*q*K, K*gr*L, L*rs*M, M*st*N, etc., serán iguales y, por tanto, análogas a tiempos y fuerzas gravitatorias iguales. Pero el área AB*q*K (por el Corolario 3 del Lema VII, y el Lema VIII del Libro I) es al área B*kq* como K*q* a ½*kq* o como AC a ½AK, esto es, como la fuerza de la gravedad a la resistencia en la mitad del tiempo primero. Y por un argumento similar las áreas *g*KL*r*, *r*LM*s*, *s*MN*t*, etc., son a las áreas *qklr*, *rlms*, *smnt*, etc., como las fuerzas de gravedad a las resistencias en la mitad de los tiempos segundo, tercero, cuarto, etc. Por lo tanto, puesto que las áreas iguales BAK*g*, *g*KL*r*, *r*LM*s*, *s*MN*t*, etc., son análogas a las fuerzas de gravedad, las áreas B*kq*, *qklr*, *rlms*, *smnt*, etc., serán análogas a las resistencias en la mitad de cada tiempo, esto es (por hipótesis) a las velocidades y, por ello, análogas a los espacios recorridos. Tómense las sumas de las cantidades análogas y las áreas B*kq*, B*lr*, B*ms*, B*nt*, etc., serán análogas a los espacios totales recorridos; lo mismo que las áreas AB*q*K, AB*r*L, AB*s*M, AB*r*N, etc., a los tiempos. Por tanto, un cuerpo mientras desciende describe en un tiempo cualquiera AB*r*L el espacio B*lr*, y en un tiempo L*rt*N un espacio *rlnt*. Q. E. D. Semejante es la demostración para un movimiento ascendente. Q. E. D.



COROLARIO 1. Por tanto, la mayor velocidad que puede adquirir un cuerpo al caer es a la velocidad adquirida en un tiempo dado como la fuerza dada de la gravedad por la que es urgido continuamente es a la fuerza de resistencia que se le opone al final de dicho tiempo.

COROLARIO 2. Y si el tiempo aumenta en progresión aritmética, la suma de dicha velocidad máxima y de la velocidad en ascenso, así como su diferencia en descenso, decrece en progresión geométrica.

COROLARIO 3. Y también las diferencias de los espacios recorridos en iguales diferencias de tiempos decrecen en la misma progresión geométrica.

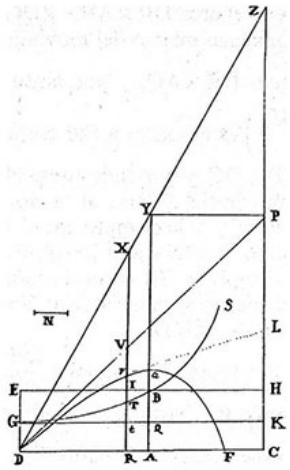
COROLARIO 4. El espacio recorrido por el cuerpo es la diferencia de dos espacios, uno de los cuales es como el tiempo tomado desde el inicio de la caída, y el otro,

como la velocidad, los cuales, por cierto, son iguales entre sí al inicio del descenso.

PROPOSICIÓN IV. PROBLEMA II[1]

Supuesto que la fuerza de la gravedad sea uniforme en un medio homogéneo cualquiera y que tienda perpendicularmente hacia el plano horizontal; definir en dicho medio el movimiento de un proyectil que sufre una resistencia proporcional a la velocidad.

Parta un proyectil desde un lugar cualquiera D según una recta cualquiera DP, y represente la longitud DP su velocidad al comienzo del movimiento. Desde el punto P sobre la horizontal DC descienda la perpendicular PC, y córtese DC en A de modo que DA sea a AC como la resistencia del medio originada al comienzo del movimiento hacia arriba es a la fuerza de la gravedad; o (lo que es lo mismo) de modo que el rectángulo comprendido entre DA y DP sea al rectángulo entre AC y CP como la resistencia total al comienzo del movimiento es a la fuerza de la gravedad. Con asíntotas DC, CP, trácese una hipérbola GTBS secante de las perpendiculares DG, AB, en G y B; y complétese el paralelogramo DGKC, cuyo lado GK corte a AB en Q. Tómese una línea N que esté en razón a QB como DC a CP; y desde un punto cualquiera R de la recta DC elévese la perpendicular RT, que encuentre a la hipérbola en T y a las rectas EH, GK, DP, en I, *t* y V; sobre esta perpendicular tómese V*r* igual a $\frac{t\text{GH}}{\text{N}}$ o, lo que es lo mismo, tómese R*r* igual a $\frac{\text{GTIE}}{\text{N}}$; y el proyectil, en el tiempo DRTG alcanzará el punto *r* describiendo la curva D*ra*F, que siempre toca el punto *r*, alcanzando la altura máxima *a* en la perpendicular AB, y aproximándose después siempre a la asíntota PC. Y su velocidad en un punto cualquiera *r* es como la tangente de la curva *rl* Q. E. I.



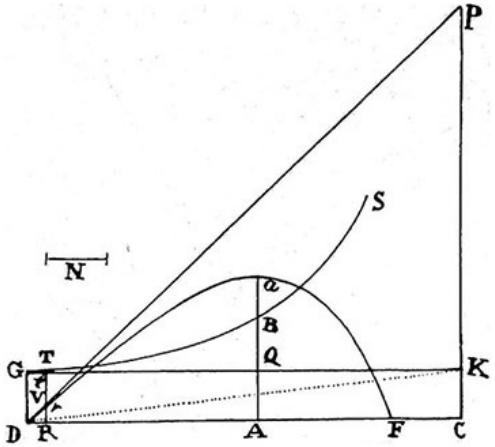
Pues N es a QB como DC a CP o DR a RV, y por tanto RV es igual a $\frac{DR \times QB}{N}$, y R*r* $\left(\text{esto es } R\textit{r} - V\textit{r} \text{ o } \frac{DR \times QB - \textit{t}GT}{N}\right)$ es igual a $\frac{DR \times AB - RDGT}{N}$. Ahora represéntese el tiempo mediante el área RDGT y (por el Corolario II de las [*Leyes*]) descompóngase el movimiento del cuerpo en dos, uno de ascenso y otro hacia el lado. Y puesto que la resistencia es como el movimiento, descompóngase ésta también en dos partes contrarias y proporcionales a las dos partes del movimiento: y por ello, la longitud descrita por el movimiento hacia el lado será (por la Proposición II de este Libro) como el área DR x AB - RDGT, esto es, como la línea R*r*. Pero al comienzo mismo del movimiento el área RDGT es igual al rectángulo DR x AQ, y por tanto dicha línea R*r* $\left(\text{o } \frac{DR \times AB - DR \times AQ}{N}\right)$ es entonces a DR como AB - AQ o QB a N, esto es, como CP a DC, y por ende como el movimiento hacia arriba al movimiento hacia el lado al inicio. Puesto que R*r* es siempre como la altura y DR siempre como la longitud: se sigue necesariamente que R*r* sea siempre a DR como la altura a la longitud, y por lo tanto que el cuerpo se mueva en la línea D*ra*F, a la cual toca siempre el punto r. Q. E. D.

COROLARIO 1. Por tanto, R*r* es igual $\frac{DR \times AB}{N} - \frac{RDGT}{N}$; y por ello, si se prolonga RT hasta X de modo que RX sea igual a $\frac{DR \times AB}{N}$; esto es, si se completa el

paralelogramo ACPY, se une DY cortando a CP en Z, y se prolonga RT hasta que toque a DY en X: X*r* será igual a $\frac{RDGT}{N}$, y por lo mismo proporcional al tiempo.

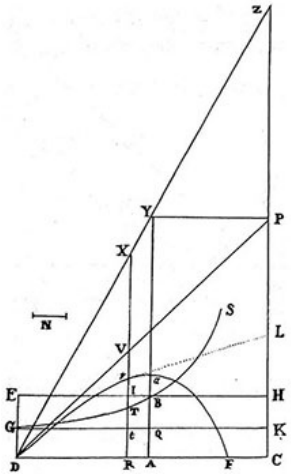
COROLARIO 2. De donde, si se toman innumerables CR o, lo que es lo mismo, innumerables ZX, en progresión geométrica; habrá otras tantas X*r* en progresión aritmética. Y de aquí que la curva D*ra*F se pueda delinear fácilmente por la tabla de logaritmos.

COROLARIO 3. Si se construye una parábola con vértice en D, con diámetro DC prolongado hacia abajo y con un *latus rectum* que sea a 2DP como la resistencia total en el mismo comienzo del movimiento a la fuerza de la gravedad: la velocidad con la que un cuerpo debe partir desde el punto D según la recta DP, para que describa en un medio uniformemente resistente la curva D*ra*F, será la misma que aquella con la que debe partir desde el mismo punto D y según la misma línea DP, para describir la parábola en un espacio sin resistencia. Pues el *latus rectum* de esta parábola en el momento inicial del movimiento es $\frac{DV^2}{V\textit{r}}$; y V*r* es $\frac{\textit{t}GT}{N}$ o también $\frac{DR \times T\textit{t}}{2N}$. Pero si se trazase una recta que fuese tangente a la hipérbola GTS en G, sería paralela a DK y, por tanto, T*t* es $\frac{CK \times DR}{DC}$, y N era $\frac{QB \times DC}{CP}$. Y por tanto V*r* es $\frac{DR^2 \times CK \times CP}{2DC^2 \times QB}$ esto es (por ser proporcionales DR y DC, DV y DP) $\frac{DV^2 \times CK \times CP}{2DP^2 \times QB}$, y el *latus rectum* $\frac{DV^2}{V\textit{r}}$ resulta ser $\frac{2DP^2 \times QB}{CK \times CP}$ esto esto es (por ser proporcionales QB y CK, DA y AC) $\frac{2DP^2 \times DA}{AC \times CP}$ y por lo tanto es a 2DP como DP x DA a CP x AC; esto es como la resistencia a la gravedad. Q. E. D.



COROLARIO 4. De aquí que si se lanza un cuerpo desde un punto D con una velocidad dada, según una recta de posición dada DP, y se da la resistencia del medio al comienzo mismo del movimiento: puede hallarse la curva D*ra*F que describirá dicho cuerpo. Pues, como es sabido, al dar la velocidad se da también el *latus rectum* de la parábola. Y tomando 2DP a dicho *latus rectum* como la fuerza de la gravedad a la de resistencia, DP está dado. Cortando entonces DC en A de modo que CP x AC esté respecto a DP x DA en la susodicha razón de la gravedad a la resistencia, se tendrá dado el punto A. Y con ello está dada la curva D*ra*F.

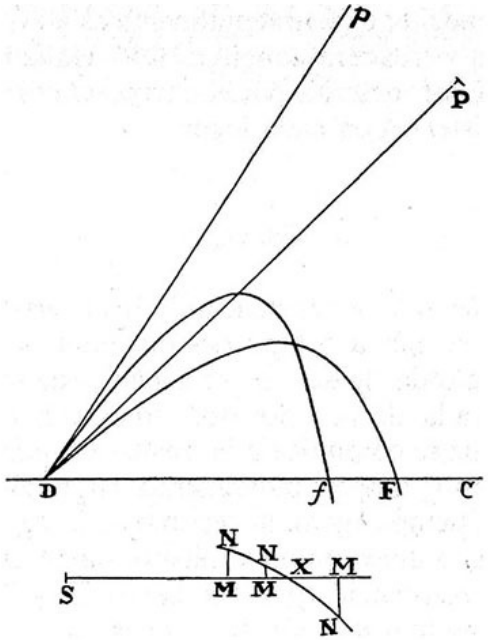
COROLARIO 5. Y, al contrario, si la curva D*ra*F está dada, también estarán dadas la velocidad del cuerpo y la resistencia del medio en cada lugar r. Puesto que por estar dada la razón CP x AC a DP x DA, viene dada también tanto la resistencia del medio al comienzo del movimiento como el *latus rectum* de la parábola: y con ello también se da la velocidad al comienzo del movimiento. Y después, a partir de la longitud de la tangente *r*L, viene dada tanto la velocidad proporcional a ella como la resistencia proporcional a la velocidad en un punto cualquiera *r*.



COROLARIO 6. Pero como la longitud 2DP es al *latus rectum* de la parábola como la gravedad a la resistencia en D, y al aumentar la velocidad aumenta en la misma razón la resistencia, mientras el *latus rectum* de la parábola aumenta en el cuadrado de dicha razón: es evidente que la longitud 2DP aumenta en dicha razón simple, y por lo tanto siempre resulta proporcional a la velocidad, y tampoco aumenta o disminuye al cambiar el ángulo CDP, a no ser que también se cambie la velocidad.

COROLARIO 7. De donde se tiene el método para determinar a partir de los fenómenos la curva D*ra*F muy aproximadamente, y después hallar la resistencia y la velocidad con la que es proyectado el cuerpo. Láncense dos cuerpos semejantes e iguales con la misma velocidad desde el punto D, según los ángulos distintos CDP, CD*p*, y sean conocidos los puntos F, *f*, donde inciden en el plano horizontal DC. Entonces, supuesta una longitud cualquiera para DP o D*p*, imagínese que la resistencia en D sea a la gravedad en una razón cualquiera, y represéntese dicha razón por una longitud cualquiera SM. Después, por cálculo, hállense a partir de la longitud supuesta DP las longitudes DF y D*f*; y de la razón $\frac{\mathrm{F}f}{\mathrm{DF}}$ hallada por cálculo réstese la misma razón hallada experimentalmente, y represéntese la diferencia mediante la perpendicular MN. Hágase lo mismo por segunda y tercera vez, imponiendo siempre una nueva razón SM de la resistencia a la gravedad, y obteniendo una nueva diferencia MN. Trácense las diferencias positivas hacia un lado de la recta SM y las negativas hacia el otro, y por los puntos N, N, N, trácese la curva regular NNN que corte a la recta SMMM en X, y SX será la verdadera razón de la resistencia a la gravedad que se trataba de hallar. La longitud DF hay que obtenerla por cálculo a partir de esta razón; y una longitud, que sea a la longitud supuesta DP como la longitud DF conocida experimentalmente es a la longitud DF así hallada, será la verdadera longitud DP.

Hallada ésta se tiene tanto la curva D*ra*F descrita por el cuerpo como la velocidad del cuerpo y la resistencia en cada lugar.



Escolio

Por lo demás, que la resistencia de los cuerpos esté en razón de la velocidad es más una hipótesis matemática que natural. En medios libres de toda rigidez, las resistencias de los cuerpos están en razón cuadrada de la velocidad. Pues con la acción de un cuerpo más veloz se comunica a la misma cantidad de un medio, en menor tiempo, mayor movimiento en razón de la mayor velocidad; y en tiempo igual, al ser mayor la cantidad de medio perturbada, se comunica un movimiento mayor en razón cuadrada; ya que la resistencia (por las Leyes II y III) es como el movimiento comunicado. Veamos, pues, qué movimientos se siguen de esta ley de resistencia.

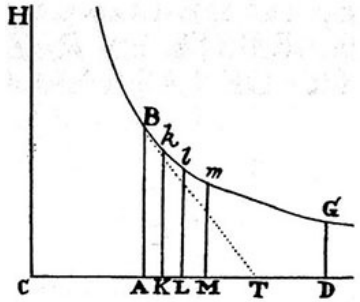
Sección II
DEL MOVIMIENTO DE LOS CUERPOS A LOS QUE SE RESISTE EN RAZÓN DEL CUADRADO DE LAS VELOCIDADES

Proposición V. Teorema III

Si se resiste a un cuerpo en razón del cuadrado de la velocidad, y el mismo se mueve sólo por la fuerza insita a través de un medio homogéneo; y los tiempos se toman en progresión geométrica que va desde los términos menores a los mayores: digo que las velocidades al comienzo de cada tiempo son como el inverso de dicha progresión; y que los espacios descritos en cada uno de los tiempos son iguales.

Pues al ser la resistencia del medio proporcional al cuadrado de la velocidad y el decremento de la velocidad proporcional a la resistencia, si se divide el tiempo en innumerables partículas iguales, los cuadrados de las velocidades al comienzo de cada tiempo serán proporcionales a las diferencias de las mismas velocidades. Tómense dichas partículas de tiempo AK, KL, LM, etc., sobre la recta CD, y elévense las perpendiculares AB, K*k*, L*l*, M*m*, hasta que encuentren a la hipérbola B*klm*G, descrita con centro en C y las asíntotas rectangulares CD, CH, en los puntos B, *k*, *l*, *m*, etc., y AB será a K*k* como CK a CA, y por partes AB - K*k* a K*k* como K*k* a CA y, en consecuencia, como AB x K*k* a AB x CA. Por lo cual, como están dados AK y AB x CA, AB - K*k* será como AB x K*k*; y por fin, cuando AB y K*k* coinciden, como AB^2. Y por un argumento similar, K*k* - L*l*, L*l* - M*m*, etc., serán como Kk^2, Ll^2, etc. Por tanto, los cuadrados de las líneas AB, K*k*, L*l*, M*m*, etcétera, son como sus diferencias y, por eso, al ser también los cuadrados de las velocidades como las diferencias de las mismas, la progresión de ambas será semejante. Demostrado esto, también se sigue que las áreas descritas con estas líneas están en progresión igual que los espacios descritos por las velocidades. Luego si la velocidad al inicio del primer tiempo AK se representa por la línea AB, y la velocidad al inicio del segundo KL se representa por la línea K*k*, y la longitud descrita en el primero por el área AK*k*B; todas las velocidades subsiguientes serán representadas por las líneas siguientes L*l*, M*m*, etc., y las longitudes descritas por las áreas K*l*, L*m*, etc. Y componiendo, si el tiempo total se representa por la suma de sus partes AM, la longitud total descrita se

representa por la suma de sus partes AM*m*B. Imagínese ahora que el tiempo AM se divide en partes AK, KL, LM, etc., de tal modo que CA, CK, CL, CM, etc., estén en progresión geométrica; aquellas partes también estarán en la misma progresión, lo mismo que las velocidades AB, K*k*, L*l*, M*m*, etc., estarán en dicha progresión inversa, y los espacios descritos A*k*, K*l*, L*m*, etc., serán iguales. Q. E. D.



COROLARIO 1. Por tanto, es evidente que si se representa el tiempo por una parte cualquiera AD de la asíntota, y la velocidad al inicio del tiempo por la ordenada AB, la velocidad al fin del tiempo se representará por la ordenada DG, y el espacio total descrito por el área hiperbólica adyacente ABDG; lo mismo que el espacio que puede recorrer un cuerpo en el tiempo AD con la primera velocidad AB en un medio no resistente, mediante el rectángulo AB x AD.

COROLARIO 2. De donde, el espacio recorrido en un medio resistente viene dado si se lo toma respecto al espacio que puede recorrerse con velocidad uniforme AB en el mismo tiempo y en medio no resistente, en la proporción del área hiperbólica ABGD al rectángulo AB x AD.

COROLARIO 3. También está dada la resistencia del medio estableciéndola, al comienzo mismo del movimiento, igual a una fuerza centrípeta uniforme, la cual pudiese generar en un cuerpo en descenso a través de un medio no resistente la velocidad AB en el tiempo AC. Pues, si se traza BT tangente a la hipérbola en B y encuentra a la asíntota en T, la recta AT será igual a la propia AC, y representará al tiempo, en el cual la resistencia inicial continuada uniformemente anularía a toda la velocidad AB.

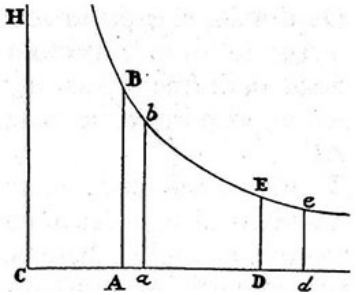
COROLARIO 4. Y de aquí que también esté dada la proporción de esta resistencia respecto a la gravedad o a otra cualquiera fuerza centrípeta dada.

COROLARIO 5. Y viceversa, si se da la proporción de la resistencia a una fuerza centrípeta dada; viene dado el tiempo AC en el que la fuerza centrípeta igual a la resistencia puede generar una velocidad AB cualquiera: y de aquí viene dado el punto B por el cual debe pasar la hipérbola con asíntotas CH, CD; al igual que el espacio ABGD que puede recorrer un cuerpo que inicia su movimiento con la susodicha velocidad AB, en un tiempo cualquiera AD y en un medio homogéneo resistente.

PROPOSICIÓN VI. TEOREMA IV

Los cuerpos esféricos homogéneos e iguales a los que se oponen resistencias como el cuadrado de las velocidades y que se mueven por su sola fuerza ínsita describen espacios iguales siempre en tiempos que sean inversamente como las velocidades iniciales, y pierden partes de las velocidades proporcionales a los todos.

Con asíntotas rectangulares CD, CH, trácese una hipérbola cualquiera B*b* E*e* que corte a las perpendiculares AB, *ab*, DE, *de*, en B, *b*, E, *e*, y representen las velocidades iniciales las perpendiculares AB, DE, etc., y los tiempos las líneas A*a*, D*d*. Pero (por hipótesis) DE es a AB como A*a* es a D*d*, y también (por la naturaleza de la hipérbola) como CA a DC; y por composición, como C*a* a C*d*. Luego las áreas AB*ba*, DE*ed*, esto es, los espacios descritos son iguales entre sí, y las velocidades iniciales AB, DE, son proporcionales a las últimas *ab*, *de*, y por ello, por división, también proporcionales a sus partes perdidas AB - *ab*, DE - *de*. Q. E. D.



PROPOSICIÓN VII. TEOREMA V

Los cuerpos esféricos a los que se resiste en razón cuadrada de la velocidad en tiempos que son directamente como los movimientos iniciales e inversamente como las resistencias iniciales, perderán partes de sus movimientos proporcionales a los todos, y describirán espacios proporcionales al producto de esos tiempos por las velocidades iniciales.

Puesto que las partes perdidas de los movimientos son como el producto de la resistencia y los tiempos. Por tanto, al ser dichas partes proporcionales a los todos, el producto de la resistencia y el tiempo deberá ser como el movimiento. Por lo cual, el tiempo será directamente como el movimiento e inversamente como la resistencia. Por lo tanto, tomadas las partículas de tiempo según dicha razón, los cuerpos perderán siempre partículas de movimiento proporcionales a los todos, y por lo mismo, retendrán velocidades siempre proporcionales a sus velocidades iniciales. Y,

por la razón de velocidades dada, siempre describirán espacios que son como las velocidades iniciales multiplicadas por los tiempos. Q. E. D.

COROLARIO 1. Por tanto, si a cuerpos igualmente veloces se los resiste en razón del cuadrado de los diámetros: los globos homogéneos que se mueven con cualquier velocidad, al describir espacios proporcionales a sus diámetros, perderán partes del movimiento proporcionales a los todos. Pues el movimiento de un globo cualquiera será como el producto de su velocidad y masa, esto es, como el producto de la velocidad y el cubo del diámetro; la resistencia (por hipótesis) será como el producto del cuadrado del diámetro y el cuadrado de la velocidad; y el tiempo (por esta Proposición) estará en razón directa de la primera razón y en razón inversa de la segunda, esto es, directamente del diámetro e inversamente de la velocidad; y por ello el espacio, proporcional al tiempo y a la velocidad, es como el diámetro.

COROLARIO 2. Si a cuerpos igualmente veloces se los resiste en razón de la potencia $^{3}/_{2}$ del diámetro: los globos homogéneos que se muevan a cualquier velocidad, al describir espacios en razón de la potencia $^{3}/_{2}$ de los diámetros, perderán partes de movimiento proporcionales a los todos.

COROLARIO 3. Y en general, si a cuerpos igualmente veloces se los resiste en razón de una potencia cualquiera del diámetro: los espacios en los cuales los globos homogéneos, al moverse con una velocidad cualquiera, perderán partes de movimiento proporcionales a los todos, serán como los cubos de los diámetros aplicados a dicha potencia. Sean dichos diámetros D y E; y si las resistencias, cuando se suponen las velocidades iguales, son D^n y E^n, los espacios en los cuales los globos, al moverse con cualquier velocidad, perderán partes de movimiento proporcionales a los todos serán como D^{3-n} y E^{3-n}. Y por lo tanto, los globos homogéneos que describan espacios proporcionales a dichos D^{3-n} y E^{3-n} retendrán velocidades en la misma mutua proporción que al principio.

COROLARIO 4. Pero si los globos no fuesen homogéneos, el espacio descrito por el globo más denso debe aumentarse en razón de la densidad. Pues el movimiento, a igual velocidad, es mayor en razón de la densidad, y el tiempo (por esta Proposición) aumenta en razón directa del movimiento, y el espacio recorrido en razón del tiempo.

COROLARIO 5. Y si los globos se mueven en distintos medios, el espacio, si el medio es más resistente, «caeteris paribus», habrá de disminuirse en razón de la resistencia. Pues el tiempo (por esta Proposición) disminuirá en razón del aumento de la resistencia, y el espacio en razón del tiempo.

LEMA II[2]

El momento de una generada es igual a los momentos de cada lado generador multiplicados por los índices de las potencias de dichos lados y sus coeficientes

continuamente.

Llamo generada a cualquier cantidad que se engendra, en aritmética por multiplicación, división y extracción de raíces de lados o términos cualesquiera; en geometría del cálculo de áreas y lados, o de extremos y medios proporcionales sin sumar ni restar. Este tipo de cantidades son productos, cocientes, raíces, rectángulos, cuadrados, cubos, lados cuadrados, lados cúbicos y similares. Considero aquí a dichas cantidades como indeterminadas y variables y como si creciesen y decreciesen con un movimiento o flujo continuo; y a sus incrementos o decrementos momentáneos es a lo que llamo momentos: de modo que los incrementos pueden considerarse como momentos añadidos o positivos y los decrementos como negativos o substraídos. Pero cuídese de no entenderlo como partículas finitas. Las partículas finitas no son momentos, sino las cantidades mismas generadas por los momentos. Han de entenderse como los mismos principios nacientes de las magnitudes finitas. Y ni siquiera se contempla en este lema la magnitud de los momentos, sino la proporción primera entre momentos nacientes. Lo mismo ocurre si en lugar de momentos se trata de las velocidades de los incrementos (que también pueden llamarse movimientos, mutaciones, fluxiones de cantidades) o bien de cualquier cantidad finita proporcional a dichas velocidades. El coeficiente de cualquier lado generador es la cantidad que resulta al aplicar la generada a dicho lado.

Por lo cual el sentido del lema es que si los momentos de unas cantidades A, B, C, etc., que aumentan o disminuyen en flujo continuo, o las velocidades de las mutaciones proporcionales a aquellos se llaman *a*, *b*, *c*, etc., el momento o mutación del rectángulo generado AB sería $aB + bA$, el momento del área generada ABC sería $aBC + bAC + cAB$: y los de las potencias generadas A^2, A^3, A^4, $A^{½}$, $A^{⅔}$, $A^{⅓}$, $A^{⅔}$, A^{-1}, A^{-2} y $A^{-½}$ serán respectivamente: $2aA$, $3aA^2$, $4aA^3$, $½aA^{-½}$, $⅔aA^{½}$, $⅓aA^{-⅔}$, $⅔aA^{-⅓}$, $-aA^{-2}$, $-2aA^{-3}$ y $-½aA^{-3/2}$; y, en general, que el momento de una potencia cualquiera $A^{n/m}$ sería $^{n}/_{m}aA^{n - m/m}$. También que el momento de la cantidad generada A^2B será $2aAB + bA^2$, que el momento de la generada $A^3B^4C^2$ será $3aA^2B^4C^2 + 4bA^3B^3C^2 + 2cA^3B^4C$, y de la generada $\frac{A^3}{B^2}$ o de A^3B^{-2} será $3aA^2B^{-2} - 2bA^3B^3$, etc. Y el Lema se demuestra del siguiente modo:

CASO 1. Un rectángulo cualquiera AB que crece con un movimiento continuo, cuando de sus lados A y B faltaban la mitad de los momentos $½a$ y $½b$, era $A - ½a$ por $B - ½b$, o también $AB - ½bA + ¼ab$; y en el momento en que ambos lados A y B quedan aumentados con las otras mitades de los momentos resulta $A + ½a$ por $B + ½b$, o $AB + ½aB + ½bA + ¼ab$. De este rectángulo réstese el anterior y quedará un exceso de $aB + bA$. Por tanto, con todos los incrementos *a* y *b* de los lados se genera el incremento $aB + bA$ del rectángulo. Q. E. D.

Caso 2. Supóngase que AB es siempre igual a G y el momento del contenido ABC, o sea, de GC (por el caso 1) será gC + cG, esto es (si se sustituyen G y g por AB y aB + bA), aBC + bAC + cAB. Y lo mismo para cualquier contenido bajo cualquier número de lados. Q. E. D.

Caso 3. Supóngase que los lados A, B, C, son siempre iguales entre sí, y el momento del rectángulo AB, esto es, de A^2, que sería aB + bA, vendrá a ser 2aA; y el de A^3, esto es, el del contenido ABC, que sería aBC + bAC + cAB, 3$a$$A^2$. Y por la misma razón el momento de una potencia cualquiera A^n será n$a$$A^{n-1}$. Q. E. D.

Caso 4. De donde, como $\frac{1}{A}$ por A sea 1, el momento de $\frac{1}{A}$ multiplicado por A, junto con $\frac{1}{A}$ multiplicado por a, será el momento de 1, es decir, nulo. Por tanto, el momento de $\frac{1}{A}$ o de A^{-1} es $\frac{-a}{A^2}$. Y, en general, puesto que $\frac{1}{A^n}$ por A^n es 1, el momento de $\frac{1}{A^n}$ multiplicado por A^n junto con $\frac{1}{A^n}$ por $na$$A^{n-1}$ será nulo. Y por lo mismo el momento de $\frac{1}{A^n}$ o de A^n será $-\frac{na}{A^{n+1}}$. Q. E. D.

Caso 5. Y, puesto que $A^{½}$ por $A^{½}$ es A, el momento de $A^{½}$ multiplicado por 2$A^{½}$ será a por el caso 3: por lo mismo el momento de $A^{½}$ será $\frac{a}{2A^{½}}$ o ½$a$$A^{-½}$. Y, en general, siendo $A^{m/n}$ igual a B, entonces A^m será igual a B^n y, por tanto, $ma$$A^{m-1}$ igual a $nb$$B^{n-1}$, o $nb$$A^{-m/n}$ y, por ello, $\frac{m}{n}a$$A^{m-n/n}$ es igual a b, es decir, igual al momento de $A^{m/n}$. Q. E. D.

Caso 6. Por tanto, el momento de cualquier cantidad generada A^mB^n es el momento de A^m multiplicado por B^n, junto con el momento de B^n multiplicado por A^m, esto es, $ma$$A^{m-1}B^n$ + $nb$$B^{n-1}A^m$; y esto tanto si los índices de las potencias m, n, son números enteros o fraccionarios, positivos o negativos. Y lo mismo ocurre para contenidos bajo muchas potencias. Q. E. D.

Corolario 1. De aquí que, cuando son cantidades continuamente proporcionales, si se da un término, los momentos de los otros términos serán como los mismos términos multiplicados por el número de intervalos entre ellos y el término dado. Sean A, B, C, D, E, F, continuamente proporcionales; y si se da el término C, los momentos del resto de los términos serán entre sí como -2A, -B, D, 2E, 3F.

Corolario 2. Y si en cuatro proporcionales se dan los dos medios, los momentos de los extremos serán como los propios extremos. Lo mismo hay que entender respecto a los lados de un rectángulo dado cualquiera.

Corolario 3. Y si se da la suma o la diferencia de dos cuadrados, los momentos de los lados serán inversamente como los lados.

Escolio

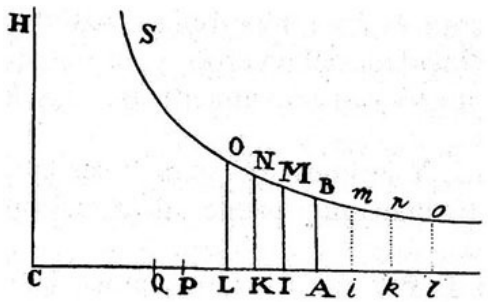
En una carta que escribí al Sr. Collins el 10 de diciembre de 1672 una vez que describí el método de tangentes que sospechaba ser el mismo que el de *Sluse*, no publicado todavía en aquella fecha, añadí: «Esto es un caso particular o mejor un Corolario de un método general que se extiende sin farragosos cálculos, no sólo al trazado de tangentes a cualquier línea curva geométrica o mecánica, o relacionadas de cualquier forma a otras líneas rectas o curvas, sino también a la resolución de otras clases de problemas más abstrusos sobre curvaturas, áreas, longitudes, centros de gravedad de curvas, etc., y tampoco (como el método de máximos y mínimos de *Hudden*) se limita a las solas ecuaciones que se hallan libres de cantidades irracionales. He combinado dicho método con este otro mediante el cual resuelvo las ecuaciones reduciéndolas a series infinitas». Hasta aquí la carta. Y estas últimas palabras se refieren a un tratado que había escrito sobre estos temas en el año 1671. Mas, el fundamento de dicho método general se halla contenido en el Lema anterior.

Proposición viii. Teorema vi[3]

Si un cuerpo, en un medio uniforme, bajo la acción uniforme de la gravedad, asciende o desciende en línea recta, y se divide la totalidad del espacio recorrido en partes iguales, y se averiguan las fuerzas absolutas al principio de cada una de las partes (sumando la resistencia del medio a la fuerza de la gravedad cuando el cuerpo asciende y restándola cuando desciende), digo que dichas fuerzas absolutas se hallan en progresión geométrica.

Represéntese la fuerza de la gravedad por una línea dada AC; la resistencia por una línea indefinida AK; la fuerza absoluta en el descenso del cuerpo por la diferencia KC; la velocidad del cuerpo por la línea AP, que sea media proporcional entre AK y AC, y por tanto como la raíz cuadrada de la resistencia; el incremento de la resistencia producido en una partícula de tiempo dada por el segmento KL, y el incremento simultáneo de velocidad por el segmento PQ; y con centro en C y con las asíntotas rectangulares CA, CH trácese una hipérbola BNS que corte a las perpendiculares AB, KN, LO, en los puntos B, N, O. Puesto que AK es como AP^2, el

momento KL de éste será como el momento 2APQ de aquél, esto es, como AP x KC; pues el incremento PQ de la velocidad (por la Ley II del movimiento) es proporcional a la fuerza generatriz KC. Multiplicando la razón de KL por la razón de KN se tendrá el rectángulo KL x KN que será como AP x KC x KN; esto es, por estar dado el rectángulo KC x KN, como AP. Pero la razón última del área hiperbólica KNOL al rectángulo KL x KN cuando coinciden los puntos K y L es la de igualdad. Luego dicha área hiperbólica evanescente es como AP. Por tanto, toda el área hiperbólica ABOL se compone de partículas KNOL siempre proporcionales a la velocidad AP, y en consecuencia es proporcional al espacio descrito con dicha velocidad. Divídase ahora la susodicha área en partes iguales ABMI, IMNK, KNOL, etcétera, y las fuerzas absolutas AC, IC, KC, LC, etc., estarán en proporción geométrica. Q. E. D. Y por el mismo argumento, cuando el cuerpo asciende, tomando, hacia el lado contrario del punto A, las áreas iguales AB*mi*, *imnk*, *knol*, etc., se verá que las fuerzas absolutas AC, *i*C, *k*C, *l*C, etc., son continuamente proporcionales. Y por lo tanto, si en el ascenso y en el descenso se toman iguales todos los espacios, todas las fuerzas absolutas *l*C, *k*C, *i*C, AC, IC, KC, LC, etc., serán continuamente proporcionales. Q. E. D.



COROLARIO 1. De aquí que si se representa el espacio recorrido mediante el área hiperbólica ABNK, la fuerza de la gravedad, la velocidad del cuerpo y la resistencia del medio pueden representarse respectivamente por las líneas AC, AP y AK; y viceversa.

COROLARIO 2. Y la línea AC viene a ser la representación de la velocidad máxima que puede alcanzar un cuerpo en un descenso infinito.

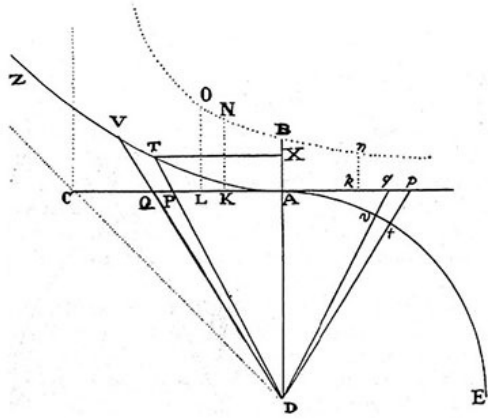
COROLARIO 3. Por tanto, si se conoce la resistencia de un medio a una velocidad dada, se hallará la velocidad máxima tomando a ésta con relación a dicha velocidad dada como la raíz cuadrada de la razón entre la fuerza de la gravedad y la resistencia dada del medio ya conocida.

PROPOSICIÓN IX. TEOREMA VII

Supuesto lo ya demostrado, digo que si las tangentes de los ángulos de un sector circular y de un sector hiperbólico se toman proporcionalmente a las velocidades, siendo el radio de una magnitud determinada, el tiempo total de ascenso hasta el punto más alto será como el sector del círculo, y el tiempo total de descenso desde el punto más alto como el sector de la hipérbola.

Trácese AD perpendicular e igual a la recta AC que representa a la fuerza de la gravedad. Con centro en D y semidiámetro AD, trácese tanto el cuadrante de círculo A*t*E, como la hipérbola rectangular AVZ que tiene como eje AX, el vértice principal en A y la asíntota DC. Únanse D*p*, DP y el sector circular A*t*D será como el tiempo total de ascenso al punto más alto; y el sector hiperbólico ATD como el tiempo total de descenso desde el punto más alto, mientras las tangentes A*p*, AP de los sectores sean como las velocidades.

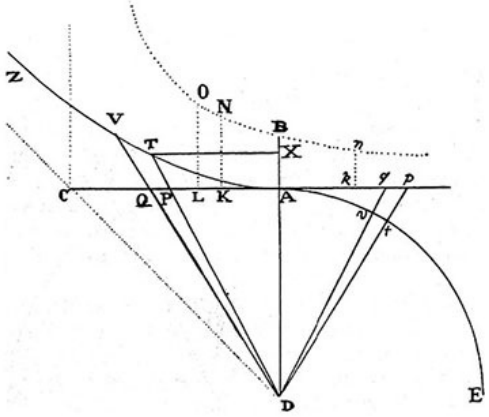
CASO 1. Trácese D*vq* de modo que corte los momentos del sector AD*t* y del triángulo AD*p*, o partículas mínimas descritas al mismo tiempo *t*D*v*, y *q*D*p*. Como dichas partículas, por tener el mígalo D común, son como el cuadrado de los lados, la partícula *t*D*v* será como $\frac{q\mathrm{D}p \times t\mathrm{D}^2}{p\mathrm{D}^2}$, esto es, por estar dada *t*D, como $\frac{q\mathrm{D}p}{p\mathrm{D}^2}$. Pero $p\mathrm{D}^2$ es $\mathrm{AD}^2 + \mathrm{A}p^2$, es decir, AD^2 + AD x A*k*, o, AD x C*k*, y *q*D*p* es ½ AD x *pq*. Por lo tanto, *t*D*v*, la partícula del sector, es como *pq*/C*k*; esto es, directamente como el decremento mínimo de la velocidad *pq*, e inversamente como la fuerza C*k* que hace disminuir la velocidad; y por tanto como la partícula de tiempo correspondiente al decremento de la velocidad. Y, componiendo, la suma de todas las partículas *t*D*v* en el sector AD*t* vendrá a resultar como la suma de las partículas de tiempo correspondientes a cada partícula *pq* perdida de la velocidad decreciente A*p*, hasta que dicha velocidad, disminuyendo hasta anularse, desaparezca; esto es, el sector completo AD*t* es como el tiempo total de ascenso hasta el punto más alto. Q. E. D.



CASO 2. Trácese DQV tal que corte las partículas mínimas TDV y PDQ tanto del sector DAV como del triángulo DAQ; y dichas partículas serán entre sí como DT^2 a DP^2, esto es (si TX y AP son paralelas) como DX^2 a DA^2 o TX^2 a AP^2 y, por partes, como $DX^2 - TX^2$ a $DA^2 - AP^2$. Pero, por la naturaleza de la hipérbola, $DX^2 - TX^2$ es AD^2, y por hipótesis, AP^2 es AD x AK. Luego, las partículas son entre sí como AD^2 a AD^2 - AD x AK; esto es, como AD a AD - AK, o también, AC a CK; y por ello la partícula TDV del sector es $\frac{\text{PDQ x AC}}{\text{CK}}$; y por estar dadas AC y AD, como PQ/CK, esto es, directamente como el incremento de la velocidad e inversamente como la fuerza que genera el incremento. Y, componiendo, se tendrá la suma de las partículas de tiempo en las cuales se generan todas las partículas PQ de la velocidad AP, que vendrá a ser como la suma de las partículas del sector ATD, esto es, el tiempo total como el sector total. Q. E. D.

COROLARIO 1. De aquí que, si AB es igual a la cuarta parte de AC, el espacio recorrido por un cuerpo que cae en un tiempo cualquiera será al espacio que puede recorrer dicho cuerpo con la velocidad máxima AC, moviéndose uniformemente durante el mismo tiempo, como el área ABNK, que representa el espacio recorrido en la caída, al área ATD, que representa el tiempo. Pues, dado que AC es a AP como AP a AK, LK será (por el corolario 1 del Lema II de este libro) a PQ como 2AK a AP, esto es, como 2AP a AC, y por ello LK a ½PQ como AP a ¼AC o AB; pero KN es a AC o AD como AB a CK; y por tanto, «*ex aequo*», LKNO es a DPQ como AP a CK. Pero DPQ era a DTV como CK a AC. Luego «*ex aequo*» otra vez, LKNO es a DTV como AB a AC; esto es, como la velocidad del cuerpo que cae a la velocidad máxima que puede adquirir el cuerpo al caer. Puesto que los momentos LKNO y DTV de las áreas ABNK y ATD son como las velocidades, todas las partes de dichas áreas generadas en los mismos tiempos, serán como los espacios recorridos en los mismos

tiempos y, por lo mismo, las áreas totales ABNK y ATD generadas desde el principio serán como los espacios totales recorridos desde el comienzo del descenso. Q. E. D.



COROLARIO 2. Y lo mismo se sigue para el espacio recorrido en el ascenso. A saber, que dicho espacio total es al espacio recorrido con velocidad uniforme AC en el mismo tiempo, como el área AB*nk* al sector AD*t*.

COROLARIO 3. La velocidad del cuerpo que cae en el tiempo A I D es a la velocidad que adquiriría en el mismo tiempo en un espacio sin resistencia, como el triángulo APD al sector hiperbólico ATD. Pues la velocidad en un medio no resistente sería como el tiempo ATD, mientras en un medio resistente es como AP, esto es, como el triángulo APD. Y estas velocidades al comienzo del descenso son iguales entre sí, como lo son las áreas ATD, APD.

COROLARIO 4. Por la misma razón la velocidad en el ascenso es a la velocidad con la cual un cuerpo en un espacio no resistente perdería todo su movimiento ascendente, como el triángulo A*p*D al sector circular A*t*D, o como la recta A*p* al arco A*t*.

COROLARIO 5. Por lo tanto, el tiempo en el cual un cuerpo, cayendo en un medio resistente, adquiriría la velocidad AP es al tiempo en el cual adquiriría la velocidad máxima AC cayendo en un espacio no resistente, como el sector ADT al triángulo ADC: y el tiempo en el cual perdería ascendiendo en un medio resistente la velocidad A*p* será al tiempo en el cual puede perder la misma velocidad ascendiendo en un espacio no resistente, como el arco A*t* a su tangente A*p*.

COROLARIO 6. De aquí que, dado el tiempo, está dado el espacio recorrido en el ascenso o en el descenso. Pues la velocidad máxima de un cuerpo que desciende hasta el infinito está dada (por los Corolarios 2 y 3 del Teorema VI del Libro II) y con ello se da el tiempo en el cual el cuerpo puede alcanzar dicha velocidad cayendo en un espacio no resistente. Y tomando el sector ADT o AD*t* respecto al triángulo ADC

en la razón del tiempo dado al tiempo recién hallado, se tendrán tanto la velocidad AP o A*p* como el área ABNK o AB*nk* que es al sector ADT o AD*t* como el espacio buscado al espacio que, en un tiempo dado, y con la velocidad máxima ya antes hallada, podría recorrer uniformemente.

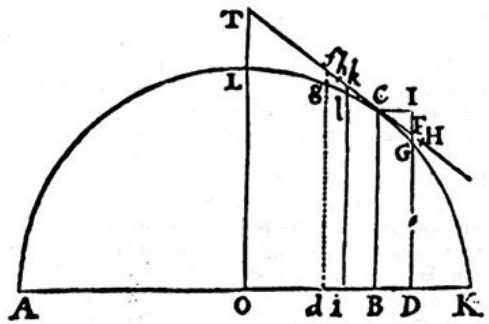
COROLARIO 7. Y en sentido inverso, a partir de los espacios dados de ascenso o descenso AB*nk* o ABNK, se tendrá el tiempo AD*t* o ADT.

PROPOSICIÓN X. PROBLEMA III[4]

Suponiendo que la fuerza uniforme de la gravedad tiende directamente hacia el plano del horizonte, y que la resistencia del medio es como la densidad del mismo y el cuadrado de la velocidad conjuntamente: hallar tanto la densidad del medio en cada punto capaz de obligar al cuerpo a moverse en una curva dada cualquiera, como la velocidad del cuerpo y la resistencia del medio en cada punto.

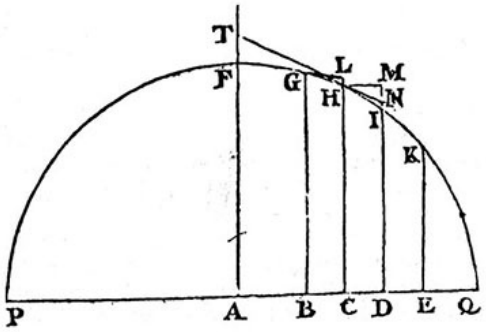
Sea PQ un plano perpendicular al plano de la figura; PFHQ una línea curva que corta a dicho plano en los puntos P y Q; G, H, I, K, cuatro lugares del cuerpo que recorre dicha curva desde F hacia Q; y GB, HC, ID, KE, cuatro paralelas ordenadas trazadas desde dichos puntos sobre el plano horizontal y que caen sobre la línea horizontal PQ en los puntos B, C, D, E; y sean iguales las distancias BC, CD, DE, comprendidas entre esas ordenadas. Desde los puntos G y H trácense las rectas GL, HN, tangentes a la curva en G y H, y cortando a las ordenadas CH, DI, prolongadas hacia arriba en L y N, y complétese el paralelogramo HCDM. Y los tiempos en los cuales el cuerpo describe los arcos GH, HI, serán como la raíz cuadrada de las alturas LH, NI que describiría el cuerpo en dichos tiempos, cayendo desde las tangentes; y las velocidades serán como las longitudes recorridas GH, HI, directamente e inversamente como los tiempos. Represéntense los tiempos por T y *t* y las velocidades por GH/T y HI/*t*; y el decremento de la velocidad ocurrido en el tiempo *t* represéntese por $\frac{GH}{T} - \frac{HI}{t}$. Este decremento surge de la resistencia que retarda al cuerpo y de la gravedad que lo acelera. La gravedad genera, en un cuerpo que cae y al caer recorre el espacio NI, una velocidad capaz de hacer al cuerpo recorrer un espacio doble de aquel en el mismo tiempo, como demostró GALILEO; esto es, 2NI/*t*; pero en el cuerpo que describe el arco HI solamente aumenta en dicho arco la longitud HI - HN o $\frac{MI \times NI}{HI}$; y, por lo mismo, tan sólo genera la velocidad $\frac{2MI \times NI}{t \times HI}$. Añádase esta velocidad al decremento antedicho, y se tendrá el decremento de la velocidad

debido a la sola resistencia, a saber, $\frac{GH}{T} - \frac{HI}{t} + \frac{2MI \times NI}{t \times HI}$. Y por tanto, puesto que en el mismo tiempo la gravedad genera en un cuerpo que cae la velocidad $\frac{2NI}{t}$, la resistencia será a la gravedad como $\frac{GH}{T} - \frac{HI}{t} + \frac{2MI \times NI}{t \times HI}$ a $\frac{2NI}{t}$ o como $\frac{t \times GH}{T} - HI + \frac{2MI \times NI}{HI}$ a 2NI.



Y ahora, en lugar de las abscisas CB, CD, CE, escríbase -*o*, *o*, 2*o*. Por la ordenada CH escríbase P, y por MI una serie cualquiera como Q*o* + R*oo* + So^3 + etc. Y todos los términos de la serie posteriores al primero, a saber, R*oo* + So^3 + etc., serán NI, y las ordenadas DI, EK y BG serán P - Q*o* - R*oo* - So^3 - etc., P - 2Q*o* - 4R*oo* - 8So^3 - etc., y P + Q*o* - R*oo* + So^3 - etc., respectivamente. Elevando al cuadrado las diferencias de las ordenadas BG - CH y CH - DI y sumando a los cuadrados resultantes los cuadrados de las propias BC, CD se obtendrán *oo* + QQ*oo* - 2QRo^3 + etc., y $\overline{oo + QQoo - 2QRo^3 + \text{etc.}}$, cuadrados de los arcos GH, HI, cuyas raíces $o\sqrt{1 + QQ} - \frac{QRoo}{\sqrt{1 + QQ}}$ y $o\sqrt{1 + QQ} + \frac{QRoo}{\sqrt{1 + QQ}}$ son los arcos GH y HI. Además, si de la ordenada CH se resta la semisuma de las ordenadas BG y DI, y de la ordenada DI se resta la semisuma de las ordenadas CH y EK quedarán Roo y R*oo* + 3So^3, sagitas de los arcos GI y HK. Pero éstas son proporcionales a los segmentos LH y NI y, por tanto, como el cuadrado de los tiempos infinitamente pequeños T y *t*; y de aquí que la razón $\frac{t}{T}$ es $\sqrt{\frac{R + 3So}{R}}$ o $\frac{R + \frac{3}{2}So}{R}$; mientras que $\frac{t \times GH}{T} - HI + \frac{2MI \times NI}{HI}$, una vez sustituido los valores de *t*/T, GH, HI, MI y NI por los ahora hallados, resulta

$\frac{3Soo}{2R}$ x $\sqrt{1 + QQ}$. Y puesto que 2NI es 2R*oo*, la resistencia será ahora a la gravedad como $\frac{3Soo}{2R}$ x $\sqrt{1 + QQ}$ a 2R*oo*, esto es, como 3S$\sqrt{1 + QQ}$ a 4RR.



Y la velocidad es aquella con la cual un cuerpo partiendo de un punto cualquiera H según la tangente HN podría moverse descendiendo en el vacío por una parábola de diámetro HC y «latus rectum» $\frac{HN^2}{NI}$ o $\frac{1 + QQ}{R}$.

Y la resistencia es como la densidad del medio y el cuadrado de la velocidad conjuntamente, y por tanto la densidad del medio es directamente como la resistencia e inversamente como el cuadrado de la velocidad, esto es, como $\frac{3S\sqrt{1 + QQ}}{4RR}$ directamente y $\frac{1 + QQ}{R}$ inversamente, esto es, como $\frac{S}{R\sqrt{1 + QQ}}$. Q. E. I.

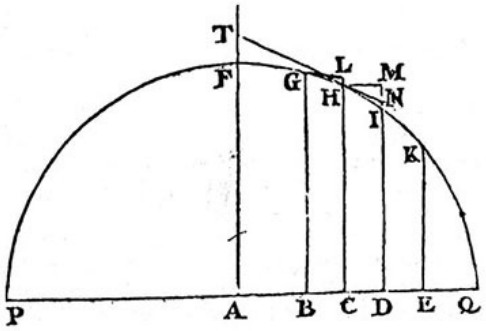
COROLARIO 1. Si la tangente HN se prolonga a ambos lados de modo que llegue a cortar a una ordenada cualquiera AF en T, $\frac{HT}{AC}$ será igual a $\sqrt{1 + QQ}$, y por tanto en lo que antecede se puede escribir en lugar de $\sqrt{1 + QQ}$. Por lo cual la resistencia será a la gravedad como 3S x HT a 4RR x AC, la velocidad será como $\frac{HT}{AC\sqrt{R}}$ y la densidad del medio como $\frac{S \text{ x } AC}{R \text{ x } HT}$.

COROLARIO 2. Y por ello, si se definiese la línea curva PFHQ mediante la relación

entre la base o abscisa AC y la ordenada CH, como es habitual, y el valor de la ordenada se resuelve en una serie convergente, el problema se resuelve rápidamente mediante los primeros términos de la serie, como en los ejemplos siguientes.

EJEMPLO 1. Sea la línea PFHQ un semicírculo descrito sobre el diámetro PQ, y hállese la densidad del medio que obligue a un proyectil a moverse en dicha línea.

Bisecado el diámetro PQ en A, llámese a AQ, *n*; a AC, *a*; a CH, *e*; y a CD, *o*: y DI^2 o AQ^2 - AD^2 será igual a: *nn* - *aa* - *2ao* - *oo*, o también, *ee* - 2*ao* - *oo*, y al extraer la raíz según nuestro método, resultará



$$DI = e - \frac{ao}{e} - \frac{oo}{2e} - \frac{aaoo}{2e^3} - \frac{ao^3}{2e^3} - \frac{a^3o^3}{2e^5} - \text{etc.}$$

Aquí sustituyase *ee* + *aa* por *nn* y resultará

$$DI = e - \frac{ao}{e} - \frac{nnoo}{2e^3} - \frac{anno^3}{2e^5} - \text{etc.}$$

En esta serie distingo entre los términos sucesivos del modo siguiente. Llamo término primero a aquel en el que no se da la cantidad pequeña *o*; segundo a aquel en el que dicha cantidad es de una sola dimensión; tercero al que la contiene con dos; cuarto al que la contiene con tres; y así indefinidamente. Y el término primero, que aquí es *e*, denotará siempre la longitud de la ordenada CH incidente en el comienzo de la cantidad indefinida *o*. El término segundo, que aquí es $\frac{ao}{e}$, denotará la diferencia entre CH y DN, esto es, el segmento MN, que queda cortado al completar el paralelogramo HCDM, y por lo tanto determina siempre la posición de la tangente HN; como en este caso tomando MN a HM como $\frac{ao}{e}$ a *o*, o también *a* a *e*. El tercer

término que es aquí $\frac{nnoo}{2e^3}$, denotara el segmento IN que se halla entre la tangente y la curva, y por tanto determina el ángulo de contacto IHN o la curvatura que tiene la línea curva en H. Si dicho segmento IN es de magnitud finita será designada por el término tercero junto con los que le siguen hasta el infinito. Pero si dicho segmento disminuye hasta el infinito, los términos siguientes resultarán infinitamente más pequeños que el tercero y, por tanto, pueden despreciarse. El término cuarto determina la variación de la curvatura, el quinto la variación de la variación y así sucesivamente. De lo cual, de paso, se desprende la utilidad no desdeñable de estas series en la solución de problemas que dependen de tangentes y de la curvatura de las curvas.

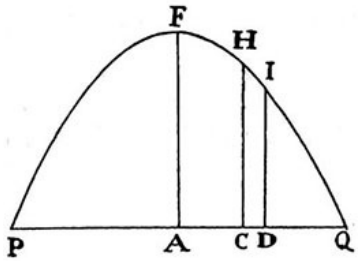
Compárese ahora la serie

$$e - \frac{ao}{e} - \frac{nnoo}{2e^3} - \frac{annno^3}{2e^5} - \text{etc.}$$

con la serie P - Q*o* - R*oo* - So^3 - etc. y después en lugar de P, Q, R y S escríbase e, $\frac{a}{e}$, $\frac{nn}{2e^3}$, $\frac{ann}{2e^5}$ y en lugar de $\sqrt{1 + QQ}$ escríbase $\sqrt{1 + \frac{aa}{ee}}$ o $\frac{n}{e}$, y la densidad del medio resultará como $\frac{a}{ne}$, esto es (por estar *n* dada) como $\frac{a}{e}$ o como $\frac{AC}{CH}$, es decir, como la longitud de la tangente HT que termina en el semidiámetro AF que se eleva perpendicularmente sobre PQ; mientras la resistencia será a la gravedad como 3a a 2n, esto es, como 3AC al diámetro del círculo PQ; la velocidad en cambio será $\sqrt{CH}$. Por tanto, si el cuerpo parte del punto F, con la velocidad precisa, según una línea paralela a PQ, y la densidad del medio en cada punto H es como la longitud HT de la tangente, y la resistencia es también en cada punto H respecto a la fuerza de la gravedad como 3AC a PQ, dicho cuerpo describirá el cuadrante FHQ del círculo. Q. E. I.

Pero si dicho cuerpo partiese del punto P según una línea perpendicular a PQ, y comenzase a moverse en el arco del semicírculo PFQ, habría que tomar AC o *a* hacia el otro lado del centro A y por tanto habría que cambiar su signo y escribir -*a* en lugar de +*a*. Con ello resultaría la densidad del medio como $-\frac{a}{e}$. Pero la naturaleza no admite una densidad negativa o, lo que es lo mismo, que acelere el movimiento: y por lo tanto no puede acontecer naturalmente que un cuerpo ascendiendo desde P describa el cuadrante PF del círculo. Para esto el cuerpo tendría que ser acelerado por un medio impelente, no ser retardado por un medio resistente.

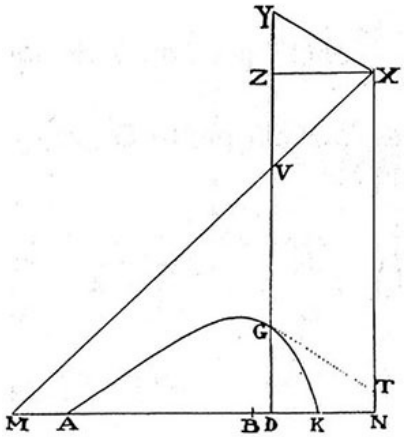
EJEMPLO 2. Sea la línea PFQ una parábola que tiene su eje AF perpendicular a la línea del horizonte PQ, y hállese la densidad del medio que haga que un proyectil se mueva por ella.



De la naturaleza de la parábola se sigue que el rectángulo PDQ es igual al rectángulo comprendido bajo la ordenada DI y una cierta recta dada: esto es, si llamamos a dicha recta *b*, a PC, *a*; a PQ, *c*; a CH, *e*; y a CD, *o*; el rectángulo (*a* + *o*) x (*c* - *a* - *o*), o, *ac* - *aa* - *2ao* + *co* - *oo* es igual a *b* x DI, y por tanto, DI es igual a $\frac{ac - aa}{b} + \frac{c - 2a}{b}o - \frac{oo}{b}$.

Ahora habrá que escribir el segundo término de esta serie $\frac{c - 2a}{b}$ x *o*, por Q*o*, mientras el tercero $\frac{oo}{b}$, lo sera en lugar de R*oo*. Pero como no hay más términos, el coeficiente S del cuarto desaparecerá, con lo cual la cantidad $\frac{S}{R\sqrt{1 + QQ}}$ a laque es proporcional la densidad, se anula. Y por consiguiente, un proyectil se moverá en una parábola cuando la densidad del medio sea nula, como ya demostró Galileo. Q. E. I.

EJEMPLO 3. Sea la línea AGK una hipérbola que tiene como asíntota a NX perpendicular al plano horizontal AK; y se trata de hallar la densidad del medio que haga a un proyectil moverse en dicha línea.



Sea MX la otra asíntota que corta a la prolongación de la ordenada DG en V; y por la naturaleza de la hipérbola vendrá dado el rectángulo XV x VG. También estará dada la razón de DN a VX, y por tanto está dado también el rectángulo DN x VG. Sea éste *bb*: y completando el paralelogramo DNXZ, escríbase por BN, *a*; por BD, o; por NX, *c*; y la razón dada de VZ a ZX o DN supongamos que es $\frac{m}{n}$. Y entonces DN será igual a $a - o$, VG igual $\frac{bb}{a - o}$, VZ igual a $\frac{m}{n}(a - o)$ y GD o NX - VZ - VG igual a

$$c - \frac{m}{n}\,\mathrm{a} + \frac{m}{n}\,o - \frac{bb}{a - o}\,,$$

Resuélvase el término $\frac{bb}{a - o}$ en la serie convergente

$$\frac{bb}{a} + \frac{bb}{aa}\,o + \frac{bb}{a^3}\,oo + \frac{bb}{a^4}\,o^3, \text{ etc.;}$$

y GD resultará igual a

$$c - \frac{m}{n}\,\mathrm{a} - \frac{bb}{a}\,o + \frac{m}{n}\,o - \frac{bb}{aa}\,o - \frac{bb}{a^3}\,o^2 - \frac{bb}{a^4}\,o^3, \text{ etc.}$$

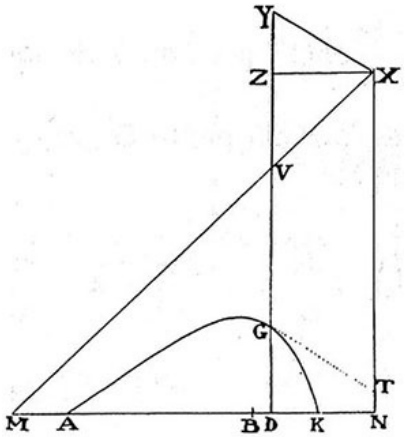
El segundo término, $\frac{m}{n}o - \frac{bb}{aa}o$, de esta serie ha de tomarse por Q*o*, el tercero con el

signo cambiado, $\frac{bb}{a^3}o^2$, por Ro^2, el cuarto también con el signo cambiado, $\frac{bb}{a^4}o^3$ por So^3, y sus coeficientes $\frac{m}{n} - \frac{bb}{aa}$, $\frac{bb}{a^3}$ y $\frac{bb}{a^4}$ deben ponerse en la regla anterior en lugar de Q, R y S. Una vez hecho esto, la densidad del medio resultará como

$$\frac{{}^{bb}\!/_{a^4}}{{}^{bb}\!/_{\mathrm{a}^3}\sqrt{1 + {}^{mm}\!/_{nn} - {}^{2mbb}\!/_{naa} + {}^{b^4}\!/_{a^4}}}$$

o como, $\frac{1}{\sqrt{aa + {}^{mm}\!/_{nn}aa - {}^{2mbb}\!/_{n} + {}^{b^4}\!/_{aa}}}$, esto es, si sobre VZ se toma VY igual a VG, como $\frac{1}{\mathrm{XY}}$. Porque *aa* y $\frac{mm}{nn}aa - \frac{2mbb}{n} + \frac{b^4}{aa}$ son los cuadrados de XZ y de ZY. Luego la razón de la resistencia a la gravedad es la de 3XY a 2TG; y la velocidad es aquella con la que un cuerpo recorrería una parábola que tuviese el vértice G, el diámetro DG, y el «latus rectum» $\frac{\mathrm{XY}^2}{\mathrm{VG}}$. Supóngase también que las densidades del medio en cada punto G son inversamente como las distancias XY, y que la resistencia en un punto G es a la gravedad como 3XY a 2YG; y un cuerpo lanzado desde el punto A con la velocidad adecuada describirá la hipérbola dicha AGK. Q. E. I.

EJEMPLO 4. Supóngase indefinidamente que la línea AGK es una hipérbola con centro en X, descrita entre las asíntotas MX, NX, de tal modo que, construido el rectángulo XZDN, cuyo lado ZD corte a la hipérbola en G y a su asíntota en V, y fuese VG inversamente como ZX o DN a una potencia cualquiera DN^n, cuyo índice es el número *n*; y se busca la densidad del medio con la cual un proyectil recorra dicha curva.



Escríbanse A, O, C, en lugar de BN, BD, NX respectivamente, y sea VZ a XZ o DN como *d* a *e*, y VG igual a $\frac{bb}{DN^3}$; y DN será igual a A - O, VG= $\frac{bb}{(A - O)^n}$, VZ= $\frac{d}{e}$(A - O), y GD o NX - VZ - VG igual a C - $\frac{d}{e}$A + $\frac{d}{e}$O - $\frac{bb}{(A - O)^n}$. Resuélvase el termino $\frac{bb}{(A - O)^n}$ en la serie infinita

$$\frac{bb}{A^n} + \frac{nbb}{A^{n+1}}\,O + \frac{nn + n}{2A^{n+2}}\,bbO^2 + \frac{n^3 + 3nn + 2n}{6A^{n+3}}\,bbO^3, \text{ etc.}$$

y GD será igual a

$$C - \frac{d}{e}\,A - \frac{bb}{A^n} + \frac{d}{e}\,O - \frac{nbb}{A^{n+1}}\,O$$

$$- \frac{+nn + n}{2A^{n+2}}\,bbO^2 - \frac{+n + 3nn + 2n}{6A^{n+3}}\,bbO^3$$

etcétera. El segundo término, $\frac{d}{e}O - \frac{nbb}{A^{n+1}}O$, de esta serie habrá de tomarse por Q*o*, mientras el tercero, $\frac{nbb + n}{2A^{n+2}}bbO^2$, lo será por R*o*^{2}, el cuarto, $\frac{n^3 + 3nn + 2n}{6A^{n+3}}bbO^3$, por

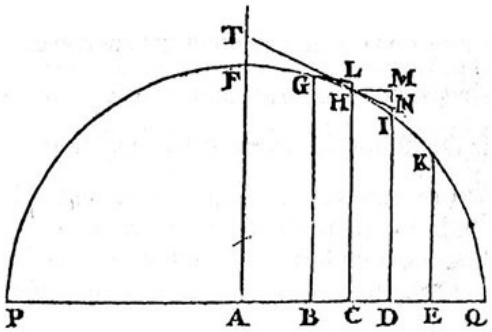
S*o*^{3}. Y de aquí que la densidad del medio, $\frac{S}{R\sqrt{1 + QQ}}$, en un punto G cualquiera resulte

$$\frac{n + 2}{3\sqrt{a^2 + {}^{dd}\!/_{ee}A^2 - {}^{2dnbb}\!/_{eA^n}A + {}^{nnb^4}\!/_{A^{2n}}}}$$

y por lo tanto, si sobre VZ se toma VY igual a *n* x VG, dicha densidad es inversamente como XY. Porque A^2 y $\frac{dd}{ee}A^2 - \frac{2dnbb}{eA^n}A + \frac{nnb^4}{A^{2n}}$ son los cuadrados de las propias XZ y ZY. Y la resistencia en el mismo punto G se hace respecto a la gravedad como 3S x $\frac{XY}{A}$ a 4RR, es decir, como XY a $\frac{2nn + 2n}{n + 2}$VG. Y la velocidad es allí la misma que aquella con la cual el cuerpo lanzado se movería en una parábola con vértice en G, diámetro GD y «latus rectum» $\frac{1 + QQ}{R}$ o $\frac{2XY^2}{(nn + n) \text{ x } VG}$. Q. E. I.

Escolio

Por la misma razón por la que en el Corolario primero resulta la densidad del medio como $\frac{S \text{ x } AC}{R \text{ x } HT}$, si la resistencia se supone como una potencia cualquiera V^n de la velocidad V, la densidad del medio resultará como $\frac{S}{R^{4 - n/2}}$ x $\left(\frac{AC}{HT}\right)^{n - 1}$. Y por lo tanto, si la curva puede ser establecida de modo que responda a la razón de $\frac{S}{R^{4 - n/2}}$ a $\left(\frac{HT}{AC}\right)^{n - 1}$, o $\frac{S^2}{R^{4 - n}}$ a $(1 + QQ)^{n - 1}$, el cuerpo se moverá en dicha curva en un medio uniforme con resistencia que sea como la potencia V^n de la velocidad. Pero volvamos a curvas más sencillas.



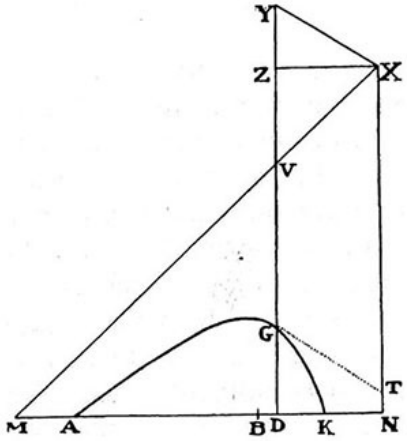
Puesto que el movimiento no ocurre en parábolas salvo en medios no resistentes, mientras que en las hipérbolas descritas aquí ocurre con resistencia continua, es evidente que la línea descrita por un proyectil en un medio uniformemente resistente se acerca más a estas hipérbolas que a una parábola. La línea es ciertamente de naturaleza hiperbólica, pero tal que en el vértice dista más de las asíntotas y en las partes lejanas del vértice se acerca más a las asíntotas de lo que se refleja en las hipérbolas aquí descritas. Sin embargo, la diferencia entre éstas y aquélla no es tan acusada como para impedir el uso adecuado de unas por otra en la práctica. Y quizá sean más útiles éstas que una hipérbola más exacta y a la vez más compleja. Por lo demás pueden utilizarse como sigue.

Complétese el paralelogramo HYGT, y la recta GT tocará a la hipérbola en G, y por tanto la densidad del medio en G es inversamente como la tangente GT, y la velocidad en ese punto como $\frac{GT^2}{GV}$, la resistencia, en cambio, respecto a la fuerza de la gravedad es como GT a $\frac{2nn + 2n}{n + 2}$ x GV.

Por tanto, si un cuerpo lanzado desde el punto A, según la línea AH describiere la hipérbola AGK, y prolongada AH cortase a la asíntota NX en H, y trazada AI paralela a la anterior cortase a la otra asíntota MX en I; la densidad del medio será en A inversamente como AH, y la velocidad del cuerpo como $\sqrt{\frac{AH^2}{AI}}$, y la resistencia respecto a la gravedad en ese punto como AH a $\frac{2nn + 2n}{n + 2}$ x AI. De donde se deducen las siguientes reglas.

Regla 1. Si se mantienen tanto la densidad del medio en A como la velocidad con que se lanza el cuerpo, y se cambia el ángulo NAH, se mantendrán las longitudes AH, AI, HX. Y por tanto, si dichas longitudes fuesen halladas en algún caso, podrá

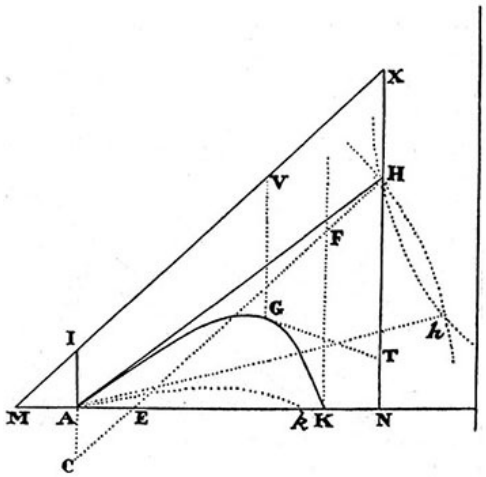
determinarse la hipérbola sin dificultad a partir de un ángulo NAH cualquiera dado.



Regla 2. Si se mantienen el ángulo NAH y la densidad del medio en A, pero se cambia la velocidad con que se proyecta al cuerpo, se mantendrá la longitud AH, y variará IA en razón inversa del cuadrado de la velocidad.

Regla 3. Si se mantienen el ángulo NAH, la velocidad del cuerpo en A y la gravedad aceleratriz, mientras se aumenta la proporción de la resistencia en A respecto a la gravedad motriz según una razón cualquiera, la proporción de AH a AI aumentará en la misma razón, pero se mantendrá el «latus rectum» de la parábola susodicha y la proporcionalidad a él de la longitud $\frac{AH^2}{AI}$; y por tanto, AH disminuye en esa proporción, mientras AI disminuye como el cuadrado de dicha razón. Pero la proporción de la resistencia al peso aumenta, tanto si a igual tamaño la gravedad específica es menor, como si la densidad del medio es mayor, o si la resistencia, a menor tamaño, disminuye en menor razón que el peso.

Regla 4. Puesto que la densidad del medio es mayor cerca del vértice de la hipérbola que en el punto A, para obtener la densidad media debe hallarse la razón de la menor de las tangentes GT a la tangente AH, y aumentar la densidad en A en una razón algo mayor que la de la semisuma de estas tangentes respecto a la menor de las tangentes GT.

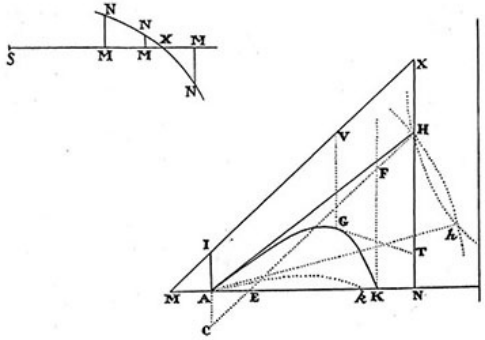


REGLA 5. Si se dan las longitudes AH, AI y hay que trazar la figura AGK; prolónguese HN hasta X, de modo que HX sea a AI como $n + 1$ a 1, y con centro en X y asíntotas MX, NX trácese la hipérbola por el punto A, de tal suerte que AI sea a cualquiera VG como XV^n a XI^n.

REGLA 6. Cuanto mayor es el número *n*, tanto más exactas son estas hipérbolas al ascender el cuerpo desde A y tanto menos exactas en su descenso hacia K; y viceversa. La hipérbola cónica goza de una razón media y es más simple que las otras. Por lo cual, si la hipérbola es de este tipo y hay que hallar el punto K, en el cual el cuerpo lanzado incide sobre una recta cualquiera AN que pasa por el punto A: la recta AN prolongada corte a las asíntotas MX, NX, en M y N, y tómese NK igual a AM.

REGLA 7. Y de aquí se sigue un método expeditivo para determinar esta hipérbola partiendo de fenómenos. Láncense dos cuerpos semejantes e iguales, con la misma velocidad, con ángulos distintos AHK, *h*A*k*, que caigan sobre el plano horizontal en K y *k*; y anótese la proporción de AK a A*k*. Sea ésta *d* a *e*. Levantando ahora una perpendicular AI de una longitud cualquiera, señálese en cualquier forma la longitud AH o A*h*, y desde esto, mediante la Regla 6, obténgase gráficamente la longitud de AK y de A*k*. Si la razón de AK a A*k* es la misma que la de *d* a *e*, la longitud AH se tomó correctamente. Pero si es menor, tómese sobre la recta indefinida SM la longitud SM igual a la longitud asumida AH y elévese la perpendicular MN igual a la diferencia de las razones $\frac{AK}{Ak} - \frac{d}{e}$ multiplicada por una recta dada cualquiera. Con método semejante, a partir de la asunción de varias longitudes AH se tendrán varios puntos N, y por todos ellos se hará pasar una curva regular NNXN que corta en X a la recta SMMM. Asúmase por fin AH igual a la abscisa SX; de nuevo determínese

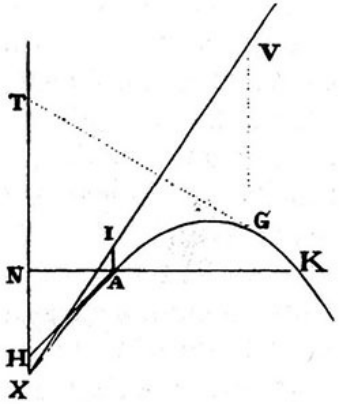
desde aquí la longitud AK; y las longitudes, que sean respecto a la asumida longitud AI y a esta última AH como la longitud AK conocida experimentalmente a la longitud AK últimamente hallada, serán las verdaderas longitudes AI y AH que era preciso hallar. Pero una vez dadas éstas, también estará dada la resistencia del medio en el lugar A, toda vez que ésta será a la fuerza de la gravedad como AH a 2AI. Y aumentando la densidad del medio según la Regla 4, si la resistencia así hallada aumenta en la misma proporción, será aún más exacta.



REGLA 8. Halladas las longitudes AH, HX; si ahora se busca la posición de la recta AH, según la cual, un proyectil lanzado con aquella velocidad, incidirá sobre un punto cualquiera K, elévense por los puntos A y K las rectas AC, KF perpendiculares al horizonte, de ellas AC en dirección hacia abajo y que sea igual a AI o a ½HX. Con asíntotas AK, KF trácese una hipérbola cuya conjugada pase por el punto C, y con centro en A y distancia AH trácese un círculo que corte a dicha hipérbola en el punto H; y el proyectil lanzado según la recta AH caerá en el punto K. Q. E. I. Porque el punto H, debido a la longitud dada AH, está ubicado en un punto del círculo descrito. Trácese CH de modo que encuentre a AK y KF en E y F respectivamente; y por ser paralelas CH y MX, y por ser iguales AC y AI, AE será igual a AM, y por tanto también a KN. Pero CE es a AE como FH a KN, y por tanto CE y FH son iguales. Luego el punto H cae sobre la hipérbola descrita con asíntotas AK, KF, cuya conjugada pasa por el punto C, y por lo tanto se halla en la intersección de esta hipérbola y el círculo descrito. Q. E. D. Pero hay que observar que esta operación resulta lo mismo, tanto si la recta AKN es paralela al horizonte, como si está inclinada respecto a él en algún ángulo; y que de las dos intersecciones H, *h*, surgen dos ángulos NAH, NA*h*; y que en la práctica mecánica basta con describir un solo círculo y después aplicar una línea indefinida CH sobre el punto C de modo que su sección comprendida entre el círculo y la recta FK sea igual a la parte CE situada entre el punto C y la recta AK.

Lo dicho sobre las hipérbolas se aplica fácilmente a las parábolas. Efectivamente, si

XAGK representa una parábola a la cual XV es tangente en el vértice X, y las ordenadas IA, VG fuesen como unas potencias cualesquiera XI^n, XV^n de las abscisas XI, XV; trácense XT, GT, AH, de las cuales XT sea paralela a VG, y GT, AH sean tangentes de la parábola en G y A: y un cuerpo lanzado desde un lugar cualquiera A, según la línea AH prolongada, y con la velocidad adecuada, describirá esta parábola, siempre que la densidad del medio en cada lugar G sea inversamente como la tangente GT. La velocidad en G será aquella con la cual se desplazaría el proyectil en un espacio no resistente en una parábola cónica de vértice G, diámetro prolongado hacia abajo VG, y «latus rectum» $\frac{2GT^2}{(nn - n) \times VG}$. Y la resistencia en G será a la fuerza de la gravedad como GT a $\frac{2nn - 2n}{n - 2}$ VG. De donde si NAK representa una línea horizontal y se mantienen tanto la densidad del medio en A como la velocidad con la cual se lanza al cuerpo, mientras varía el ángulo NAH, las longitudes AH, AI, HX, permanecerán, y de aquí se tiene dado el vértice X de la parábola, la posición de la recta XI, y, tomando VG a IA como XV^n a XI^n, se tienen dados los puntos todos G de la parábola por los cuales pasará el proyectil.

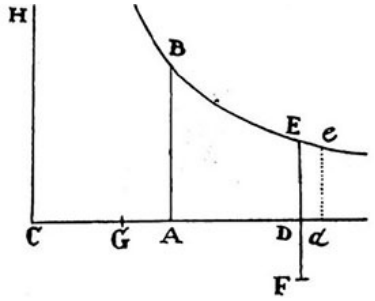


Sección III
SOBRE EL MOVIMIENTO DE LOS CUERPOS A LOS QUE SE RESISTE PARTE EN RAZÓN DE LAS VELOCIDADES Y PARTE EN RAZÓN DEL CUADRADO DE DICHA RAZÓN

Proposición XI. Teorema VIII

Si se resiste a un cuerpo parte en razón de su velocidad y parte en razón del cuadrado de su velocidad, y se mueve en un medio uniforme con su sola fuerza ínsita, mientras los tiempos se toman en progresión aritmética; las cantidades inversamente proporcionales a las velocidades, aumentadas en una cantidad dada, estarán en progresión geométrica.

Con centro en C y asíntotas rectangulares CAD*d* y CH, trácese la hipérbola BE*e*, y AB, DE, *de*, sean paralelas a la asíntota CH. Sobre la asíntota CD sean dados los puntos A, G: y si el tiempo se representa por el área hiperbólica uniformemente creciente ABED, digo que la velocidad se puede representar por la longitud DF, cuya inversa GD junto con la línea dada CG compone la longitud CD que crece en progresión geométrica.



Pues sea DE*ed* el área mínima correspondiente a un incremento mínimo de tiempo, y D*d* será inversamente como DE, y tanto directamente como CD. Pero el decremento

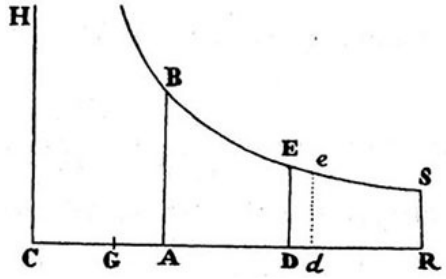
$\frac{1}{GD}$ del mismo, que (por el Lema II de este Libro) es $\frac{Dd}{GD^2}$, sera como $\frac{CD}{GD^2}$ o $\frac{CG + GD}{GD^2}$, esto es, como $\frac{1}{GD} + \frac{CG}{GD^2}$. Por tanto, al crecer uniformemente el tiempo ABED por la adición de las partículas dadas ED*de*, decrece $\frac{1}{GD}$ junto con la velocidad en la misma razón. Pues el decremento de la velocidad es como la resistencia, esto es (por hipótesis), como la suma de dos cantidades, una de las cuales es como la velocidad y la otra como el cuadrado de la velocidad; y el decremento del propio $\frac{1}{GD}$ es como la suma de las cantidades $\frac{1}{GD}$ y $\frac{CG}{GD^2}$, de las cuales la primera es la misma $\frac{1}{GD}$, y la otra, $\frac{CG}{GD}$, es como $\frac{1}{GD^2}$: por tanto, $\frac{1}{GD}$, por tener el mismo decremento, es como la velocidad. Y si la cantidad GD, inversamente proporcional a $\frac{1}{GD}$, es aumentada en una cantidad dada CG; la suma CD, creciendo uniformemente el tiempo ABED, crecerá en progresión geométrica. Q. E. D.

COROLARIO 1. Por lo tanto, si dados los puntos A, G, se representa el tiempo por el área hiperbólica ABED, la velocidad puede representarse por $\frac{1}{GD}$, inverso de GD.

COROLARIO 2. Y tomando GA a GD como el inverso de la velocidad al comienzo al inverso de la velocidad al final de un tiempo cualquiera ABED, se hallará el punto G. Hallado éste, se puede hallar la velocidad para cualquier otro tiempo dado.

PROPOSICIÓN XII. TEOREMA IX

Con los mismos supuestos, digo que si los espacios descritos se toman en progresión aritmética, las velocidades, aumentadas en una cantidad dada cualquiera, estarán en progresión geométrica.



Sobre la asíntota CD sea dado el punto R, y elevada la perpendicular RS que encuentre a la hipérbola en S, represéntese el espacio descrito mediante el área hiperbólica RSED; y la velocidad será entonces como la longitud GD, que con la cantidad dada CG compone la longitud CD decreciente en progresión geométrica, mientras el espacio RSED crece en progresión aritmética.

Puesto que, al estar dado el incremento ED*de* del espacio, el segmento D*d*, que es el decremento de la propia GD, será como el inverso de ED y por tanto directamente como CD, esto es como la suma del mismo GD y de la longitud dada CG. Pero el decremento de la velocidad, en el tiempo inversamente proporcional a ella, en el cual es descrita la partícula dada de espacio D*de*E, es una cantidad que es como la resistencia y el tiempo conjuntamente, esto es, directamente como la suma de dos cantidades, de las cuales la una es como la velocidad, la otra como el cuadrado de la velocidad, e inversamente como la velocidad; y por lo tanto, directamente como la suma de dos cantidades, de las cuales la una está dada y la otra es como la velocidad. Por consiguiente, tanto el decremento de la velocidad como el de la línea GD es como una cantidad dada y una cantidad decreciente conjuntamente, y al ser análogos los decrementos, las cantidades decrecientes siempre serán análogas; a saber, la velocidad y la línea GD. Q. E. D.

COROLARIO 1. Si la velocidad se representa por la línea GD, el espacio descrito será como el área hiperbólica DESR.

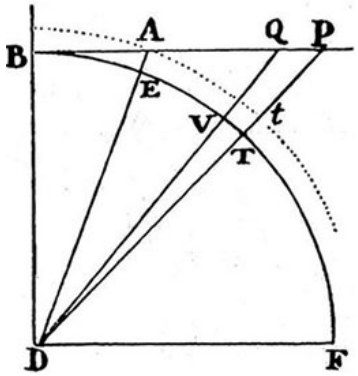
COROLARIO 2. Y si se supone en cualquier parte el punto R, se hallará el punto G tomando GR a GD como es la velocidad inicial a la velocidad tras recorrer un espacio cualquiera RSED. Hallado el punto G, se obtiene el espacio a partir de una velocidad dada y viceversa.

COROLARIO 3. De donde, puesto que (por la Proposición XI) se obtiene la velocidad a partir de un tiempo dado, y por esta Proposición se obtiene el espacio a partir de una velocidad dada; de un tiempo dado se obtendrá el espacio y viceversa.

PROPOSICIÓN XIII. TEOREMA X

Supuesto que un cuerpo, atraído hacia abajo por una gravedad uniforme, asciende o

desciende por una recta, y que sufre una resistencia en parte proporcional a la velocidad y en parte proporcional al cuadrado de la velocidad: digo que, si se trazan rectas paralelas a los diámetros de un círculo y de una hipérbola, las cuales pasen por los extremos de los diámetros conjugados, y las velocidades fuesen como ciertos segmentos de las paralelas trazados desde un punto dado; los tiempos serán como los sectores de las áreas separados por rectas trazados desde el centro hasta los extremos de los segmentos; y viceversa.

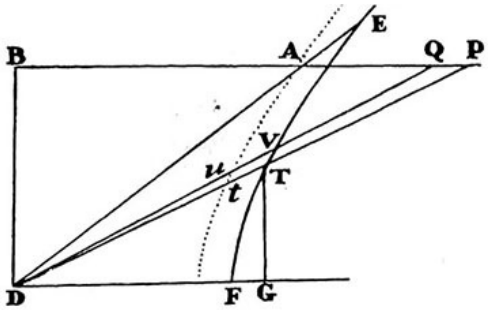


CASO 1. Supongamos primero que el cuerpo asciende, y con centro en D y un semidiámetro cualquiera DB, trácese el cuadrante del círculo BETF, y por el extremo B del semidiámetro DB trácese la recta indefinida BAP paralela al semidiámetro DF. Sobre ella sea dado el punto A, y tómese el segmento AP proporcional a la velocidad. Y puesto que una parte de la resistencia es como la velocidad y otra parte como el cuadrado de la velocidad, la resistencia total será como $AP^2 + 2BAP$. Únanse DA, DP cortando al círculo en E y T, y represéntese la gravedad por DA^2 de modo que la gravedad sea a la resistencia como DA^2 a $AP^2 + 2BAP$: y el tiempo de todo el ascenso será como el sector EDT del círculo.

Pues, trácese DVQ, separando el momento PQ de la velocidad AP y el momento DTV del sector DET correspondiente a un momento dado de tiempo; y este decremento PQ de la velocidad será como la suma de las fuerzas de la gravedad DA^2 y de la resistencia $AP^2 + 2BAP$, esto es (por la Proposición XII del Libro II de los Elementos), como DP^2. Luego el área DPQ, proporcional a PQ, es como DP^2, y el área DTV, que es al área DPQ como DT^2 a DP^2, es como DT^2 ya dada. Por tanto, el área EDT decrece uniformemente, al modo del tiempo futuro, mediante la pérdida de partículas DTV dadas, y por tanto es proporcional al tiempo de ascenso total. Q. E. D.

CASO 2. Si la velocidad en el ascenso del cuerpo se representa por la longitud AP como antes, y se supone la resistencia como $AP^2 + 2BAP$, y la gravedad fuese menor que la representable por DA^2, tómese entonces la longitud de BD tal que $AB^2 - BD^2$

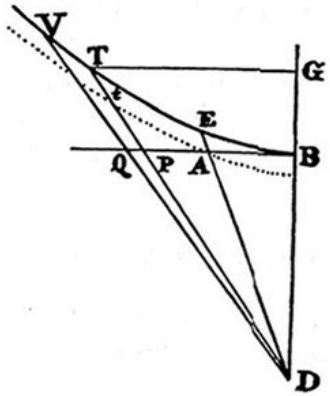
sea proporcional a la gravedad, y DF sea perpendicular e igual a DB, trazando por el vértice F la hipérbola FTVE, de cuyo semidiámetro sean conjugadas DB y DF, la cual hipérbola corta a DA en E, y a DP, DQ, en T y V; y el tiempo de todo el ascenso será como el sector TDE de la hipérbola.



Pues el decremento PQ de la velocidad, ocurrido en una partícula dada de tiempo, es como la suma de la resistencia $AP^2 + 2BAP$ y la gravedad $AB^2 - BD^2$, esto es, como $BP^2 - BD^2$. Pero el área DTV es al área DPQ como DT^2 a DP^2; y por tanto, si sobre DF cae una perpendicular GT, como GT^2 o $GD^2 - DF$ a BD^2, y como GD^2 a BP^2, y por partes, como DF^2 a $BP^2 - BD^2$. Por lo cual, dado que el área DPQ es como PQ, esto es, como $BP^2 - BD^2$, el área DTV será como la ya dada DF^2. Por consiguiente, el área EDT decrece uniformemente en cada partícula igual de tiempo por sustracción de otras tantas partículas dadas DTV, y por lo mismo es proporcional al tiempo. Q. E. D.

CASO 3. Sea AP la velocidad de descenso de un cuerpo, y $AP^2 + 2BAP$ la resistencia, y $BD^2 - AB^2$ la fuerza de la gravedad, siendo recto el ángulo DBA. Y si con centro en D, vértice principal F, se traza la hipérbola rectangular BETV que corte a las prolongaciones de DA, DP y DQ en E, T y V; el sector DET de esta hipérbola será como el tiempo total de descenso.

Pues el incremento PQ de la velocidad, y el área DPQ proporcional a él, es como el exceso de la gravedad sobre la resistencia, esto es, como $BD^2 - AB^2 - 2BAP - AP^2$, o $BD^2 - BP^2$. Y el área DTV es al área DPQ como DT^2 a DP^2, y por tanto, como GT^2 o $GD^2 - BD^2$ a BP^2 y como GD^2 a BD^2 y por partes como BD^2 a $BD^2 - BP^2$. Por lo tanto, dado que el área DPQ es como $BD^2 - BP^2$, el área DTV será como el ya dado BD^2. Crece, pues, el área EDT uniformemente en cada partícula igual de tiempo según se vaya añadiendo cada partícula dada DTV, y por tanto es proporcional al tiempo de descenso. Q. E. D.



COROLARIO. Si con centro en D y semidiámetro DA se traza por el vértice A el arco A*t*, semejante al arco ET, que subtienda similarmente al ángulo ADT: la velocidad AP será a la velocidad, que podría perder subiendo o adquirir bajando dicho cuerpo en el tiempo EDT en un espacio no resistente, como el área del triángulo DAP al área del sector DA*t*; y por tanto estará dada para un tiempo dado. Pues la velocidad, en un medio no resistente, es proporcional al tiempo y, por ello, a dicho sector; en un medio resistente es como el triángulo; y en ambos, cuando es muy pequeña, se acerca a la razón de igualdad, como ocurre con el sector y el triángulo.

ESCOLIO

El caso podría demostrarse lo mismo en el ascenso de un cuerpo, cuando la fuerza de la gravedad es menor que la que es representable por DA^2 o $AB^2 + BD^2$ y mayor que la representable por $AB^2 - BD^2$, y debe representarse por AB^2. Pero vayamos a otras cosas.

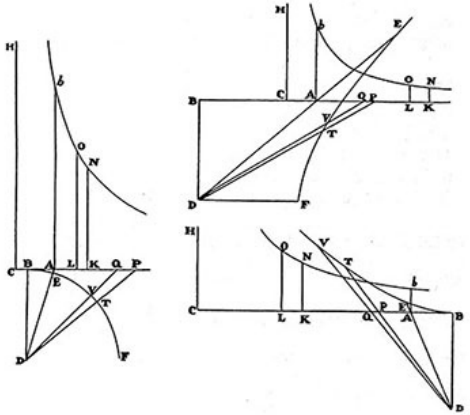
PROPOSICIÓN XIV. TEOREMA XI[5]

Con los mismos supuestos digo que el espacio descrito en el ascenso o descenso es como la diferencia del área con la que se representa el tiempo, y alguna otra área aumentada o disminuida en progresión aritmética, siempre que las fuerzas compuestas de la resistencia y la gravedad se tomen en progresión geométrica.

Tómese AC (en las tres figuras últimas) proporcional a la gravedad, y AK a la resistencia. Pero tómense hacia el mismo lado del punto A si el cuerpo desciende, y de no ser así en lados opuestos. Elévese A*b* de modo que sea a DB como DB^2 a

4BAC: V una vez trazada la hipérbola *b*N respecto a las asíntotas rectangulares CK, CH, y elevada KN perpendicular a CK, el área A*b*NK aumentará o disminuirá en progresión aritmética, siempre que las fuerzas CK se tomen en progresión geométrica. Por tanto digo que la distancia del cuerpo respecto a su máxima altitud es como el exceso del área A*b*NK sobre el área DET.

Pues, dado que la resistencia es como AK, esto es, como $AP^2 + 2BAP$, supuesta una cantidad dada cualquiera Z, supongamos que AK es igual a $\frac{AP^2 + 2BAP}{Z}$; y (por el Lema II de este Libro), el momento KL de AK será igual a $\frac{2APQ + 2BA \times PQ}{Z}$ o a $\frac{2BPQ}{Z}$, y el momento KLON del área A*b*NK será igual a $\frac{2BPQ \times LO}{Z}$ o a $\frac{BPQ \times BD^3}{2Z \times CK \times AB}$.



CASO 1. Ahora si el cuerpo está ascendiendo, sea la gravedad como $AB^2 + BD^2$ y siendo BET un círculo (en la primera figura), la línea AC que es proporcional a la gravedad, será $\frac{AB^2 + BD^2}{Z}$, y DP^2 o $AP^2 + 2BAP + AB^2 + BD^2$ será $AK \times Z + AC \times Z$ o también $CK \times Z$; y por tanto, el área DTV será al área DPQ como DT^2 o DB^2 a $CK \times Z$.

CASO 2. Pero si el cuerpo asciende y la gravedad es como $AB^2 - BD^2$, la línea AC (en

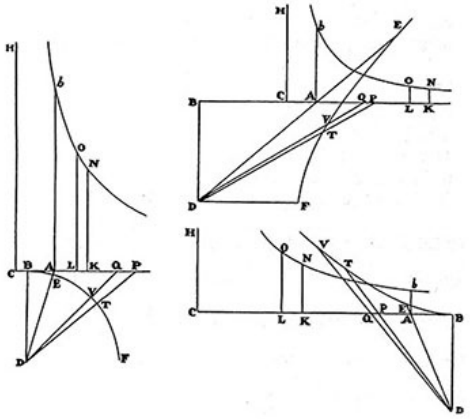
la segunda figura) será $\frac{AB^2 - BD^2}{Z}$, y DT^2 será a DP^2 como DF^2 o DB^2 a $BP^2 - BD^2$ o $AP^2 + 2BAP + AB^2 - BD^2$, esto es, como a AK x Z + AC x Z, o CK x Z. Y por tanto, el área DTV será al área DPQ como DB^2 a CK x Z.

CASO 3. Y, por la misma razón, si el cuerpo desciende, y por tanto la gravedad es como $BD^2 - AB^2$, y la línea AC (en la tercera figura) es igual a $\frac{BD^2 - AB^2}{Z}$, el área DTV será al área DPQ como DB^2 a CK x Z, como antes.

Ahora bien, al estar estas áreas siempre en dicha razón, si el; área DTV, mediante la cual se representa el momento de tiempo siempre igual a ella, es sustituida por un rectángulo cualquiera determinado, tal como BD x *m*, el área DPQ, esto es, ½BD x PQ, será a BD x *m* como CK x Z a BD^2. Y de aquí que PQ x BD^3 sea igual a 2BD x *m* x CK x Z, y el momento KLON del área A*b*NK, hallado antes, sea $\frac{BP \times BD \times m}{AB}$. Réstese del área DET el momento DTV, o sea, BD x *m*, y quedará $\frac{AP \times BD \times m}{AB}$. Luego, la diferencia de los momentos, esto es, el momento de la diferencia de áreas, es igual a $\frac{AP \times BD \times m}{AB}$, y por tanto, por estar dado $\frac{BD \times m}{AB}$, como la velocidad AP, esto es, como el momento de espacio descrito por el cuerpo en el ascenso o en el descenso. Y, en consecuencia, la diferencia de áreas y dicho espacio, al crecer o decrecer en momentos proporcionales que comienzan y se desvanecen a la vez, son proporcionales. Q. E. D.

COROLARIO. Si se llama M a la longitud que resulta de dividir el área DET por la línea BD, y se toma otra longitud V que esté en razón de la longitud M como la de la línea DA a la línea DE: el espacio descrito por un cuerpo en todo el ascenso o el descenso en un medio resistente será al espacio descrito por un cuerpo en el mismo tiempo cayendo desde un estado de reposo, como la diferencia de las áreas mencionadas a $\frac{BD \times V^2}{AB}$; y por tanto, para un tiempo dado, está dado. Pues, el espacio en un medio no resistente es como el cuadrado del tiempo, o como V^2; y, por estar dadas BD y AB, como $\frac{BD \times V^2}{AB}$. Esta área es igual al área $\frac{DA^2 + BD \times M^2}{DE^2 \times AB}$, y el momento de M

es *n*; y por tanto el momento de dicha área es $\frac{DA^2 \times BD \times 2M \times m}{DE^2 \times AB}$. Pero este momento es al momento de la diferencia de las áreas antedichas DET y A*b*NK, o sea, a $\frac{AP + BD \times m}{AB}$, como $\frac{DA^2 \times BD \times M}{DE^2}$ a ½BD x AP, o como $\frac{DA^2}{DE^2}$ x DET a DAP, por consiguiente, cuando las áreas DET y DAP son mínimas, están en razón de igualdad. Pues el área $\frac{BD \times V^2}{AB}$ y la diferencia de las áreas DET y A*b*NK, cuando todas estas áreas son mínimas, tienen iguales momentos, y por lo mismo, son iguales. De donde las velocidades, y por tanto también los espacios recorridos en ambos medios en el mismo tiempo tienden a ser iguales al principio del descenso o al final del ascenso, y por tanto son entonces entre sí como el área $\frac{BD \times V^2}{AB}$ y la diferencia de las áreas DET y A*b*NK; y además, puesto que el espacio en un medio no resistente es siempre como $\frac{BD \times V^2}{AB}$ mientras el espacio es un medio resistente es siempre como la diferencia de las áreas DET y A*b*NK: es preciso que los espacios en ambos medios, descritos en tiempos iguales cualesquiera, sean entre sí como dicha área $\frac{BD \times V^2}{AB}$ y la diferencia de las áreas DET y A*b*NK. Q. E. D.



ESCOLIO[6]

En los fluidos la resistencia de los cuerpos esféricos procede en parte de la

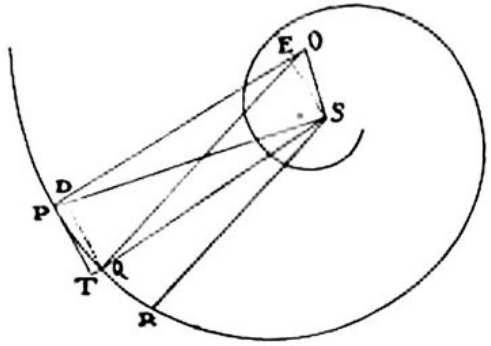
tenacidad, en parte del rozamiento y en parte de la densidad del medio. Y la parte de la resistencia procedente de la densidad del fluido es como dijimos, como el cuadrado de la velocidad; la parte procedente de la tenacidad del fluido es uniforme, o sea, como el momento del tiempo; por lo mismo se podría pasar a considerar el movimiento de los cuerpos a los cuales se resiste en parte con una fuerza uniforme, o sea, en razón de los momentos de tiempo, y parte en razón cuadrada de la velocidad. Pero basta haber abierto la puerta para esta especulación en las Proposiciones VIII y IX precedentes y sus Corolarios. Efectivamente, en dichas Proposiciones se puede sustituir la resistencia uniforme a un cuerpo ascendente debida a la gravedad por la resistencia uniforme procedente de la tenacidad del medio, cuando el cuerpo se mueve por la sola fuerza ínsita; y cuando el cuerpo asciende por una recta habrá que añadir esta resistencia uniforme a la fuerza de gravedad; mientras que habrá que restarla, cuando el cuerpo desciende por una recta. También se podría pasar al movimiento de cuerpos a los que se resiste en parte uniformemente, parte en razón de la velocidad y parte en tazón cuadrada de la velocidad. He abierto un camino en las Proposiciones precedentes XIII y XIV, en las que también puede sustituirse la resistencia uniforme debida a la tenacidad del medio en lugar de la fuerza de la gravedad, o componerse con ella como antes. Pero paso a otras cosas.

Sección IV
DEL MOVIMIENTO CIRCULAR DE LOS CUERPOS EN MEDIOS RESISTENTES

LEMA III

Sea PQR una espiral que corte todos los radios SP, SQ, SR, etc., en ángulos iguales. Trácese la línea recta PT tangente a la espiral en un punto cualquiera P y que corte al radio SQ en T; y, elevadas las perpendiculares PO, QO a la espiral que concurren en O, únase SO. Digo que si los puntos P y Q se aproximan hasta coincidir, el ángulo PSO resulta recto, y la razón última del rectángulo TQ x 2PS a PQ^2 será la razón de igualdad.

Puesto que al restar de los ángulos rectos OPQ, OQR, los ángulos iguales SPQ, SQR, quedarán los ángulos iguales OPS, OQS. Luego el círculo que pasa por los puntos O, S, P pasará también por el punto Q. Coincidan los puntos P y Q, y este círculo será tangente a la espiral en el lugar de coincidencia de PQ, y por tanto cortará perpendicularmente a la recta OP. Resultará, pues, OP diámetro de este círculo y el ángulo OSP recto sobre el semicírculo. Q. E. D.



Sobre OP trácense las perpendiculares QD, SE y las últimas razones de las líneas serán como sigue: TQ a PD como TS o PS a PE, o sea, 2PO a 2PS; y también PD a PQ como PQ a 2PO; y permutando, TQ a PQ como PQ a 2PS. De donde resulta que

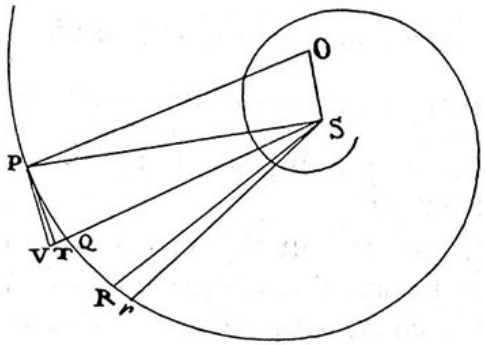
PQ^2 es igual a TQ x 2PS. Q. E. D.

Proposición xv. Teorema xii

Si la densidad de un medio en cada punto es inversamente como la distancia del punto a un centro inmóvil, y la fuerza centrípeta es como el cuadrado de la densidad: digo que un cuerpo puede girar en una espiral que corte en un ángulo dado a todos los radios trazados desde dicho centro.

Dense los supuestos del Lema anterior, y prolónguese SQ hasta V, de modo que SV sea igual a SP. En un tiempo cualquiera y en medio resistente, describa el cuerpo el arco mínimo PQ, y en doble tiempo el arco mínimo PR; y los decrementos de estos arcos procedentes de la resistencia, o las diferencias respecto a los arcos que describirían en los mismos tiempos en un medio no resistente, serán entre sí como los cuadrados de los tiempos en que se generan: por tanto, el decremento del arco PQ es la cuarta parte del decremento del arco PR. De donde también, si se toma el área PSQ igual al área QS*r*, el decremento del arco PQ será igual a la mitad del segmento R*r*; y por tanto, la fuerza de resistencia y la fuerza centrípeta son entre sí como los mismos segmentos ½R*r* y TQ que generan a la vez. Puesto que la fuerza centrípeta, con la que el cuerpo es urgido en P, es inversamente como SP^2, y (por el Lema x del Libro i) el segmento TQ, generado por dicha fuerza, está en razón compuesta de la razón de dicha fuerza y de la razón del cuadrado del tiempo en que se describe el arco PQ (pues en este caso desprecio la resistencia por ser infinitamente más pequeña que la fuerza centrípeta), TQ x SP^2, esto es (por el Lema anterior), ½PQ^2 x SP, será como el cuadrado del tiempo, y por tanto el tiempo es como PQ x $\sqrt{SP}$; y la velocidad del cuerpo con la que describe en dicho tiempo el arco PQ como $\frac{PQ}{PQ \times \sqrt{SP}}$ o $\frac{1}{\sqrt{SP}}$, esto es, como el inverso de la raíz cuadrada de SP. Y por la misma razón, la velocidad con la que se describe el arco QR es como el inverso de la raíz cuadrada del propio SQ. Pero dichos arcos PQ y QR son como las velocidades descriptoras entre ellas, es decir, como la raíz cuadrada de SQ a SP, o sea, como SQ a $\sqrt{SP \times SQ}$, y por ser iguales los ángulos SPQ, SQ*r* y las áreas PSQ, QS*r* también iguales, el arco PQ es al arco Q*r* como SQ a SP. Tómense las diferencias de consecuentes proporcionales y resultará el arco PQ al arco R*r* como SQ a SP - $\sqrt{SP \times SQ}$, es decir, ½VQ. Pues, al coincidir los puntos P y Q, la razón última de SP - $\sqrt{SP \times SQ}$ a ½VQ es la de igualdad. Dado que el decremento del arco PQ originado por la resistencia, o el doble del mismo R*r*, es como la resistencia y el cuadrado del tiempo conjuntamente, la

resistencia será como $\frac{Rr}{PQ^2 \times SP}$. Pero PQ era a R*r* como SQ a ½VQ, y de aquí que $\frac{Rr}{PQ^2 \times SP}$ sea como $\frac{\frac{1}{2}VQ}{PQ \times SP \times SQ}$ o como $\frac{\frac{1}{2}OS}{OP \times SP^2}$.



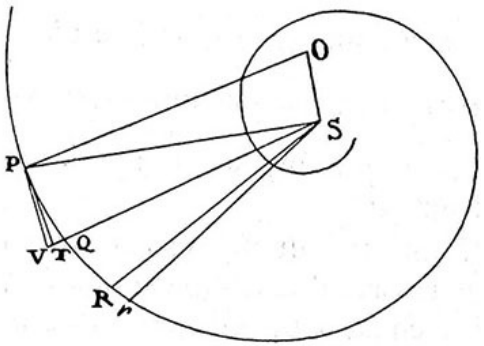
Pues cuando los puntos P y Q coinciden, coinciden también SP y SQ, y el ángulo PVQ se hace recto; y por ser semejantes los triángulos PVQ, PSO, resulta PQ a ½VQ como OP a ½OS. Luego $\frac{OS}{OP \times SP^2}$ es, entonces, como la resistencia, esto es, como la densidad del medio en P y como el cuadrado de la velocidad conjuntamente. Réstese la razón del cuadrado de la velocidad, o sea, $\frac{1}{SP}$ y quedará la densidad del medio en P como $\frac{OS}{OP \times SP}$. Dése, pues, una espiral, y por estar dada la razón de OS a OP, la densidad del medio en P será como $\frac{1}{SP}$. Por tanto, en un medio cuya densidad sea inversamente proporcional a la distancia al centro SP, un cuerpo puede girar en dicha espiral. Q. E. D.

Corolario 1. En un punto cualquiera P la velocidad será siempre aquella con la que un cuerpo, en un medio no resistente y la misma fuerza centrípeta, giraría en un círculo a la misma distancia SP del centro.

Corolario 2. Si la distancia SP está dada, la densidad del medio es como $\frac{OS}{OP}$, pero si no se da dicha distancia es como $\frac{OS}{OP \times SP}$. Y por ello una espiral puede

acomodarse a cualquiera densidad de medio.

COROLARIO 3. La fuerza de resistencia en un punto cualquiera P es a la fuerza centrípeta en el mismo punto como ½OS a OP. Pues tales fuerzas son entre sí como ½R*r* y TQ, o como $\frac{¼VQ \times PQ}{SQ}$ y $\frac{½PQ^2}{SP}$, es decir, como ½VQ y OP. Dada, pues, una espiral, está dada la proporción de la resistencia respecto a la fuerza centrípeta y, viceversa, dada dicha proporción, está dada la espiral.



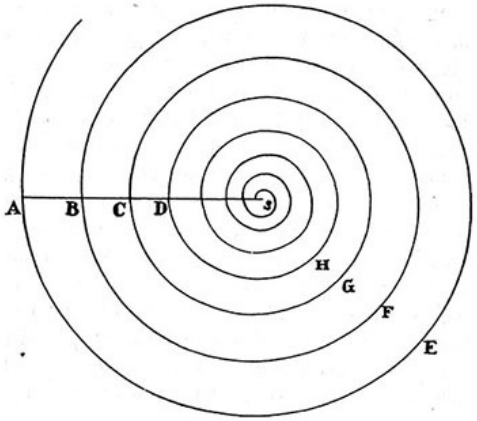
COROLARIO 4. Por consiguiente, el cuerpo no puede girar en esta espiral a no ser que la fuerza de resistencia sea menor que la mitad de la fuerza centrípeta. Sea la resistencia igual a la mitad de la fuerza centrípeta, y la espiral coincidirá con la línea recta PS, y el cuerpo descenderá hacia el centro por dicha recta con una velocidad que será a la velocidad con la que, según hemos probado antes para una parábola (Teorema X del Libro I) ocurre el descenso en un medio no resistente, como la raíz cuadrada de la razón de la unidad al número dos. Y los tiempos de descenso serán aquí como el inverso de las velocidades, y por tanto están dados.

COROLARIO 5. Y puesto que a distancias iguales del centro la velocidad es igual en la espiral PQR y en la recta SP, y la longitud de la espiral a la longitud de la recta se halla en una razón dada, a saber, en la razón de OP a OS, el tiempo de descenso por la espiral será al tiempo de descenso por la recta SP en esa misma razón dada y, por tanto, también dado.

COROLARIO 6. Si con centro en S y con dos distancias dadas cualesquiera se trazan dos círculos, y manteniendo estos círculos se hace cambiar en alguna forma el ángulo que forma la espiral con el radio PS, el número de revoluciones que puede completar un cuerpo en el interior de las circunferencias de dichos círculos caminando en espiral de una circunferencia a otra es como $\frac{PS}{OS}$, o como la tangente del ángulo que

forma la espiral con el radio PS, mientras el tiempo de dichas revoluciones será como $\frac{OP}{OS}$, es decir, como la secante del mismo ángulo, o inversamente proporcional a la densidad del medio.

COROLARIO 7. Si un cuerpo, en un medio con densidad inversamente proporcional a la distancia de cada punto al centro, gira por una curva cualquiera AEB en torno a dicho centro, y corta al primer radio AS en B con el mismo ángulo con que lo hiciese antes en A, y esto con una velocidad que sea a la velocidad anterior en A en razón inversa de la raíz cuadrada de las distancias al centro (esto es, como AS a la media proporcional entre AS y BS) dicho cuerpo describirá innumerables revoluciones semejantes BFC, CGD, etc. y con sus intersecciones dividirá el radio AS en partes continuamente proporcionales AS, BS, CS, DS, etc. Mas, los tiempos de las revoluciones serán directamente como los perímetros de las órbitas AEB, BFG, CGD, etc. e inversamente como las velocidades en A, B, C; esto es, como $AS^{3/2}$, $BS^{3/2}$, $CS^{3/2}$. Y el tiempo total empleado por el cuerpo para llegar al centro será al tiempo de la primera revolución como la suma de todas las continuamente proporcionales hasta el infinito $AS^{3/2}$, $BS^{3/2}$, $CS^{3/2}$ al primer término $AS^{3/2}$; esto es, como dicho primer término $AS^{3/2}$ a la diferencia de los dos primeros $AS^{3/2}$ - $BS^{3/2}$, o como $\frac{3}{2}$AS a AB, muy aproximadamente. De donde dicho tiempo total puede hallarse sin dificultad.



COROLARIO 8. De lo cual puede obtenerse con relativa facilidad el movimiento de cuerpos en medios de densidad uniforme o que se atenga a cualquiera otra ley determinada. Con centro en S, y distancias continuamente proporcionales SA, SB, SC, etc., trácense círculos cualesquiera, y establézcase que los tiempos de revolución entre los perímetros de dos cualesquiera de estos círculos en el medio del que hablé antes es al tiempo de revolución en el medio propuesto entre los mismos círculos, como la densidad media del medio propuesto entre los dos círculos a la densidad

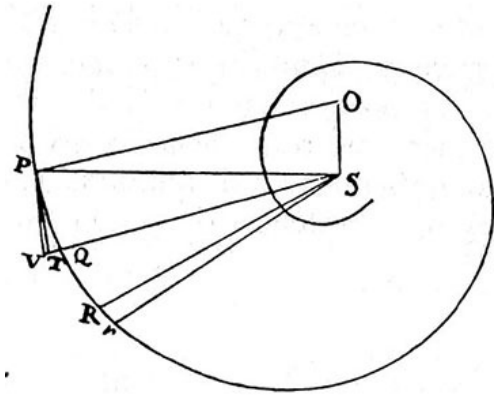
media del medio del que se trató más arriba entre los mismos círculos muy aproximadamente: y asimismo que la secante del ángulo con que la espiral antes determinada corta al radio AS en el medio arriba considerado está en la misma razón respecto a la secante del ángulo con que esta nueva espiral corta al mismo radio en el medio ahora propuesto; y también que los números totales de revoluciones entre dichos dos círculos son muy aproximadamente como las tangentes de sus ángulos. Si esto ocurre continuamente entre cada par de círculos, el movimiento se continuará a través de todos los círculos. Y de este modo no será difícil imaginar cómo y en qué tiempos deberán girar los cuerpos en cualquier medio regular.

COROLARIO 9. Y aunque los movimientos excéntricos ocurran en espirales próximas a la forma oval, no obstante, si imaginamos cada revolución de dichas espirales distante de las siguientes con los mismos intervalos y acercándose al centro en los mismos grados que la espiral anteriormente descrita, comprenderemos también cómo ocurren los movimientos de los cuerpos en semejantes espirales.

PROPOSICIÓN XVI. TEOREMA XIII

Si la densidad de un medio es en cada punto inversamente proporcional a la distancia de esos puntos a un centro inmóvil, y la fuerza centrípeta es inversamente proporcional a una potencia cualquiera de dicha distancia: digo que un cuerpo puede girar en una espiral que corta a todos los radios trazados desde dicho centro con un ángulo dado.

Se demuestra del mismo modo que la Proposición anterior. Pues si la fuerza centrípeta es en P inversamente proporcional a cualquiera potencia SP^{n+1} de la distancia SP, cuyo índice es $n + 1$, se deduce, como más arriba, que el tiempo en el cual el cuerpo describe un arco PQ será como PQ x $PS^{½n}$ y la resistencia en P como $\frac{Rr}{PQ^2 \times SP^n}$, o también, como $\frac{(1 - ½n) \times VQ}{PQ \times SP^n \times SQ}$ y, por tanto, como $\frac{(1 - ½n) \times OS}{OP \times SP^{n+1}}$, esto es, por estar dado $\frac{(1 - ½n) \times OS}{OP}$, inversamente como SP^{n+1}. Y, por lo mismo, puesto que la velocidad es inversamente como $SP^{½n}$, la densidad en P será inversamente proporcional a SP.



COROLARIO 1. La resistencia es a la fuerza centrípeta como (1 - ½n) x OS a OP.

COROLARIO 2. Si la fuerza centrípeta fuese como SP^3, entonces (1 - ½n) = 0; y, por tanto, la resistencia y la densidad del medio serán nulas, como en la Proposición IX del Libro I.

COROLARIO 3. Si la fuerza centrípeta es inversamente como una potencia cualquiera del radio SP cuyo índice sea mayor que 3, la resistencia positiva se transforma en negativa.

ESCOLIO

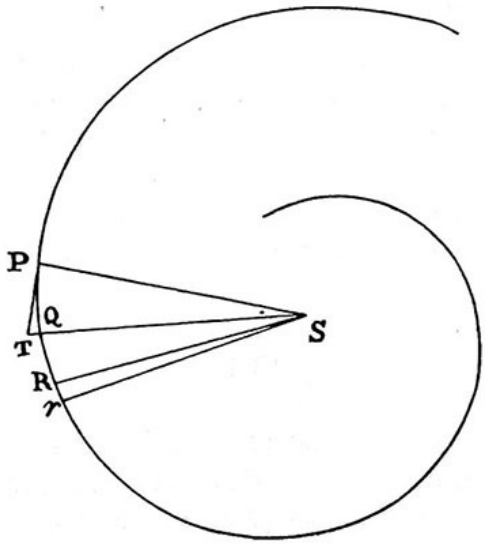
Por lo demás, esta Proposición y las anteriores, que se refieren a medios con densidades desiguales, deben entenderse referidas al movimiento de cuerpos tan pequeños que la densidad del medio mayor hacia un lado del cuerpo que hacia el otro, no venga a ser digna de consideración. También supongo que la resistencia, «caeteris paribus», es proporcional a la densidad. Por lo cual, para medios cuya fuerza de resistencia no sea como la densidad, deberá la densidad aumentarse o disminuirse de modo que quede suprimido el exceso de resistencia o suplida su falta.

PROPOSICIÓN XVII. PROBLEMA IV

Hallar tanto la fuerza centrípeta como la resistencia del medio que permitan a un cuerpo, dada la ley de la velocidad, girar en una espiral dada.

Sea dicha espiral PQR. De la velocidad con la cual el cuerpo recorre el arco mínimo PQ se tendrá dado el tiempo, y de la altura TQ, que es como la fuerza centrípeta y el cuadrado del tiempo, se tendrá dada la fuerza. Después, de la diferencia RSr de áreas PSQ, QSR, completadas en partículas iguales de tiempo, se

obtendrá la retardación del cuerpo, y a partir de ésta se hallarán la resistencia y la densidad del medio.



Proposición xviii. Problema v

Dada la ley de la fuerza centrípeta, hallar la densidad del medio en cada punto, con la cual describa un cuerpo una espiral dada.

Hay que hallar la velocidad en cada punto a partir de la fuerza centrípeta, y después, hay que hallar la densidad del medio a partir de la retardación de la velocidad, como en la Proposición anterior.

Pero ya indiqué el método para tratar estos problemas en la Proposición décima de este Libro y en el Lema segundo; y no deseo entretener más al lector en estas abstrusas disquisiciones. Ahora habrá que añadir algo sobre las fuerzas de los cuerpos que se desplazan y sobre la densidad y resistencia de los medios en los cuales ocurren los movimientos vistos hasta ahora y otros similares a ellos.

Sección v
SOBRE LA DENSIDAD Y COMPRESIÓN DE LOS FLUIDOS, Y SOBRE HIDROSTÁTICA

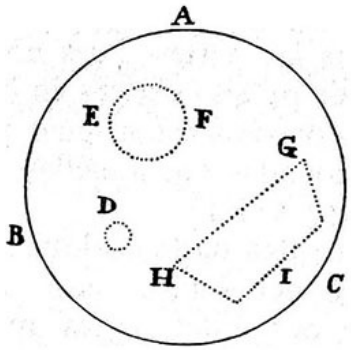
Definición de fluido

Fluido es todo cuerpo cuyas partes ceden a la aplicación de cualquier fuerza, y, al ceder, se mueven entre sí con facilidad.

Proposición xix. Teorema xiv

Todas las partes de un fluido homogéneo e inmóvil encerrado en un vaso cualquiera inmóvil y comprimido por todos lados (sin tomar en cuenta ni la condensación ni la gravedad ni fuerza centrípeta alguna) soportan igual presión por todas partes y permanecen en sus lugares sin que dicha presión origine movimiento alguno.

Caso 1. Introdúzcase un fluido en el vaso esférico ABC y comprímase uniformemente por todas partes: digo que ninguna parte del mismo se moverá debido a la tal presión. Pues, si una parte D se moviese, sería preciso que todas las partes del mismo situadas a la misma distancia del centro se moviesen a la vez con un movimiento similar, y esto por ser semejante e igual la presión de todas, y se supone excluido todo otro movimiento, salvo el procedente de la presión. Pero no pueden acercarse más hacia el centro, salvo que el fluido sea condensado hacia el centro; contra la hipótesis. Tampoco pueden alejarse de él más, a no ser que el Huido se condense hacia la circunferencia; también contra la hipótesis. Tampoco pueden, manteniendo su distancia al centro, moverse en dirección alguna, porque, por la misma razón, se moverían en dirección contraria; pues la misma parte no puede moverse en direcciones opuestas en el mismo momento. Luego ninguna parte del fluido se moverá de su lugar. Q. E. D.



Caso 2. Ahora digo que todas las partes esféricas de este fluido soportan presiones iguales por todos lados. Pues, sea EF una parte esférica del fluido, y si no recibiese igual presión por todas partes, auméntese la presión menor hasta que resulte presionada igual por todas partes; y sus partes, por el Caso primero, permanecerán en su lugar. Pero antes de aumentar la presión, también permanecían en su lugar, por el mismo primer Caso, y con la presión añadida se moverían de su lugar, por la definición de fluido. Pero estas dos afirmaciones son contradictorias. Luego era falsa la afirmación de que la esfera EF no padecía igual presión por todos lados. Q. E. D.

Caso 3. Además digo que la presión de las diversas partes esféricas es igual. Pues las partes esféricas contiguas se presionan mutuamente e igualmente en los puntos de contacto, por la Ley iii del Movimiento. Pero, por el Caso segundo, soportan igual fuerza de presión por todos lados. Luego dos partes esféricas no contiguas soportarán la misma fuerza, toda vez que una parte esférica intermedia puede tocar a ambas. Q. E. D.

Caso 4. Ahora digo que todas las partes de un fluido soportan igual presión en todas las partes. Pues dos partes cualesquiera pueden ser tocadas por partes esféricas en cualquier punto, y ahí dichas partes esféricas presionan por igual, por el Caso tercero, y a la inversa son presionadas por aquéllas por igual, por la Ley tercera del Movimiento. Q. E. D.

Caso 5. Puesto que una parte cualquiera GHI del fluido está encerrada en el resto del fluido como en un recipiente y es presionada por igual por todas partes, mientras las partes del mismo se presionan entre sí por igual y están entre sí en reposo; es evidente que las partes de cualquier GHI de un fluido, al estar igualmente presionadas por todos lados, se presionan entre sí por igual y reposan entre sí. Q. E. D.

Caso 6. Por tanto, si dicho fluido está encerrado en un vaso no rígido, y no fuese presionado por igual por todos lados, cederá ante la presión más fuerte, por la definición de fluidez.

Caso 7. Y por lo mismo, un fluido en un vaso rígido no soportará una presión más fuerte por un lado que por otro, sino que cederá ante ella, y esto

instantáneamente, porque el lado rígido del vaso no sigue al líquido que cede. Pero al ceder presionará sobre el lado opuesto, y así la presión tiende a igualarse por doquier. Y puesto que el fluido trata de apartarse instantáneamente de la parte más presionada, mientras es soportado por la resistencia del vaso en el lado contrario, la presión se restablece instantáneamente por todos lados por igual sin movimiento local, con lo que las partes del fluido, por el Caso quintó, se presionan entre sí por igual y reposan entre sí. Q. E. D.

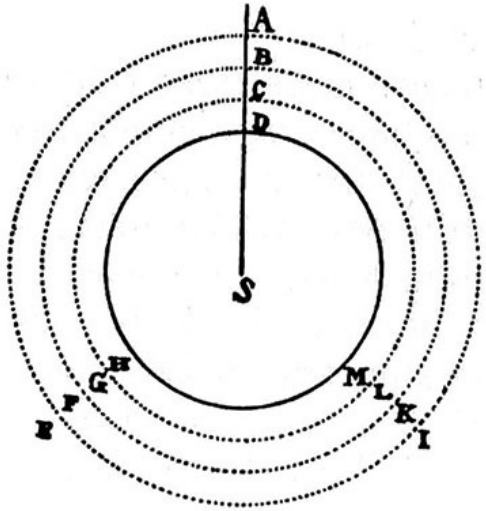
Corolario. De donde, tampoco podrá cambiarse el movimiento de las partes del fluido entre sí, mediante una presión efectuada sobre la superficie externa del fluido, salvo en la medida en que la superficie cambie de forma en algún punto, o en que las partes del fluido oprimiéndose entre ellas más o menos intensamente se deslicen entre ellas con mayor o menor facilidad.

Proposición xx. Teorema xv

Si cada parte de un fluido esférico, homogéneo a distancias iguales del centro, que descansa sobre un fondo esférico concéntrico, gravita hacia el centro del todo, el fondo soportará el peso de un cilindro cuya base fuese igual a la superficie del fondo y cuya altura fuese igual a la altura del fluido sobrepuesto.

Sea DHM la superficie del fondo, y AEI la superficie superior del fluido. Divídase el fluido en esferas concéntricas de igual grosor, mediante innumerables superficies esféricas BFK, CGL; imagínese que la fuerza de la gravedad actúe únicamente sobre la superficie exterior de cada esfera y que las acciones son iguales sobre cada parte igual de todas las superficies. Entonces, la superficie más externa AE sufre la sola presión de su fuerza de gravedad, que presiona no sólo sobre las partes todas de la superficie superior sino que también (por la Proposición xix) presiona igualmente en razón de su extensión sobre la superficie segunda BFK. Esta segunda superficie BFK sufre la presión de su propia fuerza de gravedad, además de la presión añadida de la primera, lo que duplica la presión. La tercera superficie CGL padece una presión triple, esto es, la de su propia fuerza de gravedad junto con la presión de la anterior que padece en función de su propia extensión. Y de igual modo, la superficie cuarta padece una presión cuádruple, la quinta quíntuple, y así sucesivamente. Por consiguiente, la presión que padece cada superficie, no es como el volumen de fluido que reposa encima de ella, sino como el número de esferas que hay hasta la superficie superior del fluido; y es igual a la gravedad del orbe inferior multiplicada por el número de orbes: esto es, a la gravedad de un sólido que tiene razón última de igualdad con el cilindro susodicho (siempre que aumente el número de orbes y disminuya su grosor infinitamente, de manera que la acción de la gravedad desde la superficie inferior hasta la superior se haga continua). Luego la superficie

inferior soporta el peso del cilindro definido arriba. Q. E. D. Y, por la misma razón, será evidente la proposición cuando la gravedad decrece en una razón cualquiera dada desde el centro, lo mismo que cuando el fluido es más raro arriba y más denso abajo. Q. E. D.



COROLARIO 1. Por tanto, el fondo no soporta todo el peso del fluido que hay encima, sino sólo aquella parte del peso que se describe en la Proposición; el peso restante es soportado por la forma abovedada del fluido.

COROLARIO 2. A distancias iguales del centro la cantidad de presión es siempre la misma, tanto si la presión sobre la superficie es paralela al horizonte como si es perpendicular u oblicua; tanto si el fluido, rebosando por encima de la superficie presionada, asciende perpendicularmente en línea recta como si serpea oblicuamente por sinuosas cavidades y canales, sean regulares o irregulares, anchos o angostos. Se infiere que la presión en nada cambia por estas circunstancias al aplicar la demostración de este Teorema a cada uno de los casos de fluidos.

COROLARIO 3. Por la misma demostración se infiere también (por la Proposición XIX) que las partes de un fluido pesado no adquieren, por la presión del peso que soportan, ningún movimiento entre ellas; si se excluye el movimiento debido a la condensación.

COROLARIO 4. Y, por lo tanto, si otro cuerpo de la misma gravedad específica, que no admita condensación, se sumergiese en este fluido, tampoco adquirirá movimiento alguno por causa del peso que incide sobre él: no descenderá, no ascenderá ni se verá obligado a cambiar su forma. Si es esférico permanecerá esférico, pese a la presión; si es -cuadrado permanecerá cuadrado: y esto tanto si es blando como si es muy fluido, tanto si flota libremente en el fluido como si reposa en el fondo. Pues toda parte

perteneciente al fluido posee la misma razón que el cuerpo sumergido, y la razón de todos los cuerpos sumergidos, que tengan la misma figura y gravedad específica siendo de la misma magnitud, es la misma también. Si un cuerpo sumergido, conservando su peso, se licuase y tomase la forma del fluido, tal cuerpo, si antes ascendía o descendía o tomaba una forma nueva debido a la presión, también ahora ascenderá o descenderá o habrá de tomar una forma nueva: y esto precisamente porque permanecen su gravedad y las demás causas de los movimientos. Pero (por el Caso 5 de la Proposición XIX) ahora reposaría y conservaría su figura. Luego también antes.

COROLARIO 5. Por consiguiente, el cuerpo específicamente más grave que el fluido contiguo a él se hundirá y el específicamente más leve ascenderá, y se seguirá el movimiento y cambio de figura que puedan deberse al exceso o defecto de gravedad susodichos. Pues tales exceso o defecto son la razón del impulso que urge al cuerpo, que de otro modo está en equilibrio con las partes del fluido; y puede compararse con el exceso o defecto de peso en uno de los dos platillos de la balanza.

COROLARIO 6. Por tanto, hay una doble gravedad en los cuerpos ubicados en un fluido; una verdadera y absoluta, otra aparente, común y relativa. Gravedad absoluta es la fuerza total con la que el cuerpo tiende hacia abajo; relativa y común es el exceso de gravedad con el cual el cuerpo tiende hacia abajo más que el fluido que lo rodea. Con el primer tipo de gravedad las partes de los fluidos y de todos los cuerpos gravitan hacia los lugares que ocupan; y por ello componen, con sus pesos juntos, el peso del todo. Pues el todo completo es grave, como se puede ver en los vasos llenos de líquido; y el peso del todo es igual a los pesos de todas las partes y, por tanto, se compone de ellas. Con el otro tipo de gravedad los cuerpos no gravitan hacia sus lugares, esto es, al concurrir no se sobrecargan, sino que impidiendo los mutuos intentos de descenso reposan en sus lugares. Las cosas que están en el aire y no sobrecargan comúnmente no se juzgan pesadas. Las que sobrecargan sí lo son, en la medida en que no son sostenidas por el peso del aire. Para el común de la gente los pesos no son sino los excesos de peso verdadero por encima del peso del aire. De aquí que vulgarmente se llamen leves a las cosas que son menos graves y, al ceder ante la sobrecarga del aire, se elevan hacia arriba. Son relativamente leves, pero no absolutamente, toda vez que en el vacío descienden. Así como en el agua los cuerpos, que por su mayor o menor gravedad descienden o ascienden, son relativa y aparentemente graves o leves, y su gravedad o levedad relativa y aparente es el exceso o defecto con el que su verdadera gravedad o supera a la gravedad del agua o es superada por ella. Pero las cosas que no descienden sobrecargando ni ascienden cediendo a una sobrecarga, pese a que con sus pesos verdaderos aumenten el peso del todo, no gravitan en el agua relativamente y para el sentir común. Pues la demostración de estos casos es similar.

COROLARIO 7. Lo demostrado para la gravedad, tiene valor para otras fuerzas centrípetas cualesquiera.

COROLARIO 8. Por consiguiente, si el medio, en el cual se mueve un cuerpo cualquiera, estuviese sometido a la acción de su propia gravedad o de otra fuerza centrípeta cualquiera, y el cuerpo padece esta misma acción más fuertemente, la diferencia de fuerzas es aquella fuerza motriz que hemos considerado en las Proposiciones anteriores como fuerza centrípeta. Si, en cambio, el cuerpo sufre esa acción más levemente, la diferencia de fuerzas debe considerarse como fuerza centrífuga.

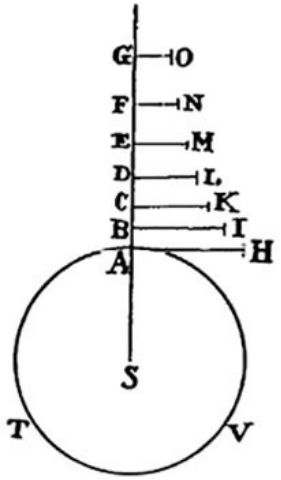
COROLARIO 9. Pero como los fluidos, al presionar sobre los cuerpos inmersos en ellos, no cambian sus figuras externas, también se sigue (por el Corolario de la Proposición XIX) que no cambiarán la ubicación de las partes internas entre sí: y por ende, si se sumergen animales, y toda sensación procede del movimiento de las partes, ni sufrirán daño los cuerpos sumergidos ni tendrán excitación sensorial alguna, salvo en la medida en que estos cuerpos pudieran ser objeto de condensación por compresión. Y lo mismo ocurre con cualquier sistema de cuerpos contenido en un fluido que lo comprima. Todas las partes del sistema sufrirán los mismos movimientos que si se hallasen en el vacío y retuviesen únicamente su gravedad, salvo en la medida en que el fluido o bien resista algo a sus movimientos o bien las obligue con su compresión a aglutinarse.

PROPOSICIÓN XXI. TEOREMA XVI

Sea la densidad de un fluido cualquiera proporcional a la compresión y sean sus partes atraídas hacia abajo por una fuerza centrípeta inversamente proporcional a las distancias de esas partes al centro: y digo que, si tales distancias se toman continuamente proporcionales, las densidades del fluido a esas mismas distancias también serán continuamente proporcionales.

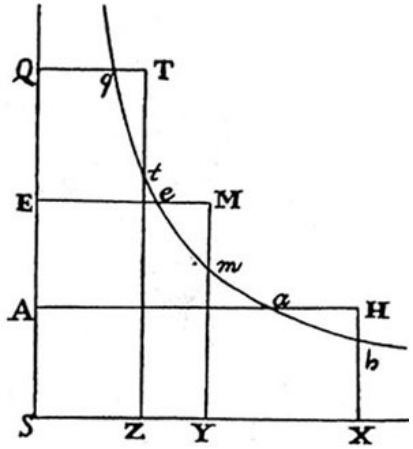
Represente ATV el fondo esférico sobre el que descansa el fluido, S el centro, SA, SB, SD, SE, SF, etc., distancias continuamente proporcionales. Elévense las perpendiculares AH, BI, CK, DL, EM, FN, etc., que sean como las densidades del medio en los lugares A, B, C, D, E, F; y las gravedades específicas en esos puntos serán como $\frac{AH}{AS}$, $\frac{BI}{BS}$, $\frac{CK}{CS}$, etc., o, lo que es lo mismo, como $\frac{AH}{AB}$, $\frac{BI}{BC}$, $\frac{CK}{CD}$, etc. Imaginemos, en primer lugar, que estas gravedades son uniformemente continuas de A a B, de B a C, de C a D, etc., ocurriendo los decrementos por grados en los puntos B, C, D, etc. Multiplicadas estas gravedades por las alturas AB, BC, CD, etc., se obtendrán las presiones AH, BI, CK, etc., con las cuales está urgido el fondo ATV (según el Teorema XV). Por tanto, la partícula A soporta todas las presiones AH, BI, CK, DL, y así sucesivamente hasta el infinito; y la partícula B soportará todas menos

la primera AH; la partícula C soportará a todas menos las dos primeras AH, BI; y así sucesivamente: y en consecuencia, la densidad AH de la partícula primera A es a la densidad BI de la segunda partícula B como la suma de todos los AH + BI + CK + DL, hasta el infinito, a la suma de todos los BI + CK + DL, etc. Y la densidad BI de la segunda partícula B es a la densidad CK de la tercera C como la suma de BI + CK + DL, etc., a la suma de CK + DL, etc. Por tanto, dichas sumas son proporcionales a sus diferencias AH, BI, CK, etc., y por ello continuamente proporcionales (por el Lema I de este libro) y, por tanto, las diferencias AH, BI, CK, etc., proporcionales a las sumas, son también continuamente proporcionales. Por lo cual, al ser las densidades en los puntos A, B, C, etc., como AH, BI, CK, etc., éstas serán también continuamente proporcionales. Procédase por tramos, y por corresponder a las distancias SA, SC, SE, continuamente proporcionales, serán también las densidades AH, CK, EM, continuamente proporcionales. Y por la misma razón, para distancias cualesquiera SA, SD, SG, continuamente proporcionales, las densidades AH, DL, GO, serán continuamente proporcionales. Júntense ahora los puntos A, B, C, D, E, etc., de modo que la progresión de las gravedades específicas desde el fondo A hasta la cima del fluido resulte continua, y para distancias cualesquiera SA, SD, SG, continuamente proporcionales, las densidades AH, DL, GO, al ser siempre continuamente proporcionales, permanecerán también ahora continuamente proporcionales. Q. E. D.



COROLARIO. De aquí que si se tiene dada la densidad en dos puntos, como A y E, puede inferirse la densidad del fluido en otro punto cualquiera Q. Trácese, con centro en S y asíntotas rectas SQ, SX, una hipérbola que corte a las perpendiculares AH, EM, QT, en *a*, *e*, *q*, lo mismo que a las perpendiculares HX, MY, TZ, trazadas sobre la asíntota SX, en *h*, *m*, *t*. Sea el área Y*mt*Z respecto al área dada Y*mh*X como el área

dada E*eq*Q al área dada E*ea*A; y la línea Z*t* prolongada cortará a la línea QT proporcional a la densidad. Pues si las líneas SA, SE, SQ, son continuamente proporcionales, las áreas E*eq*Q, E*ea*A serán iguales y, por tanto, las áreas proporcionales a ellas, Y*mt*Z, X*hm*Y, serán iguales también, y las líneas SX, SY, SZ, esto es, AH, EM, QT, necesariamente serán continuamente proporcionales. Y si las líneas SA, SE, SQ, adquieren otro orden cualquiera en la serie de proporciones continuas, las líneas AH, EM, QT, por la proporcionalidad de las áreas hiperbólicas, adquirirán el mismo orden en otra serie de proporcionalidades continuas.

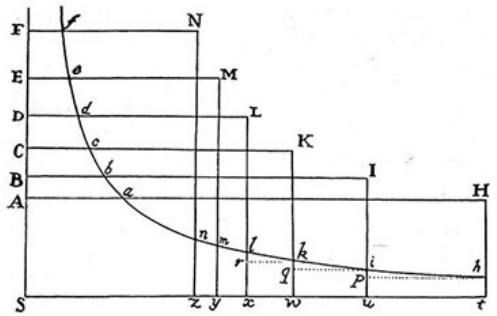


PROPOSICIÓN XXII. TEOREMA XVII

Sea la densidad de un fluido cualquiera proporcional a la compresión y sus partes sean atraídas hacia ahajo por una gravedad inversamente proporcional a los cuadrados de sus distancias al centro: digo que si las distancias se toman en proporción armónica, las densidades del fluido a esas distancias estarán en progresión geométrica.

Sea S el centro y SA, SB, SC, SD, SE las distancias en progresión geométrica. Elévense las perpendiculares AH, BI, CK, etc., que sean como las densidades de los fluidos en los puntos A, B, C, D, E, etc., y sus gravedades específicas en esos puntos serán como $\frac{AH}{SA^2}$, $\frac{BI}{SB^2}$, $\frac{CK}{SC^2}$, etc. Supongamos que estas gravedades son uniformemente continuas, la primera desde A hasta B, la segunda desde B hasta C, la tercera desde C hasta D, etc. Y éstas multiplicadas por las alturas AB, BC, CD, DE, etc., o, lo que es lo mismo, por las distancias SA, SB, SC, etc., proporcionales a

dichas alturas, darán las representaciones de las presiones $\frac{AH}{SA}$, $\frac{BI}{SB}$, $\frac{CK}{SC}$, etc. Por lo cual, ya que las densidades son como las sumas de estas presiones, las diferencias de las densidades AH - BI, BI - CK, etc., serán como las diferencias de las sumas $\frac{AH}{SA}$, $\frac{BI}{SB}$, $\frac{CK}{SC}$, etc. Con centro en S y asíntotas SA, S*x*, trácese una hipérbola que corte a las perpendiculares AH, BI, CK, etc., en *a*, *b*, *c*, etc., así como a H*t*, I*u*, K*w*, perpendiculares trazadas sobre la asíntota S*x*, en *h*, *i*, *k*; y las diferencias de densidades *tu*, *uw*, etc., serán como, $\frac{AH \times th}{SA}$, $\frac{BI \times ui}{SB}$, etc. Y los rectángulos *tu* x *th*, *uw* x *ui*, etc., o *tp*, *uq*, etc., serán como $\frac{AH \times th}{SA}$, $\frac{BI \times ui}{SB}$, etc., esto es, como A*a*, B*b*, etc. Pues, por la naturaleza de la hipérbola, SA es a AH o S*t* como *th* a A*a*, y por tanto $\frac{AH \times th}{SA}$ es igual a A*a*. Y por igual razonamiento $\frac{BI \times ui}{SB}$ es igual a B*b*, etc. Pero A*a*, B*b*, C*c*, etc., son continuamente proporcionales y, por tanto, proporcionales a sus diferencias A*a* - B*b*, B*b* - C*c*, etc.; y por ello, los rectángulos *tp*, *uq*, etcétera, son proporcionales a las susodichas diferencias; al igual que las sumas de los rectángulos *tp* + *uq* o *tp* + *uq* + *wr* lo son a las sumas de las diferencias A*a* - C*c* o A*a* - D*d*. Si se suponen muchos términos de éstos, la suma de todas las diferencias, pongamos A*a* - F*f*, será proporcional a la suma de todos los rectángulos, pongamos *zthn*. Auméntese el número de términos y disminúyanse las distancias de los puntos A. B, C etc., hasta el infinito, y dichos rectángulos resultarán iguales al área hiperbólica *zthn* y, por tanto, la diferencia A*a* - F*f* es proporcional a esta área. Ahora tómense unas distancias cualesquiera, como SA, SD, SF en progresión armónica, y las diferencias A*a* - D*d*, D*d* - F*f* serán iguales; y por ello, las áreas *thlx*, *xlnz* proporcionales a dichas diferencias, serán iguales entre sí, y las densidades S*t*, S*x*, S*z*, esto es, AH, DL, FN, continuamente proporcionales. Q. E. D.



COROLARIO. De aquí que, si se dan dos densidades cualesquiera, AH y BI, de un fluido, estará dada el área *thiu*, correspondiente a su diferencia *tu*; y de ahí se hallará la densidad FN para cualquier altura SF, tomando el área *thnz* respecto al área susodicha dada *thiu* como es la diferencia de A*a* - F*f* a la diferencia de A*a* - B*b*.

ESCOLIO

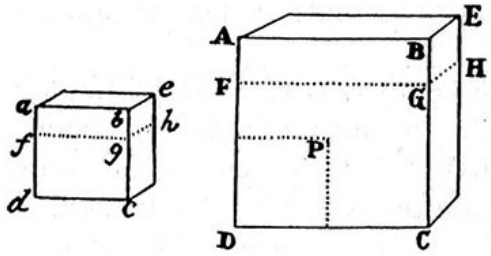
Por un razonamiento similar se puede probar que, si la gravedad de las partículas de un fluido disminuye en razón cúbica de las distancias al centro mientras los inversos del cuadrado de las distancias SA, SB, SC, etc. $\left(\text{o sea, } \frac{SA^3}{SA^2}, \frac{SA^3}{SB^2}, \frac{SA^3}{SC^2}\right)$ se toman en progresión aritmética, las densidades AH, BI, CK, etc., estarán en progresión geométrica. Y si la gravedad disminuye en razón de la cuarta potencia de las distancias y los inversos de los cubos de las distancias $\left(\text{como, } \frac{SA^4}{SA^3}, \frac{SA^4}{SB^3}, \frac{SA^4}{SC^3}\right)$, etc. se toman en progresión aritmética, las densidades AH, BI, CK, etc., estarán en progresión geométrica. Y así hasta el infinito. Y también, si la gravedad de las partículas de un fluido es la misma a todas las distancias, y las distancias están en progresión aritmética, las densidades estarán en progresión geométrica, como descubrió el eminente *Edmund Halley*. Si la gravedad es como la distancia y los cuadrados de las distancias se hallan en progresión aritmética, las densidades estarán en progresión geométrica. Y así hasta el infinito. Esto ocurre así cuando la densidad del fluido condensado por compresión es como la fuerza de compresión, o, lo que es igual, cuando el espacio ocupado por el fluido es inversamente como dicha fuerza. Se pueden imaginar otras leyes de condensación, tales como que el cubo de la fuerza de condensación es como la cuarta potencia de la densidad, o que el cubo de la fuerza es igual que la cuarta potencia de la densidad. En tal caso, si la gravedad es inversamente como el cuadrado de la distancia al centro, la densidad será inversamente como el cubo de la distancia. Imagínese que el cubo de la fuerza de compresión sea como la quinta potencia de la densidad, y si la gravedad es inversamente como el cuadrado de las distancias, la densidad será inversamente proporcional a la potencia ½ de la distancia. Supongamos que la fuerza de compresión sea como el cuadrado de la densidad, y la gravedad inversamente proporcional al cuadrado de la distancia, la densidad será entonces inversamente proporcional a la distancia. Sería demasiado prolijo repasar todos los casos. Pero consta experimentalmente que la densidad del aire es como la fuerza de compresión exacta o al menos muy aproximadamente: y, por tanto, la densidad del aire en la

atmósfera de la tierra es como el peso de todo el aire que sobrecarga, esto es, como la altura del mercurio en el barómetro.

PROPOSICIÓN XXIII. TEOREMA XVIII[7]

Si la densidad de un fluido compuesto de partículas que huyen unas de otras es como la compresión, las fuerzas centrífugas de las partículas son inversamente proporcionales a las distancias de sus centros. Y viceversa, las partículas que se rehuyen con fuerzas inversamente proporcionales a las distancias de sus centros componen un fluido elástico cuya densidad es proporcional a la compresión.

Supongamos que el fluido está contenido en un espacio cúbico ACE, y de éste, por compresión, se reduce a otro espacio cúbico menor *ace;* y las distancias entre las partículas que tienen Ubicación semejante entre sí en uno y otro espacio, serán como los lados AB, *ab* de los cubos; y las densidades de uno y otro medio serán inversamente proporcionales a los espacios que los contienen, AB^3 y ab^3. Sobre el lado plano ABCD del cubo mayor tómese el cuadrado DP, igual al lado plano del cubo menor *db*; y por hipótesis, la presión con la cual actúa el cuadrado DP sobre el fluido contenido será a la presión con la que el susodicho cuadrado *db* actúa sobre el fluido contenido como las densidades de los medios entre sí, esto es como ab^3 a AB^3. Pero la presión con la cual el cuadrado DB actúa sobre el fluido contenido es a la presión con la cual el cuadrado DP actúa sobre el mismo contenido como el cuadrado DB al cuadrado DP, esto es, como AB^2 a ab^2. Por consiguiente, la presión con la cual el cuadrado DB actúa sobre el fluido es a la presión con la cual actúa sobre el fluido el cuadrado *db* como *ab* a AB. Trazando por el interior de los cubos los planos FGH, *fgh*, quede el fluido separado en dos partes, y éstas se oprimirán entre sí con las mismas fuerzas con las cuales son presionadas por los planos AC, *ac*, esto es, en la proporción de *ab* a AB: y por ello, las fuerzas centrífugas que responden a estas presiones, están en la misma proporción, puesto que el número de partículas es el mismo y la disposición semejante en ambos cubos, las fuerzas con las cuales todas las partículas actúan sobre todas según los planos FGH, *fgh*, son como las fuerzas que cada una ejerce sobre cada una. Luego las fuerzas que cada una ejerce sobre cada una según el plano FGH en el cubo mayor son a las fuerzas que cada una ejerce sobre cada una según el plano *fgh* en el cubo menor como *ab* a AB, esto es, inversamente proporcionales a las distancias mutuas de las partículas. Q. E. D.



Y viceversa, si las fuerzas de cada partícula son inversamente proporcionales a la distancia, es decir, inversamente proporcionales a los lados AB, *ab*, de los cubos; las sumas de las fuerzas estarán en la misma proporción, y las presiones de los lados DB, *db*, serán como las sumas de las fuerzas; y la presión del cuadrado DP será a la presión del lado DB como ab^2 a AB^2. Y, por consiguiente, la presión del cuadrado DP será a la presión del lado *db* como ab^3 a AB^3, es decir, la fuerza de compresión a la fuerza de compresión como la densidad a la densidad. Q. E. D.

Escolio

Y por un razonamiento semejante, si las fuerzas centrífugas de las partículas son inversamente proporcionales a los cuadrados de las distancias entre centros, los cubos de las fuerzas de compresión serán como la cuarta potencia de las densidades. Si las fuerzas centrífugas son inversamente proporcionales a la tercera o cuarta potencia de las distancias, los cubos de las fuerzas de compresión serán como la quinta o sexta potencia de las densidades. Y de modo general, si D representa la distancia, E la densidad del fluido comprimido y las fuerzas centrífugas son inversamente proporcionales a una potencia cualquiera de la distancia D cuyo índice sea *n* las fuerzas de compresión serán como las raíces cúbicas de la potencia de E^{n+2}, cuyo índice sea $n + 2$: y viceversa. Pero esto ha de entenderse respecto a fuerzas centrífugas de partículas que limitan con partículas próximas o que no se expanden mucho más allá. Tenemos un ejemplo en los cuerpos magnéticos. El poder de atracción de los cuales termina casi en los cuerpos contiguos de su propio género. La fuerza magnética disminuye por la interposición de una lámina de hierro, y casi termina en la lámina. Pues los cuerpos ulteriores son atraídos más por la lámina que por el imán. Del mismo modo, si unas partículas repelen a otras partículas de su especie que están próximas, mientras que no ejercen fuerza alguna sobre otras partículas más lejanas, es de partículas de este género de las que se componen los fluidos sobre los que se ha tratado en esta Proposición. Porque si la fuerza de una partícula cualquiera se propagase hasta el infinito, se necesitaría una fuerza mayor para alcanzar una condensación igual de una mayor cantidad de fluido. Pero es una cuestión física la de si los fluidos elásticos constan de partículas que se repelen

mutuamente. Por nuestra parte hemos demostrado matemáticamente la propiedad de los fluidos compuestos de tales partículas, para dar a los filósofos un asidero al tratar esta cuestión.

Sección VI
DEL MOVIMIENTO Y RESISTENCIA DE CUERPOS PENDULARES

Proposición XXIV. Teorema XIX

Las cantidades de materia de cuerpos pendulares cuyos centros de oscilación están a igual distancia del centro de suspensión se hallan en razón compuesta de la razón del peso y de la razón del cuadrado de los tiempos de oscilación en el vacío.

Pues la velocidad que sobre una materia dada puede engendrar una fuerza dada en un tiempo dado, es directamente como la fuerza y el tiempo e inversamente como la materia. Cuanto mayores son la fuerza o el tiempo, o menor la materia, tanto mayor será la velocidad generada. Esto es evidente por la segunda Ley del Movimiento. Ahora bien, si los péndulos son de igual longitud, las fuerzas motrices en puntos equidistantes de la perpendicular son como los pesos: y por ello, si dos cuerpos oscilando describen arcos iguales y dichos arcos se dividen en partes iguales, por ser los tiempos en que los cuerpos describen cada parte de los arcos correspondientes como los tiempos totales de oscilación, las velocidades en cada parte correspondiente de oscilación serán entre sí como las fuerzas motrices y los tiempos totales de oscilación directamente e inversamente como las cantidades de materia: y por tanto, las cantidades de materia serán directamente como las fuerzas y los tiempos de oscilación e inversamente como las velocidades. Pero las velocidades son inversamente como los tiempos y, por consiguiente, los tiempos directamente y las velocidades inversamente son como los cuadrados de los tiempos, por lo cual las cantidades de materia son como las fuerzas motrices y los cuadrados de los tiempos, es decir, como los pesos y el cuadrado de los tiempos. Q. E. D.

Corolario 1. Por tanto, si los tiempos son iguales, las cantidades de materia en cada cuerpo serán como los pesos.

Corolario 2. Si los pesos son iguales, las cantidades de materia serán como los cuadrados de los tiempos.

Corolario 3. Si las cantidades de materia son iguales, los pesos serán inversamente proporcionales a los cuadrados de los tiempos.

Corolario 4. De donde, al ser, «caeteris paribus», los cuadrados de los tiempos

como las longitudes de los péndulos, si son iguales los tiempos y las cantidades de materia, los pesos serán como la longitud de los péndulos.

Corolario 5. Y en general, la cantidad de materia pendular es directamente proporcional al peso y al cuadrado del tiempo, e inversamente a la longitud del péndulo.

Corolario 6. Pero en medio no resistente la cantidad de materia pendular es directamente como el peso relativo y el cuadrado del tiempo e inversamente como la longitud del péndulo. Pues el peso relativo es la fuerza motriz del cuerpo en un medio grave cualquiera, como expliqué más arriba; y por tanto, cumple el mismo papel en un medio no resistente que el peso absoluto en el vacío.

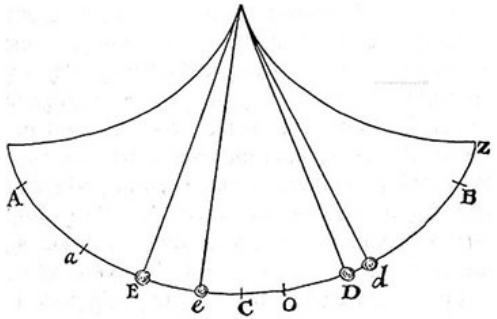
Corolario 7. Y de aquí se obtiene la razón tanto para comparar los cuerpos entre sí en cuanto a la cantidad de materia en cada uno, como para comparar los pesos del mismo cuerpo en distintos lugares, para conocer las variaciones de la gravedad. Pero, mediante experimentos realizados con la mayor precisión siempre hallé que la cantidad de materia en cada cuerpo es proporcional a su peso.

Proposición XXV. Teorema XX

Los cuerpos pendulares a los cuales, en un medio cualquiera, se resiste en razón de los momentos de tiempo y los cuerpos pendulares que se mueven en un medio no resistente de la misma gravedad especifica, completan en el mismo tiempo las oscilaciones en una cicloide y describen en tiempos iguales partes proporcionales de arcos.

Sea AB el arco de la cicloide que el cuerpo D describe al oscilar en un tiempo cualquiera y en medio no resistente. Sea bisecado en C de modo que C resulte su punto inferior; y la fuerza aceleratriz que empuja al cuerpo en un punto cualquiera D, o *d*, o E, será como la longitud del arco CD, o C*d*, o CE. Represéntese esa fuerza por dicho arco, y al ser la resistencia como el momento de tiempo, y por ello estar dada, represéntese ésta por una parte dada del arco cicloide CO, y tómese el arco O*d* respecto al arco CD en la misma proporción que tiene el arco OB respecto al arco CB: y la fuerza que empuja al Cuerpo en *d* en medio resistente, al ser el exceso de la fuerza C*d* sobre la resistencia CO, vendrá representada por el arco O*d*, y por eso será a la fuerza que empuja al cuerpo D en un medio no resistente en el punto D, como el arco O*d* al arco CD; por tanto, también en el punto B como el arco OB al arco CB. De suerte que si dos cuerpos D, *d*, parten del punto B empujados por estas fuerzas, al ser las fuerzas al comienzo como los arcos CB y OB, las velocidades iniciales y los arcos inicialmente descritos estarán en la misma razón. Sean dichos arcos BD y B*d* y los arcos restantes CD, O*d*, estarán en la misma proporción. Por consiguiente, las fuerzas proporcionales a los propios CD, O*d* seguirán siendo proporcionales en la

misma razón que al principio y, por tanto, los cuerpos seguirán describiendo arcos simultáneamente en la misma proporción. Por lo cual, las fuerzas, las velocidades y los restantes arcos CD, O*d*, siempre serán como los arcos totales CB, OB, y por tanto dichos arcos restantes siempre serán descritos simultáneamente. Por lo cual los dos cuerpos D, *d*, llegarán a la vez a los puntos C y O, uno, en un medio no resistente, al punto C, mientras el otro, en un medio resistente, al punto O. Mas, como las velocidades en C y O son como los arcos CB, OB, los arcos que describen los cuerpos después, al marchar a la vez, serán arcos que estén en la misma razón. Sean estos CE, O*e*. La fuerza que retarda al cuerpo D en un medio no resistente en el punto E es como CE, y la fuerza que retarda al cuerpo *d* en un medio resistente en el punto *e* es como la suma de la fuerza C*e* y de la resistencia CO, es decir, como O*e*; por tanto, las fuerzas que retardan a los cuerpos son como los arcos CB, OB, proporcionales a los arcos CE, O*e*; y por lo mismo las velocidades, retardadas en esa razón dada, se mantienen en esa misma razón dada. Por consiguiente, las velocidades y los arcos descritos con ellas siempre están entre sí en la susodicha razón de los arcos CB y OB; y por ello, si se toman los arcos totales AB, *a*B, en la misma razón, los cuerpos D, *d*, describirán esos arcos simultáneamente, y en los puntos A, *a*, perderán a la vez todo su movimiento. Y, por consiguiente, las oscilaciones totales son isócronas y cualesquiera partes de arcos, BD, B*d*, o BE, B*e* descritas a la vez son proporcionales a los arcos totales BA, B*a*. Q. E. D.



COROLARIO. Por lo cual, el movimiento más rápido en un medio resistente no ocurre en el punto más bajo C, sino que se encuentra en el punto O, en el cual es bisecado el arco total descrito *a*B. Y el cuerpo al ir desde allí hasta *a* se va retardando en el mismo grado en que antes se aceleraba al descender desde B hasta O.

PROPOSICIÓN XXVI. TEOREMA XXI

Las oscilaciones en una cicloide de los cuerpos pendulares, a los que se resiste en razón de las velocidades, son isócronas.

Pues si dos cuerpos igualmente distantes de sus centros de suspensión describen oscilando arcos desiguales y las velocidades en las partes correspondientes de arcos son entre sí como los arcos totales, las resistencias, proporcionales a las velocidades, serán también entre sí como los propios arcos. Por lo cual, si estas resistencias se suman o restan a las fuerzas motrices originadas por la gravedad que son como los propios arcos, las sumas o las diferencias serán entre sí en la misma razón de los arcos: y como los incrementos o decrementos de las velocidades son como estas sumas o diferencias, las velocidades serán siempre como los arcos totales: por consiguiente, las velocidades, si en una ocasión son como los arcos totales, permanecerán siempre en esa proporción. Pero al principio del movimiento, cuando los cuerpos comienzan a descender y describir dichos arcos, las fuerzas, al ser proporcionales a los arcos, generarán velocidades proporcionales a los arcos. Luego las velocidades siempre serán como los arcos totales a describir, y por consiguiente, dichos arcos serán descritos en el mismo tiempo. Q. E. D.

PROPOSICIÓN XXVII. TEOREMA XXII

Si se resiste a los cuerpos pendulares en razón del cuadrado de las velocidades, las diferencias entre los tiempos de oscilación en un medio resistente y los tiempos de oscilación en un medio no resistente de la misma gravedad específica serán aproximadamente proporcionales a los arcos descritos al oscilar.

Pues, si con péndulos iguales se describen en un medio resistente los arcos desiguales A, B; la resistencia del cuerpo en el arco A será a la resistencia del cuerpo en la parte correspondiente del arco B como el cuadrado de las velocidades, esto es, aproximadamente como AA a BB. Si la resistencia en el arco B fuese a la resistencia en el arco A como AB a AA, los tiempos en los arcos A y B serían iguales, por la Proposición anterior. Por consiguiente, la resistencia AA en el arco A, o AB en el arco B, produce el exceso de tiempo en el arco A sobre el tiempo en medio no resistente; y la resistencia BB produce el exceso de tiempo en el arco B sobre el tiempo en medio no resistente. Pero tales excesos son como las fuerzas productoras muy aproximadamente, esto es, como los arcos A y B. Q. E. D.

COROLARIO 1. Por ello, a partir de los tiempos de oscilaciones realizadas en arcos desiguales y en medios resistentes se puede conocer el tiempo de oscilación en un medio no resistente de la misma gravedad específica. Pues la diferencia de tiempos será al exceso de tiempo en el arco menor sobre el tiempo en medio no resistente como la diferencia de los arcos al arco menor.

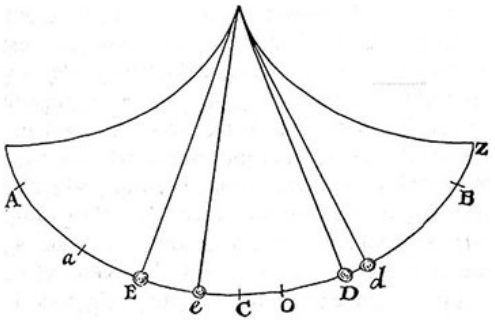
COROLARIO 2. Las oscilaciones más cortas son más isócronas, y las muy cortas se realizan aproximadamente en los mismos tiempos que en un medio no resistente. Pero los tiempos de aquellas que se realizan en arcos más grandes son un poco

mayores, debido a que la resistencia al descender el cuerpo, por la que el tiempo se alarga, es mayor con relación a la longitud descrita en el descenso que la resistencia en el ascenso subsiguiente gracias a la cual el tiempo se acorta. Pero los tiempos de oscilación, tanto breves como largas, parecen alargarse un poco por el movimiento del medio. Pues a los cuerpos que se retardan los resiste algo menos en razón de su velocidad, mientras que a los que se aceleran los resiste algo más que a aquellos que se producen uniformemente: y esto por cuanto que el medio, desplazándose con el movimiento recibido de los cuerpos hacia el mismo lado, se agita más en el primer caso y menos en el segundo; y por ello colabora más o menos con los cuerpos en movimiento. De suerte que en el descenso resiste más a los péndulos y en el ascenso menos que en razón de su velocidad, con lo que por una y otra causa el tiempo resulta prolongado.

Proposición xxviii. Teorema xxiii

Si a un cuerpo pendular que oscila por una cicloide se le resiste en razón de los momentos de tiempo, su resistencia será a la fuerza de la gravedad como el exceso del arco descrito en el descenso total sobre el arco de ascenso descrito subsiguientemente al doble de la longitud del péndulo.

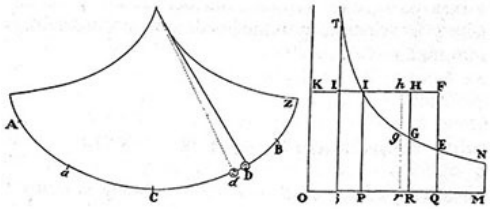
Sea BC el arco descrito en el descenso, C*a* el arco descrito en el ascenso, y A*a* la diferencia entre los arcos: y manteniendo las cosas construidas y demostradas en la Proposición xxv, la fuerza que actúa sobre el cuerpo en oscilación en un punto cualquiera D será a la fuerza de resistencia como el arco CD al arco CO, que es la mitad de la dicha diferencia A*a*. Y por ello, la fuerza que actúa sobre el cuerpo oscilante al principio de la cicloide o en el punto más alto, esto es, la fuerza de la gravedad será a la resistencia como el arco cicloide comprendido entre el punto más alto y el punto más bajo C al arco CO; es decir (si los arcos se duplican) como el arco de toda la cicloide, o el doble de la longitud del péndulo, al arco A*a*. Q. E. D.



Proposición xxix. Problema vi

Supuesto que a un cuerpo oscilante en una cicloide se le resiste en razón del cuadrado de la velocidad, hallar la resistencia en cada punto.

Sea B*a* el arco descrito con la oscilación completa, y sea C el punto más bajo de la cicloide y CZ la mitad del arco cicloide total, igual a la longitud del péndulo; y tratemos de hallar la resistencia del cuerpo en un punto cualquiera D. Córtese la recta infinita OQ en los puntos O, S, P, Q, con la condición de que (si se elevan las perpendiculares OK, ST, PI, QE, y con centro en O y asíntotas OK, OQ se traza la hipérbola TIGE que corte a las perpendiculares ST, PI, QE, en T, I, E, y por el punto I se traza KF paralela a la asíntota OQ y que corte a la asíntota OK en K y a las perpendiculares ST y QE en L y F) el área hiperbólica PIEQ sea al área hiperbólica PITS como el arco BC descrito en el descenso del cuerpo al arco C*a* descrito en el ascenso, y el área IEF al área ILT como OQ a OS. Después, con la perpendicular MN sepárese el área hiperbólica PINM que sea al área hiperbólica PIEQ como el arco CZ al arco BC descrito en el descenso. Y si con la perpendicular RG se segrega el área hiperbólica PIGR que sea al área PIEQ como un arco cualquiera CD al arco BC descrito en el descenso completo; la resistencia en el punto D será a la fuerza de la gravedad como el área $\frac{OR}{OQ}$ IEF - IGH al área PINM.

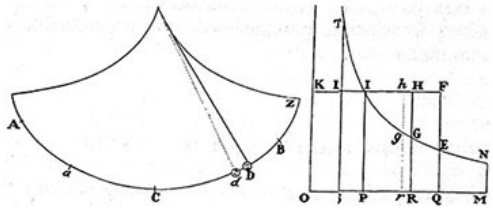


Y si las fuerzas procedentes de la gravedad que actúan sobre el cuerpo en los puntos Z, B, D, a, son como los arcos CZ, CB, CD, C*a*, y dichos arcos son como las áreas PINM, PIEQ, PIGR, PITS; entonces, represéntense tanto los arcos como las fuerzas por estas áreas. Además, sea D*d* un espacio muy pequeño recorrido por el cuerpo en descenso y represéntese por el área mínima RG*gr*, comprendida entre las paralelas RG, *rg*; y prolónguese *rg* hasta *h* de suerte que GH*hg* y RG*gr* sean los decrementos simultáneos de las áreas IGH, PIGR. El incremento GH*hg* - $\frac{Rr}{OQ}$ IEF, o R*r* x HG - $\frac{Rr}{OQ}$ IEF del área $\frac{OR}{OQ}$ IEF - IGH será al decremento RG*gr*, o R*r* x RG del área PIGR

como HG - $\frac{IEF}{OQ}$ a RG; y por tanto, como OR x HG - $\frac{OR}{OQ}$ IEF a OR x GR, o OP x PI, es decir (por ser iguales OR x HG, OR x HR - OR x GR, ORHK - OPIK, PIHR y PIGR + IGH) como PIGR + IGH - $\frac{OR}{OQ}$ IEF a OPIK. Luego, si el área $\frac{OR}{OQ}$ IEF - IGH se llama Y, y el decremento RG*gr* del área PIGR se hallase dado, el incremento del área Y sería como PIGR - Y.

Porque si V designase la fuerza procedente de la gravedad proporcional al arco a describir CD y que actúa sobre el cuerpo en D, y se escribe R para representar la resistencia, V - R será la fuerza total que actúa sobre el cuerpo en D. Y así el incremento de la velocidad es como V - R y la partícula de tiempo en que se produce conjuntamente: pero la propia velocidad es directamente como el incremento contemporáneo de espacio recorrido e inversamente como la misma partícula de tiempo. De donde, al ser, por hipótesis, la resistencia como el cuadrado de la velocidad, el incremento de la resistencia (por el Lema II) será como la velocidad y el incremento de la velocidad conjuntamente, es decir, como el momento de espacio y V - R conjuntamente; y por consiguiente, si se da el momento de espacio, como V - R; de suerte que, si por la fuerza V escribimos su representación PIGR, y la resistencia R se representa por otra área cualquiera Z, vendrá a ser como PIGR - Z.

Por tanto, al decrecer continuamente el área PIGR por la sustracción de momentos dados, crecen las áreas, Y, en razón de PIGR - Y, y Z en razón de PIGR - Z. Y por consiguiente, si las áreas Y y Z comienzan a la vez y son iguales al principio, continuarán siendo iguales por la adición de momentos iguales, de igual modo que después desaparecerán a la vez al decrecer por momentos iguales. Y viceversa, si empiezan y se desvanecen a la vez, tendrán momentos iguales y serán siempre iguales: y esto además porque, si la resistencia Z aumenta, la velocidad junto con el arco aquel C*a*, que se describe en si ascenso del cuerpo, disminuirán; y al acercarse el punto en que cesan todo movimiento y toda resistencia al punto C, la resistencia se desvanece más rápidamente que el área Y. Lo contrario habrá de ocurrir cuando se disminuye la resistencia.



Ahora bien, el área Z comienza y acaba cuando la resistencia es nula, es decir, al principio del movimiento, cuando el arco CD se iguala al arco CB y la recta RG

incide sobre la recta QE, y al final del movimiento, cuando el arco CD se iguala al arco C*a* y RG incide sobre la recta ST. Y el área Y, o sea $\frac{OR}{OQ}$ EIF - IGH, comienza y acaba también cuando es nula, y por ello cuando $\frac{OR}{OQ}$ IEF y IGH son iguales: es decir (por construcción) cuando la recta RG incide sucesivamente sobre las rectas QE y ST. Y por tanto, dichas áreas comienzan y se desvanecen a la vez, por lo que siempre son iguales. En consecuencia el área $\frac{OR}{OQ}$ IEF - IGH es igual al área Z, por la cual se representa la resistencia, y por ello es al área PINM por la que se representa la gravedad, como la resistencia a la gravedad. Q. E. D.

COROLARIO 1. Por tanto, la resistencia en el punto más bajo C es a la fuerza de la gravedad como el área $\frac{OP}{OQ}$ IEF al área PINM.

COROLARIO 2. En cambio se hace máxima cuando el área PIHR es al área IEF como OR a OQ. Pues en tal caso su momento (a saber PIGR - Y) resulta nulo.

COROLARIO 3. De aquí se obtiene también la velocidad en cada punto: ya que es como la raíz cuadrada de la resistencia, y al comienzo mismo del movimiento es igual a la velocidad del cuerpo oscilante en la misma cicloide sin resistencia alguna.

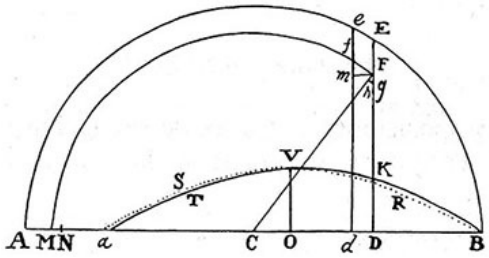
Por otra parte, dada la dificultad del cálculo para hallar la resistencia y la velocidad por medio de esta proposición, parece oportuno añadir la siguiente.

PROPOSICIÓN XXX. TEOREMA XXIV

Si la recta aB fuese igual al arco de la cicloide descrito por un cuerpo oscilante, y por cada uno de sus puntos D se elevan las perpendiculares DK, tales que sean a la longitud del péndulo como la resistencia del cuerpo en los puntos correspondientes del arco a la fuerza de la gravedad: digo que la diferencia entre el arco descrito en todo el descenso y el arco descrito en todo el ascenso subsiguiente multiplicada por la semisuma de ambos arcos será igual al área BKa ocupada por todas las perpendiculares DK.

Represéntese, pues, el arco de la cicloide descrito en una oscilación completa mediante la susodicha recta *a*B, igual a él, y el arco que se hubiere descrito en el vacío mediante la longitud AB. Al bisecar AB en C, el punto C representará el punto inferior de la cicloide, y CD será como la fuerza procedente de la gravedad que actúa sobre el cuerpo en D según la tangente de la cicloide, y tendrá respecto a la longitud del péndulo la misma razón que tiene la fuerza en D respecto a la fuerza de la

gravedad. Represéntese, pues, dicha fuerza por la longitud CD y la fuerza de la gravedad mediante la longitud del péndulo, y si sobre DE se toma DK en la misma razón respecto a la longitud del péndulo que la de la resistencia respecto a la gravedad, DK representará a la resistencia. Con centro en C y distancia CA o CB constrúyase el semicírculo BE*e*A. Y el cuerpo recorra en un tiempo mínimo el espacio D*d*, y elevando las perpendiculares DE, *de* que toquen a la circunferencia en E y e, éstas serán como las velocidades que adquiriría el cuerpo cayendo en el vacío desde el punto B al alcanzar los puntos D y *a*. Esto es evidente (por la Proposición LII del Libro I). Represéntense, pues, estas velocidades por las susodichas perpendiculares DE, *de*; y sea DF la velocidad que adquiere en D al caer desde B en un medio resistente. Y si con centro en C y distancia CF se traza el círculo F*f*M que corte a las rectas *de* y AB en *f* y M, será M el punto hasta el que ascendería sin resto de resistencia, y *df* la velocidad que alcanzaría en *d*. De donde también, si F*g* representa el momento de velocidad que pierde el cuerpo D al describir el espacio mínimo D*d* por causa de la resistencia, y si se toma CN igual a C*g*, será N el punto hasta el cual ascienda el cuerpo sin resto de resistencia, y MN será el decremento de ascenso debido a la pérdida antedicha de velocidad. Sobre *df* trácese la perpendicular F*m*, y el decremento F*g* de la velocidad DF debido a la resistencia DK será al incremento *fm* de la misma velocidad debido a la fuerza CD como la fuerza generatriz DK a la fuerza generatriz CD. Pero por ser semejantes los triángulos F*mf*, F*hg*, FDC, resulta que *fm* es a F*m* o D*d* como CD a DF; y por tanto F*g* a D*d* como DK a DF. Y además, F*h* a F*g* como DF a CF; y por transposición, F*h* o MN a D*d* como DK a CF o CM; y por consiguiente, la suma de todos los MN x CM será igual a la suma de todos los D*d* x DK. En el punto móvil M supóngase siempre levantada una ordenada rectangular igual a la indeterminada CM, que moviéndose continuamente es multiplicada por toda la longitud A*a*; y el trapecio resultante de dicho movimiento, o su igual, el rectángulo A*a* x ½*a*B será igual a la suma de todos los D*d* x DK, es decir, al área BKVT*a*. Q. E. D.



COROLARIO. De aquí que, de la ley de la resistencia y de la diferencia A*a* entre los arcos C*a*, CB, se pueda inferir la proporción de la resistencia respecto a la gravedad muy aproximadamente.

Pues si la resistencia DK es uniforme, la figura BKT*a* será un rectángulo

comprendido entre B*a* y DK, y DK será igual a ½A*a*. Por lo cual, al ser DK la representación de la resistencia, y la longitud del péndulo la de la gravedad, la resistencia será a la gravedad como ½A*a* es a la longitud del péndulo; tal y como se demostró en la Proposición XXVIII.

Si la resistencia fuese como la velocidad, la figura BKTa será muy aproximadamente una elipse. Pues si el cuerpo, en un medio no resistente, describiera con una oscilación completa la longitud BA, la velocidad en un punto cualquiera D sería como la ordenada DE aplicada al diámetro AB del círculo descrito. Por consiguiente, dado que B*a* en un medio resistente y BA en un medio no resistente son descritas casi en los mismos tiempos, por eso mismo las velocidades en cada punto de B*a* son muy aproximadamente a las velocidades en los puntos correspondientes de la longitud BA como B*a* a BA; la velocidad en el punto D en un medio resistente será como la ordenada aplicada al diámetro B*a* del círculo o elipse descritos con él; y por ello, la figura BKVT*a* será aproximadamente una elipse. Al suponer la resistencia proporcional a la gravedad, OV será la representación de la resistencia en el punto medio O; y la elipse BRVS*a*, descrita con centro en O y semiejes OB, OV, será aproximadamente igual a la figura BKVT*a*, o a su igual rectangular A*a* x BO. Por tanto, A*a* x BO es a OV x BO como el área de dicha elipse a OV x BA; es decir, A*a* es a OV como el área del semicírculo al cuadrado del radio, o también, como 11 a 17 aproximadamente: y por consiguiente $^{7}/_{11}$A*a* es a la longitud del péndulo como la resistencia en O del cuerpo oscilante a su gravedad.

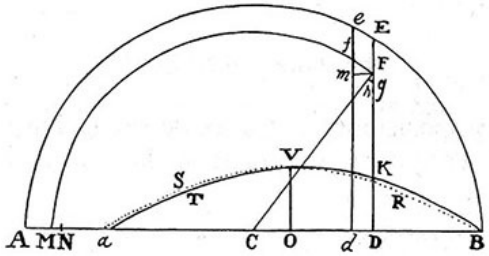
Pero si la resistencia DK fuese como el cuadrado de la velocidad, la figura BKVT*a* sería casi una parábola de vértice en V y eje OV, y, por tanto, igual al rectángulo comprendido entre $^{2}/_{3}$B*a* y OV muy aproximadamente. Por consiguiente, el rectángulo $^{2}/_{3}$B*a* x A*a* es igual al comprendido entre $^{2}/_{3}$B*a* y OV, y por ello, OV es igual a ¼A*a*: por lo cual la resistencia en O de un cuerpo oscilante es a su gravedad como ¼A*a* a la longitud del péndulo.

Y considero que estas conclusiones son más que suficientemente exactas en la práctica. Porque, al coincidir la elipse o la parábola BRVS*a* con la figura BKVT*a* en el punto medio V, si esta figura excede hacia una parte BRV o hacia otra VS*a* a la otra figura, será más pequeña que ella hacia la otra parte, y de ese modo se igualará muy aproximadamente con ella.

PROPOSICIÓN XXXI. TEOREMA XXV

Si la resistencia de un cuerpo oscilante en cada parte proporcional de los arcos descritos es aumentada o disminuida en una razón dada, la diferencia entre el arco descrito en el descenso y el descrito en el subsiguiente ascenso aumentará o disminuirá en la misma razón.

Pues dicha diferencia procede de la retardación del péndulo debida a la resistencia del medio, y es, por tanto, como la retardación total y como la resistencia retardante, proporcional a aquélla. En la Proposición anterior el rectángulo comprendido bajo la recta ½*a*B y la diferencia A*a* de los dichos arcos CB, C*a* era igual al área BKT*a*. Y dicha área, si se mantiene la longitud *a*B, aumenta o disminuye en razón de las ordenadas aplicadas DK; es decir, en razón de la resistencia y, por ello, es como la longitud *a*B y la resistencia conjuntamente. Y por lo mismo, el rectángulo comprendido entre A*a* y ½*a*B es como *a*B y la resistencia conjuntamente y, por lo tanto, A*a* es como la resistencia. Q. E. D.



COROLARIO 1. De donde, si la resistencia es como la velocidad, la diferencia de arcos en el mismo medio será como el arco total descrito, y viceversa.

COROLARIO 2. Si la resistencia fuese como el cuadrado de la velocidad, dicha diferencia sería como el cuadrado del arco total, y viceversa.

COROLARIO 3. Y en general, si la resistencia es como el cubo u otra potencia cualquiera de la velocidad, la diferencia será como esa potencia del arco total, y viceversa.

COROLARIO 4. Y si la resistencia es en parte proporcional a la velocidad y en parte proporcional al cuadrado de la velocidad, la diferencia será en parte como el arco total y en parte como el cuadrado del mismo: y viceversa. La misma será la ley y la razón de la resistencia respecto a la velocidad que la ley y razón de la dicha diferencia con respecto a la longitud del arco.

COROLARIO 5. Y, por consiguiente, si para un péndulo que describe sucesivamente arcos desiguales es posible hallar la razón del incremento o decremento de esta diferencia con respecto a la longitud del arco descrito, se tendrá también la razón del incremento o decremento de la resistencia respecto a una velocidad mayor o menor.

ESCOLIO GENERAL[8]

De estas proposiciones, a través de oscilaciones de péndulos en medios cualesquiera, se puede hallar la resistencia de los medios. Por mi parte he investigado la resistencia del aire con los experimentos siguientes. Suspendí con un hilo delgado

de un clavo bien firme a una bola de madera de $57^{7}/_{22}$ onzas *Romanas* de peso y de un diámetro de 6⅞ pulgadas de *Londres* y de modo que entre el clavo y el centro de oscilación de la bola la distancia fuese de 10½ pies. Sobre el hilo señalé un punto a diez pies y una pulgada del centro de suspensión, y a la altura de dicho punto situé una regla dividida en pulgadas que permitiesen anotar las longitudes de los arcos descritos por el péndulo. Después conté las oscilaciones en las que la bola perdía la octava parte de su movimiento. Si el péndulo se apartaba de la perpendicular hasta una distancia de dos pulgadas y desde allí se soltaba, de modo que en su descenso total recorriese un arco de dos pulgadas y con toda la primera oscilación un arco de casi cuatro pulgadas, aquél, al cabo de 164 oscilaciones, había perdido la octava parte de su movimiento, de suerte que en el último ascenso recorría un arco de una pulgada y tres cuartos. Si en el primer descenso describió un arco de cuatro pulgadas, perdió la octava parte de su movimiento en 121 oscilaciones, de suerte que en la última ascensión describió un arco de 3½ pulgadas. Si en el primer descenso describió un arco de ocho pulgadas, o dieciséis, o treinta y dos, o sesenta y cuatro, perdió la octava parte de su movimiento con 69 oscilaciones, o 35½, o 18½, o $9^{2}/_{3}$, respectivamente. Por tanto, la diferencia entre los arcos descritos con el primer descenso y el último ascenso era en el caso primero, segundo, tercero, cuarto, quinto, sexto de ¼, ½, 1, 2, 4, 8, pulgadas, respectivamente. Divídanse estas diferencias por el número de oscilaciones en cada caso y, para una oscilación media en la que se describió un arco de 3¾, 7½, 15, 30, 60 y 120 pulgadas, la diferencia de los arcos descritos en el descenso y el subsiguiente ascenso será de $^{1}/_{656}$, $^{1}/_{242}$, $^{1}/_{69}$, $^{4}/_{71}$, $^{8}/_{37}$, $^{24}/_{29}$ partes de pulgada, respectivamente. Ahora bien, estas diferencias, en las oscilaciones mayores, son aproximadamente como los cuadrados de los arcos descritos, mientras que en las menores, son un poco mayores que en dicha proporción; y por tanto (por el Corolario 2 de la Proposición XXXI de este Libro), la resistencia de la bola, cuando se mueve con mayor celeridad, es aproximadamente como el cuadrado de la velocidad; y cuando se mueve más lentamente es algo mayor que en dicha proporción.

Ahora llamemos V a la velocidad máxima en una oscilación cualquiera, y sean A, B, C unas cantidades dadas, y supongamos que la diferencia de los arcos sea $AV + BV^{3/2} + CV^{2}$. Puesto que las velocidades máximas son en la cicloide como las mitades de los arcos descritos en la oscilación, mientras en el círculo son como las cuerdas de los dichos semiarcos y, por tanto, a iguales arcos sean mayores en la cicloide que en el círculo y ello en razón de los semiarcos a sus cuerdas; mientras por otra parte, los tiempos en el círculo son mayores que en la cicloide y ello en razón inversa de la velocidad, es evidente que las diferencias de los arcos (que son como la resistencia y el cuadrado del tiempo conjuntamente) han de ser las mismas aproximadamente en una y otra curva: pues deberán las susodichas diferencias aumentar en la cicloide, a la vez que la resistencia, casi como la razón cuadrada del arco a la cuerda, por estar aumentada la velocidad en dicha razón simple; y disminuir, junto con el cuadrado del

tiempo, en la misma razón cuadrada. Para trasladar esto a la cicloide, se han de tomar las mismas diferencias de arcos que fueron observadas en el círculo, y las velocidades máximas análogas a los arcos medios o enteros, es decir, análogos a los números ½, 1, 2, 4, 8, 16. Entonces, para los casos segundo, cuarto y sexto, escribamos en lugar de V los números 1, 4 y 16; y la diferencia de arcos resultará $\frac{½}{121}$ = A + B + C, para el segundo caso; $\frac{2}{35½}$ = 4A + 8B + 16C, para el cuarto caso; y $\frac{8}{9⅔}$ = 16A + 64B + 256C, para el sexto caso. Una vez que estas ecuaciones han sido resueltas por la debida reducción, resultará A = 0,0000916, B = 0,0010847, y C = 0,0029558. Por tanto, la diferencia de los arcos es como $0,0000916V + 0,0010847V^{3/2} + 0,0029558V^2$; y por consiguiente, dado que (por el Corolario de la Proposición xxx aplicado a este caso) la resistencia del globo en la mitad del arco de oscilación, cuando la velocidad es V, es a su peso como $\frac{7}{11}AV + \frac{7}{10}BV^{3/2} + ¾CV^2$ es a la longitud del péndulo, si se escriben en lugar de A, B, y C los números hallados, la resistencia del globo será a su peso como $0,0000583V + 0,0007593V^{3/2} + 0,0022169V^2$ a la longitud del péndulo entre el centro de suspensión y la regla, es decir, a 121 pulgadas. Y puesto que V en el segundo caso representa a 1, en el cuarto a 4 y en el sexto a 16, la resistencia será al peso del globo en el segundo caso como 0,0030345 a 121, en el cuarto como 0,041748 a 121 y en el sexto como 0,61705 a 121.

El arco descrito por el punto marcado en el hilo era en el caso sexto de 120 - $\frac{8}{9⅔}$, o $119\frac{5}{29}$ pulgadas. Y puesto que el radio es de 121 pulgadas y la longitud del péndulo entre el punto de suspensión y el centro de la bola era de 126 pulgadas, el arco descrito por el centro del globo era de $124\frac{3}{31}$ pulgadas. Dado que la velocidad máxima del cuerpo oscilante, por la resistencia del aire, no coincide con el punto inferior del arco descrito, sino que ocurre casi en el punto medio del arco total, ésta será casi la misma que si el globo en todo su descenso en un medio no resistente describiese la mitad de dicho arco, mitad equivalente a $62\frac{3}{62}$ pulgadas, y esto en una cicloide, a la que antes hemos reducido el movimiento del péndulo: y por consiguiente, dicha velocidad será igual a la velocidad que adquiriría el globo cayendo perpendicularmente y describiendo en su caída una altura igual al seno verso de aquel arco. Pero en la cicloide dicho seno verso es a este arco de $62\frac{3}{62}$ como el propio arco a 252, doble de la longitud del péndulo y, por tanto, igual a 15,278 pulgadas. Por lo cual, la velocidad es la misma que la que adquiriría el cuerpo cayendo y describiendo en su caída un espacio de 15,278 pulgadas. Luego a semejante velocidad el cuerpo sufre una resistencia que es a su peso como 0,61705 a 121, o (si se tiene en cuenta solamente aquella parte de resistencia que es como el

cuadrado de la velocidad) como 0,56752 a 121.

Por medio de experimentos hidrostáticos hallé que el peso de esta bola de madera era al peso de un globo de agua de la misma magnitud como 55 a 97: y por tanto, al estar 121 respecto a 213,4 en la misma proporción, la resistencia del globo de agua que se desplaza con la velocidad antedicha será a su peso como 0,56752 a 213,4, es decir, como 1 a $376\frac{1}{50}$. Y por tanto, dado que el peso del globo de agua generaría toda esa velocidad en el globo que cae en el mismo tiempo en el cual dicho globo con velocidad uniformemente continua describiría un espacio de 30,556 pulgadas, es evidente que la fuerza de resistencia uniformemente continua durante ese mismo tiempo podría quitar una velocidad menor en razón de 1 a $376\frac{1}{50}$, es decir, $\frac{1}{376\frac{1}{50}}$ parte de la velocidad total. Y por consiguiente, en el tiempo en que el globo describiría, con dicha velocidad uniformemente continua, la longitud de su semidiámetro, o sea $3\frac{7}{16}$ pulgadas, en ese tiempo perdería $\frac{1}{3342}$ parte de su movimiento.

También contaba las oscilaciones en las que el péndulo perdía la cuarta parte de su movimiento. En la tabla siguiente los números superiores significan la longitud del arco descrito en el primer descenso, expresada en pulgadas y partes de pulgada; los números intermedios significan la longitud del arco descrito en el último ascenso; en último lugar aparecen los números de las oscilaciones. He reseñado el experimento por resultar más exacto que el que sólo cuenta la pérdida de la octava parte de movimiento. Trate de calcularlo el que quiera.

Primer descenso ……	2	4	8	16	32	64
Último ascenso ……	1½	3	6	12	24	48
Número de oscilaciones ……	374	272	162½	83⅓	$41\frac{2}{3}$	$22\frac{2}{3}$

Suspendí después del mismo hilo una bola de plomo de dos pulgadas de diámetro y de 26¼ onzas *romanas* de peso, de suerte que entre el centro del globo y el punto de suspensión hubiese una distancia de 10½ pies, y conté las oscilaciones en que se perdía una parte dada del movimiento. La primera tabla de las siguientes muestra el número de oscilaciones en que se perdió la octava parte del movimiento total; la segunda muestra el número de oscilaciones en que se perdió la cuarta parte del mismo.

Primer descenso ……	1	2	4	8	16	32	64
Último ascenso ……	⅞	7/4	3½	7	14	28	56
Número de oscilaciones							

.....	226	228	193	140	90½	53	30
Primer descenso	1	2	4	8	16	32	64
Último ascenso	¾	1½	3	6	12	24	48
Número de oscilaciones	510	518	420	318	204	121	70

Seleccionando en la tabla primera las observaciones tercera, quinta y séptima y expresando las velocidades máximas en estas observaciones por los números 1, 4, 16, respectivamente, y en general por V como antes, resultará $\frac{½}{193}$ = A + B + C para la tercera; $\frac{2}{90½}$ = 4A + 8B + 16C, para la quinta; y $\frac{8}{30}$ = 16A + 64B + 256C, para la séptima. Una vez resueltas estas ecuaciones dan, A = 0,001414, B = 0,000297, y C = 0,000879. Y de ello se desprende que la resistencia del globo moviéndose con la velocidad V estará respecto a su peso de 26¼ onzas en la misma razón que la que hay entre $0{,}0009V + 0{,}000208V^{3/2} + 0{,}000659V^2$ y la longitud de 121 pulgadas del péndulo. Y si consideramos solamente aquella parte de resistencia que es como el cuadrado de la velocidad, ésta será al peso del globo como $0{,}000659V^2$ a 121 pulgadas. Pero esta parte de resistencia era en el experimento primero respecto al peso de la bola de madera de $57\frac{7}{22}$ onzas como $0{,}002217V^2$ a 121: de aquí que la resistencia del globo de madera a la resistencia del globo de plomo (siendo iguales sus velocidades) sea como $57\frac{7}{22}$ multiplicado por 0,002217 a 26¼ multiplicado por 0,000659, es decir, como 7⅓ a 1. Los diámetros de los dos globos eran de 6⅞ y 2 pulgadas, y los cuadrados de éstos son entre sí como 47¼ y 4, o $11\frac{13}{16}$ y 1 muy aproximadamente. Luego la resistencia de los globos a igual velocidad estaba en menor proporción que la razón del cuadrado de los diámetros. Pero no hemos considerado aún la resistencia del hilo, que ciertamente era grande, y que debe restarse de la resistencia hallada para los péndulos. No pude definirla con precisión, pero encontré que era, no obstante, mayor que la tercera parte de la resistencia total del péndulo menor; y de ello inferí que las resistencias de los globos, descontada la resistencia del hilo, están aproximadamente como el cuadrado de los diámetros. Pues, la razón 7⅓ - ⅓ a 1 - ⅓, o 10½ a 1 no dista mucho de la razón cuadrada de los diámetros $11\frac{13}{16}$ a 1.

Puesto que la resistencia del hilo en globos mayores es de menor importancia, traté también de probar con un globo de 18¾ pulgadas de diámetro. La longitud del péndulo entre el punto de suspensión y el centro de oscilación era de 122½ pulgadas, y entre el punto de suspensión y el nudo del hilo de 109½ pulgadas. El arco descrito

por el nudo en el primer descenso del péndulo fue de 32 pulgadas. El arco descrito por el mismo nudo en el último ascenso, después de cinco oscilaciones, fue de 28 pulgadas. La suma de arcos, o el arco total descrito en una oscilación media fue de 60 pulgadas. La diferencia de arcos de 4 pulgadas. Su décima parte o la diferencia entre el descenso y el ascenso en una oscilación media fue de ⅖ de pulgada. Y como el radio 109½ es al radio 122½ así también el arco total de 60 pulgadas descrito por el nudo en una oscilación media es al arco total de 67⅛ descrito por el centro del globo en una oscilación media; y así también la diferencia de y es a la nueva diferencia de 0,4475. Si se aumentase la longitud del péndulo, manteniendo la longitud del arco descrito, en la razón de 126 a 122½, el tiempo de oscilación aumentaría y la velocidad del péndulo disminuiría como la raíz cuadrada de aquella razón, pero, en cambio, se conservaría la diferencia de 0,4475 entre los arcos descritos en el descenso y consiguiente ascenso. Pero si el arco descrito aumentase en razón de $124\frac{3}{31}$ a 67⅛, la diferencia de 0,4475 aumentaría como el cuadrado de dicha razón y vendría a ser de 1,5295. Esto sería así, bajo la hipótesis de que la resistencia del péndulo fuese como el cuadrado de la velocidad. Luego, si el péndulo describiese un arco total de $124\frac{3}{31}$ pulgadas y su longitud entre el punto de suspensión y el centro de oscilación fuese de 126 pulgadas, la diferencia de arcos descritos en el descenso y consiguiente ascenso sería de 1,5295 pulgadas. Y esta diferencia multiplicada por el peso del globo-péndulo, que era de 208 onzas, da un producto de 318,136. Ahora bien, cuando el péndulo primero, hecho de una bola de madera, describía con su centro de oscilación, que distaba del punto de suspensión 126 pulgadas, un arco total de $124\frac{3}{31}$ pulgadas, la diferencia de arcos descritos en el descenso y el ascenso fue de $\frac{126}{121}$ x $\frac{8}{9\frac{2}{3}}$ que multiplicada por el peso del globo, que era de $57\frac{7}{22}$ onzas da un producto de 49,396. Multipliqué estas diferencias por los pesos de los globos para hallar sus resistencias. Pues, las diferencias proceden de las resistencias y son directamente como las resistencias e inversamente como los pesos. Las resistencias son, por tanto, como los números 318,136 y 49,396. Pero la parte de resistencia del globo menor, que es como el cuadrado de la velocidad, resultaba ser con respecto a la resistencia total como 0,56752 a 0,61675, es decir, como 45,453 a 49,396; mientras que la parte de resistencia correspondiente del globo mayor venía a ser casi igual a la resistencia total del mismo y, por tanto, dichas partes son como 318,136 y 45,453 muy aproximadamente, es decir, como 7 y 1. Pero los diámetros de los globos son 18¾ y 6⅞; y sus cuadrados $351\frac{9}{16}$ y $47\frac{17}{64}$ son como 7,438 y 1, es decir, aproximadamente como las resistencias 7 y 1 de los globos. La diferencia de razones no es mayor que la que pudo provenir de la resistencia del hilo. Por consiguiente, aquellas partes de las resistencias que, a globos iguales, son como los cuadrados de las velocidades, son también, a velocidades iguales, como los cuadrados de los diámetros de los globos.

también, a velocidades iguales, como los cuadrados de los diámetros de los globos.

Por lo demás, el mayor de los globos que utilicé en estos experimentos no era perfectamente esférico y por ello en el cálculo aquí presentado prescindí de minucias en razón de la brevedad, sin poner mayor empeño en la exactitud en un experimento que no era bastante exacto. Y por eso quisiera, al depender de esto la demostración del vacío, que se realizaran estos experimentos con globos más numerosos, más grandes y más exactos. Si se toman los globos en progresión geométrica, cuyos diámetros sean, pongamos de 4, 8, 16, 32, pulgadas, se inferirá de la progresión de los experimentos qué es lo que debería acontecer con globos aún mayores.

Ahora bien, para comparar las resistencias de distintos fluidos entre sí, hice lo siguiente. Preparé una caja de madera de cuatro pies de larga y de un pie de ancha y otro de alta. Desprovista de tapa la llené de agua de la fuente, e hice que unos péndulos sumergidos en ella se moviesen oscilando. Un globo de plomo de 166⅙ onzas de peso y de 3⅝ pulgadas de diámetro se movía como se expone en la tabla siguiente, siendo además la longitud del péndulo desde el punto de suspensión hasta un punto señalado en el hilo de 126 pulgadas, hasta el centro de oscilación de 134⅜ pulgadas.

Arco del primer descenso descrito por el punto señalado en el hilo, en pulgadas	64	32	16	8	4	2	1	½	¼
Arco del último ascenso, en pulgadas	48	24	12	6	3	1½	¾	⅜	$^{3}/_{16}$
Diferencia de arcos proporcional al movimiento perdido, en pulgadas	16	8	4	2	1	½	¼	⅛	$^{1}/_{16}$
Número de oscilaciones en el agua			$^{39}/_{60}$	1⅕	3	7	$11^{1}/_{14}$	$12^{2}/_{3}$	13⅓
Número de oscilaciones en el aire	85½		287	535					

En el experimento de la cuarta columna se perdieron iguales movimientos con 535 oscilaciones en el aire y con 1⅕ en el agua. Ciertamente las oscilaciones en el aire eran algo más veloces que en el agua. Pero si las oscilaciones en el agua se acelerasen de modo que los movimientos de los péndulos fuesen igualmente veloces en ambos medios, se mantendría el mismo número de oscilaciones de 1⅕ en el agua, en las cuales se perdería el mismo movimiento que antes, por el hecho de haber aumentado la resistencia y a la vez haber disminuido el cuadrado del tiempo según el cuadrado de esa misma razón. Luego con iguales velocidades de los péndulos se han perdido iguales movimientos con 535 oscilaciones en el aire y con 1⅕ en el agua;

a 1⅕. Esta es la proporción de las resistencias en el caso de la columna cuarta.

Ahora llamemos $AV + CV^2$ a la diferencia de arcos descritos en el descenso y subsiguiente ascenso por el globo moviéndose con la máxima velocidad; y puesto que la velocidad máxima en el caso de la columna cuarta es a la velocidad máxima en el caso de la columna primera como 1 a 8, y la susodicha diferencia de arcos en el caso de la columna cuarta es a la diferencia de arcos en el caso de la columna primera como $\frac{2}{535}$ a $\frac{16}{85¼}$, o como 85½ a 4280, escribamos en estos casos 1 y 8 por las velocidades, y 85½ y 4280 por las diferencias de arcos, y resultará A + C = 85½, y 8A + 64C = 4280, o A + 8C = 535; y resolviendo las ecuaciones resultará de aquí que 7C = 449½, y C = $64^{3}/_{14}$, y A = $21^{2}/_{7}$; y por tanto, la resistencia, al ser como $^{7}/_{11}AV + ¾V^2$, será como $13^{6}/_{11} + 48^{9}/_{56}V^2$. Por lo cual, en el caso de la cuarta columna, donde la velocidad era 1, la resistencia total es a la parte de ella misma que es proporcional al cuadrado de la velocidad como $13^{6}/_{11} + 48^{9}/_{56}$, o $61^{12}/_{17}$ a $48^{9}/_{56}$; y por esto, la resistencia del péndulo en el agua es a la susodicha parte de resistencia en el aire que es proporcional al cuadrado de la velocidad y que solamente entra en consideración en los movimientos más veloces, como $61^{13}/_{17}$ a $48^{9}/_{56}$ y 535 a 1⅕ conjuntamente, es decir, como 571 a 1. Si se hubiese sumergido el hilo entero del péndulo que oscilaba en el agua, su resistencia aún habría sido mayor; hasta el punto que aquella resistencia del péndulo oscilante en el agua que es proporcional al cuadrado de la velocidad y que solamente merece consideración en los cuerpos más veloces, será a la resistencia del péndulo mismo completo, moviéndose con la misma velocidad en el aire, aproximadamente como 850 a 1, es decir, aproximadamente como la densidad del agua a la densidad del aire.

En este cálculo también se debería considerar aquella parte de resistencia del péndulo en el agua que fuese como el cuadrado de la velocidad, pero (aunque parezca extraño) la resistencia en el agua aumentaba más que en razón del cuadrado de la velocidad. Buscando la causa de ello, vine a dar en que la caja era demasiado pequeña para el tamaño del globo del péndulo e impedía demasiado el movimiento del agua en retroceso por su estrechez. Pues, si se sumergía un globo pendular de una pulgada de diámetro, la resistencia aumentaba aproximadamente como el cuadrado de la velocidad. Ensayaba esto construyendo un péndulo de dos globos, de los cuales el menor y más bajo oscilase en el agua, mientras el mayor y más alto estaba sujeto por el hilo encima de ella, y oscilando en el aire ayudaba al movimiento del péndulo y lo hacía más rápido. Y los experimentos realizados de este modo tuvieron los resultados recogidos en la tabla siguiente.

Arco descrito en el primer descenso	16	8	4	2	1	½	¼
Arco descrito en el último ascenso	12	6	3	1½	¾	⅜	$^{3}/_{16}$

Diferencia de arcos proporcionales al movimiento perdido ……	4	2	1	½	¼	⅛	1/16
Número de oscilaciones ……	3⅜	6½	12½	21⅕	34	53	62⅕

También hice, comparando entre sí las resistencias de los medios, que péndulos de hierro oscilasen en mercurio. La longitud del hilo de hierro era de casi tres pies y el diámetro del globo pendular de casi un tercio de pulgada. Otro globo de plomo estaba sujeto al hilo inmediatamente encima del mercurio para que el movimiento del péndulo continuase durante más tiempo. Entonces llenaba un recipiente, de casi tres libras de capacidad de mercurio, alternativamente con mercurio y agua común, para hallar así la proporción de resistencias al péndulo oscilante sucesivamente en uno y otro fluido: y la resistencia en el mercurio resultó ser a la del agua aproximadamente como 13 ó 14 a 1, es decir, como la densidad del mercurio a la del agua. Cuando ponía un globo pendular algo mayor, como de un tercio o de dos tercios de pulgada de diámetro, la resistencia del mercurio venía a ser a la del agua casi como la de 12 ó 10 a 1. Pero es más fiable el experimento anterior, ya que en este último el vaso era demasiado pequeño para el tamaño del globo sumergido. Al aumentar el globo se debería ampliar también el vaso. Había pensado en repetir experimentos de este tipo utilizando recipientes mayores y con líquidos de metales fundidos o bien con otros tanto calientes como fríos: pero no hay tiempo de hacer todos los experimentos, y por los ya descritos consta suficientemente que la resistencia de los cuerpos velozmente movidos es proporcional aproximadamente a la densidad de los fluidos en que se mueven. No digo exactamente. Pues, los fluidos más tenaces, a igual densidad, sin duda resisten más que los más líquidos, como el aceite frío más que el caliente, el caliente más que el agua de lluvia, el agua más que el espíritu de vino. Pero en los líquidos que suelen ser bastante fluidos, como en el aire, en agua dulce o salada, en espíritus de vino, trementina, y sales, en aceite caliente y limpio por destilación de residuos, en aceite de vitriolo y en mercurio, en metales fundidos y otros por el estilo que son tan fluidos que permiten, al ser agitados en los vasos, conservar durante largo tiempo el movimiento, y cuando se derraman se producen gotas fácilmente, no me cabe duda de que la regla dada resulte bastante exacta: sobre todo si los experimentos se hicieren con cuerpos pendulares mayores y movidos más velozmente.

Y por fin, dada la opinión de algunos, según la cual existe un cierto medio etéreo y muy sutil que impregna los poros e intersticios de todos los cuerpos y además debe dar origen a alguna resistencia debida al tal medio fluyente por los poros de los cuerpos, con el fin de averiguar si la resistencia que experimentamos en los cuerpos en movimiento se halla toda en la superficie externa de los mismos o también las partes internas en sus propias superficies la experimentan de modo perceptible, diseñé el siguiente experimento. Con un hilo de once pies de largo suspendí de un gancho de acero, con un anillo de acero también, una caja redonda de abeto para construir un péndulo de dicha longitud. El gancho superior estaba muy afilado por su

construir un péndulo de dicha longitud. El gancho superior estaba muy afilado por su cara cóncava, para que el anillo con su arco superior fijado al filo se moviese con toda libertad. El hilo estaba atado al arco inferior. Una vez hecho el péndulo de esta manera, lo separé de la perpendicular hasta una distancia de casi seis pies, y esto según un plano perpendicular al filo del gancho, para que el anillo, al oscilar el péndulo, no resbalase sobre el filo del gancho hacia un lado o hacia otro. Pues el punto de suspensión en que el anillo toca al gancho, debe permanecer inmóvil. Señalé exactamente el lugar hasta el cual separé al péndulo, y después de soltarlo, señalé los tres puntos hasta los que regresó al final de las oscilaciones primera, segunda y tercera. Repetí esto muchas veces para determinar lo más exactamente posible dichos puntos. Después llené la caja de plomo y de los metales más pesados que tenía a mano. Pero antes pesé la caja vacía, lo mismo que el hilo que la rodeaba y la mitad del resto del hilo que se extendía entre el gancho y la caja suspendida. Pues el hilo tendido actúa siempre con la mitad de su peso sobre el péndulo separado de la perpendicular. A este peso añadí el peso del aire contenido en la caja. El peso total vino a ser aproximadamente como una setenta y ochava parte del peso de la caja llena de metales. Entonces, como la caja llena de metales estiraba el hilo con su peso, aumentando la longitud del péndulo, reduje el hilo para que, al oscilar, tuviese la misma longitud que antes. Después, separando el péndulo hasta el primer punto señalado y dejándolo caer, conté las oscilaciones, unas setenta y siete, hasta que la caja regresó al punto señalado en segundo lugar, y otras tantas hasta que llegó a la tercera señal, y de nuevo otras tantas hasta que llegó a la cuarta. De donde, concluyo que la resistencia total de la caja llena no guarda mayor proporción respecto a la caja vacía que 78 a 77. Pues si las resistencias de ambas fuesen iguales, la caja llena, debido a que su fuerza ínsita es setenta y ocho veces mayor que la fuerza ínsita de la caja vacía, debería conservar su movimiento de oscilación durante tanto más tiempo y, por tanto, retornar hasta los puntos señalados tras 78 oscilaciones completas. Pero regresó a ellos tras 77 oscilaciones completas.

Represente A la resistencia de la caja en su superficie externa y B la resistencia de la caja vacía en sus partes interiores; y si la resistencia de los cuerpos igualmente veloces en sus partes internas es como la materia o el número de partículas que resisten, 78B será la resistencia de la caja llena en sus partes internas y, por tanto, la resistencia total A + B de la caja vacía será a la resistencia total A + 78B de la caja llena como 77 a 78, y parcialmente, A + B a 77B, como 77 a 1, y por ello, A + B a B como 77 x 77 a 1, y parcialmente, A a B como 5928 a 1. Por consiguiente, la resistencia de la caja vacía en sus partes internas es más de cinco mil veces menor que su resistencia en la superficie externa. Argumentamos así bajo la hipótesis de que la mayor resistencia de la caja llena no obedece a otra causa oculta sino sólo a la acción sobre el metal incluido de algún fluido sutil.

He expuesto este experimento de memoria. Pues he perdido los papeles en que lo había anotado. Por ello he tenido que omitir algunas fracciones numéricas que

Pero ahora no hay tiempo de intentarlo otra vez. La primera vez, por utilizar un gancho no seguro, la caja llena se retardaba más rápidamente. Al buscar la causa, hallé que el gancho, débil, cedía al peso de la caja y cediendo a sus oscilaciones se doblaba hacia todas partes. Preparé entonces un gancho firme, que garantizase la inmovilidad del punto de suspensión, y después todo ocurrió como arriba queda descrito.

Sección VII[9]
SOBRE EL MOVIMIENTO DE FLUIDOS Y LA RESISTENCIA DE PROYECTILES

Proposición XXXII. Teorema XXVI

Si dos sistemas semejantes de cuerpos constan de igual número de partículas y las partículas correspondientes son semejantes y proporcionales una a una en cada sistema y están situadas de manera semejante entre sí y además tienen entre sí una razón dada de densidad, y además empiezan a moverse de manera semejante entre sí en tiempos proporcionales (las de un sistema entre sí y las del otro entre ellas) y las que pertenecen a un sistema no se tocan entre ellas más que en el momento de reflexión, ni se atraen ni repelen más que con las fuerzas aceleratrices que sean inversamente proporcionales al diámetro de las partículas correspondientes y directamente al cuadrado de las velocidades: digo que las partículas de los sistemas continuarán moviéndose de manera semejante entre ellas y con tiempos proporcionales.

Digo que cuerpos iguales en sitios semejantes se mueven de modo semejante entre ellos en tiempos proporcionales, si sus situaciones mutuas al final de dichos movimientos son siempre semejantes: tal y como si se comparasen las partículas de un sistema con las partículas correspondientes de otro. De donde los tiempos, en los cuales se describen por las partículas correspondientes partes semejantes y proporcionales de figuras semejantes, serán proporcionales también. Por consiguiente, si se diesen dos sistemas de esta clase, las partículas correspondientes, debido a la semejanza de los movimientos iniciales, continuarán moviéndose con movimientos semejantes hasta que choquen entre sí. Pues si no actúa sobre ellas fuerza alguna, proseguirán uniformemente en línea recta, por la Ley I del Movimiento. Si se perturban mutuamente con alguna fuerza, y ésta es inversamente como el diámetro de las partículas correspondientes y directamente como el cuadrado de las velocidades, puesto que las situaciones de las partículas son semejantes y las fuerzas proporcionales, las fuerzas totales que actúan sobre las partículas correspondientes, que resultan compuestas de las fuerzas singulares en acción (por el Corolario II de las Leyes) tendrán las mismas direcciones, como si se dirigiesen a

centros similarmente situados entre las partículas; y dichas fuerzas totales serán entre sí, como las fuerzas componentes singulares, es decir, inversamente como los diámetros de las partículas correspondientes y directamente como los cuadrados de las velocidades: por lo cual, harán que las partículas correspondientes continúen describiendo figuras semejantes. Esto será así (por los Corolarios 1 y 8 de la Proposición IV del Libro I) si los centros susodichos están en reposo. Pero si están en movimiento, gracias a la semejanza de las traslaciones, sus situaciones permanecen semejantes entre las partículas de los sistemas, y producirán cambios semejantes en las figuras que describen las partículas. Por tanto, los movimientos de partículas correspondientes y semejantes serán semejantes hasta sus primeros choques y, por ser semejantes los choques, y semejantes las reflexiones, también después serán (por lo ya mostrado) semejantes los movimientos entre sí hasta que de nuevo choquen entre ellas, y así sucesivamente hasta el infinito. Q. E. D.

COROLARIO 1. De aquí que si dos cuerpos cualesquiera, semejantes entre sí y semejantemente situados respecto a las partículas correspondientes de los sistemas, empiezan a moverse entre éstas de modo semejante y en tiempos proporcionales y sus magnitudes y densidades son mutuamente como las magnitudes y densidades de las partículas correspondientes, continuarán moviéndose semejantemente en tiempos proporcionales. Pues la razón de las partes mayores de uno y otro sistema es la misma que la de las partículas de los mismos.

COROLARIO 2. Y si todas las partes semejantes y semejantemente dispuestas de los sistemas reposan entre ellas, y dos de ellas, mayores que las otras, y mutuamente correspondientes entre sí en uno y otro sistema, empiezan a moverse según líneas semejantemente situadas y con movimientos semejantes en cualquiera dirección, producirán en las otras partes de los sistemas movimientos semejantes, movimientos que continuarán de modo semejante entre las partes en tiempos proporcionales; y describirán, por ello, espacios proporcionales a sus diámetros.

PROPOSICIÓN XXXIII. TEOREMA XXVII

Con los mismos supuestos, digo que las partes mayores de los sistemas padecen una resistencia que está en razón compuesta del cuadrado de sus velocidades, del cuadrado de sus diámetros y de la razón simple de la densidad de las partes de los sistemas.

Pues la resistencia procede en parte de las fuerzas centrípetas o centrífugas, con las cuales las partículas de los sistemas se agitan mutuamente, parte de las colisiones y reflexiones de las partículas y las partes mayores. Las resistencias del primer género son entre sí como las fuerzas totales motrices que las producen, es decir, como las fuerzas totales aceleratrices y las cantidades de materia en las partes

correspondientes; es decir (por hipótesis), directamente como el cuadrado de las velocidades e inversamente como las distancias de las partículas correspondientes y directamente como las cantidades de materia en las partículas correspondientes; y por lo mismo, como las distancias de las partículas de un sistema son a las distancias correspondientes de las partículas del otro como el diámetro de la partícula o de una parte en el primer sistema al diámetro de la partícula o parte correspondiente en el otro, y la cantidad de materia es como las densidades de las partes y el cubo de los diámetros, las resistencias son entre sí como los cuadrados de las velocidades y los cuadrados de los diámetros y las densidades de las partes de los sistemas. Q. E. D. Las resistencias de la segunda especie son como el número de reflexiones correspondientes y sus fuerzas conjuntamente. Pero el número de reflexiones son entre sí directamente como las velocidades de las partes correspondientes, e inversamente como los espacios entre las reflexiones de las mismas. Y las fuerzas de las reflexiones son como las velocidades, las magnitudes y las densidades de las partes correspondientes conjuntamente, es decir, como las velocidades, los cubos de los diámetros y las densidades de las partes. Y reunidas todas estas razones, las resistencias de las partes correspondientes son entre sí como los cuadrados de las velocidades y los cuadrados de los diámetros y las densidades de las partes conjuntamente. Q. E. D.

COROLARIO 1. Por consiguiente, si tales sistemas fuesen dos fluidos elásticos, como el aire, y sus partes están en reposo entre ellas, y dos cuerpos semejantes y proporcionales en volumen y densidad a las partes de los fluidos y, además, situados semejantemente entre esas partes, fuesen lanzados en cualquier dirección según líneas semejantemente dispuestas, mientras las fuerzas aceleratrices con las cuales se agitan mutuamente las partículas de los fluidos son inversamente como los diámetros de los cuerpos lanzados y directamente como los cuadrados de las velocidades, dichos cuerpos en tiempos proporcionales excitarán movimientos semejantes en los fluidos y describirán espacios semejantes y proporcionales a sus diámetros.

COROLARIO 2. Por lo cual, en dicho fluido un proyectil veloz sufre una resistencia que es aproximadamente como el cuadrado de su velocidad. Pues si las fuerzas que agitan entre sí a las partículas distantes se aumentasen como el cuadrado de la velocidad, la resistencia estaría en dicha razón cuadrada exactamente; y por tanto, en un medio cuyas partes distantes entre sí no se agitan con fuerza alguna, la resistencia es exactamente como el cuadrado de la velocidad. Sean, pues, A, B, C, tres medios que constan de partes semejantes e iguales y dispuestas regularmente según distancias iguales. Las partes de los medios A y B repélanse entre sí con fuerzas que sean también entre sí como T y V, mientras que las partes del medio C carecen por completo de estas fuerzas. Y si cuatro cuerpos iguales D, E, F, G se movieran en estos medios, los dos primeros D y E en los dos medios primeros A y B, y los otros dos F y G en el tercer medio C, y la velocidad del cuerpo D es a la velocidad del cuerpo E, y la velocidad del cuerpo F a la del cuerpo G como la raíz cuadrada de las fuerzas T a

las fuerzas V, la resistencia del cuerpo D será a la resistencia del cuerpo E, y la resistencia del cuerpo F a la del cuerpo G, como el cuadrado de las velocidades; y por tanto, la resistencia del cuerpo D será a la resistencia del cuerpo F como la resistencia del cuerpo E a la del cuerpo G. Tengan los cuerpos D y F las mismas velocidades, así como los cuerpos E y G; y aumentando las velocidades de los cuerpos D y F en una razón cualquiera, y disminuyendo las fuerzas de las partículas del medio B según el cuadrado de la misma razón, el medio B se aproximará cuanto se quiera a la forma y condición del medio C, y por lo mismo, las resistencias de los cuerpos iguales y a iguales velocidades E y G en tales medios se aproximarán constantemente a la igualdad, de suerte que su diferencia resulte al final menor que una dada cualquiera. Por tanto, cuando las resistencias de los cuerpos D y F sean entre sí como las resistencias de los cuerpos E y G, también éstas se aproximan a la razón de igualdad. Pues las resistencias de los cuerpos D y F, cuando se mueven muy velozmente, son aproximadamente iguales: y por ello, cuando la resistencia del cuerpo F sea como el cuadrado de la velocidad, la resistencia del cuerpo D estará aproximadamente en esa razón.

COROLARIO 3. La resistencia de un cuerpo moviéndose muy velozmente en un fluido elástico es casi la misma que si las partes del fluido careciesen de sus fuerzas centrífugas y no se repeliesen entre sí; ello siempre que la fuerza elástica del fluido proceda de las fuerzas centrífugas de las partículas y la velocidad sea tan grande que las fuerzas no tengan tiempo suficiente para actuar.

COROLARIO 4. Puesto que las resistencias de cuerpos semejantes y con velocidades iguales, en medios cuyas partes distantes no se repelen entre sí, son como los cuadrados de los diámetros; las resistencias de cuerpos a iguales velocidades y con movimiento muy veloz en un fluido elástico son también muy aproximadamente como los cuadrados de los diámetros.

COROLARIO 5. Y puesto que cuerpos semejantes, iguales y con velocidades iguales, en medios de la misma densidad y cuyas partículas no se repelen unas a otras, tanto si dichas partículas son muchas y pequeñas como si son pocas y grandes, tropezarán con igual cantidad de materia en tiempos iguales, y trasmitirán a ésta igual cantidad de movimiento, y a su vez (por la Ley tercera del Movimiento) padecerán de ella la misma cantidad de reacción, es decir, serán resistidos por igual: también es evidente que, en fluidos elásticos de igual densidad, cuando se mueven a grandísima velocidad, sus resistencias serán aproximadamente las mismas; tanto si dichos fluidos están formados de partículas más gruesas como si lo están de las más sutiles imaginables. La resistencia de los proyectiles con movimientos muy rápidos no disminuye mucho por la sutileza del medio.

COROLARIO 6. Todas estas cosas son así en fluidos cuya fuerza elástica tiene origen en las fuerzas centrífugas de las partículas. Pero si dicha fuerza obedeciese a otra causa, como la expansión de partículas al modo de la lana o las ramas de los árboles, o a otra causa cualquiera, por la cual los movimientos de las partículas entre

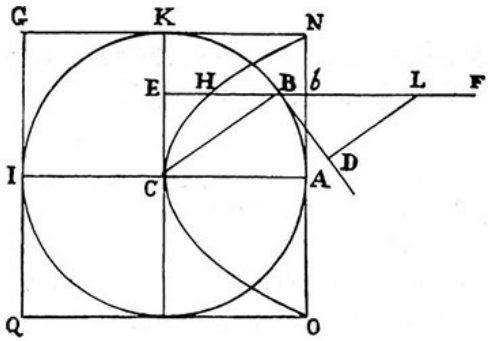
sí resulten menos libres, la resistencia, por la menor fluidez del medio, sería mayor que en los Corolarios anteriores[10].

PROPOSICIÓN XXXIV. TEOREMA XXVIII[11]

Si un globo y un cilindro de diámetros iguales se mueven con iguales velocidades en la dirección del eje del cilindro, en un medio raro y que consta de partículas iguales y libremente dispuestas a distancias mutuas iguales entre ellas: la resistencia del globo será la mitad de la resistencia del cilindro.

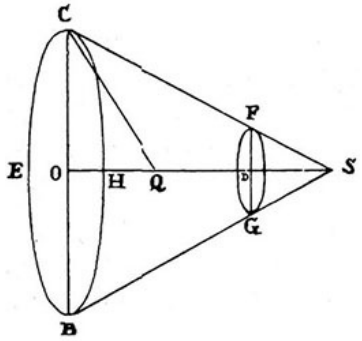
Puesto que la acción del medio sobre un cuerpo es idéntica (por el Corolario V de las Leyes) tanto en el caso de que el cuerpo se mueva en un medo en reposo como si las partículas del medio chocan a la misma velocidad contra el cuerpo en reposo, consideremos al cuerpo en reposo y veamos con qué fuerza es urgido por el medio en movimiento. Sea, pues, ABKI un cuerpo esférico descrito con centro en C y semidiámetro CA, e incidan sobre dicho cuerpo esférico las partículas del medio con una velocidad dada y según rectas paralelas a AC; y sea FB una recta de éstas. Tómese sobre ella LB, igual al semidiámetro CB, y trácese BD tangente a la esfera en B. Sobre KC y BD caigan las perpendiculares BE, LD, y la fuerza con la cual una partícula del medio, incidiendo oblicuamente según la recta FB, golpeará al globo en B, será a la fuerza con la que la misma partícula golpearía perpendicular mente en *b* al cilindro ONGQ descrito con eje ACI en torno al globo, como LD a LB, o BE a BC. Efectivamente, la eficacia de esta fuerza para mover el globo según la dirección FB o AC de su incidencia es a su eficacia para moverlo según la dirección de su propia determinación, es decir, según la dirección de la recta BC con la que actúa directamente sobre el globo como BE a BC. Y, por ambas razones conjuntas, la eficacia de una partícula incidente oblicuamente según la recta FB sobre un globo para moverlo en la dirección de su incidencia es a la eficacia de la misma partícula incidente perpendicularmente según la misma recta sobre un cilindro para moverlo en la misma dirección, como BE^2 a BC^2. Por lo tanto, si sobre *b*E, que es perpendicular a la base circular NAO del cilindro e igual al radio AC se toma *b*H igual a $\frac{BE^2}{CB}$, resultará que *b*H es a *b*E como el efecto de la partícula sobre el globo al efecto de la partícula sobre el cilindro. Y por lo mismo, el sólido que comprende a todas las líneas *b*H será al sólido que comprende a todas las líneas *b*E, como el efecto de todas las partículas sobre el globo al efecto de todas las partículas sobre el cilindro. Pero el primer sólido es un paraboloide descrito con vértice en C, eje CA y «latus rectum» CA; mientras el segundo es un cilindro circunscrito al paraboloide: y es sabido que un paraboloide es la mitad del cilindro circunscrito. Luego la fuerza total del medio

sobre el globo es la mitad de la fuerza total del medio sobre el cilindro. Y por ello, si las partículas del medio reposasen y el globo y el cilindro se moviesen con velocidades iguales, la resistencia del globo sería la mitad que la resistencia del cilindro. Q. E. D.



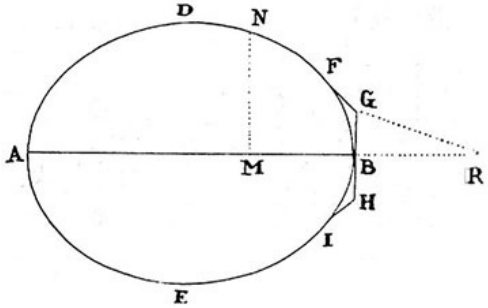
Escolio

Con el mismo método pueden compararse entre sí otras figuras en lo relativo a la resistencia y encontrar aquellas que sean más adecuadas para mantener sus movimientos en medios resistentes. Así, si hubiese que construir, con la base circular CEBH, que tiene su centro en O y radio OC, y con la altura OD, un tronco de cono CBGF que fuese el de menor resistencia de todos los troncos de cono construidos con la misma base y altura al desplazarse según la dirección de su eje hacia D, biséquese la altura OD en Q y prolónguese OQ hasta S de modo que QS sea igual a QC, y S será el vértice del cono cuyo tronco se busca.



Y, de paso, puesto que el ángulo CSB siempre es agudo, se infiere que si el sólido ADBE fuese generado por circunvolución de la figura elíptica u oval ADBE en torno a su eje AB, mientras la figura generatriz es tocada por tres rectas FG, GH, HI, en los

puntos F, B e I, bajo la condición de que GH sea perpendicular al eje en el punto de contacto B, y FG, HI y la propia GH encierren ángulos FGB, BHI de 135 grados, el sólido generado por la circunvolución de la figura ADFGHIE sobre su mismo eje AB sufrirá menor resistencia que el sólido anterior; siempre que ambos se muevan en la dirección de su eje AB y vaya por delante el extremo B de ambos. Proposición esta que no creo haya de ser inútil para la construcción de naves.



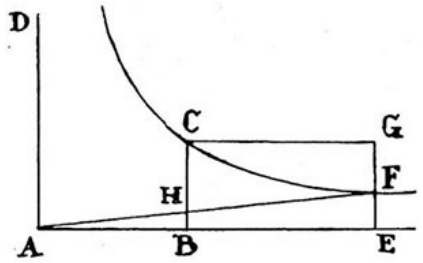
Pero si la figura DNFG fuese una curva tal que, si de su punto N cualquiera se hace descender NM perpendicular al eje AB, y desde un punto dado G se traza la recta GR que sea paralela a la tangente de la figura en N, y corte a la prolongación del eje en R, MN vendría a ser a GR como GR^3 a 4BR x GB^2; el sólido, descrito por la revolución de esta figura sobre su eje AB, moviéndose en el susodicho medio raro desde A hacia B, sufre menor resistencia que cualquier otro sólido circular descrito con las mismas longitud y anchura.

Proposición xxxv. Problema vii

Si un medio raro está constituido de partículas mínimas en reposo y colocadas libremente a iguales distancias entre ellas, hallar la resistencia de un globo que se desplaza uniformemente en dicho medio.

Caso 1. Imagínese que un cilindro descrito con el mismo diámetro y la misma altura se desplaza a la misma velocidad según la línea de su eje en el mismo medio. Y supongamos que las partículas del medio sobre las que inciden el globo o el cilindro retrocedan con la máxima fuerza de reflexión. Y, puesto que la resistencia del globo (por la última Proposición) es la mitad de la resistencia del cilindro y el globo es al cilindro como dos a tres y el cilindro al chocar perpendicularmente con las partículas y reflejándolas así de modo máximo, les comunica el doble de su propia velocidad, el cilindro, en el tiempo en que recorre desplazándose uniformemente la longitud de su semieje, comunicará a las partículas un movimiento que será al movimiento total del

cilindro como la densidad del medio a la densidad del cilindro; y el globo, en el tiempo en que recorre toda la longitud de su diámetro desplazándose uniformemente, comunicará el mismo movimiento a las partículas; y en el tiempo en que recorre los dos tercios de su diámetro, comunicará a las partículas un movimiento que será al movimiento total del globo como la densidad del medio a la densidad del globo. Y por tanto, el globo sufre una resistencia que será a la fuerza con la cual puede suprimirse o generarse todo su movimiento en el tiempo en que puede recorrer, desplazándose uniformemente, las dos terceras partes de su diámetro, como la densidad del medio a la densidad del globo.



Caso 2. Supongamos que las partículas del medio incidentes sobre el globo o sobre el cilindro no son reflejadas; entonces el cilindro, al incidir perpendicularmente sobre las partículas, les comunicará simplemente su propia velocidad, y por ello sufre una resistencia que será la mitad de la del caso anterior, y la resistencia del globo también será la mitad que la anterior.

Caso 3. Supongamos que las partículas del medio se reflejen del globo con una fuerza que no es la máxima ni tampoco nula, sino una fuerza media; en tal caso la resistencia del globo estará en la misma razón media entre la resistencia del primer caso y la del segundo. Q. E. I.

Corolario 1. De aquí que si la dureza del globo y de las partículas es infinita y, por tanto, están desprovistos de toda fuerza elástica y, por tanto, también carecen de toda fuerza de reflexión, la resistencia del globo será a la fuerza con la cual puede suprimirse o generarse todo su movimiento en el tiempo en que recorre cuatro terceras partes de su diámetro, como la densidad del medio a la densidad del globo.

Corolario 2. La resistencia del globo, «caeteris paribus», es como el cuadrado de la velocidad.

Corolario 3. La resistencia del globo, «caeteris paribus», es como el cuadrado del diámetro.

Corolario 4. La resistencia del globo, «caeteris paribus», es como la densidad del medio.

Corolario 5. La resistencia del globo está en razón compuesta del cuadrado de la velocidad, del cuadrado del diámetro y de la densidad del medio.

Corolario 6. Y el movimiento del globo con su resistencia puede representarse como sigue: sea AB el tiempo en el cual el globo puede perder todo su movimiento a causa de su resistencia uniformemente continua. Sobre AB elévense las perpendiculares AD, BC. Y sea BC dicho movimiento total, mientras por el punto C y con asíntotas AD, AB se traza la hipérbola CF. Prolónguese AB hasta un punto cualquiera E. Elévese la perpendicular EF que toque a la hipérbola en F. Complétese el paralelogramo CBEG, y trácese AF que corte a BC en H. Y si el globo en un tiempo cualquiera BE, con su movimiento inicial BC continuado uniformemente, recorre en un medio no resistente el espacio representado por el área del paralelogramo CBEG, el mismo globo en un medio resistente describirá el espacio CBEF representado por el área de la hipérbola, y su movimiento al final de aquel tiempo vendrá representado por la ordenada de la hipérbola EF, una vez perdida su parte FG de movimiento. Y su resistencia, al final del mismo tiempo, vendrá representada por la longitud BH, una vez perdida la parte CH de resistencia. Todo esto es evidente por los Corolarios 1 y 3 de la Proposición v del Libro ii.

Corolario 7. De aquí que si un globo, debido a su resistencia uniformemente continua R, perdiese todo su movimiento M en el tiempo T, el mismo globo en el tiempo t en un medio resistente debido a una resistencia decreciente en razón del cuadrado de la velocidad, perderá de su movimiento M la parte $\frac{t\mathrm{M}}{\mathrm{T}+t}$, reteniendo la parte $\frac{\mathrm{TM}}{\mathrm{T}+t}$; y describirá un espacio que será al espacio descrito con el movimiento uniforme M en el tiempo t como el logaritmo del numero $\frac{\mathrm{T}+t}{\mathrm{T}}$ multiplicado por el número 2,302585092994 es al número $\frac{t}{\mathrm{T}}$ dado que el área hiperbólica BCFE es al rectángulo BCGE en esta misma proporción.

Escolio

En esta Proposición he expuesto la resistencia y retardación de proyectiles esféricos en medios no continuos, y he mostrado que esta resistencia es a la fuerza con la cual podría suprimirse o generarse todo el movimiento del globo en el tiempo en que el globo describiría con velocidad uniformemente continua las dos terceras partes de su diámetro, como la densidad del medio a la densidad del globo, con tal de que tanto el globo como las partículas del medio sean sumamente elásticas y gocen de la máxima fuerza de reflexión: y también que esta fuerza es el doble menor cuando

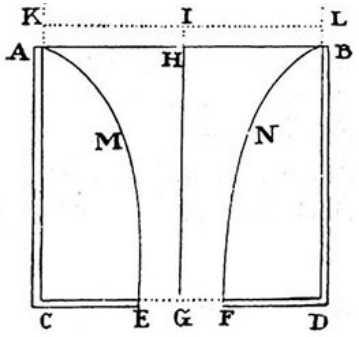
el globo y las partículas del medio son infinitamente duras y prácticamente carentes de fuerza de reflexión. Pero en medios continuos como es el agua, el aceite caliente, el mercurio, en los cuales el globo no incide inmediatamente sobre todas las partículas del fluido generadoras de resistencia, sino que más bien presiona solamente sobre las partículas más próximas y éstas presionan a las siguientes y éstas a las siguientes, la resistencia es todavía otra mitad menor. Efectivamente, en esta clase de medios muy fluidos el globo padece una resistencia que es a la fuerza con la cual se podría suprimir o generar todo su movimiento en el tiempo en el cual, con el susodicho movimiento continuo uniforme, recorrería ocho terceras partes de su diámetro, como la densidad del medio a la densidad del globo. Es lo que trataremos de mostrar en lo que sigue.

Proposición xxxvi. Problema viii[12]

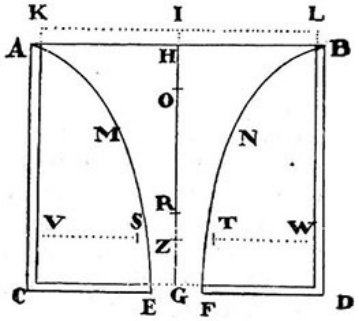
Definir el movimiento del agua que fluye por un orificio hecho en el fondo de un vaso cilíndrico.

Sea ACDB un vaso cilíndrico, AB su orificio superior, CD el fondo paralelo al horizonte, EF un orificio circular en mitad del fondo, G el centro del orificio y GH el eje del cilindro perpendicular al horizonte. Y supongamos que un cilindro de hielo APQB de la misma anchura que la cavidad del vaso y con el mismo eje desciende con movimiento uniforme y continuo y que sus partes en cuanto tocan la superficie AB se licúan y convertidas en agua discurren en el vaso por su gravedad y forman al caer una catarata o columna de agua ABNFEM y atravesando el orificio EF lo llenan completamente. La velocidad uniforme del hielo que cae y la del agua contigua en el círculo AB sea la misma que la que tendría el agua cayendo y describiendo en su caída la altura IH; y estén IH y HG sobre la misma línea, y por el punto I trácese la recta KL paralela al horizonte y que toca a los lados del hielo en K y L. Y la velocidad del agua que cae por el orificio EF será aquella que adquiriría el agua cayendo desde I y describiendo en su caída la altura IG. Y por consiguiente, por los teoremas de *Galileo*, IG será a IH como el cuadrado de la velocidad del agua que cae por el orificio a la velocidad del agua en el círculo AB, es decir, como el cuadrado del círculo AB al círculo EF; pues estos círculos son inversamente como las velocidades de las aguas que llenándolos completamente pasan a su través en el mismo tiempo y en igual cantidad. Se trata aquí de la velocidad del agua hacia el plano horizontal. Y los movimientos paralelos al horizonte con los cuales se aproximan mutuamente las partes del agua que cae, al no proceder de la gravedad, no cambiarán el movimiento perpendicular al horizonte surgido de la gravedad, y no se consideran, por tanto, aquí. Ciertamente suponemos que las partes de agua cohesionan entre sí un tanto y que, mientras caen, por esa cohesión, se aproximan unas a otras mediante movimientos

horizontales de modo que constituyan una sola catarata y no se dividan en muchas. Pero el movimiento paralelo al horizonte, proveniente de la susodicha cohesión, no lo consideramos aquí.



Caso 1. Ahora supongamos que toda la cavidad del vaso alrededor del agua que cae ABNFEM está llena de hielo, de modo que el agua caiga a través del hielo como si fuera a través de un embudo. Si el agua ni siquiera tocase al hielo, o lo que es lo mismo, si lo toca pero, por lo sumamente pulido de la superficie del hielo, resbala libremente y sin resistencia alguna, caerá por el orificio EF con la misma velocidad que antes, y todo el peso de la columna de agua ABNFEM presionará para generar su caída como antes, mientras el fondo del vaso soportará el peso del hielo que rodea a la columna.



Ahora licúese el hielo del vaso, y el flujo del agua, en cuanto a la velocidad, permanecerá como antes. No será menor, porque el hielo hecho agua tratará de salir: tampoco mayor, porque el hielo hecho agua no puede descender si no es impidiendo el descenso de otra agua igual a su propio descenso. La misma fuerza debe generar la misma velocidad del agua que fluye.

Pero el orificio del fondo del vaso, debido a los movimientos oblicuos de las partículas del agua que fluye, debe ser algo mayor que antes. Pues las partículas de agua ya no pasan todas por el orificio discurriendo perpendicularmente, sino que,

concurriendo por doquier desde los costados del vaso y convergiendo sobre el orificio, lo atraviesan con movimientos oblicuos, y curvando su trayectoria hacia abajo confluyen en la vena de agua que cae, que es algo más delgada bajo el orificio que en el orificio mismo, siendo su diámetro, respecto al diámetro del orificio como 5 a 6, o como 5½ a 6½ aproximadamente, si he medido bien los diámetros. Preparé una lámina plana y muy delgada perforada en el centro, siendo el orificio circular de un diámetro de cinco octavos de pulgada. Para que la vena de agua fluyente no se acelerase al caer y al acelerarse se hiciese más estrecha, coloqué esta lámina, no en el fondo sino en un costado del vaso, de tal modo que la susodicha vena brotase según una línea paralela al horizonte. Después, una vez lleno el vaso de agua, abrí el orificio para que saliese el agua; y el diámetro de la vena, a una distancia de casi media pulgada del orificio, medido con la máxima exactitud, vino a ser de $^{21}/_{40}$ partes de pulgada. Por tanto, el diámetro del orificio circular era al diámetro de la vena de agua aproximadamente como 25 a 21. Por consiguiente, el agua, al atravesar el orificio, converge desde todas partes, y después de salir del vaso, se hace más estrecha con la convergencia, y por el estrechamiento se hace más rápida, hasta alcanzar una distancia de media pulgada del orificio, y a esa distancia se torna más acelerada y estrecha que en el orificio mismo en la razón de 25 x 25 a 21 x 21 o de 17 a 12 aproximadamente, es decir, aproximadamente como la raíz cuadrada de dos a la unidad. Pero consta experimentalmente que la cantidad de agua que fluye en un tiempo dado por el orificio circular efectuado en el fondo del vaso es aquella que con la velocidad antedicha debería fluir en el mismo tiempo, no por aquel orificio, sino por el orificio circular cuyo diámetro es al diámetro de aquel orificio como 21 a 25. Y por tanto, la susodicha agua que cae tiene una velocidad hacia abajo en el mismo orificio que es la misma que adquiriría un grave cayendo y describiendo en su caída la semialtura del agua contenida en el vaso muy aproximadamente. Pero, después de salir del vaso, se acelera convergiendo hasta que llega a una distancia del orificio casi igual al diámetro del orificio, y adquiere una velocidad mayor casi en la proporción de la raíz cuadrada de dos a uno; velocidad que, ciertamente, casi adquiriría un grave cayendo y recorriendo en su caída la altura total del agua contenida en el vaso.

En lo sucesivo, por tanto, designemos al diámetro de la vena mediante aquel orificio menor, al cual hemos llamado EF. Y supóngase que se traza un plano VW encima y paralelo al del orificio EF y a una distancia igual al diámetro del orificio, dotado también de un orificio ST mayor que el primero, y así caiga por él una vena de agua que llene completamente el orificio inferior EF y que además tenga un diámetro de una proporción respecto al del orificio inferior como 25 a 21 aproximadamente. De este modo la vena de agua pasará perpendicularmente por el orificio inferior, y la cantidad de agua fluyente, según el tamaño de este orificio, será aproximadamente aquella que exige la solución del problema. En cambio, el espacio comprendido entre los dos planos y la vena de agua que cae puede considerarse como fondo del vaso. Pero para que la solución del problema sea más simple y más

matemática, es más conveniente considerar solamente al plano inferior como fondo del vaso y suponer que el agua, que caía por el hielo o por el embudo y salía del vaso por el orificio EF practicado en el plano inferior, conserva continuamente su movimiento, mientras el hielo está en reposo. Por tanto, sea en lo sucesivo ST el diámetro del orificio descrito con centro en Z a cuyo través fluye del vaso la catarata cuando el agua toda del vaso es fluida. Y sea EF el diámetro del orificio que el chorro, al pasar, llena completamente, tanto si el agua sale del vaso a través del orificio superior ST, como si cae a través del hielo a modo de embudo. Y sea el diámetro del orificio superior ST al del inferior EF aproximadamente como 25 a 21, y la distancia perpendicular entre los planos de los orificios sea igual al diámetro del orificio menor EF. En tal caso, la velocidad del agua que sale del vaso por el orificio ST será en el mismo orificio hacia abajo la misma que la que adquiriría un cuerpo cayendo desde la mitad de la altura IZ: mientras que la velocidad de la catarata que cae de ambos será en el orificio EF la que adquiriría un cuerpo que cayese de la altura total IG.

CASO 2. Si el orificio EF no estuviese en el centro del fondo del vaso, sino que el fondo estuviese perforado en otro lugar, el agua saldrá con la misma velocidad que antes, siempre que el orificio sea del mismo tamaño. Pues, un grave desciende ciertamente en un tiempo mayor hasta la misma profundidad siguiendo una línea oblicua que siguiendo la perpendicular, pero al bajar adquiere la misma velocidad en uno y otro caso, como demostró *Galileo*.

CASO 3. La misma es la velocidad del agua que fluye por un orificio abierto en un costado del vaso. Pues si el orificio fuese pequeño, de modo que la distancia entre las superficies AB y KL sea imperceptible para los sentidos y la vena de agua que sale horizontalmente venga a tener una forma parabólica, del «latus rectum» de esta parábola se infiere que la velocidad del agua que discurre será aquella que adquiriría un cuerpo al caer desde la altura HG o IG del agua retenida en el vaso. Efectivamente, hecho el experimento, hallé que si la altura del agua retenida en el vaso era de veinte pulgadas sobre el orificio y la altura del orificio sobre el plano horizontal era también de 20 pulgadas, la vena de agua que fluía caía sobre dicho plano a la distancia de casi 37 pulgadas de la perpendicular trazada desde el orificio sobre el plano horizontal. Pues, sin resistencia, la vena debería haber caído sobre el plano a una distancia de 40 pulgadas, siendo el «latus rectum» de la vena parabólica de 80 pulgadas.

CASO 4. E incluso si el agua fluyente se orienta hacia arriba, saldrá con la misma velocidad. Pues la pequeña vena de agua fluyente asciende con un movimiento perpendicular hasta la altura GH o GI del agua retenida en el vaso, salvo en la medida en que su ascenso es un tanto impedido por la resistencia del aire; y por tanto, sale con la velocidad que adquiriría cayendo desde dicha altura. Cada partícula del agua retenida sufre igual presión por todas partes (por la Proposición XIX del Libro II) y al ceder a la presión, se desplaza hacia todas partes con igual impulso, tanto si cae por el

orificio en el fondo del vaso, como si fluye horizontalmente a través de un orificio lateral, o si penetra en un canal y después asciende por un pequeño orificio abierto en la parte superior del canal. Que la velocidad con que fluye el agua es la que se ha establecido en esta Proposición no sólo se infiere racionalmente, sino que también se evidencia mediante los bien conocidos experimentos que acabamos de describir.

Caso 5. La velocidad del flujo de agua es la misma tanto si la forma del orificio es circular como si es cuadrada, triangular u otra cualquiera igual a la circular. Porque la velocidad del agua fluyente no depende de la figura del orificio sino que surge de su altura por debajo del plano KL.

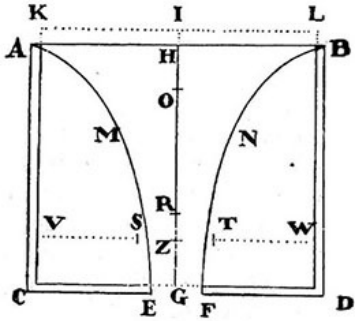
Caso 6. Si la parte inferior del vaso ABDC se sumerge en agua estancada y la altura del agua estancada sobre el fondo del vaso fuera GR, la velocidad con la que el agua que está en el vaso saldría por el orificio EF hacia el agua estancada sería la misma que la que adquiriría el agua cayendo desde la altura IR y recorriéndola en su caída. Pues el peso de toda el agua del vaso que está por debajo de la superficie del agua estancada, será sostenido en equilibrio por el peso del agua estancada, y por ello en nada acelerará el movimiento del agua que desciende en el vaso. También puede mostrarse este caso mediante experimentos, midiendo el tiempo en que el agua fluye.

Corolario 1. Por ello, si la altura CA del agua se prolonga hasta K, de modo que AK sea a CK como el cuadrado de la razón del área de un orificio abierto en cualquier parte del fondo al área del círculo AB; la velocidad del agua fluyente será igual a la velocidad que adquiriría el agua cayendo y recorriendo en su caída la altura KC.

Corolario 2. Y la fuerza con la cual se puede generar el movimiento total del agua fluyente es igual al peso de la columna cilíndrica de agua cuya base es el orificio EF y cuya altura es 2GI o 2CK. Pues el agua que sale, en el tiempo en que iguala a dicha columna, adquiriría cayendo por su propio peso desde la altura GI la velocidad con que sale.

Corolario 3. El peso de toda el agua contenida en el vaso ABDC es a la parte del peso que presiona para que salga el agua como la suma de los círculos AB y EF al doble del círculo EF. Pues sea IO la media proporcional entre IH e IG; y el agua que sale por el orificio EF en el tiempo en que una gota cayendo desde I puede recorrer la altura IG, es igual a un cilindro cuya base es el círculo EF y cuya altura es 2IG, es decir, a un cilindro cuya base es el círculo AB y cuya altura es 210, pues el círculo EF es al círculo AB como la raíz cuadrada de la altura IH a la de la altura IG, es decir, en razón simple de la media proporcional IO a la altura IG; y en el tiempo en que la gota cayendo desde I describiría la altura IH, el agua que sale será igual a un cilindro cuya base es el círculo AB y cuya altura es 2IH; y en el tiempo en que la gota cayendo desde I por H hasta G describe la diferencia de alturas HG, el agua que sale, es decir, toda el agua contenida en el sólido ABNFEM, será igual a la diferencia de los cilindros, es decir, al cilindro cuya base es AB y cuya altura es 2HO. Y por consiguiente, toda el agua contenida en el vaso ABDC es al agua toda que cae

contenida en el sólido ABNFEM como HG a 2HO, es decir, como HO + OG a 2HO, o IH + IO a 2IH. Pero el peso de toda el agua contenida en el sólido ABNFEM presiona para que salga el agua; y por tanto, el peso de toda el agua contenida en el vaso es a la parte de peso que presiona para que salga el agua como IH + IO a 2H, y por tanto, como la suma de los círculos EF y AB al doble del círculo EF.



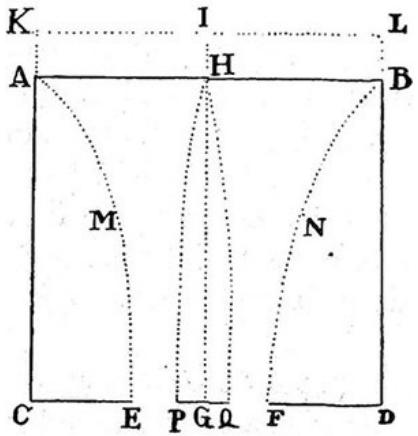
Corolario 4. Y por ello el peso de toda el agua contenida en el vaso ABDC es a aquella parte del peso que soporta el fondo del vaso, como la suma de los círculos AB y EF a la diferencia de los propios círculos.

Corolario 5. Y la parte de peso soportada por el fondo del vaso es a la parte de peso que presiona sobre la salida del agua, como la diferencia de los círculos AB y EF al doble del círculo menor, o como el área del fondo al doble del orificio.

Corolario 6. Mas, la parte de peso que presiona solamente sobre el fondo es al peso de toda el agua que descansa perpendicularmente sobre el fondo como el círculo AB a la suma de los círculos AB y EF, o como el círculo AB al exceso del duplo del círculo AB sobre el fondo. Pues la parte de peso, que presiona sola sobre el suelo, es al peso de toda el agua del vaso, como la diferencia de los círculos AB y EF a la suma de esos círculos, por el Corolario 4; y el peso de toda el agua contenida en el vaso es al peso de toda el agua que descansa perpendicularmente sobre el fondo como el círculo AB a la diferencia de los círculos AB y EF. Y por tanto, permutando adecuadamente, la parte de peso que sólo urge sobre el fondo, es al peso de toda el agua que descansa sobre el fondo perpendicularmente, como el círculo AB a la suma de los círculos AB y EF o exceso del duplo del círculo AB sobre el fondo.

Corolario 7. Si en el centro del orificio EF se coloca el pequeño círculo PQ, descrito en torno al centro G y paralelo al horizonte, el peso del agua sostenida por dicho pequeño círculo es mayor que el peso de la tercera parte del cilindro de agua cuya base fuese el pequeño círculo y la altura GH. Pues, sea ABNFEM la catarata o columna de agua que cae con eje GH como más arriba, e imagínese que toda el agua contenida en el vaso se congelase, tanto en torno de la catarata como sobre el pequeño círculo, por no ser necesaria su fluidez para un descenso inmediato y rápido del agua. Y sea PHQ la columna de agua congelada sobre el pequeño círculo, con

vértice en H y altura GH. Imagínese que semejante catarata cae con todo su peso, y no se apoya sobre PHQ ni presiona en absoluto, sino que libremente y sin fricción se va deslizando, salvo acaso en el vértice mismo del hielo donde la catarata al comenzar a caer puede ser cóncava. Y, puesto que, el agua congelada AMEC, ANFD que rodea a la catarata es convexa en su superficie interior AME, BNF hacia la catarata que cae, también la columna PHQ será convexa hacia la catarata, y por tanto, mayor que el cono cuya base es el pequeño círculo PQ y la altura GH, es decir, mayor que la tercera parte del cilindro descrito con tales base y altura. Sostiene, pues, dicho pequeño círculo el peso de esta columna, es decir, un peso que es mayor que el peso del cono o mayor que la tercera parte de aquel cilindro.



Corolario 8. El peso de agua que el círculo muy pequeño PQ soporta parece ser menor que el peso de las dos terceras partes del cilindro de agua cuya base es dicho círculo y la altura HG. Pues, supuesto lo que hemos dicho antes, imagínese que se describe un semiesferoide cuya base es el susodicho pequeño círculo y cuyo semieje o altura es HG. Esta figura será igual a dos terceras partes de aquel cilindro y comprenderá a la columna de agua congelada PHQ cuyo peso sostiene el pequeño círculo. Pues, como el movimiento del agua sea directo en gran medida, la superficie externa de la dicha columna concurrirá con la base PQ en ángulo un tanto agudo, toda vez que el agua se acelera continuamente al caer y, debido a esa aceleración se va haciendo más estrecha. Por tanto, puesto que el ángulo es menor que un recto, las partes inferiores de la columna estarán dentro del semiesferoide. Y en sus partes superiores también será aguda o puntiaguda, pues de lo contrario el movimiento horizontal del agua habría de ser infinitamente más veloz en el vértice que su movimiento hacia el plano horizontal. Y cuanto menor sea el círculo PQ, más agudo será el vértice de la columna, y si el círculo disminuye hasta el infinito, el ángulo PHQ disminuirá hasta el infinito, por lo cual la columna estará dentro del semiesferoide. Por consiguiente, la columna es menor que el semiesferoide, o que las

dos terceras partes del cilindro cuya base es el pequeño círculo y cuya altura es GH. Así el pequeño círculo soporta una fuerza de agua igual al peso de la columna, mientras el peso del agua que la rodea presiona para que el agua salga por el orificio.

Corolario 9. El peso de agua que soporta el pequeño círculo PQ muy reducido es aproximadamente igual al peso de un cilindro de agua cuya base fuese dicho pequeño círculo y su altura ½GH. Pues este peso es una media aritmética de los pesos del cono y el semiesferoide mencionado más arriba. Pero si el pequeño círculo no fuese muy reducido, sino más bien aumentase hasta igualar al orificio EF, soportaría el peso de toda el agua que reposa perpendicularmente sobre él, es decir, el peso del cilindro de agua cuya base es el pequeño círculo y cuya altura es GH.

Corolario 10. Y (hasta donde me es posible precisarlo) el peso que soporta este pequeño círculo es siempre al peso de un cilindro de agua cuya base sea el pequeño círculo y su altura ½GH como EF^2 es a EF^2 - ½PQ^2, o como el círculo EF al exceso de dicho círculo sobre la mitad del pequeño círculo, muy aproximadamente.

Lema iv

Si un cilindro se desplaza uniformemente en la dirección de su longitud, la resistencia que sufre no cambia en absoluto aumentando o disminuyendo la longitud y, por tanto, es igual a la resistencia que sufriría un círculo descrito con el mismo diámetro y que se desplazase a la misma velocidad según una línea recta perpendicular a su plano.

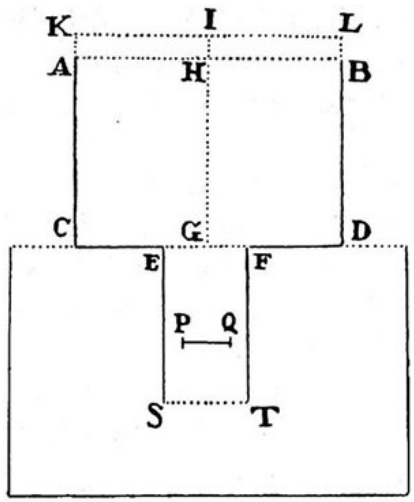
Pues los lados no se oponen en absoluto al movimiento y un cilindro se transforma en un círculo cuando su longitud disminuye hasta el infinito.

Proposición xxxvii. Teorema xxix[13]

Si un cilindro se desplaza uniformemente según la dirección de su longitud en un fluido comprimido, ilimitado y no elástico, la resistencia debida a la magnitud de su sección transversal es a la fuerza con la cual se generaría o anularía todo su movimiento en el tiempo en que recorre el cuádruplo de su longitud como la densidad del medio a la densidad del cilindro muy aproximadamente.

Pues sea ABDC un vaso que toca con su fondo CD la superficie del agua estancada; hágase que el agua salga de dicho vaso hacia el agua estancada por el canal cilíndrico EFTS, perpendicular al horizonte, sitúese el pequeño círculo PQ paralelo al horizonte en una posición cualquiera en medio del canal y prolónguese CA hasta K de modo que AK sea a CK como el cuadrado del exceso del orificio del canal

EF sobre el pequeño círculo PQ al círculo AB. Y es evidente (por el Caso 5, el Caso 6 y el Corolario 1 de la Proposición XXXVI) que la velocidad del agua que pasa por el anillo comprendido entre el pequeño círculo y los costados del recipiente será aquella que adquiriría el agua cayendo y describiendo en su caída la altura KC o IG.



Y (por el Corolario 10 de la Proposición XXXVI) si la anchura del vaso es infinita, de modo que la breve línea HI desaparece y las alturas IG, HG se hacen iguales, la fuerza del agua que cae sobre el pequeño círculo será al peso del cilindro cuya base es el mismo pequeño círculo y cuya altura es ½IG, como EF^2 a EF^2 - ½PQ^2 muy aproximadamente. Pues la fuerza del agua que fluye con movimiento uniforme por todo el canal será siempre igual sobre el pequeño círculo en cualquier parte del canal en que se halle colocado.

Ciérrense ahora los orificios EF, ST del canal, hágase ascender al pequeño círculo en el fluido comprimido por doquier y obligue con su ascenso al agua de arriba a descender por el espacio anular comprendido entre el pequeño círculo y los costados del canal: y la velocidad de ascenso del pequeño círculo será a la velocidad de descenso del agua como la diferencia entre los círculos EF y PQ al círculo PQ, y la velocidad de ascenso del pequeño círculo a la suma de las velocidades, es decir, a la velocidad relativa de descenso del agua con la cual discurre al lado del pequeño círculo ascendente, como la diferencia de los círculos EF y PQ al círculo EF, o sea, como EF^2 - PQ^2 a EF^2. Sea dicha velocidad relativa igual a la velocidad con la cual se vio más arriba que el agua pasaba por el mismo espacio anular mientras el pequeño círculo permanecía inmóvil, es decir, a la velocidad que adquiriría el agua cayendo y describiendo en su caída la altura IG: en tal caso, la fuerza del agua sobre el pequeño círculo ascendente será la misma que antes (por el Corolario V de estas Leyes), es

decir, la resistencia del pequeño círculo ascendente será al peso del cilindro de agua cuya base sea el pequeño círculo y su altura ½IG como EF^2 a EF^2 - ½PQ^2 muy aproximadamente. Pero la velocidad del pequeño círculo será a la velocidad que adquiriría el agua cayendo y describiendo con su caída la altura IG, como EF^2 - PQ^2 a EF^2.

Auméntese la anchura del canal hasta el infinito, y las susodichas razones entre EF^2 - PQ^2 y EF^2 y entre EF^2 y EF^2 - ½PQ^2 llegarán en último término a ser razones de igualdad. Y, por consiguiente, la velocidad del pequeño círculo será ahora aquella que adquiriría el agua cayendo y describiendo con su caída la altura IG, mientras su resistencia vendrá a ser igual al peso del cilindro cuya base sea el pequeño círculo y su altura la mitad de la altura IG, desde la cual deberá caer el cilindro para adquirir la velocidad con que asciende el pequeño círculo; y a esta velocidad, en el tiempo de su caída, describirá el cuádruplo de su longitud. En cambio, la resistencia del cilindro, desplazándose a esta velocidad en la dirección de su longitud, es la misma que la resistencia del pequeño círculo (por el Lema IV) y en consecuencia, es igual a la fuerza con la cual se generaría su movimiento mientras describe el cuádruplo de su longitud muy aproximadamente.

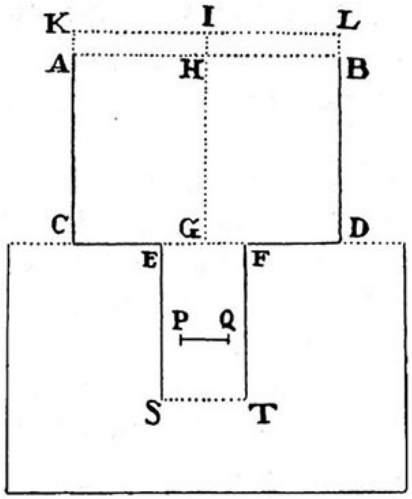
Si la longitud del cilindro se aumentase o disminuyese, su movimiento al igual que el tiempo en que describe el cuádruplo de su longitud, aumentarán o disminuirán en la misma proporción; y por tanto, la fuerza con la cual se generaría o suprimiría el movimiento aumentado o disminuido, en tiempo correlativamente aumentado o disminuido, no sufrirá cambio alguno; y por ende, también ahora es igual a la resistencia del cilindro, pues también esta permanece sin cambios según el Lema IV.

Si la densidad del cilindro aumenta o disminuye, su movimiento, al igual que la fuerza con la cual en el mismo tiempo se generaría o suprimiría el movimiento, aumentará o disminuirá en la misma proporción. Por consiguiente, la resistencia de todo el cilindro será a la fuerza con la cual se generaría o se suprimiría todo su movimiento en el tiempo que tarda en recorrer el cuádruplo de su longitud, como la densidad del medio a la densidad del cilindro muy aproximadamente. Q. E. D.

Pero un fluido ha de ser comprimido para que sea continuo, y debe ser continuo y no elástico para que toda la presión debida a su compresión pueda propagarse instantáneamente y no cambie la resistencia actuando para ello de modo igual sobre todas las partes del cuerpo movido. La presión debida al movimiento del cuerpo se consume en generar el movimiento de las partes del fluido y crea la resistencia. Pero la presión que surge de la compresión del fluido, aunque sea fuerte, si se propaga instantáneamente, no genera movimiento alguno en las partes de un fluido continuo, no provoca cambio alguno de movimiento; y por ello ni aumenta ni disminuye la resistencia. Ciertamente, la acción del fluido debida a su compresión no puede ser en las partes posteriores más fuerte que en las anteriores de un cuerpo movido, y por ello la resistencia descrita en esta Proposición no puede disminuirse: y no será más fuerte en las partes anteriores que en las posteriores, siempre que la propagación de la

misma sea infinitamente más rápida que el movimiento del cuerpo presionado. Pero será infinitamente más rápida y se propagará instantáneamente siempre que el fluido sea continuo y no elástico.

COROLARIO 1. Las resistencias que sufren cilindros que se desplazan en la dirección de sus longitudes uniformemente y en medios infinitos están en razón compuesta del cuadrado de las velocidades, del cuadrado de los diámetros y de la densidad del medio.



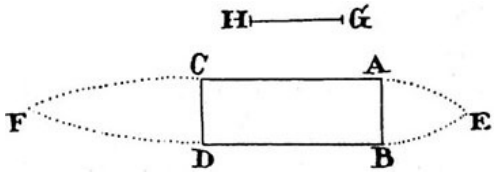
COROLARIO 2. Si la anchura del canal no se aumenta hasta el infinito, sino que el cilindro avanza según la dirección de su longitud en el medio contenido en reposo, coincidiendo además el eje del cilindro con el eje del canal, su resistencia vendrá a ser, respecto a la fuerza con la cual se generaría o suprimiría todo su movimiento, en el tiempo en que describe el cuádruplo de su longitud, como la razón compuesta de la razón EF^2 a EF^2 - ½PQ^2, del cuadrado de la razón EF^2 a EF^2 - PQ^2 y de la razón de la densidad del medio a la densidad del cilindro.

COROLARIO 3. Con los mismos supuestos y también que la longitud L sea al cuádruplo de la longitud del cilindro en la razón compuesta de la razón de EF^2 - ½PQ^2 a EF^2 y del cuadrado de la razón de EF^2 - PQ^2 a EF^2, la resistencia del cilindro será a la fuerza con la que se generaría o suprimiría todo su movimiento en el tiempo en que describe la longitud L, como la densidad del medio a la densidad del cilindro.

ESCOLIO

En esta Proposición hemos investigado la resistencia que procede solamente de la

magnitud de la sección transversal del cilindro, prescindiendo de la parte de resistencia que puede surgir de la oblicuidad de los movimientos. Ya que, al igual que en el Caso 1 de la Proposición XXXVI, la oblicuidad de los movimientos, con los cuales las partes del agua en el vaso convergían de todos lados sobre el orificio EF, impedía la salida del agua por el orificio, del mismo modo en esta Proposición, la oblicuidad de los movimientos con los cuales las partes del agua presionadas por el extremo anterior del cilindro ceden a la presión y divergen en todas direcciones, retardan sus traslaciones por los alrededores de dicho extremo anterior hacia las partes posteriores del cilindro y hace que el fluido se remueva hasta distancias mayores y aumente la resistencia, y esto casi en la misma razón en la que disminuye la salida de agua del vaso, es decir, casi como el cuadrado de la razón de 25 a 21. Y, al igual que en el caso primero de la susodicha Proposición, hacíamos que las partes de agua pasasen a caño lleno y perpendicularmente por el orificio EF, suponiendo que toda el agua del vaso en torno a la catarata estaba congelada, y cuyo movimiento oblicuo e inútil permanecía inmóvil, del mismo modo en esta Proposición, para eliminar la oblicuidad de los movimientos, y para que las partes del agua, cediendo con un movimiento directo y breve, faciliten al máximo el tránsito del cilindro y únicamente permanezca la resistencia debida a la magnitud de la sección transversal, resistencia que sólo puede disminuir disminuyendo el diámetro del cilindro, habremos de suponer que las partes del fluido cuyos movimientos oblicuos e inútiles crean resistencia, están en reposo entre sí en ambos extremos del cilindro, cohesionadas entre sí y unidas al cilindro. Sea ABCD un rectángulo y sean AE y BE dos arcos parabólicos descritos con eje AB y con un «latus rectum» tal que sea al espacio HG, que habría de ser descrito por el cilindro en su caída para adquirir la velocidad con que se mueve, como HG a ½AB. Y sean también CF y DF otros dos arcos parabólicos descritos con eje CD y con «latus rectum» cuatro veces mayor que el anterior; y por revolución de la figura en torno al eje EF se generará un sólido cuya parte intermedia ABDC será el cilindro del que tratamos, y las partes extremas ABE y CDF contienen las partes del fluido en reposo entre sí, y condensadas en dos cuerpos rígidos adheridos al cilindro por ambas partes como si fueran cabeza y cola. En tal caso, la resistencia de este sólido EACFDB, desplazándose según la longitud de su eje FE en dirección a E, será aproximadamente la misma que la que hemos determinado en esta Proposición, es decir, aquella que estaría respecto a la fuerza con la cual se generaría o suprimiría todo el movimiento del cilindro en el tiempo en que con dicho movimiento uniformemente continuo recorrería la longitud 4AC, como la densidad del fluido a la densidad del cilindro, muy aproximadamente. Y, por el Corolario 7 de la Proposición XXXVI, esta resistencia no puede ser menor respecto a dicha fuerza que la razón de 2 a 3.



Lema v

Si una esfera, un cilindro y un esferoide de igual anchura se colocan sucesivamente en medio del canal cilíndrico de tal modo que el eje de los mismos coincida con el eje del canal, dichos cuerpos impedirán por igual el flujo del agua a través del canal.

Pues los espacios entre el canal y el cilindro, la esfera y el esferoide, a través de los cuales pasa el agua, son iguales: y el agua pasa igual por espacios iguales.

Esto es así bajo la hipótesis de que toda el agua que está sobre el cilindro, sobre la esfera o sobre el esferoide y cuya fluidez no es necesaria para que el agua pase con toda rapidez, se congele, como ya expuse en el Corolario 7 de la Proposición xxxvi.

Lema vi

Con los mismos supuestos, los susodichos cuerpos son igualmente urgidos por el agua que fluye por el canal.

Se sigue del Lema v y de la tercera Ley. Pues el agua y los cuerpos actúan entre sí mutua e igualmente.

Lema vii

Si el agua del canal está en reposo y los susodichos cuerpos se desplazan por el canal a velocidades iguales y en direcciones opuestas, sus resistencias serán iguales entre sí.

Se sigue del Lema anterior, pues los movimientos relativos siguen siendo iguales entre sí.

Escolio

El mismo es el caso para los cuerpos todos convexos y redondos cuyos ejes

coinciden con el eje del canal. Alguna diferencia podría surgir de la fricción mayor o menor, pero en estos Lemas suponemos que los cuerpos son sumamente pulidos y que la tenacidad del medio y la fricción son nulas y también que las partes del fluido, que con sus movimientos oblicuos y superfluos pudieran perturbar, impedir y retardar el flujo del agua por el canal, reposan entre ellas como constreñidas por la congelación y pegadas a las partes anteriores y posteriores de los cuerpos, tal y como he expuesto en el Escolio de la Proposición anterior. Pues en lo que sigue se trata de la resistencia mínima que pueden tener los cuerpos redondos descritos con una sección transversal máxima dada.

Los cuerpos flotantes en fluidos, cuando se desplazan en línea recta, hacen que el fluido ascienda por la parte delantera y descienda por la parte posterior, sobre todo si las figuras son obtusas; y por ello, padecen una resistencia algo mayor que si la cabeza y la cola fuesen agudas. Y los cuerpos que se mueven en fluidos elásticos, si son obtusos por delante y por detrás, condensan el fluido un poco más por delante y lo relajan algo más por la parte de atrás; y por ello sufren una resistencia algo mayor que si fuesen agudos por la cabeza y por la cola. Pero nosotros en estos Lemas no tratamos de fluidos elásticos, sino de los no elásticos, tampoco de los cuerpos que flotan en los fluidos, sino de los hundidos en los mismos. Y cuando se conoce la resistencia de los cuerpos en fluidos no elásticos, hay que aumentar esta resistencia un poco, tanto para los fluidos elásticos, como el aire, como para las superficies de fluidos estancados, como los mares y los lagos.

Proposición xxxviii. Teorema xxx

La resistencia de un globo, que se desplaza uniformemente en un fluido comprimido, infinito y no elástico, es a la fuerza, que generaría o suprimiría todo su movimiento en el mismo tiempo en que recorre ocho terceras partes de su diámetro, como la densidad del fluido a la densidad del globo muy aproximadamente.

Pues el globo es al cilindro circunscrito como dos a tres; y por tanto aquella fuerza que suprimiese todo el movimiento del cilindro en el tiempo en que éste recorre una longitud de cuatro diámetros, suprimirá todo el movimiento del globo en el tiempo en que éste recorra dos terceras partes de dicha longitud, es decir, ocho terceras partes del propio diámetro. Pero la resistencia del cilindro es a esta fuerza muy aproximadamente como la densidad del fluido a la densidad del cilindro o del globo por la Proposición xxxvii, y la resistencia del globo es igual a la resistencia del cilindro, por los Lemas v, vi y viii. Q. E. D.

Corolario 1. Las resistencias de los globos, en medios comprimidos infinitos, están en razón compuesta del cuadrado de la velocidad, del cuadrado del diámetro y de la densidad de los medios.

COROLARIO 2. La máxima velocidad con la cual un globo puede descender por su peso relativo en un fluido resistente es aquella que puede adquirir el mismo globo, con el mismo peso, cayendo sin resistencia y describiendo en su caída un espacio que sea a las cuatro terceras partes de su diámetro como la densidad del globo a la densidad del fluido. Pues el globo en el tiempo de su caída, con la velocidad adquirida al caer, describirá un espacio que resultará ser a ocho tercios de su diámetro, como la densidad del globo a la densidad del fluido; y la fuerza del peso generadora de este movimiento será a la fuerza que generaría este mismo movimiento en el tiempo en que el globo recorrería las ocho terceras partes de su diámetro a la misma velocidad, como la densidad del fluido a la densidad del globo: y por consiguiente, por esta Proposición, la fuerza del peso será igual a la fuerza de resistencia y, por ello, no puede acelerar al globo.

COROLARIO 3. Dadas la densidad del globo y su velocidad al principio del movimiento, así como la densidad del fluido comprimido en reposo en que el globo se mueve, están dados en todo momento, tanto la velocidad del globo como su resistencia, al igual que el espacio descrito por él, por el Corolario 7 de la Proposición XXXV.

COROLARIO 4. Un globo moviéndose en un fluido comprimido en reposo y de su misma densidad, perderá la mitad de su movimiento antes de que haya recorrido una longitud de dos diámetros propios, por el mismo Corolario 7.

PROPOSICIÓN XXXIX. TEOREMA XXXI

La resistencia de un globo que se desplaza por un fluido encerrado en un canal cilíndrico y uniformemente comprimido, es a la fuerza con la cual se generaría o suprimiría todo su movimiento en el tiempo en que recorrería ocho terceras partes de su diámetro, como la razón compuesta de la razón del orificio del canal al exceso de este orificio sobre la mitad del círculo máximo del globo, del cuadrado de la razón del orificio del canal al exceso de este orificio sobre el círculo máximo del globo y de la razón de la densidad del fluido a la densidad del globo, muy aproximadamente.

Es evidente por el Corolario 2 de la Proposición XXXVII, mientras que la demostración sigue los pasos de la demostración de la Proposición anterior.

ESCOLIO

En las dos últimas Proposiciones (lo mismo que en el Lema V) he supuesto que toda el agua que precede al globo y cuya fluidez aumenta la resistencia del globo, se halla congelada. Pero si toda dicha agua está líquida, aumentará la resistencia un

tanto. Aunque dicho aumento en estas Proposiciones será pequeño y puede despreciarse, dado que la superficie convexa del globo produce casi todo el efecto del hielo.

PROPOSICIÓN XL. PROBLEMA IX

Hallar a partir de los fenómenos la resistencia de un globo que se desplaza en un medio comprimido muy fluido.

Sea A el peso del globo en el vacío, B su peso en un medio resistente, D el diámetro del globo, F el espacio que sea a ⅔D como la densidad del globo a la densidad del medio, es decir, como A a A - B, G el tiempo en el cual el globo cayendo sin resistencia con el peso B describe el espacio F, y H la velocidad que adquiere el globo con esta su caída. Y H será la velocidad máxima con que el globo, con su peso B, puede descender en un medio resistente, por el Corolario 2 de la Proposición XXXVIII, y la resistencia que sufre el globo, descendiendo con dicha velocidad, será igual a su peso B, mientras que la resistencia que sufre a otra velocidad cualquiera será al peso B como la razón cuadrada de esta velocidad a la susodicha velocidad máxima H, por el Corolario 1 de la Proposición XXXVIII.

Esta es la resistencia que se debe a la inercia de la materia del fluido. Pero la que se debe a la elasticidad, tenacidad y fricción de sus partes se hallará como sigue.

Déjese caer el globo de modo que descienda en el fluido por su peso B; y sea P el tiempo de caída y sea medido en segundos si el tiempo G está dado en segundos. Hállese el número absoluto N correspondiente al logaritmo 0,4342944819 $\frac{2P}{G}$, y sea L el logaritmo del número $\frac{N+1}{N}$, y la velocidad adquirida al caer será $\frac{N-1}{N+1}$H, y la altura descrita será $\frac{2PF}{G}$ - 1,3862943611F + 4,605170186LF. Si el fluido es suficientemente profundo, se puede despreciar el término 4,605170186LF, y la altura descrita será muy aproximadamente la de $\frac{2PF}{G}$ - 1,3862943611F. Todo esto se sigue de la Proposición novena del Libro segundo y de sus Corolarios en la hipótesis de que el globo no sufra otra resistencia que la debida a la inercia de la materia. Pero si sufriese otra resistencia además de ésta, el descenso será más lento, y de la retardación se inferirá la cantidad de esa resistencia.

Para conocer más fácilmente la velocidad y el descenso de un cuerpo que cae en un

fluido, he compuesto la siguiente tabla, cuya primera columna expresa los tiempos de caída, la segunda las velocidades adquiridas en la caída, siendo la velocidad máxima 100000000, la tercera representa los espacios descritos al caer en dichos tiempos, siendo 2F el espacio descrito por el cuerpo en el tiempo G y con la velocidad máxima, y la cuarta representa los espacios descritos en los mismos tiempos máxima. Los números de la cuarta columna son $\frac{2PF}{G}$, y sustrayendo el número 1,3862944 - 4,6051702L, se hallarán los números de la tercera columna, números estos que hay que multiplicar por el espacio F para hallar los espacios descritos al caer. A estas anteriores se ha añadido una quinta columna que contiene los espacios descritos por el cuerpo en estos mismos tiempos cayendo en el vacío con la fuerza de su peso de comparación B.

Tiempos P	*Velocidades de caída en el fluido*	*Espacios descritos cayendo en el fluido*	*Espacios descritos con movimiento máximo*	*Espacios descritos cayendo en el vacío*
0,001G	99999 29/30	0,000001F	0,002F	0,000001F
0,01G	999967	0,0001F	0,02F	0,0001F
0,1G	9966799	0,0099834F	0,2F	0,01F
0,2G	19737532	0,0397361F	0,4F	0,04F
0,3G	29131261	0,0886815F	0,6F	0,09 F
0,4G	37994896	0,1559070F	0,8F	0,16F
0,5G	46211716	0,2402290F	1,0F	0,25F
0,6G	53704957	0,3402706F	1,2F	0,36F
0,7G	60436778	0,4545405F	1,4F	0,49 F
0,8G	66403677	0,5815071F	1,6F	0,64F
0,9G	71629787	0,7196609F	1,8F	0,81F
1G	76159416	0,8675617F	2F	1F
2G	96402758	2,6500055F	4F	4F
3G	99505475	4,6186570F	6F	9F
4G	99932930	6,6143765F	8F	16F
5G	99990920	8,6137964F	10F	25F
6G	99998771	10,6137179F	12F	36F
7G	99999834	12,6137073F	14F	49 F
8G	99999980	14,6137059F	16F	64F
9G	99999997	16,6137057F	18F	81F
10G	99999999 3/5	18,6137056F	20F	100F

Escolio

Con objeto de investigar mediante experimentos las resistencias de los fluidos me procuré un vaso de madera cuadrado y con longitud y latitud interiores de nueve pulgadas de pie *londinense* y con una profundidad de nueve pies y medio, y lo llené de agua de lluvia; y con globos hechos de cera conteniendo plomo, anoté los tiempos de descenso de los globos, siendo la altura del descenso de 112 pulgadas. El pie

cúbico sólido *londinense* contiene 76 libras *romanas* de agua de lluvia, y la pulgada sólida cúbica de este pie contiene $^{19}/_{36}$ onzas de esa libra, o también 253⅓ granos; y un globo de agua de una pulgada de diámetro 132,645 granos en el aire, o también 132,8 granos en el vacío; y otro globo cualquiera será como el exceso de su peso en el vacío sobre su peso en el agua.

Experimento 1. Un globo, cuyo peso era de 156¼ granos en el aire y de 77 granos en el agua describió la altura completa de 112 pulgadas en un tiempo de cuatro segundos. Y repetido el experimento, el globo de nuevo cayó en el mismo tiempo de cuatro segundos.

El peso del globo en el vacío es de $156^{13}/_{38}$ granos, y el exceso de este peso sobre el peso del globo en el agua es de $79^{13}/_{18}$ granos. De ello se sigue que el diámetro del globo es de 0,84224 partes de pulgada. Pero la densidad del agua a la densidad del globo es como dicho exceso al peso del globo en el vacío, igual que las ocho terceras partes del diámetro del globo (es decir, 2,24597 pulgadas) al espacio 2F, que será, por tanto, de 4,4256 pulgadas.

El globo en un segundo, con todo su peso de $156^{13}/_{38}$ granos cayendo en el vacío, describe 193⅓ pulgadas; y en el tiempo G, que es a un segundo como la raíz cuadrada del espacio F, es decir, como 2,2128 pulgadas a 95,219 pulgadas, describirá 2,2128 pulgadas, y adquirirá la velocidad máxima H con la que puede descender en el agua. Por tanto, el tiempo G es de 0,15244 segundos. Y en este tiempo G, con la susodicha velocidad máxima H, el globo describirá el espacio 2F de 4,4256 pulgadas; y por tanto, en un tiempo de 4 segundos describirá un espacio de 116,1245 pulgadas. Réstese el espacio de 1,3862944F o de 3,0676 pulgadas, y quedará un espacio de 113,0569 pulgadas, que describirá el globo cayendo en agua en un vaso muy ancho y en un tiempo de 4 segundos. Este espacio, por la estrechez del vaso de madera antedicho, debe disminuirse en una razón que está compuesta de la razón de la raíz cuadrada del orificio del vaso al exceso de este orificio sobre el semicírculo máximo del globo y de la razón simple del mismo orificio a su exceso sobre el círculo máximo del globo, o sea, en la razón de 1 a 0,9914. Hecho esto, se tendrá un espacio de 112,08 pulgadas, que según la teoría debería haber descrito el globo cayendo en el agua en dicho vaso de madera en un tiempo de 4 segundos, muy aproximadamente. Y según el experimento describió 112 pulgadas.

Experimento 2. Tres globos iguales, cuyos pesos eran de 76⅓ granos cada uno en el aire, y de $5^{1}/_{16}$ granos en el agua, se dejaron caer sucesivamente; y cada uno cayó en el agua en el tiempo de 15 segundos, describiendo en su caída la altura de 112 pulgadas.

Haciendo un cálculo, resulta un peso del globo en el vacío de $76^{5}/_{12}$ granos, el exceso de este peso sobre el peso en el agua de $71^{17}/_{48}$ granos, el diámetro del globo de 0,81296 pulgadas, las ocho terceras partes de este diámetro 2,16789 pulgadas, el espacio 2F de 2,3217 pulgadas, el espacio que describirá el globo cayendo con un

peso de $5^{1}/_{16}$ granos sin resistencia en el tiempo de un segundo será de 12,808 pulgadas, y el tiempo G será 0″,301056. Por tanto, el globo, con la máxima velocidad que puede recibir de la fuerza de un peso de $5^{1}/_{16}$ granos y con la cual descendería, recorrerá un espacio de 2,3217 pulgadas en el tiempo de 0″,301056, y en un tiempo de 15″, recorrerá un espacio de 115,678 pulgadas. Réstese el espacio 1,3862944F, o sea, 1,609 pulgadas y quedará un espacio de 114,069 pulgadas, que debería recorrer el globo en el mismo tiempo al caer en un vaso muy ancho. Por la estrechez de nuestro vaso debe restarse un espacio de aproximadamente 0,895 pulgadas. Y entonces quedará un espacio de 113,174 pulgadas, que es lo que, según la teoría, debió recorrer cayendo en este vaso en un tiempo de 15 segundos. Pero describió según el experimento 112 pulgadas. La diferencia es inapreciable.

EXPERIMENTO 3. Tres globos iguales, cuyos pesos eran de 121 granos cada uno en el aire, y de 1 grano en el agua, se dejaron caer sucesivamente; y caían en el agua en tiempos de 46, 47 y 50 segundos, describiendo la altura de 112 pulgadas.

Según la teoría estos globos debieron caer en un tiempo de 40 segundos aproximadamente. Creo que no es determinable por qué cayeron más lentamente, si fue debido a la menor proporción de resistencia que se sigue de la fuerza de inercia en los movimientos lentos respecto a la resistencia que surge de otras causas, o si más bien ha de atribuirse a algunas burbujas adheridas al globo, o a la rarefacción de la cera por el calor tanto del ambiente como de la mano que dejaba caer el globo, o quizá también a errores insensibles al pesar los globos en agua. Por todo ello el peso del globo en agua debe ser de varios granos para que el experimento sea cierto y digno de confianza.

EXPERIMENTO 4. Inicié los experimentos descritos hasta aquí, con el fin de investigar las resistencias de los fluidos, antes de elaborar la teoría expuesta en las Proposiciones recién expuestas. Después, para examinar la teoría recién hallada, me procuré un vaso de madera de una anchura interna de $8^{2}/_{3}$ pulgadas y de una profundidad de 15⅓ pies. A continuación construí cuatro globos de cera con plomo incluido, cada uno de un peso de 139¼ granos en el aire, y de 7⅛ granos en el agua. Y los deje caer de modo que pudiera medir los tiempos de caída en el agua mediante un péndulo oscilante ajustado a oscilaciones de medios segundos. Los globos, mientras los pesaba y después caían, estaban fríos y permanecían fríos durante algún tiempo; porque el calor enrarece un tanto la cera y por esa rarefacción el peso del globo en el agua disminuye, y la cera enrarecida no se reduce instantáneamente por el frío a la densidad anterior. Antes de dejarlos caer los sumergía completamente en el agua, para que su descenso inicial no se viese acelerado por el peso de alguna parte situada fuera del agua. Y cuando estaban completamente sumergidos y en reposo, eran soltados con todo cuidado para que no recibiesen impulso alguno procedente de la mano que los soltaba. Cayeron, pues, sucesivamente en los tiempos de 47½, 48½, 50 y 51 oscilaciones, describiendo una altura de 15 pies y 2 pulgadas. Pero el ambiente estaba un poco más frío que cuando se pesaban los globos, de modo que

repetí el experimento otro día, y los globos cayeron en tiempos de 49, 49½, 50 y 53 oscilaciones, y un tercero en tiempos de 49½, 50, 51 y 53 oscilaciones. Y repetido varias veces el experimento, los globos cayeron en la mayor parte de las veces en tiempos de 49½ y 50 oscilaciones. Cuando cayeron más lentamente sospecho que fueron retardados por los choques con los costados del vaso.

Haciendo ahora el cálculo según la teoría, resulta un peso del globo en el vacío de 139⅖ granos. El exceso de este peso sobre el peso del globo en el agua de $132^{11}/_{40}$ granos. El diámetro del globo de 0,99868 pulgadas. Las ocho terceras partes del diámetro 2,66315 pulgadas. El espacio 2F de 2,8066 pulgadas. El espacio que describe el globo cayendo sin resistencia con el peso de 7⅛ granos en el tiempo de un segundo de 9,88164 pulgadas. Y el tiempo G de 0″,376843. Por lo tanto, el globo, con la velocidad máxima con que puede descender en el agua con la fuerza del peso de 7⅛ granos en el tiempo de 0″,376843 describe un espacio de 2,8066 pulgadas, y en el tiempo de 1″ un espacio de 7,44766 pulgadas, y en el tiempo de 25″, o de 50 oscilaciones, un espacio de 186,1915 pulgadas. Réstese el espacio 1,386294F, o 1,9454 pulgadas, y quedará un espacio de 184,2461 pulgadas, que describirá el globo en el mismo tiempo en un vaso muy ancho. Por la estrechez de nuestro vaso disminúyase este espacio en una razón que esté compuesta ce la raíz cuadrada de la razón del orificio del vaso al exceso de este orificio sobre el semicírculo máximo del globo, y de la razón simple del mismo orificio a su exceso sobre el círculo máximo del globo; y se tendrá un espacio de 181,86 pulgadas, que debió describir el globo aproximadamente, según la teoría, en el tiempo de las 50 oscilaciones. Y describió, según el experimento, 182 pulgadas en el tiempo de 49½ o de 50 oscilaciones.

EXPERIMENTO 5. Cuatro globos de 154⅜ granos de peso en el aire y de 21½ granos en el agua, soltados muchas veces, caían en tiempos de 28½, 29, 29½ y 30 oscilaciones, y algunas veces de 31, 32 y 33, describiendo una altura de 15 pies y 2 pulgadas.

Según la teoría debieron caer en un tiempo de 29 oscilaciones muy aproximadamente.

EXPERIMENTO 6. Cinco globos de 212⅜ granos de peso en el aire y de 79½ en el agua, soltados muchas veces, caían en tiempos de 15, 15½, 16, 17 y 18 oscilaciones, describiendo una altura de 15 pies y 2 pulgadas.

Según la teoría debieron caer en un tiempo muy aproximado al de 15 oscilaciones.

EXPERIMENTO 7. Cuatro globos de 293⅜ granos de peso en el aire y de 35⅞ granos en el agua, soltados muchas veces, caían en tiempos de 29½, 30, 30½, 31, 32 y 33 oscilaciones describiendo una altura de 15 pies y 1½ pulgadas.

Según la teoría debieron caer en un tiempo muy aproximado al de 28 oscilaciones.

Indagando la causa de que globos del mismo peso y magnitud cayeran unos más lentos y otros más rápidos, di con la siguiente; los globos, tan pronto como se

soltaban y empezaban a caer, oscilaban en torno a sus centros, cayendo por delante el lado que por casualidad fuese más pesado, generando un movimiento de oscilación. Pero, el globo por sus oscilaciones comunica al agua un movimiento mayor que si descendiese sin oscilar; y al comunicarlo pierde una parte de su propio movimiento con el cual debía descender: y según la mayor o menor oscilación, se retarda más o menos. Pero además, el globo siempre se desvía del lado que desciende por oscilación, y desviándose se aproxima a los costados del vaso y, a veces, choca con ellos. Y esta oscilación es más fuerte en los globos más graves, y en los más grandes agita más el agua. Por lo cual, para que la oscilación de los globos fuese menor, construí nuevos globos de cera y plomo, incrustando el plomo en un costado del globo cerca de su superficie; y solté el globo de tal forma que el lado más grave en la medida de lo posible, estuviese debajo desde el principio del descenso. Así, las oscilaciones resultantes son mucho menores que antes, y los globos cayeron en tiempos mucho menos desiguales, como en los experimentos siguientes.

EXPERIMENTO 8. Cuatro globos de 139 granos de peso en el aire y 6½ en el agua, soltados muchas veces, cayeron en tiempos de no más de 52 oscilaciones y no menos de 50, y la mayor parte de las veces en tiempos de casi 51, describiendo una altura de 182 pulgadas.

Según la teoría debieron caer en un tiempo de casi 52 oscilaciones.

EXPERIMENTO 9. Cuatro globos de 273¼ granos en el aire y de 140¾ en el agua, soltados muchas veces, cayeron en tiempos, no menores de 12 oscilaciones y no mayores de 13, describiendo una altura de 182 pulgadas.

Según la teoría estos globos debieron caer en un tiempo muy aproximado a 11⅓ oscilaciones.

EXPERIMENTO 10. Cuatro globos de un peso de 384 granos en el aire y 119½ en el agua, soltados muchas veces, caían en tiempos de 17⅐, 18, 18½ y 19 oscilaciones, describiendo una altura de 181½ pulgadas. Y cuando cayeron en un tiempo de 19 oscilaciones, pude oír algunas veces su golpear sobre los costados del vaso antes de que llegasen al fondo.

Pero según la teoría debieron caer en un tiempo de $15\frac{5}{9}$ oscilaciones muy aproximadamente.

EXPERIMENTO 11. Tres globos iguales de un peso de 48 granos en el aire y $3\frac{29}{32}$ en el agua, soltados muchas veces, cayeron en tiempos de 43½, 44, 44½, 45 y 46 oscilaciones, y la mayor parte de las veces 44 y 45 oscilaciones, describiendo una altura de 182½ pulgadas aproximadamente.

Según la teoría debieron caer en un tiempo de $46\frac{5}{9}$ oscilaciones muy aproximadamente.

EXPERIMENTO 12. Tres globos iguales de un peso de 141 granos en el aire y de 4⅜ en el agua, soltados algunas veces, cayeron en tiempos de 61, 62, 63, 64 y 65 oscilaciones, describiendo una altura de 182 pulgadas.

Y según la teoría debieron caer en un tiempo de 64½ oscilaciones muy aproximadamente.

Con estos experimentos se pone en evidencia que cuando los globos cayeron más lentamente, como en los Experimentos segundo, cuarto, quinto, octavo, undécimo y duodécimo, la teoría representa correctamente los tiempos de caída; pero cuando los globos cayeron más velozmente, como en los Experimentos sexto, noveno y décimo, se da una resistencia algo mayor que en razón del cuadrado de la velocidad. Pues los globos, mientras caen, oscilan un tanto; y esta oscilación en los globos más ligeros y que caen más lentamente cesa muy pronto, por ser los movimientos más débiles; pero en los más graves y mayores dura más tiempo por ser los movimientos más fuertes, y solamente tras muchas oscilaciones puede ser detenido por el agua que los rodea. Además, cuanto más veloces son los globos tanto menos son presionados en su parte posterior por el fluido; y si la velocidad se aumentase continuamente, al final acabarían por dejar a sus espaldas un espacio vacío, salvo que la compresión del fluido se aumentase a la vez. Pues la compresión del fluido (por las Proposiciones XXXII y XXXIII) debe aumentarse en razón del cuadrado de la velocidad para que la resistencia siga en la misma razón cuadrada. Al no ocurrir esto, los globos más veloces son presionados por detrás un poco menos, y por la falta de esta presión, su resistencia se torna un poco mayor que en razón del cuadrado de la velocidad.

Por consiguiente, la teoría se ajusta a los fenómenos de caídas de cuerpos en el agua, y sólo falta examinar los fenómenos de las caídas en el aire.

EXPERIMENTO 13. En la ciudad de *Londres* en el mes de junio del año 1710 desde la cúpula de la iglesia de *San Pablo* se soltaron dos globos a la vez, uno lleno de mercurio y el otro lleno de aire; y al caer describían una altura de 220 pies *londinenses*. Una tabla de madera estaba suspendida por uno de sus extremos por unas bisagras de hierro y por el otro se apoyaba en un soporte de madera; y los dos globos colocados encima de la tabla se dejaban caer a la vez, tirando del soporte de madera mediante un cable de hierro lanzado hasta el suelo, de suerte que la tabla sostenida solamente por las bisagras de hierro girase sobre ellas, y en el mismo instante un péndulo que oscilaba al segundo con un tirón sobre el susodicho cable de hierro la hacía caer y empezaba a oscilar. Los diámetros y los pesos de los globos y los tiempos de caída se exponen en la tabla siguiente.

Por lo demás, los tiempos observados deben corregirse. Pues los globos de mercurio (por la teoría de *Galileo*) describirán en cuatro segundos 257 pies *londinenses*, y 220 pies en sólo 3″ 42‴. Efectivamente la tabla de madera, al quitar el soporte, giraba más lentamente que lo que se requería y con la lentitud de su giro impedía la caída de los globos en el inicio de la misma. Pues los globos reposaban casi en el centro de la tabla y hasta estaban un poco más cerca del eje de la misma que del soporte. Y por eso, los tiempos de caída se alargaron casi 18 minutos terceros, y ahora hay que corregir restando dichos minutos terceros, sobre todo en los globos mayores que por causa de su mayor diámetro reposaban un poco más sobre la tabla

giratoria. Una vez hecho esto, los tiempos en que cayeron los seis globos mayores resultan ser 8″ 12‴, 7″ 42‴, 7″ 42‴, 7″ 57‴, 8″ 12‴ y 7″ 42‴.

Globos llenos de mercurio			Globos llenos de aire		
Pesos	Diámetros	Tiempos de caída	Pesos	Diámetros	Tiempos de caída
granos	*pulgadas*	*segundos*	*granos*	*pulgadas*	*segundos*
908	0,8	4	510	5,1	8½
983	0,8	4-	642	5,2	8
866	0,8	4	599	5,1	8
747	0,75	4+	515	5,0	8¼
808	0,75	4	483	5,0	8½
784	0,75	4+	641	5,2	8

Por tanto, el quinto de los globos llenos de aire, construido con un diámetro de cinco pulgadas y un peso de 483 granos, cayó en un tiempo de 8″ 12‴, describiendo una altura de 220 pies. El peso del agua igual a este globo es de 16 600 granos; y el peso del aire igual al dicho globo es de $\frac{16600}{860}$ granos o también de $19\frac{3}{6}$ granos y, por tanto, el peso del globo en el vacío es de $502\frac{3}{10}$ granos, y este peso es al peso del aire igual al globo como $502\frac{3}{10}$ a $19\frac{3}{10}$, y como 2F a las ocho terceras partes del diámetro del globo, es decir, a 13⅓ pulgadas. De donde 2F resulta ser de 28 pies y 11 pulgadas. El globo cayendo en el vacío con todo su peso de $502\frac{3}{10}$ granos en el tiempo de un segundo describirá 193⅓ pulgadas, como antes, y con el peso de 483 granos describirá 185,905 pulgadas, y con el mismo peso de 483 granos también describirá en el vacío el espacio F, o 14 pies y 5½ pulgadas en el tiempo de 57‴ 58′′′, y adquiere la máxima velocidad con que puede descender en el aire. Con esta velocidad el globo, en el tiempo de 8″ 12‴, describirá un espacio de 245 pies y 5⅓ pulgadas. Réstese 1,3863F, o 20 pies y ½ pulgada, y quedarán 225 pies y 5 pulgadas. Por lo tanto, el globo, en un tiempo de 8″ 12‴ debió describir en su caída, según la teoría, el susodicho espacio. Y describió, según el experimento, un espacio de 220 pies. La diferencia es inapreciable.

Aplicando cálculos análogos a los otros globos llenos de aire, construí la tabla siguiente.

Pesos de los globos	Diámetros	Tiempo de caída desde 220 pies		Espacios a describir según teoría		Excesos sobre cálculo	
granos	*pulgadas*	*segundos*	*terceros*	*pies*	*pulgadas*	*pies*	*pulgadas*
510	5,1	8	12	226	11	6	11
642	5,2	7	42	230	9	10	9
599	5,1	7	42	227	10	7	10
515	5	7	57	224	5	4	5
483	5	8	12	225	5	5	5
641	5,2	7	42	230	7	10	7

EXPERIMENTO 14. En el año 1719, en el mes de julio, el *Sr. Desaguliers* de nuevo inició experimentos de éstos construyendo esferas con vejigas de puercos mediante

una esfera cóncava de madera en la que se introducían humedecidas y luego tenían que llenar inflándolas con aire; una vez secas y extraídas se las dejó caer del punto más alto de la linterna de la cúpula del susodicho templo, o sea de una altura de 272 pies, y a la vez se soltaba un globo de plomo de casi dos libras *romanas*. Mientras tanto, algunas personas situadas en lo más alto del templo, donde los globos eran soltados, anotaban los tiempos totales de caída, a la vez que otras personas en el suelo anotaban la diferencia de tiempos entre la caída del globo de plomo y la caída de la vejiga. Los tiempos se medían mediante péndulos que oscilaban en medios segundos. De los que estaban en el suelo, uno tenía un reloj con muelle que vibraba cuatro veces por segundo; otro tenía otra máquina de péndulo hecha con precisión que también vibraba cuatro veces por segundo. Y una máquina similar tenía uno de los que estaban en lo alto del templo. Y estas máquinas estaban hechas de modo que sus movimientos se iniciaban o se paraban a voluntad. Pues bien, el globo de plomo caía en un tiempo de cuatro segundos y cuarto aproximadamente. Añadiendo este tiempo a la diferencia antedicha de tiempos, se obtenía el tiempo total de caída de la vejiga. Los tiempos en los cuales cayeron las cinco vejigas la primera vez después de la caída del globo de plomo eran, 14¾ segundos, 12¾ segundos, 14⅝ segundos, 17¾ segundos y 16⅞ segundos, mientras que en la segunda vez fueron de 14½ segundos, 14¼ segundos, 14 segundos, 19 segundos y 16¾ segundos. Añádase el tiempo de cuatro segundos y cuarto en el cual cayó el globo de plomo y los tiempos totales en los cuales las cinco vejigas vinieron a caer fueron, en la primera vez 19, 17, 18⅞, 22 y 21⅛ segundos respectivamente, y en la segunda vez 18¾, 18½, 18¼, 23¼ y 21 segundos. Los tiempos observados en lo alto del templo eran la primera vez de 19¾, 17¼, 18¾, 22⅛ y 21⅝ segundos, mientras que la segunda vez eran de 19, 18⅝, 18⅜, 24 y 21¼ segundos. Por lo demás, las vejigas no caían siempre en línea recta, sino que algunas veces revoloteaban y oscilaban de aquí para allá mientras caían. Y con estos movimientos los tiempos de caída se alargaron y aumentaron, a veces, en medio segundo y, a veces, en un segundo completo. Caían más directamente las vejigas segunda y cuarta la primera vez; y la primera y la tercera la segunda vez. La vejiga quinta era rugosa y por causa de las rugosidades se retardaba un tanto. Obtenía los diámetros de las vejigas de sus circunferencias medidas con un hilo muy fino que daba dos vueltas. Y comparé la teoría con los experimentos en la tabla siguiente, suponiendo que la densidad del aire es a la del agua de lluvia como 1 a 860 y calculando los espacios que los globos debían recorrer en su caída según la teoría.

Peso de las vejigas	Diámetro de las vejigas	Tiempo de caída desde 272 pies	Espacios a recorrer según teoría		Diferencias entre teoría y experimentos	
granos	*pulgadas*	*segundos*	*pies*	*pulgadas*	*pies*	*pulgadas*
128	5,28	19	271	11	-0	1
156	5,19	17	272	0½	+0	0½
137½	5,3	18½	272	7	+0	7
97½	5,26	22	277	4	+5	4
99⅛	5	21⅛	282	0	+10	0

Por tanto, la resistencia de los globos en movimiento, tanto en el aire como en el agua, es adecuada y correctamente representada por nuestra teoría y es proporcional a la densidad de los fluidos a iguales velocidades y magnitudes de los globos.

En el Escolio subsiguiente a la Sección sexta hemos mostrado mediante los experimentos con péndulos que las resistencias de globos iguales e igualmente veloces en el aire, en el agua, y en el mercurio son como las densidades de los fluidos. Aquí mostramos lo mismo de modo más preciso mediante experimentos con cuerpos que caen en el aire y en el agua. Pues los péndulos en cada oscilación suscitan un movimiento en el fluido siempre contrario al movimiento de regreso del péndulo, y con la resistencia debida a este movimiento así como con la resistencia del hilo del que se suspendía el péndulo, la resistencia total del mismo venía a ser mayor que la resultante de los experimentos de cuerpos cayendo. Pues por los experimentos de péndulos expuestos en dicho Escolio, un globo de la misma densidad que el agua, describiendo en el aire la longitud de su movimiento, sólo tendría que perder de su movimiento $\frac{1}{3342}$ parte. Pero según la teoría expuesta en esta Sección séptima confirmada con los experimentos de cuerpos que caen, el mismo globo describiendo la misma longitud sólo debería perder de su movimiento $\frac{1}{4586}$ parte, en el supuesto de que la densidad del agua fuese a la densidad del aire como 860 a 1. Por consiguiente, las resistencias resultaron mayores por los experimentos de los péndulos (por las causas ya dichas) que por los experimentos de los globos que caen, y esto casi en una razón de 4 a 3. Sin embargo, puesto que las resistencias de los péndulos oscilantes en aire, agua o mercurio se ven incrementadas igualmente por iguales causas, la proporción de las resistencias en tales medios se podrá representar con bastante exactitud tanto mediante experimentos con péndulos como con experimentos con cuerpos que caen. Y de ahí se puede concluir que la resistencia de los cuerpos en cualesquiera fluidos, incluso en los más fluidos, siendo iguales las demás circunstancias, son como las densidades de los fluidos.

Una vez establecido esto, se puede ahora tratar de qué parte de su movimiento perderá aproximadamente un globo en un tiempo dado, si se proyecta a través de un fluido cualquiera. Sea D el diámetro del globo, V su velocidad al inicio del movimiento, y T el tiempo en el cual el globo, con la velocidad V, describirá en el vacío un espacio que sea al espacio $^{8}/_{3}$D como la densidad del globo a la densidad del fluido: y el globo proyectado en dicho fluido, en otro tiempo cualquiera *t*, perderá la parte $\frac{tV}{T+t}$, manteniendo la parte $\frac{TV}{T+t}$, y describirá un espacio que será al descrito

en el mismo tiempo con velocidad uniforme en el vacío como el logaritmo del número $\frac{T+t}{T}$, multiplicado por el número 2,302585093, es al número $\frac{t}{T}$, por el Corolario 7 de la Proposición XXXV. En los movimientos lentos la resistencia puede ser algo menor, por cuanto la figura del globo resulte algo más apta para el movimiento que la de un cilindro descrito con el mismo diámetro. En movimientos veloces, la resistencia puede ser algo mayor, toda vez que la elasticidad y la compresión del fluido no aumentan en razón del cuadrado de la velocidad. Pero no me detengo en estos detalles.

Y aunque el aire, el agua, el mercurio y otros fluidos semejantes, por división hasta el infinito de sus partes, se hiciesen más sutiles y viniesen a ser medios infinitamente fluidos, no por eso resistirían menos a los globos proyectados. Pues la resistencia de que se trata en las Proposiciones anteriores procede de la inercia de la materia; y la inercia de la materia es esencial a los cuerpos y proporcional siempre a la cantidad de materia. Ciertamente se puede disminuir la resistencia que procede de la tenacidad y fricción de las partes mediante la división de las partes de un fluido: pero la cantidad de materia no disminuye por la división de sus partes; y manteniéndose la cantidad de materia, se mantiene su fuerza de inercia a la cual es siempre proporcional la resistencia de la que aquí se trata. Para que dicha resistencia disminuya debe disminuir la cantidad de materia en los espacios a cuyo través se mueven los cuerpos. Y por eso, los espacios celestes, por los que los globos de los planetas y los cometas se mueven siempre libremente en todas direcciones y sin ninguna disminución sensible de movimiento, carecen de todo fluido corpóreo, salvo acaso algunos vapores sumamente tenues y los rayos de luz.

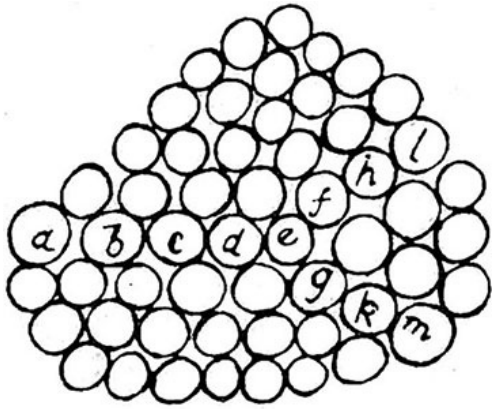
También los proyectiles al desplazarse suscitan movimiento en los fluidos, y este movimiento surge del exceso de presión del fluido en las partes anteriores del proyectil sobre la presión en sus partes posteriores y no puede ser menor en medios infinitamente fluidos que en el aire, en el agua, y en el mercurio según la densidad de la materia de cada uno. Ahora bien, este exceso de presión, en razón de su cantidad, no sólo suscita movimiento en el fluido, sino que también actúa sobre el proyectil retardando su movimiento: y por eso, la resistencia en todo fluido es como el movimiento suscitado en el fluido por el proyectil, y no puede ser menor en el éter más sutil en razón de la densidad del éter, que en el aire, en el agua, o en el mercurio, en razón de las densidades de dichos fluidos.

Sección VIII
SOBRE EL MOVIMIENTO QUE SE PROPAGA POR LOS FLUIDOS

Proposición XLI. Teorema XXXII

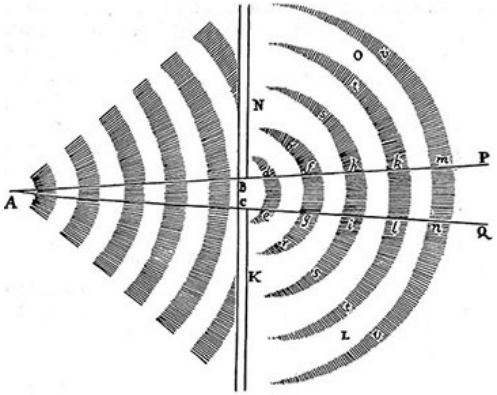
La presión no se propaga a través de un fluido según líneas rectas salvo cuando las partículas del fluido están colocadas directamente en tal sentido.

Si las partículas *a, b, c, d, e*, se hallasen colocadas en línea recta, la presión puede, ciertamente, propagarse directamente desde *a* hasta *e*; pero la partícula *e* urgirá a las partículas colocadas oblicuamente *f* y *g* de modo oblicuo, y dichas partículas *f* y *g* no mantendrán la presión recibida, salvo que sean soportadas por las partículas *h* y *k*; pero en tanto en que son sostenidas, presionan sobre las que sostienen; y éstas, a su vez, no soportarán la presión a no ser que sean sostenidas por las siguientes *l* y *m* y las presionen también, y así sucesivamente hasta el infinito; y una vez que empieza a propagarse oblicuamente, si incidiere en partículas ulteriores que no se hallan colocadas en línea recta, de nuevo divergirá; y esto tantas veces cuantas incidiese sobre partículas no situadas exactamente en línea recta. Q. E. D.



Corolario. Si una parte cualquiera de la presión que se propaga desde un punto

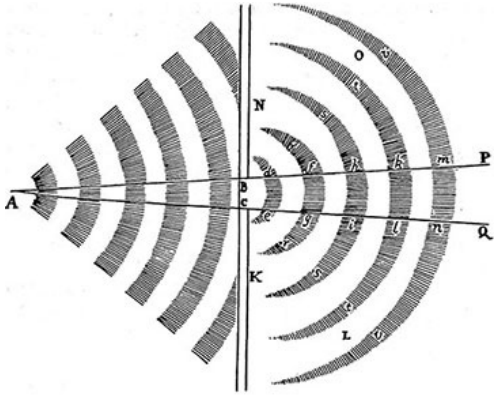
dado fuese obstruida por un obstáculo, la parte restante que no es interceptada, divergirá por el espacio ulterior al obstáculo. Cosa que puede también demostrarse como sigue. Sea una presión que se propaga desde el punto A hacia cualquier parte y ello, si es posible, según líneas rectas, y con el obstáculo NBCK, perforado en BC, obstrúyase toda ella, salvo la parte coniforme APQ, que pasa por el orificio circular BC. Sepárese en troncos el cono APQ mediante los planos transversales *de, fg, hi*; en tal caso, mientras el cono ABC, propagando la presión, impulsa al tronco de cono ulterior *degf* sobre la superficie *de*, y el tronco anterior empuja al tronco siguiente *fghi* sobre la superficie *fg*, y dicho tronco empujará sobre el tercer tronco, y así sucesivamente hasta el infinito; está claro (por la Ley tercera del Movimiento) que el primer tronco *defg*, debido a la reacción del tronco segundo *fghi*, sufrirá una presión sobre la superficie *fg* igual a la presión que efectúa sobre el dicho tronco segundo. Luego el tronco *degf* es comprimido por ambos lados entre el cono A*de* y el tronco *fhig*, y por consiguiente (por el Corolario 6 de la Proposición XIX) no puede conservar su figura a no ser que sea comprimido por todas partes con la misma fuerza. Luego intentará ceder hacia los lados *df, eg* con un ímpetu igual a aquel con que es comprimido en las superficies *de, fg;* y en esos lugares (al no ser rígido, sino completamente fluido) se saldrá y se expandirá, salvo que el fluido de entorno sea tal que impida este intento. Por lo cual, con este intento de salir, presionará sobre el fluido de entorno tanto sobre los lados *df, eg* como sobre el tronco *fghi* con idéntico ímpetu; y por ello, la presión no se propagará menos hacia los lados *df, eg* por ambos espacios NO, KL que lo que se propague desde la superficie *fg* hacia PQ. Q. E. D.



Proposición XLII. Teorema XXXIII

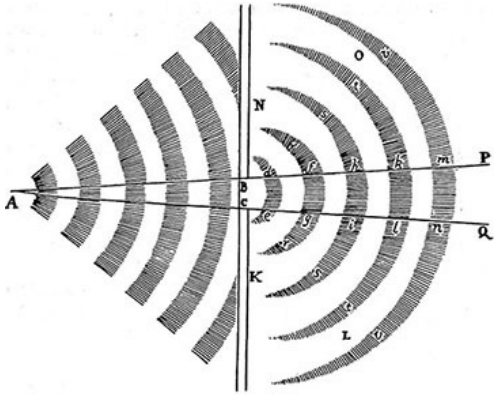
Todo movimiento propagado a través de un fluido diverge de la trayectoria recta en espacios inmóviles.

CASO 1. Propáguese el movimiento desde el punto A a través del orificio BC, y vaya, si es posible, hacia el espacio cónico BCQP según líneas rectas divergentes desde el punto A. Y supongamos en primer lugar que este movimiento fuese de ondas en la superficie de agua estancada. Y sean *de, fg, hi, kl*, etc., las crestas de cada una de las ondas, separadas entre sí por los valles correspondientes. En tal caso, puesto que el agua en las crestas de las ondas está más alta que en las partes en reposo LK, NO del fluido, descenderá desde las crestas *e, g, i, l*, etc., *d, f, h, k*, etc., por un lado y por otro hacia KL, NO; y también descenderá desde dichos lugares en reposo hacia los valles de las ondas. Con el primero de los descensos las crestas de las ondas, y con el segundo los valles se expanden hacia uno y otro lado y se propagan hacia KL y NO. Y puesto que el movimiento de las ondas desde A hacia PQ ocurre mediante el descenso continuo de las crestas hacia los valles siguientes, y por lo mismo, no ocurre con mayor velocidad que la velocidad de descenso; y también el descenso del agua por un lado y por otro hacia KL y NO debe ocurrir con esa misma velocidad; y así la dilatación de las ondas por uno y otro lado hacia KL y NO avanzará con la misma velocidad con que avanza la propia onda desde A hacia PQ en línea recta. Por tanto, todo el espacio de ambos lados hacia KL y NO será ocupado por las ondas dilatadas *rfgr, skis, tklt, vmnv*, etc. Q. E. D. Cualquiera puede experimentar en agua estancada que esto es así.



CASO 2. Ahora supongamos que *de, fg, hi, kl, mn*, representan impulsos propagados desde el punto A a través de un medio elástico sucesivamente. Imaginemos que los impulsos se propagan por medio de sucesivas condensaciones y rarefacciones del medio, como si la parte más densa del impulso ocupase la superficie esférica descrita en torno al centro A, y entre impulsos sucesivos mediasen intervalos iguales. Las líneas *de, fg, hi, kl*, etc., representan las partes más densas de los impulsos, propagadas a través del orificio BC. Y puesto que es allí el medio más denso que en los espacios laterales hacia KL y NO, se dilatará tanto hacia esos espacios KL y NO por ambos lados como hacia los intervalos más enrarecidos de las

pulsaciones; de este modo el medio que va resultando cada vez más raro tras los intervalos y más denso tras las pulsaciones, participará de sus movimientos. Y dado que el movimiento progresivo de los impulsos procede de la relajación continua de las partes más densas que anteceden a los valles más enrarecidos; y los impulsos deben relajarse hacia ambos lados de las partes del medio en reposo KL y NO casi con la misma velocidad; y dichos impulsos se dilatarán casi con la misma velocidad por todas partes en los espacios inmóviles KL y NO, con la que se propagan en directo desde el centro A; por tanto, ocuparán todo el espacio KLON. Q. E. D. Esto lo experimentamos en los sonidos, que se oyen tanto con un monte interpuesto como, si entran en una habitación por una ventana y se extienden por todas partes de la habitación, se oyen en todas las esquinas, y no por reflexión en las paredes opuestas, sino propagándose directamente desde la ventana; en tanto que es posible juzgar por los sentidos.



CASO 3. Finalmente, supongamos que un movimiento de cualquier clase se propaga desde A a través del orificio BC: y puesto que esta propagación no ocurre si no es en la medida en que las partes del medio más cercanas al centro A urgen y agitan a las partes siguientes, y las partes que son agitadas son fluidas y, por tanto, retroceden por doquier hacia las regiones en que son menos presionadas: tales partes retrocederán hacia las partes del medio que están todas en reposo, tanto las laterales KL y NO, como las frontales PQ, y de este modo todo el movimiento, tan pronto como haya pasado por el orificio BC, empezará a dilatarse y a propagarse desde allí directamente hacia todas partes, como si fuera desde un principio y centro. Q. E. D.

PROPOSICIÓN XLIII. TEOREMA XXXIV

Todo cuerpo vibrante en un medio elástico propagará el movimiento de sus pulsaciones en directo hacia todas partes; en cambio, en un medio no elástico,

excitará un movimiento circular.

CASO 1. Por cuanto que las partes del cuerpo vibrante que van y vienen alternativamente, con su ida urgirán y empujarán a las partes del medio que están inmediatas a ellas, y al urgirías las presionarán y las condensarán; y después con su retorno dejarán a las partes presionadas retroceder y expandirse. Por tanto, las partes del medio próximas al cuerpo vibrante irán y regresarán alternativamente, a la par de las partes del dicho cuerpo vibrante: y por la misma razón que las partes de este cuerpo agitaban a estas partes del medio, éstas a su vez perturbadas por vibraciones semejantes agitarán a las partes inmediatas a ellas, y éstas agitadas igualmente agitarán a las siguientes y así, sucesivamente, hasta el infinito. Y al igual que las primeras partes del medio, al ir se condensan y al regresar se relajan, de igual modo las restantes partes se condensan cuando van y se expanden cuando vuelven. Y, por tanto, no todas irán y volverán a la vez (pues así conservarían distancias fijas entre ellas y no podrían enrarecerse y condensarse alternativamente) sino que acercándose entre ellas cuando se condensan y separándose cuando se enrarecen, unas de ellas irán cuando las otras vuelvan: y esto alternativamente hasta el infinito. Pero las partes que van y al ir se condensan, por su movimiento progresivo, con el cual coinciden en los obstáculos, son los impulsos; y por ello, los impulso sucesivos se propagan desde todo cuerpo vibrante en directo; y esto casi con iguales intervalos entre ellos, por ser iguales los intervalos de tiempo en los cuales el cuerpo con sus vibraciones excita cada impulso. Y aunque las partes del cuerpo vibrante vayan y vengan según una dirección cierta y fija, no obstante los impulsa propagados desde allí a través del medio se extenderán hacia lo: lados por la Proposición anterior; y desde dicho cuerpo vibrante como desde un centro común y bajo la forma de superficies cai esféricas y concéntricas se irán propagando en todas direcciones. Cosa de la que tenemos un ejemplo en las ondas, que al ser excitadas por un dedo vibrando, no sólo parten de allí hacia uno y oto lado según la dirección del movimiento del dedo, sino que, en forma de círculos concéntricos, ciñen inmediatamente al dedo y se propagan en todas direcciones. Pues la gravedad de las ondas suple en lugar de la fuerza elástica.

CASO 2. Que el medio no sea elástico: puesto que sus partes presionadas por las partes vibrantes del cuerpo vibrante no pueden condensarse, el movimiento se propagará instantáneamente hacia las partes donde el medio cede con facilidad, es decir, a las partes que el cuerpo vibrante dejase vacías tras él por el otro lado. Es el mismo case que el del cuerpo proyectado en un medio cualquiera. El medio, al ceder a los proyectiles, no retrocede hasta el infinito; sino que lloviéndose en círculo, se dirige hacia los espacios que el cuerpo abandona a sus espaldas. Por tanto, cuantas veces el cuerpo vibrante va hacia un lado cualquiera, el medio cede y se mueve en círculo hacia las partes que abandona el cuerpo; y cuantas veces el cuerpo regresa ál lugar inicial, el medio es expulsado dí allí para que retorne a su primer lugar. Y aunque el cuerpo vibrante no sea duro, sino flexible en todas direcciones, mientras

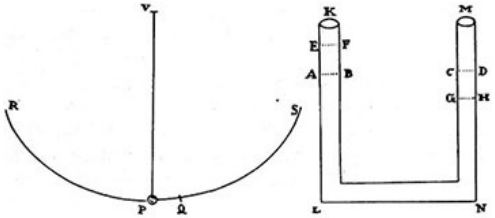
conserve una magnitud dada, y puesto que tampoco puede urgir on sus vibraciones al medio en un sentido sin que a la vez ceda ante él en otro, resultará que el medio, cediendo por los lugares por donde es presionado, se dirigirá continuamente en círculo hacia las partes que ceden ante él. Q. E. D.

COROLARIO. Desvarían, pues, quienes creen que la agitación de la llama conduce a una propagación de la presión según líneas rectas a través del medio ambiente. Semejante presión se deberá derivar no sólo de la agitación de las partes de la llama, sino también de la dilatación de todo.

PROPOSICIÓN XLIV. TEOREMA XXXV

Si el agua alternativamente asciende y desciende por tubos de brazos elevados KL, MN, y se construyese un péndulo cuya longitud entre el punto de suspensión y el centro de oscilación sea igual a la mitad de la longitud del agua qui hay en el tubo: digo que el agua ascenderá y descenderá en los mismos tiempos en los que oscile el péndulo.

Mido la longitud del agua según los ejes del tubo y de los brazos, igualándola a la suma de dichos ejes, y no tengo aquí en cuenta la resistencia del agua debida al rozamiento del tubo. Representen entonces AB y CD la altura media del agua en cada uno de los brazos, y cuando el agua en el brazo KL sube hasta la altura BF, en el brazo MN el agua descenderá hasta la altura GH. Sea P un cuerpo pendular, VP el hilo, V el punto de suspensión, RPQS la cicloide que describa el péndulo, P su punto inferior, PQ un arco igual a la altura AE. La fuerza, con la cual el movimiento del agua se acelera y se retarda alternativamente, es el exceso del peso del agua en uno de los dos brazos sobre el peso en el otro, y por ello, cuando el agua asciende en el brazo KL hasta BF, y en el otro desciende hasta GH, dicha fuerza es el doble del peso del agua EABF, y por tanto, es al peso de toda el agua como AE o PQ a VP o PR. También la fuerza con la que el peso P, en un punto cualquiera Q, es acelerado y retardado en la cicloide (por el Corolario de la Proposición LI) es a su peso total como su distancia PQ desde el lugar ínfimo P a la longitud de la cicloide PR. Por lo cual, las fuerzas motrices del agua y del péndulo que describen los espacios iguales AE, PQ, son como los pesos a mover; y por consiguiente, si el agua y el péndulo inicialmente están en reposo, dichas fuerzas los moverán igualmente en tiempos iguales y harán que vayan y vuelvan a la vez con un movimiento recíproco. Q. E. D.



COROLARIO 1. Por consiguiente, todas las subidas y bajadas del agua son isócronas, tanto si el movimiento es más intenso como si es más suave.

COROLARIO 2. Si la longitud total del agua en el tubo es de $6\frac{1}{9}$ pies *parisinos*, el agua descenderá en un segundo de tiempo y ascenderá en otro segundo; y así sucesivamente en veces alternas hasta el infinito. Pues un péndulo de $3\frac{1}{18}$ pies de longitud oscila en el tiempo de un segundo.

COROLARIO 3. Y si la longitud del agua aumentase o disminuyese, aumentará o disminuirá el tiempo de reciprocación en razón de la raíz cuadrada de la longitud.

PROPOSICIÓN XLV. TEOREMA XXXVI

La velocidad de las ondas es como la raíz cuadrada de las anchuras.

Se sigue de la construcción de la Proposición siguiente.

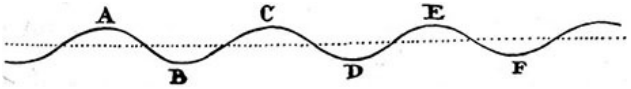
PROPOSICIÓN XLVI. PROBLEMA X

Hallar la velocidad de las ondas.

Constrúyase un péndulo cuya longitud entre el punto de suspensión y el centro de oscilación sea igual a la anchura de las ondas: y en el tiempo en el cual dicho péndulo completa cada una de sus oscilaciones, en ese mismo tiempo las ondas recorrerán casi su anchura al desplazarse.

Denomino anchura de las ondas a la medida transversal comprendida, bien entre los puntos más bajos de los valles, o bien entre los puntos más altos de las crestas. Sea ABCDEF la superficie de agua estancada, que asciende y desciende en ondas sucesivas; y sean A, C, E, etc. las cimas de las ondas y B, D, F, etc., los valles intermedios. Y puesto que el movimiento de las ondas ocurre por el sucesivo ascenso y descenso del agua, del mismo modo sus partes A, C, E, etc., que ahora están en lo más alto después estarán en lo más bajo; y la fuerza motriz por la cual las partes de arriba descienden y las de abajo ascienden, es el peso del agua elevada; dicho ascenso y descenso alternativo será análogo al movimiento recíproco del agua en el tubo y se

atendrá a las mismas leyes de tiempo: y por ello (por la Proposición XLIV) si las distancias entre los puntos más altos de las ondas A, C, E, y los puntos más bajos B, D, F, son iguales al doble de la longitud del péndulo, las partes más altas A, C, E, vendrán a ser las ínfimas en el tiempo de una oscilación, y de nuevo ascenderán en el tiempo de la oscilación siguiente. Por consiguiente, el tiempo que transcurre mientras pasa cada una de las ondas será el de dos oscilaciones; es decir, la onda describe su anchura en el tiempo en que el susodicho péndulo oscila dos veces; pero en ese tiempo un péndulo cuya longitud sea cuatro veces mayor y, por tanto, sea igual a la anchura de las ondas, oscilará una sola vez. Q. E. I.



COROLARIO 1. Por lo tanto, las ondas que tengan una anchura de $3\frac{1}{18}$ pies *parisinos* recorrerán al desplazarse la distancia de su anchura en él tiempo de un segundo; y por ello, en un tiempo de un minuto recorrerán 183⅓ pies, y en una hora recorrerán un espacio aproximado de 11 000 pies.

COROLARIO 2. Y la velocidad de ondas mayores o menores aumentará o disminuirá como la raíz cuadrada de su anchura.

Todo lo cual es así bajo la hipótesis de que las partes del agua ascienden y descienden en línea recta; pero es más verosímil que el ascenso y el descenso tengan lugar circularmente y, por tanto, mantengo que el tiempo hallado mediante esta Proposición es sólo aproximado.

PROPOSICIÓN XLVII. TEOREMA XXXVII

Al propagarse los impulsos a través de un fluido, cada una de las partículas del fluido que van y vienen con un movimiento recíproco muy breve, se aceleran y se retardan siempre de acuerdo con la ley de oscilación del péndulo.

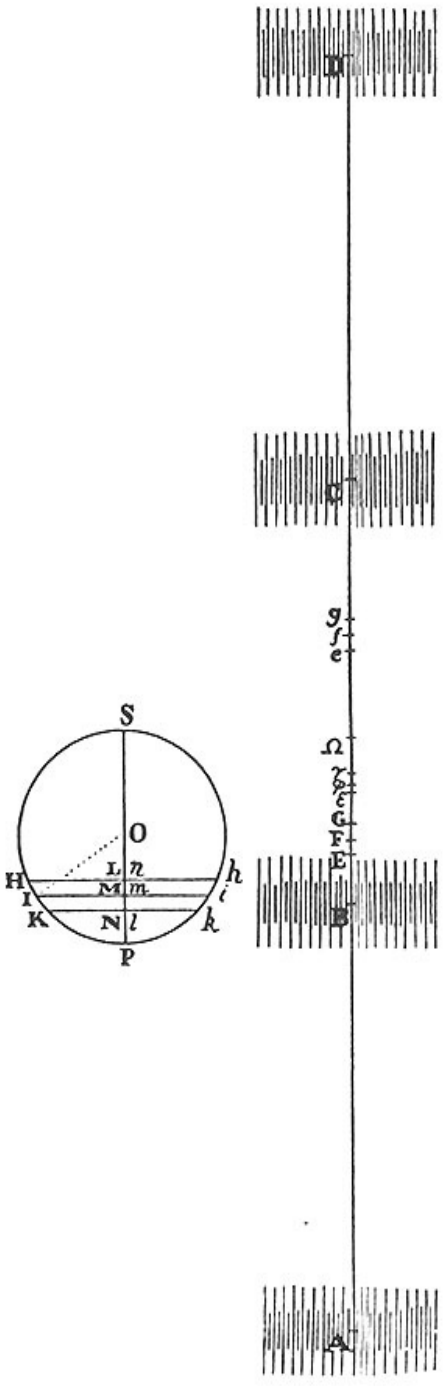
Representen AB, BC, CD, etc. distancias iguales de impulsos sucesivos; ABC la dirección del movimiento de los impulsos propagados desde A hacia B; E, F, G, representen tres puntos físicos del medio en reposo situados a distancias iguales sobre la recta AC; sean E*e*, F*f*, G*g*, espacios iguales muy breves por los cuales los susodichos puntos van y vienen en cada vibración con un movimiento recíproco; $\mathcal{E}$, φ, γ, sean unos lugares intermedios entre dichos puntos; y EF, FG unas breves líneas físicas o partes lineales del medio situadas entre dichos puntos y trasladadas sucesivamente a los lugares $\mathcal{E}\varphi$, $\varphi\gamma$ y ef, fg. Trácese la recta PS igual a la recta E*e*. Biséquese dicha recta en O y con centro en O e intervalo OP trácese el círculo SIPi. Represéntese con esta circunferencia y sus partes el tiempo total de una vibración y

sus partes proporcionales; de suerte que para un tiempo completo cualquiera PH o PHS*h*, si se traza sobre PS la perpendicular HL o *hl* y se toma E*Ɛ* igual a PL o P*l*, el punto físico E se hallará en *Ɛ*. Un punto cualquiera E, bajo esta condición, que vaya desde E por *Ɛ* hasta *e* y desde aquí regrese por *Ɛ* hasta E completará cada vibración con los mismos grados de aceleración y retardación que un péndulo oscilante. Hay que probar que cada punto físico del medio deberá agitarse con semejante movimiento. Supongamos, pues, que el medio es excitado con tal movimiento por una causa cualquiera y veamos qué se sigue de ello.

En la circunferencia PHS*h* tómense los arcos iguales Hl, IK, o *hi*, *ik* que estén en la misma razón respecto a la circunferencia entera que la que tienen las rectas iguales EF, FK respecto al intervalo completo BC de los impulsos. Trácense las perpendiculares IM, KN, o *im*, *kn*; y, puesto que los puntos E, F, G, son agitados sucesivamente por movimientos semejantes y completan sus vibraciones enteras de ida y vuelta en el tiempo en que el impulso pasa de B a C, si PH o PHS*h* fuera el tiempo desde el comienzo del movimiento del punto E, entonces PI o PHS*i* será el tiempo desde el comienzo del movimiento del punto F, y PK o PHS*k* el tiempo desde el comienzo del movimiento del punto G; y por ello, E*Ɛ*, F*φ*, G*γ*, serán respectivamente iguales a PL, PM, PN en la ida de los puntos y en el retorno de los mismos, a P*l*, P*m*, P*n*. Por consiguiente, cuando los puntos van, *Ɛγ* o EG + G*γ* - E*Ɛ* será igual a EG - LN y, cuando los puntos retornan, será igual a EG + *ln*. Pero *Ɛγ* es la anchura o expansión de la parte EG del medio en el lugar *Ɛγ*; y por tanto, la expansión de dicha parte en la ida es a su expansión media como EG - LN a EG; en cambio en el retorno es como EG + *ln* o EG + LN a EG. Por tanto, al ser LN a KH como IM al radio OP y KH a EG como la circunferencia PHS*h*P a BC, es decir, si V representase el radio de un círculo que tuviese por circunferencia un intervalo igual al del impulso BC, como OP a V; y «*ex aequo*» LN a EG como IM a V: la expansión de la parte EG o del punto físico F en el lugar *Ɛγ* será a la expansión media que dicha parte tiene en su primer lugar EG como V - IM a V a la ida y como V + *im* a V a la vuelta. De donde la fuerza elástica del punto F en el lugar *Ɛγ* es a su fuerza elástica media en el lugar EG como $\frac{1}{\text{V - IM}}$ a $\frac{1}{\text{V}}$ en la ida, y como $\frac{1}{\text{V + } im}$ a $\frac{1}{\text{V}}$ en la vuelta. Y por la misma razón las fuerzas elásticas de los puntos físicos E y G en la ida son como $\frac{1}{\text{V - HL}}$ y $\frac{1}{\text{V - KN}}$ a $\frac{1}{\text{V}}$; y la diferencia de las fuerzas es a la fuerza elástica media del medio como

$$\frac{\text{HL - KN}}{\text{VV - V x HL - V x KN + HL x KN}} \text{ a } \frac{1}{\text{V1}}$$

O sea, como $\frac{\text{HL - KN}}{\text{VV}}$ a $\frac{1}{\text{V}}$, o también como HL - KN a V, siempre que (por los estrechos límites de las vibraciones) supongamos que HL y KN sean un tanto menores que la cantidad V. Por lo cual, como la cantidad V está dada, la diferencia de fuerzas es como HL - KN, esto es (por ser proporcionales HL - KN a HK y OM a OI o OP, y estar dadas HK y OP) como OM; es decir, si se bisecase F*f* en Ω, como Ω*φ*. Y por la misma razón la diferencia de las fuerzas elásticas de los puntos físicos *Ɛ* y *γ* en el retorno de la pequeña línea física *Ɛγ* es como Ω*φ*. Pero dicha diferencia (es decir, el exceso de fuerza elástica del punto *Ɛ* sobre la fuerza elástica del punto *γ*) es la fuerza con la que la pequeña línea física intermedia del medio *Ɛγ* se acelera en la ida y se retarda en el regreso; y por tanto, la fuerza aceleratriz de la pequeña línea física *Ɛγ* es como su distancia al lugar medio Ω de la vibración. Por consiguiente, el tiempo se representa correctamente (por la Proposición XXXVIII del Libro I) mediante el arco PI; y la parte lineal del medio *Ɛγ* se mueve conforme a la ley enunciada antes, es decir, conforme a la ley del péndulo oscilante: e igual razón hay para todas las partes lineales de las que se compone el medio entero. Q. E. D.



COROLARIO. De aquí se sigue que el número de impulsos propagados es el mismo que el de vibraciones del cuerpo vibrante, y que no se multiplica con su desplazamiento. Pues la línea pequeña física $\varepsilon\gamma$ tan pronto regresa a su lugar inicial, reposa; y después tampoco se moverá, a no ser que nuevamente sea empujada, o por el ímpetu del cuerpo vibrante, o por el ímpetu de pulsaciones procedentes del cuerpo vibrante. Por tanto, reposará tan pronto cesen los impulsos propagados desde el cuerpo vibrante.

PROPOSICIÓN XLVIII. TEOREMA XXXVIII

Las velocidades de los impulsos propagados en un fluido elástico están en razón compuesta directamente de la raíz cuadrada de la fuerza elástica e inversamente de la raíz cuadrada de la densidad; siempre que la fuerza elástica se suponga proporcional a la condensación del fluido.

CASO 1. Si los medios son homogéneos y las distancias de los impulsos en estos medios son iguales entre ellas, pero el movimiento es más intenso en un medio, las contracciones y dilataciones de partes análogas serán como sus movimientos. Pero esto no es una proporción exacta. No obstante, salvo que las contracciones y dilataciones sean muy intensas, el error no será apreciable y, por tanto, puede tenerse por físicamente exacta. Pues las fuerzas elásticas motrices son como las contracciones y las dilataciones; y las velocidades generadas a la vez de partes iguales son como las fuerzas. Por consiguiente, las partes iguales y correspondientes de impulsos correspondientes completarán su ida y su vuelta a la vez por espacios proporcionales a las contracciones y dilataciones y con velocidades que son como los espacios: por lo cual los impulsos que en el tiempo de una ida y una vuelta completan desplazándose su propia longitud y alcanzan continuamente los lugares de los impulsos inmediatamente anteriores, por la igualdad de las distancias, avanzan con velocidad igual en uno y otro medio.

CASO 2. Pero si las distancias o las longitudes de los impulsos son mayores en un medio que en otro, supongamos que las partes correspondientes describen, en su ida y vuelta, espacios siempre proporcionales a las amplitudes de los impulsos cada una de las veces: entonces sus contracciones y dilataciones serán iguales. Y por tanto, si los medios son homogéneos también las fuerzas motrices elásticas por las que son agitadas con un movimiento recíproco, serán iguales. Pero la materia, que ha de ser movida por estas fuerzas, es como la amplitud de los impulsos, y en la misma razón se halla el espacio a cuyo través deben moverse cada vez en la ida y en la vuelta. Y el tiempo de una ida y vuelta está en razón compuesta de la raíz cuadrada de la materia y de la raíz cuadrada del espacio, y por ende es como el espacio. Pero los impulsos recorren el espacio de sus amplitudes en los tiempos de una ida y una vuelta, es decir,

recorren espacios proporcionales a los tiempos; y por tanto, son igualmente veloces.

Caso 3. Por consiguiente, en medios de iguales densidad y fuerza elástica, todos los impulsos son igualmente veloces. Pero si la densidad del medio o la fuerza elástica aumentasen, toda vez que la fuerza motriz aumentará en razón de la fuerza elástica y la materia a mover en razón de la densidad, el tiempo, en el cual se completen los mismos movimientos que antes, aumentará como la raíz cuadrada de la densidad, y disminuirá como la raíz cuadrada de la fuerza elástica. Y por ello la velocidad de los impulsos estará en razón compuesta de, inversamente, la raíz cuadrada de la densidad del medio y, directamente, de la raíz cuadrada de la fuerza elástica. Q. E. D.

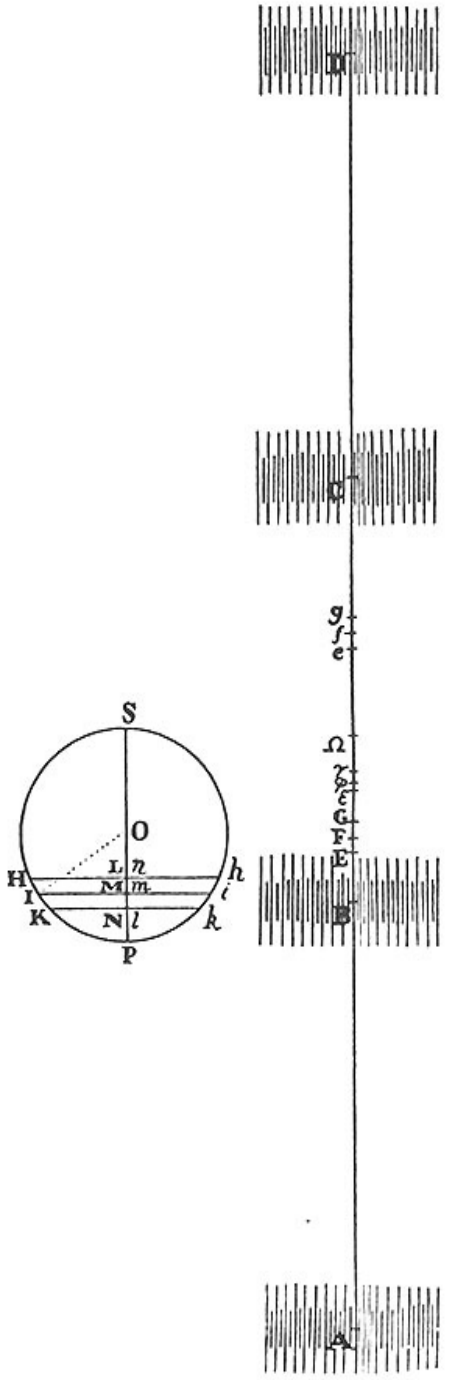
Esta Proposición estará aún más clara por la construcción de la siguiente.

Proposición xlix. Problema xi

Dadas la densidad del medio y su fuerza elástica, hallar la velocidad de los impulsos.

Imaginemos que un medio se halla comprimido por un peso superpuesto al modo del de nuestro aire; y sea A la altura de un medio homogéneo y cuyo peso sea igual al peso superpuesto y cuya densidad sea igual a la del medio comprimido en el cual se propagan los impulsos. Supongamos que se construye un péndulo cuya longitud entre el punto de suspensión y el centro de oscilación sea A: y en el tiempo en que dicho péndulo completa una oscilación entera compuesta de ida y vuelta, en ese mismo tiempo el impulso, desplazándose, completará un espacio igual a la circunferencia del círculo descrito con radio A.

Pues, manteniendo las construcciones de la Proposición xlvii, si una línea física cualquiera EF describiendo en cada vibración el espacio PS es urgida en los extremos P y S de cada ida y vuelta por una fuerza elástica igual a su peso, completará dicha línea cada una de sus vibraciones en el tiempo en que oscilaría en una cicloide cuyo perímetro fuese igual a la longitud PS: esto, además, porque fuerzas iguales, en tiempos iguales, empujan a cuerpos iguales a través de espacios iguales. Por lo cual, ya que los tiempos de las oscilaciones son como la raíz cuadrada de la longitud de los péndulos, y la longitud del péndulo es igual a la mitad del arco de la cicloide entera, el tiempo de una vibración será al tiempo de oscilación del péndulo de longitud A como la raíz cuadrada de la longitud ½PS o PO a la longitud A.



Pero la fuerza elástica por la que la pequeña línea física EG es urgida cuando está en los puntos extremos P, S, era (en la demostración de la Proposición XLVII) a toda su fuerza elástica como HL - KN a V, es decir (por caer ahora el punto K en P) como HK a V: y toda aquella fuerza, o sea, el peso superpuesto con el cual es comprimida la pequeña línea EG, es al peso de la pequeña línea como la altura A del peso superpuesto a la longitud de la pequeña línea EG; y por tanto «*ex aequo*», la fuerza que urge a la pequeña línea EG en los lugares P y S es al peso de dicha línea como HK x A a V x EG, o como PO x A a VV, pues HK era a EG como PO a V. Por tanto, toda vez que los tiempos en los cuales cuerpos iguales son impulsados a través de espacios iguales son inversamente proporcionales a la raíz cuadrada de las fuerzas, el tiempo de una vibración generada por dicha fuerza elástica será al tiempo de vibración generada por la fuerza del peso como la raíz cuadrada de VV a PO x A y, por consiguiente, al tiempo de oscilación de un péndulo de longitud A, como la raíz cuadrada de VV a PO x A y como la raíz cuadrada de PO a A conjuntamente; es decir, como la razón entera de V a A. Pero en el tiempo de una vibración compuesta de ida y vuelta, el impulso desplazándose completa su anchura BC. Luego el tiempo en el cual el impulso recorre el espacio BC es al tiempo de una oscilación compuesta de ida y vuelta como V a A, es decir, como BC a la circunferencia del círculo cuyo radio es A. Pero el tiempo en que el impulso recorre el espacio BC está en la misma razón con respecto al tiempo en que recorre una longitud igual a dicha circunferencia; y por lo mismo, en el tiempo de dicha oscilación el impulso recorre una longitud igual a dicha circunferencia. Q. E. D.

COROLARIO 1. La velocidad de los impulsos es aquella que adquieren los graves cayendo con movimiento igualmente acelerado y describiendo en su caída la mitad de la altura A. Pues en el tiempo de esta caída, cayendo con la velocidad adquirida, el impulso recorre un espacio que será igual a toda la altura A; y por lo tanto, en el tiempo de una oscilación compuesta de ida y vuelta recorrerá un espacio igual a la circunferencia del círculo trazado con radio A: porque el tiempo de caída es al tiempo de oscilación como el radio del círculo a su circunferencia.

COROLARIO 2. De donde, al ser dicha altura A directamente como la fuerza elástica del fluido e inversamente como su densidad, la velocidad de los impulsos estará en razón compuesta de la raíz cuadrada de la densidad, inversamente, y de la raíz cuadrada de la fuerza elástica, directamente.

PROPOSICIÓN L. PROBLEMA XII

Hallar las distancias de los impulsos.

Hállese el número de vibraciones del cuerpo cuya oscilación produce los impulsos en un tiempo dado. Divídase por ese número el espacio que puede recorrer

el impulso en ese mismo tiempo y el resultado será la amplitud de un impulso. Q. E. I.

ESCOLIO[14]

Las últimas Proposiciones tienen relación con el movimiento de la luz y de los sonidos. Puesto que la luz se propaga en línea recta no puede consistir en mera acción (por las Proposiciones XLI y XLII). Y los sonidos, en tanto que se producen por cuerpos vibrantes, no son otra cosa que impulsos de aire propagados, por la Proposición XLIII. Se confirma esto por las vibraciones que, si son fuertes y graves, suscitan en los cuerpos inmediatos, como los sonidos de los tambores. Pues las vibraciones rápidas y breves no se excitan con facilidad. Y es bien sabido que cualesquiera sonidos lanzados contra las cuerdas de cuerpos sonoros genera vibraciones. Y también se confirma por la velocidad de los sonidos. Puesto que los pesos específicos del agua de lluvia y del mercurio son entre sí aproximadamente como 1 a 13⅔, y cuando el mercurio alcanza en el *barómetro* una altura de 30 pulgadas *inglesas*, el peso específico del aire y del agua de lluvia se halla en la relación de 1 a 870 aproximadamente, entonces el peso específico del aire y del mercurio será aproximadamente como 1 a 11 890. Y, por tanto, cuando la altura del mercurio es de 30 pulgadas, la altura de aire uniforme cuyo peso pudiera oprimir a nuestro aire subyacente sería de 356 700 pulgadas, o de 29 725 pies *ingleses*. Y ésta es la altura misma que denominábamos A en la construcción precedente. La circunferencia del círculo descrito con radio de 29 725 pies es de 186 768 pies. Y puesto que un péndulo de 39⅕ pulgadas de largo completa como es sabido, su oscilación entera de ida y vuelta en el tiempo de dos segundos, un péndulo de 39⅕ pulgadas de largo completa, como es sabido, su completar una oscilación semejante en un tiempo de 190¾ segundos. Por tanto, en ese mismo tiempo el sonido desplazándose completará 186 768 pies, y en el tiempo de un segundo 979 pies.

Por lo demás en este cálculo no se ha tomado en cuenta el grosor de las partículas sólidas de aire a cuyo través el sonido se propaga instantáneamente. Como el peso del aire es al peso del agua como 1 a 870, y las sales son casi dos veces más densas que el agua, si suponemos que las partículas de aire son casi de la misma densidad que las partículas de agua o de sal y la rareza del aire obedece a los intervalos de las partículas, entonces el diámetro de las partículas de aire será al intervalo entre los centros de las partículas como 1 a 9 ó 10 aproximadamente, y al intervalo entre partículas como 1 a 8 ó 9. Por ello, a los 979 pies que el sonido recorrerá en un segundo según el cálculo anterior habrá que añadir $^{979}/_{9}$ pies, o aproximadamente 109 pies, por el grosor de las partículas de aire: y de este modo el sonido, en el tiempo de un segundo, recorrerá aproximadamente 1088 pies.

Añádase a esto que los vapores presentes en el aire, al ser de diferente elasticidad

y tono, poco o nada participarán en el movimiento del verdadero aire con el que se propagan los sonidos. Permaneciendo, pues, éstos en reposo, dicho movimiento se propagará por el solo verdadero aire, y esto en razón de la raíz cuadrada de la disminución de materia. De suerte que si la atmósfera constase de diez partes de verdadero aire y una de vapores, el movimiento de los sonidos sería más rápido en razón de la raíz cuadrada de 11 a 10, o casi en la razón entera de 21 a 20, que si se propagase por once partes de aire verdadero: y por ello el movimiento de los sonidos hallado antes ha de aumentarse en esta proporción. Con lo cual el sonido, en el tiempo de un segundo, recorrerá 1142 pies.

Así serán estas cosas en primavera y otoño, cuando al aire templado por el calor se enrarece y su fuerza elástica aumenta un tanto. Pero en invierno, cuando el aire se condensa por el frío y disminuye su fuerza elástica, el movimiento de los sonidos debe ser más lento en razón de la raíz cuadrada de la densidad; y viceversa, en verano debe ser más veloz.

Por otra parte, experimentalmente consta que los sonidos en un segundo recorren más o menos 1142 pies *londinenses* ó 1070 *parisinos*.

Una vez conocida la velocidad de los sonidos se obtienen también los intervalos de los impulsos. El Sr. *Sauveur* halló mediante experimentos hechos por él, que una flauta abierta de más o menos cinco pies *parisinos* de largo producía un sonido del mismo tono que el sonido de una cuerda que vibra cien veces por segundo. Por tanto, en el espacio de 1070 pies *parisinos* que recorre un sonido en un segundo hay más o menos cien impulsos; por lo cual una pulsación ocupa un espacio de casi $10^{7}/_{10}$ pies *parisinos*, es decir, casi el doble de la longitud de la flauta. Por lo que es verosímil que las amplitudes de los impulsos, en los sonidos de todas las flautas abiertas, sea igual a dos veces la longitud de las flautas.

Y por lo demás, está claro por el Corolario de la Proposición XLVII de este libro por qué cesan los sonidos cuando cesa el movimiento del cuerpo sonoro, y no se oyen por más tiempo cuando estamos muy lejos de los cuerpos sonoros que cuando estamos muy cerca. Y por qué los sonidos aumentan tanto en los tubos amplificadores está claro por los principios expuestos. Pues todo movimiento recíproco en cada regreso suele aumentar gracias a la causa generadora. Y el movimiento en los tubos que impiden la dilatación de los sonidos se pierde más tarde y se repite con más fuerza y, por tanto, es incrementado por el nuevo movimiento impreso en cada retorno. Y éstos son los principales fenómenos relativos a los sonidos.

Sección IX
SOBRE EL MOVIMIENTO CIRCULAR DE LOS FLUIDOS

Hipótesis

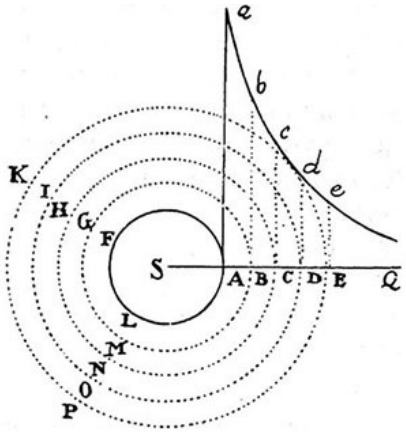
La resistencia que procede de la falta de lubricidad de las partes de un fluido, en igualdad de condiciones, es proporcional a la velocidad con que las partes del fluido se separan entre sí.

Proposición LI. Teorema XXXIX

Si un cilindro sólido infinitamente largo gira con un movimiento uniforme en torno a su eje de oposición dada en un fluido uniforme e infinito, y por su solo impulsor del fluido es obligado a girar en círculo y cada parte del fluido persevera uniformemente en su movimiento; digo que los tiempos periódicos de las partes del fluido son como sus distancias del eje del cilindro.

Sea AFL un cilindro que era uniformemente en torno de su eje S y mediante los círculos concéntricos BGM, CHN, DIO, EKP, etcétera, divídase el fluido en innumerables orbes cilíndricos, concéntricos, sólidos y del mismo grosor. Y, puesto que el fluido es homogéneo, las impresiones de unos orbes contiguos en otros (por hipótesis) serán como sus traslaciones respectivas y la superficie sobre la cual recae la impresión. Si una impresión sobre un orbe cualquiera es mayor o menor por la parte cóncava o la parte convexa, prevalecerá la impresión más fuerte, y acelerará o retardará el movimiento del orbe si se dirige hacia el mismo lado o hacia el contrario que el movimiento del mismo orbe. Por tanto, para que cada orbe persevere en su movimiento uniformemente, las impresiones en una y otra parte deben ser iguales entre sí y producirse en direcciones opuestas. Por consiguiente, como las impresiones son como las superficies contiguas y como sus traslaciones respectivas, las traslaciones serán inversamente como las superficies, es decir, inversamente como las distancias de las superficies desde el eje. Pero las diferencias de los movimientos angulares en torno al eje son como dichas traslaciones aplicadas a las distancias, o

como las traslaciones directamente y las distancias inversamente; es decir, conjuntas ambas razones, como el inverso del cuadrado de las distancias. Por lo cual, si sobre cada parte de la recta infinita SABCDEQ se eleva una perpendicular A*a*, B*b*, C*c*, D*d*, E*e*, etc., inversamente proporcional a los cuadrados de las propias SA, SB, SC, SD, SE, etc., y por los extremos de esas perpendiculares se imagina trazada una línea curva hiperbólica, las sumas de las diferencias, es decir, los movimientos angulares totales serán como las sumas correspondientes de las líneas A*a*, B*b*, C*c*, D*d*, E*e*, es decir, si se aumenta el número de orbes y se disminuye su anchura hasta el infinito para que el fluido resulte un medio uniforme, como las áreas hiperbólicas A*a*Q, B*b*Q, C*c*Q, D*d*Q, E*e*Q, etc., análogas a las susodichas sumas. Y los tiempos inversamente proporcionales a los movimientos angulares serán también inversamente proporcionales a estas áreas. Por tanto, el tiempo periódico de una partícula cualquiera D es inversamente proporcional al área D*d*Q, es decir (por las conocidas cuadraturas de curvas) directamente como la distancia SD. Q. E. D.



Corolario 1. De aquí que los movimientos angulares de las partículas del fluido sean inversamente como sus distancias al eje del cilindro, y sus velocidades absolutas iguales.

Corolario 2. Si un fluido está contenido en un vaso cilíndrico de longitud infinita que contiene en su interior a otro cilindro, y ambos giran en torno al eje común, y los tiempos de revolución son como sus semidiámetros, y cada parte del fluido permanece con su movimiento: los tiempos periódicos de cada parte serán como sus distancias al eje de los cilindros.

Corolario 3. Si a un cilindro y a un fluido movidos de este modo se les añade o quita algún movimiento angular común, puesto que con este nuevo movimiento no se altera el rozamiento mutuo de las partes del fluido, los movimientos de las partes entre ellas no sufrirán alteración. Porque las traslaciones de las partes entre ellas dependen del rozamiento. Una parte cualquiera perseverará en aquel movimiento que,

al aplicar el rozamiento por ambas partes y en sentidos contrarios, no se acelere más que se retarde.

Corolario 4: De donde, si a todo el sistema de cilindros y fluido se le suprime todo el movimiento angular del cilindro exterior, se tendrá el movimiento del fluido en el cilindro en reposo.

Corolario 5. Si permanecen en reposo el cilindro y el fluido exteriores mientras el cilindro interior gira uniformemente, se comunicará el movimiento circular al fluido y poco a poco se propagará a todo el fluido; y no dejará de aumentar hasta que cada parte del fluido adquiera el movimiento definido en el Corolario cuarto.

Corolario 6. Y puesto que el fluido intenta propagar su movimiento hasta más allá, el cilindro exterior se pondrá a girar por causa de aquel empuje, salvo que se le detenga por la fuerza; y su movimiento se irá acelerando hasta que los tiempos periódicos de ambos cilindros sean iguales entre ellos. Pero si el cilindro exterior se frenase violentamente, este tratará de retardar el movimiento del fluido; y salvo que el cilindro interior conserve su movimiento mediante alguna fuerza impresa desde fuera, acabará por hacerlo cesar poco a poco.

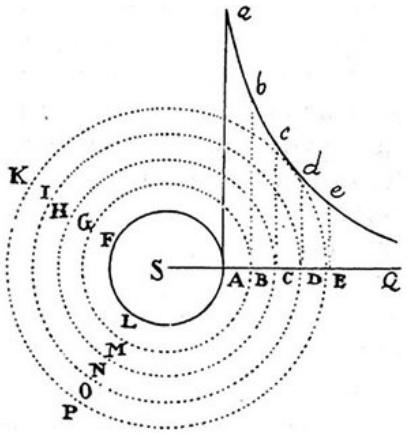
Todo esto se puede experimentar en agua profunda estancada.

Proposición lii. Teorema xl[15]

Si una esfera sólida gira con movimiento uniforme en torno a un eje de posición dada en un fluido uniforme e infinito y el fluido es obligado a girar por el solo impulso de la esfera, y cada parte del fluido persevera en su movimiento uniforme: digo que los tiempos periódicos de las partes del fluido serán como los cuadrados de las distancias al centro de la esfera.

Caso 1. Sea AFL una esfera girando uniformemente en torno al eje S y con los círculos concéntricos BGM, CHN, DIO, EKP, etc., divídase el fluido en innumerables orbes concéntricos de igual grosor. Supongamos que dichos orbes fuesen sólidos; y, al ser el fluido homogéneo, las impresiones mutuas de los orbes contiguos serán (por hipótesis) como sus traslaciones respectivas y las superficies contiguas sobre las que ocurren las impresiones. Si la impresión sobre un orbe cualquiera fuese mayor o menor por la parte cóncava o por la convexa, prevalecerá la impresión mayor, y la velocidad del orbe se acelerará o se retardará, según se dirija hacia la dirección del movimiento o hacia la dirección contraria. Por consiguiente, para que cada orbe persevere en su movimiento uniformemente, las impresiones de una y otra parte deberán igualarse entre sí y producirse en direcciones contrarias. De donde, puesto que las impresiones son como las superficies contiguas y sus traslaciones respectivas, las traslaciones serán inversamente como las superficies, es decir, como el inverso del cuadrado de la distancia de las superficies al centro. Pero las diferencias de los

movimientos angulares en torno al eje son como estas traslaciones aplicadas a las distancias, o directamente como las traslaciones e inversamente como las distancias; es decir, juntas ambas razones, como el inverso del cubo de las distancias. Por lo cual, si sobre cada parte de la recta infinita SABCDEQ se elevan las perpendiculares A*a*, B*b*, C*c*, D*d*, E*e*, etc., inversamente proporcionales a los cubos de las propias SA, SB, SC, SD, SE, etc., las sumas de las diferencias, es decir, los movimientos angulares totales serán como las correspondientes sumas de las líneas A*a*, B*b*, C*c*, D*d*, E*e*, etc., es decir (si para obtener un medio uniformemente fluido se aumenta el número de órbitas y se disminuye su grosor hasta el infinito) como las áreas hiperbólicas. A*a*Q, B*b*Q, C*c*Q, D*d*Q, E*e*Q, etc., análogas a dichas sumas. Y los tiempos periódicos inversamente proporcionales a los movimientos angulares serán también inversamente proporcionales a esas arcas. Por consiguiente, el tiempo periódico de un orbe cualquiera DIO es inversamente proporcional al área D*d*Q, es decir, por las va conocidas cuadraturas de las curvas, directamente proporcional al cuadrado de la distancia SD. Que es lo primero que quería demostrar.



Caso 2. Desde el centro de la esfera trácense muchas rectas indefinidas que formen con el eje ángulos dados y que se superen unos a otros con diferencias iguales; e imagínese que con los giros de estas rectas en torno al eje los orbes son cortados en innumerables anillos; y cada anillo tendrá cuatro anillos contiguos, uno por dentro, otro por fuera y dos a los lados. Con el rozamiento de los anillos interior y exterior un anillo cualquiera no puede ser urgido igualmente hacia direcciones contrarias, salvo que el movimiento se atenga a la ley del caso primero. Esto se sigue de la demostración del primer caso. Y por tanto, una serie cualquiera de anillos que se aleje en línea recta del globo hasta el infinito, se moverá de acuerdo con la ley del caso primero, salvo que se lo impida el rozamiento de los anillos laterales. Pero en un movimiento realizado según la susodicha ley el rozamiento de los anillos laterales es nulo y por lo mismo en nada impedirá al movimiento realizado de acuerdo con la

dicha ley. Si los anillos equidistantes del centro girasen más velozmente o más lentamente en los polos que en la eclíptica, los más lentos se acelerarán y los más veloces se retardarán gracias al mutuo rozamiento y así los tiempos periódicos siempre llegarán a ser iguales, de acuerdo con la ley del primer caso. Por consiguiente, este rozamiento no impedirá que el movimiento se realice de acuerdo con la ley del caso primero y, por tanto, dicha ley se cumplirá: es decir, los tiempos periódicos de cada anillo serán como los cuadrados de sus distancias al centro del globo. Que es lo que quería demostrar en segundo lugar.

Caso 3. Divídase ahora cada anillo, mediante secciones transversales, en innumerables partículas que constituyan una sustancia total y uniformemente fluida, y como estas secciones no afectan a la ley del movimiento circular sino que sirven solamente para la constitución del fluido, el movimiento circular perseverará como antes. Con estas secciones los minúsculos anillos no cambiarán su aspereza y su rozamiento mutuo, o lo cambiarán todos por igual. Y permaneciendo la proporción de las causas permanecerá la proporción de los efectos, es decir, la proporción de los movimientos y de los tiempos periódicos. Q. E. D. Por lo demás, como el movimiento circular, y la fuerza centrífuga nacida del mismo, es mayor en la eclíptica que en los polos, tendrá que haber una causa por la cual cada partícula se mantenga siempre en su círculo, para que la materia que se halla en la eclíptica no se aparte continuamente del centro y por la parte exterior del vórtice se vaya hasta los polos y desde allí a través del eje regrese a la eclíptica en una suerte de circulación continua.

Corolario 1. Por lo tanto, los movimientos angulares de las partes del fluido en torno al eje del globo son inversamente proporcionales al cuadrado de las distancias al centro del globo y las velocidades absolutas son inversamente proporcionales a esos cuadrados aplicados a las distancias al eje.

Corolario 2. Si un globo gira en un fluido en reposo, uniforme e infinito, con movimiento uniforme en torno a un eje de posición dada, comunicará el movimiento al fluido en forma de vórtice, y dicho movimiento se propagará poco a poco hasta el infinito; y no dejarán de acelerarse las distintas partes del fluido hasta que los tiempos periódicos de cada una no sean como los cuadrados de sus distancias al centro del globo.

Corolario 3. Puesto que las partes interiores del vórtice rozan y urgen a las partes exteriores y les comunican continuamente el movimiento con dicha acción, a la vez que esas exteriores transfieren esa misma cantidad de movimiento a otras todavía más exteriores y con esa acción conservan invariante su propia cantidad de movimiento; es evidente que el movimiento es transferido continuamente desde el centro hacia la circunferencia del vórtice y es absorbido por la infinitud de la circunferencia. La materia contenida entre dos superficies esféricas cualesquiera concéntricas al vórtice jamás se acelerará, por cuanto que transfiere siempre todo el movimiento recibido de la materia interior hacia la exterior.

Corolario 4. Por lo tanto, para que un vórtice conserve continuamente el mismo

estado de movimiento, es necesario algún principio activo del cual reciba el globo continuamente la misma cantidad de movimiento que transfiere hacia la materia. En ausencia de tal principio es necesario que el globo y las partes interiores del vórtice, al estar propagando siempre su movimiento hacia el exterior sin recibir ningún otro movimiento nuevo, acaben por retardarse poco a poco y dejen de girar.

COROLARIO 5. Si otro globo se situase en este vórtice a una cierta distancia del centro girando con una fuerza continuamente en torno a un eje de inclinación dada, su movimiento arrastraría al fluido en forma de vórtice: y primeramente este nuevo y pequeño vórtice girará a la vez que el globo en torno al centro del otro y mientras tanto su movimiento se irá alejando propagándose poco a poco hasta el infinito, al igual que el del vórtice primero. Y por la misma razón por la que el movimiento de este globo era arrastrado por el movimiento del otro, el movimiento del otro será ahora arrastrado por el movimiento de éste, de suerte que ambos globos girarán en torno a un cierto punto intermedio, alejándose uno de otro a causa de dicho movimiento circular, salvo que alguna fuerza se lo impida. Después, si cesasen las fuerzas constantes impresas, gracias a las cuales los globos perseveran en sus movimientos, y todo se remitiese a las leyes mecánicas, el movimiento de los globos iría languideciendo poco a poco (por las razones dadas en los Corolarios 3 y 4) hasta que junto con los vórtices queden finalmente en reposo.

COROLARIO 6. Si varios globos girasen continuamente en lugares dados en torno a ejes de posición dada y con unas velocidades determinadas, surgirían otros tantos vórtices expandiéndose hasta el infinito. Pues por la misma razón que un globo cualquiera propaga su movimiento hasta el infinito, cualquier otro globo propagará también su movimiento hasta el infinito, hasta que cada parte del fluido infinito se halle agitada por el movimiento resultante de las acciones de todos los globos. Por lo cual, no se podrán definir los vórtices con límites ciertos, sino que poco a poco se desplazarán entre ellos; y los globos por las mutuas interacciones de los vórtices cambiarán continuamente de lugar, como se ha explicado en el Corolario anterior; y tampoco conservarán entre ellos una posición determinada, salvo que alguna fuerza los retenga. Y al cesar las fuerzas impresas constantes sobre los globos gracias a las cuales conservaban sus movimientos, la materia, por las razones dadas en los Corolarios 3 y 4, entrará en reposo poco a poco y dejará de girar en vórtice.

COROLARIO 7. Si un fluido semejante estuviese contenido en un vaso esférico y, por la rotación uniforme de un globo situado en el centro, se le hiciese girar en un vórtice, mientras el globo y el vaso giran hacia el mismo lado sobre un mismo eje y sus tiempos periódicos son como los cuadrados de los semidiámetros: las partes del fluido no perseverarán en sus movimientos sin aceleración o retardación más que hasta que sus tiempos periódicos sean como los cuadrados de las distancias al centro del vórtice. Y ninguna otra constitución del vórtice puede ser permanente.

COROLARIO 8. Si el vaso, el fluido contenido y el globo conservan este movimiento y además girasen con un movimiento angular común en torno a un eje

dado, puesto que con este nuevo movimiento no se cambia el rozamiento de las partes del fluido entre ellas, tampoco se cambiarán los movimientos de las partes entre ellas. Pues las traslaciones de las partes entre ellas dependen del rozamiento. Cada parte perseverará en aquel movimiento que resulta de no estar más retardada por el rozamiento de un lado que acelerada por el rozamiento del otro lado.

COROLARIO 9. Por consiguiente, si el vaso está en reposo y el movimiento del globo está dado, estará dado el movimiento del fluido. Pues, imaginemos un plano que pase por el eje del globo y que gire en sentido contrario, y supongamos que la suma del tiempo de revolución de este plano y del tiempo de revolución del globo fuese al tiempo de revolución del globo como el cuadrado del semidiámetro del vaso al cuadrado del semidiámetro del globo: entonces los tiempos periódicos de las partes del fluido con respecto a dicho plano serán como los cuadrados de sus distancias al centro del globo.

COROLARIO 10. Por consiguiente, si el vaso gira en torno al mismo eje que el globo o alrededor de otro distinto con una velocidad dada, estará dado el movimiento del fluido. Porque si al sistema completo se le resta el movimiento angular del vaso, todos los movimientos permanecerán entre ellos los mismos que antes, por el Corolario 8. Y estos movimientos, por el Corolario 9, estarán dados.

COROLARIO 11. Si el vaso y el fluido están en reposo mientras que el globo gira con movimiento uniforme, el movimiento se propagará poco a poco por todo el fluido hacia el vaso, y pondrá en giro al vaso salvo que se le detenga por la fuerza, y no dejarán ni vaso ni fluido de acelerarse hasta que sus tiempos periódicos lleguen a ser iguales a los tiempos periódicos del globo. Pero si el vaso es detenido por una fuerza o gira con un movimiento constante y uniforme, el medio paulatinamente llegará al estado de movimiento establecido en los Corolarios 8, 9 y 10, y nunca permanecerá en otro estado cualquiera. Pero si después cesasen aquellas fuerzas con las cuales el vaso y el globo giran con movimientos determinados, y el sistema entero queda remitido a las leyes mecánicas, el vaso y el globo actuarán uno sobre otro por medio del fluido y no dejarán de propagar sus movimientos uno hacia otro a través del fluido hasta que sus tiempos periódicos sean iguales entre sí y el sistema entero gire a la vez al modo de un único cuerpo sólido.

ESCOLIO

En todo esto doy por supuesto que el fluido consta de una materia uniforme en cuanto a densidad y fluidez. Es así aquel en que el mismo globo, con el mismo movimiento, en el mismo espacio de tiempo, podría propagar movimientos semejantes e iguales, siempre a distancias iguales, donde quiera que el globo se halle situado. Ciertamente la materia, con su movimiento circular, trata siempre de apartarse del eje del vórtice y, por lo mismo, presiona siempre a la materia siguiente.

De esta presión nace un mayor rozamiento de las partes y mayor dificultad para su mutua separación y, por consiguiente, disminuye la fluidez de la materia. Además, si partes del fluido son más gruesas en algún sitio o mayores, la fluidez será allí menor, por ser menores en número las superficies en las cuales las partes se separarían entre ellas. En tales casos supongo que la deficiencia de fluidez se compensa con la lubricación de las partes o la lisura u otra condición cualquiera. De no ser así, la materia cuando es menos fluida se cohesionará más, será más lenta y recibirá por ello con mayor lentitud el movimiento y lo propagará más que en la razón establecida más arriba. Si la figura del vaso no fuese esférica, las partículas se moverán en líneas no circulares sino adaptadas a la figura del vaso y los tiempos periódicos serán como los cuadrados de las distancias medias al centro muy aproximadamente. En los lugares más amplios entre el centro y la circunferencia los movimientos serán más lentos y más rápidos donde son más angostos, y las partículas más rápidas tampoco tenderán hacia la circunferencia. Pues describen arcos menos curvos y el intento de separarse del centro no disminuye menos por el decremento de esta curvatura de lo que aumenta por el incremento de la velocidad. Al salir de los espacios más angostos a los más amplios se alejan un poco del centro pero con este retroceso se retardan, y pasando después de los más amplios a los más angostos se acelerarán, y así cada partícula se retardará y se acelerará alternativamente de modo continuo. Esto es así en un vaso rígido. Pues en un fluido infinito la constitución de vórtices se atiene al Corolario sexto de esta Proposición.

He tratado de investigar en esta Proposición las propiedades de los vórtices con el fin de ensayar la posibilidad de explicar de semejante modo los fenómenos celestes. Pues un fenómeno es que los tiempos periódicos de los planetas que giran en torno a Júpiter son como la potencia $^{3}/_{2}$ de su distancia al centro de Júpiter; y la misma regla vale para los planetas que giran en torno al Sol. Se cumplen estas reglas en unos y otros planetas con la mayor exactitud, hasta donde las observaciones astronómicas han podido llegar. Y por lo mismo, si dichos planetas fuesen arrastrados por vórtices que girasen en torno a Júpiter y al Sol, también deberían estos vórtices atenerse a esta ley. Pero los tiempos periódicos de las partes del vórtice han resultado estar en razón del cuadrado de las distancias al centro del movimiento y dicha razón no puede disminuirse ni reducirse a la razón de la potencia $^{3}/_{2}$ a no ser que la materia del vórtice sea más fluida cuanto más alejada del centro, o que la resistencia debida a la falta de lubrificación de las partes del fluido, por el aumento de la velocidad con la cual las partes del fluido se separan entre ellas, aumente en mayor proporción que aquella con que aumenta la velocidad. Nada de lo cual parece razonable. Las partes más gruesas y menos fluidas tenderán hacia la circunferencia, salvo que graviten hacia el centro; y, aunque propuse al principio de esta Sección para facilidad de las demostraciones la hipótesis de que la resistencia es proporcional al cuadrado de la velocidad, es verosímil, sin embargo, que dicha resistencia se halle en una proporción respecto a la velocidad menor que ésa. Supuesto esto, los tiempos periódicos de las partes del

vórtice estarán en una razón mayor que la del cuadrado de las distancias al centro del mismo. Y si los vórtices (como algunos creen) se mueven más rápidamente cerca del centro, después más lentamente hasta un límite, y de nuevo más rápidamente cerca de la circunferencia, en verdad que ni la razón de la potencia $^{3}/_{2}$ ni otra alguna cierta y determinada puede hallarse. Que vean pues los filósofos de qué manera podría explicarse mediante vórtices el susodicho fenómeno de la razón de la potencia $^{3}/_{2}$.

Proposición liii. Teorema xli

Los cuerpos que arrastrados por un vórtice giran en órbita, tienen la misma densidad que el vórtice y se mueven bajo la misma ley que sus partes en cuanto a velocidad y dirección del recorrido.

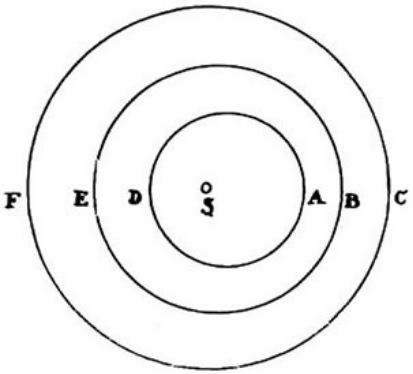
Pues si suponemos que una pequeña parte del vórtice, cuyas partículas o puntos físicos conservan entre ellos una situación dada, se congelase, dicha parte, por cuanto que ni en su densidad, ni en su fuerza ínsita, ni en su figura sufre cambio alguno, se moverá con la misma ley que antes; y al contrario, si una parte congelada y sólida del vórtice es de la misma densidad que el resto del mismo, y se convierte en fluido, se moverá con la misma ley que antes, salvo en la medida en que sus partículas ahora hechas fluidas se muevan entre ellas. Pero despreciando el movimiento de las partículas entre ellas, como si no afectase al movimiento total de progresión en nada, el movimiento total será el mismo que antes. Y el movimiento será el mismo que el de las otras partes del vórtice equidistantes del centro por cuanto que el sólido resuelto en fluido se convierte en parte del vórtice semejante a las demás partes del mismo. Luego el sólido, si es de la misma densidad que la materia del vórtice, se moverá con el mismo movimiento que las partes de éste, permaneciendo en reposo relativo respecto a la materia que lo rodea. Si fuese más denso, entonces ya tratará de alejarse más que antes del centro del vórtice, y por ello superando ahora a la fuerza del vórtice la cual lo retenía antes en su órbita como si estuviese en equilibrio, se alejará del centro y girando describirá una espiral, sin volver ya más a su órbita. Y por la misma razón, si fuese más raro se acercará al centro. Por tanto, no regresará a la misma órbita salvo que sea de la misma densidad que el fluido. Y en ese caso ya se ha mostrado que girará con la misma ley que las partes del fluido equidistantes del centro del vórtice. Q. E. D.

Corolario 1. Luego un sólido que gira en un vórtice y vuelve siempre a la misma órbita reposa relativamente en el fluido en el que flota.

Corolario 2. Y si el vórtice es de densidad uniforme el mismo cuerpo puede girar a cualquier distancia del centro del vórtice.

Escolio

Por esto, es evidente que los planetas no son arrastrados por vórtices corpóreos. Pues los planetas, de acuerdo con la hipótesis de Copérnico, arrastrados en torno al Sol giran en elipses que tienen su foco en el Sol, y con radios trazados al Sol describen áreas proporcionales a los tiempos. Pero las partes de los vórtices no pueden girar con un movimiento semejante. Representen AD, BE, CF, tres órbitas descritas en torno al Sol S, de las cuales, la exterior CF sea un círculo concéntrico con el Sol, y los afelios de las dos interiores sean A, B y los perihelios D, E. Luego un cuerpo que gire en la órbita CF al describir con un radio trazado al Sol áreas proporcionales a los tiempos, se moverá con un movimiento uniforme. Mientras que el cuerpo que gire en la órbita BE se moverá más lentamente en el afelio B y más velozmente en el perihelio E, según las leyes astronómicas; mientras que, según las leyes mecánicas, la materia del vórtice en el espacio más angosto entre A y C habrá de moverse más velozmente que en el espacio más ancho entre D y F; es decir, en el afelio más velozmente que en el perihelio. Cosas estas dos que se contradicen. Así, al comienzo del signo Virgo, donde ahora se halla el afelio de Marte, la distancia entre las órbitas de Marte y Venus es a la distancia entre las mismas órbitas al principio del signo Piscis casi como tres a dos, y por tanto la materia del vórtice entre dichas órbitas al principio de Piscis debe ser más veloz que al principio de Virgo en razón de tres a dos. Pues cuanto más estrecho es el espacio por el cual pasa la misma cantidad de materia en el mismo tiempo de una revolución, con tanta mayor velocidad debe pasar. Por consiguiente, si la Tierra reposando relativamente en esta materia celeste es arrastrada por ella y girase en torno al Sol junto con ella, su velocidad al principio de Piscis sería a su velocidad al principio de Virgo en razón de tres a dos. De donde, el movimiento diurno aparente del Sol al principio de Virgo sería superior a setenta minutos; y al principio de Piscis inferior a los cuarenta y ocho. Pero, puesto que (testigo la experiencia) dicho movimiento aparente del Sol es mayor al principio de Piscis que al principio de Virgo, por eso la Tierra es más veloz al principio de Virgo que al principio de Piscis. De suerte que la hipótesis de los vórtices está en total desacuerdo con los fenómenos y no lleva tanto a explicar cuanto a perturbar los movimientos celestes. Pero, sobre el modo de entender cómo se efectúan estos movimientos sin vórtices en espacios libres puede verse el Libro primero, y de manera más completa se mostrará ahora en el Sistema del Mundo.



Libro tercero
SOBRE EL SISTEMA DEL MUNDO

He ofrecido en los Libros anteriores principios de filosofía, aunque no tanto filosóficos cuanto meramente matemáticos, a partir de los cuales tal vez se pueda disputar sobre asuntos filosóficos. Tales son las leyes y condiciones de los movimientos y las fuerzas, que en gran medida atañen a la filosofía. Sin embargo, para que no parezcan estériles, los he ilustrado con algunos Escolios filosóficos en los que he tratado sobre aquellas cosas que son más generales y en las cuales la filosofía parece hallar mayor fundamento, tales como la densidad y resistencia de los cuerpos, los espacios vacíos de cuerpos y el movimiento de la luz y de los sonidos. Nos falta mostrar, a partir de estos mismos principios, la constitución del sistema del mundo. Sobre esto había compuesto un tercer libro según un método popular para que fuese leído por muchos. Pero quienes no hubiesen comprendido suficientemente los principios propuestos, escasamente comprenderían la fuerza de los argumentos y tampoco abandonarían los prejuicios en los cuales andaban instalados desde muchos años atrás: por lo cual, para que el tema no entrase en controversias, trasladé la suma de aquel libro a proposiciones, al estilo matemático, de suerte que sean leídas por sólo aquellos que antes hubiesen manejado los principios. Pero, dado que allí aparecen muchas proposiciones que hasta para un lector matemáticamente documentado podrían significar una gran demora, no pido a nadie que las estudie todas; será suficiente para cualquiera leer las definiciones, las leyes de los movimientos y las tres primeras secciones del Libro primero con atención, y pasar después a este Libro sobre el sistema del mundo, y consultar según le parezca las restantes proposiciones de los libros anteriores citadas en éste.

REGLAS PARA FILOSOFAR[1]

Regla Primera

No deben admitirse más causas de las cosas naturales que aquellas que sean verdaderas y suficientes para explicar sus fenómenos.

Ya dicen los filósofos: la naturaleza nada hace en vano, y vano sería hacer mediante mucho lo que se puede hacer mediante poco. Pues la Naturaleza es simple y no derrocha en superfluas causas de las cosas.

Regla II

Por ello, en tanto que sea posible, hay que asignar las mismas causas a los efectos naturales del mismo género.

Como en el caso de la respiración en el hombre y en el animal; de la caída de las piedras en *Europa* y en *América*; de la luz en el fuego de la cocina y en el Sol; de la reflexión de la luz en la Tierra y en los planetas.

Regla III

Han de considerarse cualidades de todos los cuerpos aquellas que no pueden aumentar ni disminuir y que afectan a todos los cuerpos sobre los cuales es posible hacer experimentos.

Pues las cualidades de los cuerpos sólo mediante experimentos se esclarecen, y por lo mismo se han de establecer como generales cuantas cuadran generalmente con los experimentos; y aquellas que no pueden disminuir, tampoco pueden ser suprimidas. Ciertamente no hay que fantasear temerariamente sueños en contra de la seguridad de los experimentos, ni alejarse de la analogía de la naturaleza, toda vez que ella suele ser simple y congruente consigo misma. La extensión de los cuerpos no

se nos revela si no es por los sentidos, y no se siente por todos, pero como concierne a todos los sensibles, se atribuye universalmente. Experimentamos que muchos cuerpos son duros. Pero la dureza del todo se origina de la dureza de las partes, y de aquí concluimos con razón que son duras las partículas indivisas no sólo de los cuerpos que sentimos sino también las de todos los demás. Que todos los cuerpos son impenetrables lo inferimos no de la razón sino de la sensación. Los cuerpos que manejamos resultan ser impenetrables, y de aquí concluimos que la impenetrabilidad es una propiedad de todos los cuerpos. Inferimos que todos los cuerpos son móviles y perseveran en reposo o en movimiento gracias a ciertas fuerzas (que llamamos fuerzas de inercia) a partir de estas propiedades de los cuerpos observados. La extensión, la dureza, la impenetrabilidad, la movilidad y la fuerza de inercia del todo surgen de la extensión, dureza, impenetrabilidad, movilidad y fuerza de inercia de las partes: y de ahí concluimos que todas las partes mínimas de todos los cuerpos son extensas, duras, impenetrables, móviles y dotadas de fuerza de inercia. Y este es el fundamento de toda la filosofía. Además, hemos visto por los fenómenos, que las partes divididas de los cuerpos y contiguas entre sí pueden separarse unas de otras y que las partes indivisas pueden dividirse con la razón en partes menores es cierto por la matemática. En cambio, si esas partes distinguidas matemáticamente, pero no divididas todavía, pudieran dividirse y separarse unas de otras mediante fuerzas naturales, es cosa incierta. Pero, aunque solamente constase por un solo experimento que una partícula indivisa sufriese una división al romper un cuerpo duro y sólido, concluiremos, en virtud de esta regla, que no sólo serían separables las partes divididas, sino también que las indivisas podrían ser divididas hasta el infinito.

Finalmente, si mediante experimentos y observaciones astronómicas consta universalmente que todos los cuerpos alrededor de la Tierra gravitan hacia ella, y esto según la cantidad de materia contenida en cada uno, que la Luna gravita hacia la Tierra según su cantidad de materia, y viceversa que nuestro mar gravita hacia la Luna, que todos los planetas gravitan mutuamente entre sí y que la gravedad de los cometas hacia el Sol es similar, habrá que decir, en virtud de esta regla, que todos los cuerpos gravitan entre sí. E incluso será más fuerte el argumento sobre la gravedad universal a partir de los fenómenos, que sobre la impenetrabilidad de los cuerpos: ya que de ésta no tenemos ninguna experiencia en los cuerpos celestes y tampoco observación alguna. Sin embargo no afirmo en absoluto que la gravedad sea esencial a los cuerpos. Por fuerza ínsita entiendo solamente la fuerza de inercia. Esta es inmutable. La gravedad disminuye al alejarse de la Tierra.

Regla IV

Las proposiciones obtenidas por inducción a partir de los fenómenos, pese a las hipótesis contrarias, han de ser tenidas, en filosofía experimental, por verdaderas

exacta o muy aproximadamente, hasta que aparezcan otros fenómenos que las hagan o más exactas o expuestas a excepciones.

Debe hacerse esto para evitar que el argumento de inducción sea suprimido por las hipótesis.

Fenómeno Primero[2]

Los planetas circunjoviales con radios trazados al centro de Júpiter describen áreas proporcionales a los tiempos, y sus tiempos periódicos, estando en reposo las restrellas fijas, están en razón de las distancias a los centros elevadas a la potencia $^{3}/_{2}$.

Consta por las observaciones astronómicas. Las órbitas de estos planetas no se diferencian sensiblemente de círculos concéntricos con Júpiter, y sus movimientos se perciben uniformes en dichos círculos. Y los astrónomos coinciden en que los tiempos periódicos se hallan en razón de la potencia $^{3}/_{2}$ de los semidiámetros de las órbitas; lo cual también se hace patente por la tabla siguiente.

Tiempos periódicos de los satélites de Júpiter

$1^{d}18^{h}27'\ 34''$; $3^{d}13^{h}13'\ 42''$; $7^{d}3^{h}42'\ 36''$; $16^{d}16^{h}32'\ 9''$

Distancias de los satélites al centro de Júpiter.

Según observaciones de:	1	2	3	4	
Borelli	$5^{2}/_{3}$	$8^{2}/_{3}$	14	$24^{2}/_{3}$	***Semidiámetros de Júpiter***
Townly, *con micrómetro*	5,52	8,78	13,47	24,72	
Cassini, *con telescopio*	5	8	13	23	
Cassini, *por el eclipse de los satélites*	$5^{2}/_{3}$	9	$14^{23}/_{60}$	$25^{3}/_{10}$	
Según los tiempos periódicos…	5,667	9,017	14,384	25.299	

Con excelentes micrómetros el Sr. *Pound* ha logrado determinar la elongación de los satélites de Júpiter y el diámetro de éste del modo siguiente. La máxima elongación heliocéntrica del cuarto satélite desde el centro de Júpiter se tomó con micrómetro en un telescopio de quince pies de largo, y resultó ser a la distancia media entre la Tierra y Júpiter de aproximadamente 8′, 16″. La del tercer satélite se tomó con micrómetro en un telescopio de 123 pies de largo y resultó a la misma distancia entre la Tierra y Júpiter de 4′ . 42″. Las elongaciones máximas de los otros satélites a esa misma distancia entre la Tierra y Júpiter resultan ser por los tiempos

periódicos de 2′ 56″ 47‴ y de 1′ . 51″ . 6‴.

El diámetro de Júpiter que se ha tomado muchas veces con micrómetro en un telescopio de 123 pies de largo y reducido a la distancia media entre la Tierra y Júpiter o éste y el Sol nunca ha sido mayor de 40″ ni menor que 38″ y lo más frecuente de 39″. En telescopios más pequeños este diámetro es de 40″ o de 41″. Pues la luz de Júpiter, debido a la refrangibilidad desigual, se dilata un poco, y esta dilatación está en una razón menor respecto al diámetro de Júpiter en los telescopios más largos y perfectos que en los más cortos e imperfectos. Los tiempos en los cuales los dos satélites primero y tercero discurrían por el cuerpo de Júpiter, desde el comienzo de la entrada hasta comienzo de la salida y desde la entrada completa hasta la salida completa fueron observados gracias al mismo telescopio más largo. Y el diámetro de Júpiter a su distancia media de lai Tierra resultó por el paso del primer satélite de 37⅛″ y por el tránsito del tercero de 37⅜″. También se observó el tiempo de tránsito de la sombra del primer satélite por el cuerpo de Júpiter y de ello resulta un diámetro de Júpiter a su distancia media de la Tierra de casi 37″. Supongamos que su diámetro sea muy aproximadamente de 37¼″, y las elongaciones máximas de los satélites primero, segundo, tercero y cuarto serán respectivamente iguales a 5,965; 9,494; 15,141 y 26,63 semidiámetros de Júpiter.

Fenómeno II

Los planetas circunsaturnales con radios trazados al centro de Saturno describen áreas proporcionales a los tiempos, y sus tiempos periódicos, estando en reposo las estrellas fijas, están en razón de la potencia $^{3}/_{2}$ de las distancias; al centro del mismo.

Efectivamente, *Cassini* estableció en sus observaciones las distancias de aquéllos al centro de Saturno y los tiempos periódicos del modo siguiente:

Tiempos periódicos de los satélites de Saturno.

$1^{d}. 21^{h}. 18′ . 27″$; $2^{d}. 17^{h}. 41″ . 22″$; $4^{d}. 12^{h}. 25′ . 12″$; $15^{d}. 22^{h}. 41′ . 14″$; $79^{d}. 7^{h}. 48′ . 00″$.

Distancias de los satélites al centro de Saturno en semidiámetros del anillo

Por observaciones ……	$1^{19}/_{20}$	2½	3½	8	24
Por los tiempos periódicos ……	1,93	2,47	3,45	8	23,35

La máxima elongación del cuarto satélite desde el centro de Saturno he inferido de las observaciones que suele ser de aproximadamente ocho semidiámetros. Pero la

máxima elongación de este satélite desde el centro de Saturno tomada con el mejor micrómetro en el telescopio de *Huygens* de 123 pies de largo resultó de 8,7 semidiámetros. Y de esta observación y de los tiempos periódicos, las distancias resultantes de los satélites hasta el centro de Saturno en semidiámetros de anillo son 2,1. 2,69. 3,75. 8,7 y 25,35. En el mismo telescopio el diámetro de Saturno era al diámetro del anillo como 3 a 7, y el diámetro del anillo los días 28 y 29 de mayo de 1719 resultaba de 43″. Y por lo mismo, el diámetro del anillo a la distancia media de la Tierra y Saturno es de 42″ y el diámetro de Saturno de 18″. Esto es así en los telescopios más largos y mejores, por cuanto que las magnitudes aparentes de los cuerpos celestes en los telescopios más largos tienen mayor proporción respecto a la dilatación de la luz en los extremos de dichos cuerpos que en los más cortos. Si se eliminase toda la luz difusa, el diámetro de Saturno no será mayor de 16″.

Fenómeno III

Los cinco planetas primarios, Mercurio, Venus, Marte, Júpiter y Saturno, ciñen con sus órbitas al Sol.

Que Mercurio y Venus giran en torno al Sol se demuestra por sus fases lunares. Cuando brillan con la faz llena se hallan situados al otro lado del Sol; cuando muestran media faz se hallan a la altura del Sol, y cuando su faz es anular se hallan más acá del Sol, pasando a veces por su disco en forma de manchas. También de la faz plena de Marte junto a la conjunción con el Sol y curvada en las cuadraturas se sigue con certeza que gira en torno al Sol. Lo mismo se infiere también respecto a Júpiter y a Saturno a partir de sus fases siempre llenas: pues, que éstos brillan con luz participada del Sol está claro gracias a las sombras de los satélites proyectadas sobre ellos.

Fenómeno IV

Supuestas en reposo las estrellas fijas, los tiempos periódicos de los cinco planetas primarios y el del Sol en torno a la Tierra o de la Tierra en torno al Sol están en razón de la potencia $^{3}/_{2}$ de las distancias medias al Sol.

Esta razón hallada por *Kepler* está reconocida por todos. Enes, los tiempos periódicos son los mismos y las magnitudes de las órbitas las mismas tanto si el Sol gira en torno a la Tierra como si la Tierra gira en torno al Sol. Y sobre la medida de los tiempos periódicos hay acuerdo entre todos los astrónomos. *Kepler* y *Boulliau* determinaron mediante observaciones las magnitudes de todas las órbitas con la

mayor precisión: y las distancias medias correspondientes a los tiempos periódicos no difieren sensiblemente de las distancias que ellos hallaron, y a lo sumo son intermedias entre ellas, como se puede ver en la tabla siguiente.

Tiempos periódicos de los planetas y de la Tierra en torno al Sol en relación con las estrellas fijas, en días y décimas de día.

♄	♃	♂	♁	♀	☿
10759,275	4332,514	686,9785	365,2565	224,6176	87,96

Distancias medias de los planetas y la Tierra al Sol.

	♄	♃	♂
Según *Kepler* ……	951000	519650	152350
Según *Boulliau* ……	954198	522250	152350
Según los tiempos periódicos ……	954006	520096	152369

	♁	♀	☿
Según *Kepler* ……	100000	72400	38806
Según *Boulliau* ……	100000	72398	38585
Según los tiempos periódicos ……	100000	72333	38710

No ha lugar disputa alguna sobre las distancias de Mercurio y Venus al Sol, puesto que éstas se determinan mediante sus elongaciones del Sol. Así mismo desaparece toda disputa sobre la distancia al Sol de los planetas superiores gracias a los eclipses de los satélites de Júpiter. Efectivamente, por dichos eclipses se determina la posición de la sombra proyectada por Júpiter, y con ello se tiene la longitud heliocéntrica de Júpiter. Y comparando las longitudes heliocéntrica y geocéntrica entre sí, se determina la distancia de Júpiter.

Fenómeno v

Los planetas primarios con radios trazados a la Tierra describen áreas en nada proporcionales a los tiempos; pero con radios trazados al Sol describen áreas proporcionales a los tiempos.

Pues respecto a la Tierra, ya avanzan, ya son estacionarios o bien hasta son retrógrados: pero respecto al Sol son siempre progresivos, y ello con movimiento casi uniforme, aunque un poco más rápido en el perihelio y un poco más lento en el afelio, de suerte que la descripción de las áreas resulta igual. La Proposición es archiconocida para los astrónomos, y se demuestra en Júpiter muy bien mediante los eclipses de los satélites, mediante los cuales acabamos de decir que se calculan las

distancias y longitudes heliocéntricas de ese planeta al Sol.

Fenómeno vi

La Luna con un radio trazado al centro de la Tierra describe un área, proporcional al tiempo.

Es evidente al comparar el movimiento aparente de la Luna con su diámetro aparente. Pero el movimiento lunar se perturba un tanto por causa de la fuerza solar, mas en estos fenómenos dejo a un lado estas insensibles minucias de errores.

Proposición I. Teorema I

Las fuerzas por las cuales los planetas circunjoviales son continuamente desviados de movimientos rectilíneos y retenidos en sus órbitas se dirigen hacia el centro de Júpiter y son inversamente como los cuadrados de las distancias de los lugares a dicho centro.

La primera parte de la Proposición es evidente por el fenómeno primero y por la Proposición segunda o tercera del Libro primero; y la otra parte por el fenómeno primero y el Corolario sexto de la Proposición cuarta del mismo libro.

Lo mismo se entiende respecto a los planetas que acompañan a Saturno, por el fenómeno segundo.

Proposición II. Teorema II

Las fuerzas por las cuales los planetas primarios son continuamente desviados de movimientos rectilíneos y retenidos en sus órbitas se dirigen hacia el Sol y son inversamente como los cuadrados de las distancias al centro del mismo.

La primera parte de la Proposición es evidente por el fenómeno quinto y la Proposición segunda del Libro primero: la otra parte por el fenómeno cuarto y el Corolario sexto de la Proposición cuarta del mismo Libro. Pero esta parte de la Proposición se demuestra con la mayor precisión por la inmovilidad de los afelios. Pues una mínima aberración de la razón del cuadrado debería hacer (por el Corolario 1 de la Proposición XLV del Libro I) que el movimiento de los ápsides fuese perceptible en cada revolución, y muy grande en muchas revoluciones.

Proposición III. Teorema III

La fuerza con la cual la Luna es retenida en su órbita se dirige hacia la Tierra y es

inversamente como el cuadrado de la distancia de los lugares al centro de la Tierra.

La primera parte de la afirmación es evidente por el fenómeno sexto y la Proposición segunda o tercera del Libro primero: y la segunda parte por la gran lentitud del apogeo lunar. Pues este movimiento, que es en cada revolución de tan sólo tres grados y tres minutos hacia adelante, puede despreciarse. Pues está claro (por el Corolario 1 de la Proposición XLX del Libro I) que si la distancia de la Luna al centro de la Tierra es al semidiámetro de la Tierra como D a 1, la fuerza de la que tal movimiento se originaría será inversamente proporcional a $D^{2\frac{4}{243}}$, es decir, inversamente como aquella misma potencia de D cuyo índice es $D^{2\frac{4}{243}}$, o sea, en razón inversa de un poco más que el cuadrado de la distancia, pero que es 59¾ más próxima a la razón cuadrada que a la cúbica. Y puesto que se origina de la acción del Sol (como se dirá más tarde) puede, por ello, ser ignorada ahora. La acción del Sol, en tanto que separa a la Luna de la Tierra, es muy aproximadamente como la distancia de la Luna a la Tierra; y por lo mismo (por lo expuesto en el Corolario 2 de la Proposición XLV del Libro I) es a la fuerza centrípeta de la Luna casi como 2 a 357,45, o como 1 a $178\frac{29}{40}$. Y despreciando esta minúscula fuerza del Sol, la fuerza restante con la cual la Luna es retenida en órbita será inversamente como D^2. Cosa que quedará más clara al comparar esta fuerza con la de la gravedad, como se hace en la Proposición siguiente.

Corolario. Si la fuerza centrípeta media por la que la Luna es retenida en órbita aumentase primero en la razón de $177\frac{29}{40}$ a $178\frac{29}{40}$ y después también en la razón del cuadrado del semidiámetro de la Tierra a la distancia media entre los centros de la Tierra y la Luna, se tendrá la fuerza centrípeta lunar sobre la superficie de la Tierra, supuesto que dicha fuerza descendiendo hacia la superficie de la Tierra aumente continuamente en razón inversa al cuadrado de la altura.

Proposición IV. Teorema IV[3]

La Luna gravita hacia la Tierra y es continuamente desviada del movimiento rectilíneo y retenida en su órbita por la fuerza de la gravedad.

La distancia media de la Luna a la Tierra es en las sicigias según *Ptolomeo* y los más de los astrónomos de 59 semidiámetros terrestres, según *Vendelin* y *Huygens* 60, según *Copérnico* 60⅓, según *Street* 60⅖, y según *Tycho* 56½. Pero *Tycho*, y cuantos siguen sus tablas de refracciones, al establecer las refracciones del Sol y de la Luna (totalmente contra la naturaleza de la luz) mayores que las de las estrellas fijas, y esto en unos cuatro o cinco minutos, aumentaron en otros tantos la paralaje de la Luna, es decir, casi en la duodécima o decimoquinta parte de la paralaje total. Corríjase este error y resultará una distancia de unos 60½ semidiámetros terrestres, casi la misma

asignada por los otros. Supongamos que la distancia media en las sicigias es de 60 semidiámetros; y el período lunar completo respecto a las estrellas fijas es de 27 días, 7 horas, 43 minutos, como establecen los astrónomos, y también que la circunferencia de la Tierra es de 123249600 pies *parisinos*, como han establecido los medidores *franceses*: y si se imagina que la Luna es desposeída de todo movimiento y abandonada a sí misma, de modo que, bajo la acción de toda aquella fuerza por la cual (por el Corolario de la Proposición III) es retenida en su órbita, desciende hacia la Tierra, describirá al caer en el tiempo de un minuto un espacio de $15\frac{1}{12}$ pies *parisinos*. Esto se infiere del cálculo realizado a partir de la Proposición XXXVI del Libro I o (lo que es lo mismo) del Corolario noveno de la Proposición cuarta del mismo libro. Pues el seno verso del arco que la Luna con su movimiento medio a la distancia de sesenta semidiámetros terrestres describiría en el tiempo de un minuto es aproximadamente de $15\frac{1}{12}$ pies *parisinos*, o más exactamente de 15 pies, 1 pulgada, y $1\frac{4}{9}$ líneas. Por lo cual, toda vez que dicha fuerza aumenta al acercarse a la Tierra en razón inversa del cuadrado de la distancia, y por lo mismo en la superficie terrestre será 60 x 60 veces mayor que en la Luna, dicho cuerpo cayendo con la mencionada fuerza en nuestras inmediaciones deberá describir en el tiempo de un minuto un espacio de 60 x 60 x $15\frac{1}{12}$, y en un tiempo de un segundo $15\frac{1}{12}$ pies, o más exactamente, 15 pies, 1 pulgada $1\frac{4}{9}$ líneas. Y con esa fuerza descienden de hecho los graves en la Tierra. Pues la longitud de un péndulo oscilante al ritmo de una oscilación por segundo en la latitud de *París* es de tres pies *parisinos* y 8½ líneas, como observó *Huygens*. Y la altura que recorre un grave cayendo en el tiempo de un segundo es a la mitad de la longitud del péndulo mencionado como el cuadrado de la razón de la circunferencia del círculo a su diámetro (como también indicó *Huygens*) y, por tanto, de 15 pies *parisinos*, 1 pulgada y $1\frac{7}{9}$ líneas. Y, por lo tanto, la fuerza con la cual la Luna es retenida en su órbita, si descendiera hasta la superficie terrestre, resulta igual a la fuerza de la gravedad entre nosotros, y por lo mismo (por las Reglas I y II) es esa misma fuerza, a la cual solemos llamar gravedad. Pues si la gravedad fuera distinta de ella, los cuerpos descenderían en dirección a la Tierra con doble velocidad bajo la acción de ambas fuerzas juntas, y en el tiempo de un segundo recorrerían al caer un espacio de $30\frac{1}{6}$ pies *parisinos*, contra toda la experiencia.

Se basa este cálculo en la hipótesis de que la Tierra está en reposo. Porque si la Tierra y la Luna se mueven en torno al Sol y mientras tanto además giran en torno a su centro común de gravedad, manteniéndose la ley de la gravedad, la distancia entre los centros de la Luna y la Tierra será de aproximadamente 60½ semidiámetros terrestres, como hallará quien haga el cálculo. El cálculo se puede hacer mediante la Proposición LX del Libro I.

Escolio

La demostración de la Proposición puede explicarse más ampliamente como sigue. Si girasen muchas lunas en torno a la Tierra al igual que ocurre con el sistema de Saturno o de Júpiter, sus tiempos periódicos (por un argumento de inducción) observarían la ley de los planetas hallada por Kepler, y por lo mismo sus fuerzas centrípetas serían inversamente como los cuadrados de las distancias al centro de la Tierra, según la Prop. I de este Libro. Y si la más baja de todas esas lunas fuese muy pequeña, y casi llegase a tocar las cimas de los montes más altos, la fuerza centrípeta en cuya virtud se mantiene en su órbita vendría a ser casi igual a la gravedad de los cuerpos en las cimas de dichos montes (por el cálculo precedente) y haría que esa pequeña luna, si perdiera todo el movimiento con el que se desplaza en su órbita, al faltarle la fuerza centrífuga con la que había permanecido en su órbita, caería hacia la Tierra y esto con la velocidad con la que caen los graves en esas crestas montañosas, dada la igualdad de fuerzas con que descienden. Y si la fuerza con la que desciende dicha pequeña luna más baja fuese distinta de la gravedad, y dicha luna fuera grave también hacia la Tierra como los cuerpos en lo alto de los montes, dicha luna con ambas fuerzas juntas descendería doblemente más veloz. Por lo cual, como ambas fuerzas, éstas de los cuerpos graves y aquéllas de las lunas, tiendan al centro de la Tierra y sean semejantes entre ellas e iguales, las tales fuerzas (por las Reglas I y II) tendrán la misma causa. Y por consiguiente, la fuerza por la que la Luna es retenida en su órbita será aquella misma a la cual solemos llamar gravedad: sobre todo porque si no, la pequeña luna en la cima de la montaña o carecería de gravedad o caería dos veces más rápida que lo que suelen hacer los cuerpos graves.

Proposición V. Teorema V

Los planetas circunjoviales gravitan hacia Júpiter, los circunsaturnales hacia Saturno, y los circunsolares hacia el Sol, y por la fuerza de su gravedad son continuamente desviados de movimientos rectilíneos y retenidos en órbitas curvilíneas.

Pues las revoluciones de los planetas circunjoviales en torno a Júpiter, de los circunsaturnales en torno a Saturno, y de Mercurio, Venus y el resto de los circunsolares en torno al Sol son fenómenos del mismo género que la revolución de la Luna en torno a la Tierra; y por ello (por la Regla II) dependen de causas del mismo género: sobre todo si se ha demostrado que las fuerzas, de las cuales dependen dichas revoluciones, se dirigen hacia los centros de Júpiter, Saturno y el Sol, y al apartarse de Júpiter, Saturno y el Sol decrecen con la misma ley y proporción con la que la fuerza de la gravedad decrece al alejarse de la Tierra.

Corolario 1. Luego la gravedad se da en todos los planetas. Pues nadie duda de que Venus, Mercurio y los demás planetas sean cuerpos del mismo género que Júpiter

y Saturno. Y puesto que por la Tercera Ley del movimiento toda atracción es mutua, Júpiter gravitará hacia todos sus satélites, Saturno hacia los suyos, la Tierra hacia la Luna, y el Sol hacia todos los planetas primarios.

COROLARIO 2. La gravedad que se dirige hacia cada planeta es inversamente proporcional al cuadrado de las distancias de los lugares al centro del planeta.

COROLARIO 3. Todos los planetas gravitan entre sí, por los Corolarios 1 y 2. Y de aquí que Júpiter y Saturno en la proximidad de su conjunción, atrayéndose mutuamente, perturben sus movimientos sensiblemente, el Sol perturbe los movimientos lunares, el Sol y la Luna perturben nuestros mares, como se explicará más adelante.

ESCOLIO

Hemos llamado hasta aquí centrípeta a la fuerza por la que los cuerpos celestes son retenidos en sus órbitas. Ahora ya consta que es la gravedad, y por ello la llamaremos en lo sucesivo gravedad. Pues la causa de aquella fuerza centrípeta por la que la Luna es retenida en órbita debe ser extendida a todos los planetas por las Reglas I, II y IV.

PROPOSICIÓN VI. TEOREMA VI[4]

Todos los cuerpos gravitan hacia cada planeta y sus pesos hacia un mismo planeta, a iguales distancias del centro del planeta, son proporcionales a la cantidad de materia existente en cada uno.

Hace ya tiempo que otros observaron que el descenso de todos los graves hacia la Tierra (al menos si se elimina la retardación desigual que se debe a la mínima resistencia del aire) ocurre en tiempos iguales; y se puede registrar esa igualdad de los tiempos de la manera más exacta mediante los péndulos. He tratado de examinar esto con oro, plata, plomo, vidrio, arena, sal común, madera, agua, trigo. Me servía de dos vasijas de madera, redondas e iguales. Llenaba una de madera y colgaba el mismo peso de oro (tan exacto como podía) en el centro de oscilación del otro. Las vasijas colgando de hilos iguales de once pies constituían péndulos completamente iguales en cuanto al peso, la figura y la resistencia del aire: situados juntos, iban y venían a la vez con oscilaciones iguales durante largo tiempo. Por tanto, la cantidad de materia en el oro era (por los Corolarios 1 y 6 de la Proposición XXIV del Libro II) a la cantidad de materia en la madera como la acción de la fuerza motriz sobre todo el oro a la acción de la fuerza motriz sobre toda la madera; es decir, como un peso a otro peso. Y lo mismo en los demás. En cuerpos del mismo peso la diferencia de materia

podía detectarse con estos experimentos aunque fuese de una milésima de la materia total. Y por otra parte no hay duda de que la naturaleza de la gravedad es la misma en los planetas que en la Tierra. Imagínese, pues, que estos cuerpos terrestres se elevasen hasta la órbita de la Luna y se los dejase caer junto con la Luna privada de todo movimiento, de modo que caigan a la vez hacia la Tierra; y, por lo ya expuesto antes, es cierto que en tiempos iguales describirán espacios iguales junto con la Luna y por lo mismo que son respecto a la cantidad de materia en la Luna como sus pesos al peso de la misma. Además, toda vez que los satélites de Júpiter giran en tiempos que son como la potencia $\frac{3}{2}$ de las distancias al centro de Júpiter, por lo mismo, a iguales distancias de Júpiter sus gravedades aceleratrices resultarán iguales. Por tanto, cayendo de alturas iguales en tiempos iguales describirán espacios iguales; lo mismo que ocurre con los graves en nuestra Tierra. Y por la misma razón los planetas circunsolares, lanzados desde iguales distancias del Sol, en su descenso hacia el Sol describirán espacios iguales en tiempos iguales. En cambio, las fuerzas por las que cuerpos desiguales son acelerados igualmente, son como los cuerpos, es decir, los pesos son como las cantidades de materia en los planetas. Además, que los pesos de Júpiter y de sus satélites hacia el Sol son proporcionales a la cantidad de materia existente en ellos es evidente por el movimiento sumamente regular de los satélites, por el Corolario 3 de la Proposición LXV del Libro I. Pues si alguno de ellos fuera atraído hacia el Sol en grado mayor que el debido a su cantidad de materia de lo que lo son los otros, el movimiento de los satélites (por el Corolario 2 de la Proposición LXV del Libro I) resultaría perturbado por la desigualdad de la atracción. Si, a iguales distancias del Sol, un satélite resultase más grave hacia el Sol en razón de su cantidad de materia que Júpiter en razón de la suya en una razón cualquiera dada, tal como *d* a *e* por ejemplo, la distancia entre el centro del Sol y el centro de la órbita del satélite siempre sería mayor que la distancia entre el centro del Sol y el centro de Júpiter aproximadamente como la raíz cuadrada de esa razón como hallé con un cálculo que hice. Y si el satélite fuese menos grave hacia el Sol en la proporción dicha de *d* a *e*, la distancia del centro de la órbita del satélite al Sol sería menor que la distancia del centro de Júpiter al Sol en razón de dicha raíz cuadrada. Por tanto, si a iguales distancias del Sol, la gravedad aceleratriz de un satélite cualquiera hacia el Sol fuese mayor o menor que la gravedad aceleratriz de Júpiter hacia el Sol tan sólo en una milésima parte de la gravedad total, la distancia del centro de la órbita del satélite hasta el Sol sería mayor o menor que la distancia de Júpiter hasta el Sol en 1/2000 de la distancia total, es decir, una quinta parte de la distancia del satélite exterior al centro de Júpiter: excentricidad de la órbita que sería bien sensible. Pero las órbitas de los satélites son concéntricas a Júpiter, y por lo mismo las gravedades aceleratrices de Júpiter y de los satélites hacia el Sol son iguales entre ellas. Y por la misma razón los pesos hacia el Sol de Saturno y sus compañeros, a iguales distancias del Sol, son como las cantidades de materia que hay en ellos; y los pesos hacia el Sol de la Luna y la Tierra o bien son nulos o son muy exactamente proporcionales a sus masas. Pero

alguno es el peso, por los Corolarios 1 y 3 de la Proposición V.

Y también los pesos de cada una de las partes de un planeta cualquiera hacia otro son entre sí como la materia que hay en cada parte. Porque, si unas partes gravitasen más y otras menos que en razón de su cantidad de materia, el planeta entero, según el género de partes en que abunde más, gravitará más o menos que en razón de su cantidad de materia total. Y no importa si dichas partes son internas o externas. Pues si, por ejemplo, los cuerpos terrestres que están a nuestro alrededor los elevamos imaginariamente hasta la órbita de la Luna y se comparan con el cuerpo de la misma, si sus pesos fuesen a los pesos de las partes externas de la Luna como las cantidades de materia en unos y otra, pero respecto a los pesos de las partes interiores en razón mayor o menor, tales cuerpos estarían respecto al peso total de la Luna en una proporción mayor o menor: contra lo mostrado más arriba.

COROLARIO 1. De aquí que los pesos de los cuerpos no dependan de sus formas y texturas. Pues si pudieran variar con las formas, serían mayores o menores según la variedad de las formas, a igualdad de materia, en contra totalmente de la experiencia.

COROLARIO 2. Todos los cuerpos que están alrededor de la Tierra son graves hacia ella; y los pesos de todos los que distan igual del centro de la Tierra son como las cantidades de materia en cada uno de ellos. Esta es una cualidad de todos aquellos en los que es posible hacer experimentos, y en consecuencia, por la Regla III, ha de afirmarse de todos. Si el éter u otro cuerpo cualquiera estuviese desprovisto por completo de gravedad o gravitase menos que en razón de su cantidad de materia: puesto que éste (en opinión de *Aristóteles*, *Descartes* y otros) no se diferencia de los demás cuerpos sino sólo en la forma de la materia, tal cuerpo podría mediante el cambio gradual de forma transformarse en un cuerpo de la misma condición que aquellos que gravitan lo más posible según su cantidad de materia, y viceversa, los cuerpos sumamente graves, adquiriendo gradualmente la forma de aquél, podrían perder su gravedad gradualmente. Y de esta suerte los pesos dependerían de las formas de los cuerpos y podrían variar con las formas, contra lo que se ha probado en el Corolario anterior.

COROLARIO 3. Todos los espacios no están igualmente llenos. Pues si todos los espacios estuviesen igualmente llenos, la gravedad específica del fluido con el cual se llena la región del aire, por la gran densidad de su materia, en nada cedería a la gravedad específica del mercurio, del oro o de otro cuerpo cualquiera muy denso; y por ello ni el oro, ni otro cuerpo cualquiera podría descender en el aire. Pues los cuerpos en los fluidos, salvo que sean específicamente más graves, no descienden en absoluto. Pero si la cantidad de materia en un espacio dado puede disminuir por alguna rarefacción cualquiera ¿por qué no podría disminuir infinitamente?

COROLARIO 4. Si todas las partículas sólidas de todos los cuerpos son de la misma densidad, y sin poros no son susceptibles de rarefacción, se da de hecho el vacío. Digo que son de la misma densidad aquellas cuyas fuerzas inerciales son como las magnitudes.

COROLARIO 5. La fuerza de la gravedad es de distinta naturaleza que la fuerza magnética. Pues la atracción magnética no es como la materia atraída. Unos cuerpos son más atraídos, otros menos, muchos no lo son. Y la fuerza magnética en uno y el mismo cuerpo puede aumentar y disminuir y es bastante mayor a veces respecto a la cantidad de materia que la fuerza de la gravedad, y al alejarse del imán decrece en una razón de la distancia no cuadrada sino casi cúbica, en cuanto he podido comprobar con algunos experimentos un tanto rudimentarios.

PROPOSICIÓN VII. TEOREMA VII[5]

La gravedad ocurre en todos los cuerpos y es proporcional a la cantidad de materia existente en cada uno.

Hemos probado ya que todos los planetas gravitan entre sí, y también que la gravedad hacia cada uno de ellos considerado individualmente es inversamente proporcional al cuadrado de la distancia desde cada lugar al centro del planeta. De lo cual se sigue que (por la Proposición LXIX del Libro I y sus Corolarios) la gravedad hacia todos es proporcional a la materia existente en ellos.

Por lo demás, dado que todas las partes de un planeta A gravitan hacia otro planeta B, y la gravedad de una parte cualquiera es a la gravedad del todo como la materia de la parte a la materia del todo, y para toda acción haya igual reacción (por la tercera Ley del movimiento), el planeta B gravitará a la inversa hacia todas las partes del planeta A, y su gravedad hacia cada parte será a su gravedad hacia el todo como la materia de la parte a la materia del todo. Q. E. D.

COROLARIO 1. Por consiguiente, la gravedad hacia todo el planeta surge y se compone de la gravedad hacia cada parte. De lo cual tenemos ejemplos en las atracciones magnéticas y eléctricas. Pues la atracción entera hacia el todo surge de las atracciones hacia cada parte. Para la gravedad esto se entenderá imaginando que muchos planetas menores se reúnen en un globo y constituyen uno mayor. Pues la fuerza del todo deberá originarse de las fuerzas de las partes componentes. Si alguien objeta que todos los cuerpos que nos rodean deberían gravitar entre sí según esta ley, mientras que no percibimos en absoluto una gravedad de este estilo, debo responder que la gravedad en estos cuerpos al ser respecto a la gravedad de toda la Tierra como son estos cuerpos al cuerpo de la Tierra entera, es bastante menor que la que es observable.

COROLARIO 2. La gravitación hacia cada partícula igual de un cuerpo es inversamente proporcional al cuadrado de la distancia de los lugares a las partículas. Es evidente por el Corolario 3 de la Proposición LXXIV del Libro I.

Proposición viii. Teorema viii[6]

Si la materia de dos globos que gravitan entre sí es homogénea en todos los lugares que equidistan de los centros por todos los lados, el peso de cada uno de ellos hacia el otro será inversamente como el cuadrado de la distancia entre los centros.

Después de que hube hallado que la gravedad hacia el planeta entero surge y se compone de las gravedades hacia las partes, y que es hacia cada parte inversamente proporcional a los cuadrados de las distancias a las partes, dudaba de si la dicha proporción inversa del cuadrado se cumpliría exactamente en la fuerza entera compuesta de muchas fuerzas o si sólo de modo aproximado. Pues podría ocurrir que la proporción, que en las distancias más grandes se cumplía con bastante exactitud, en las inmediaciones de la superficie del planeta, debido a las distancias desiguales de las partículas y a las desemejanzas de los lugares, variase notablemente. Mas al fin, por las Proposiciones lxxv y lxxvi del Libro i y sus Corolarios, comprendí la verdad de la Proposición de que aquí se trata.

Corolario 1. De aquí se pueden hallar y comparar entre sí los pesos de los cuerpos en los distintos planetas. Pues los pesos de cuerpos iguales que giran en círculos en torno a los planetas son (por el Corolario 2 de la Proposición iv del Libro i) directamente como los diámetros de los círculos e inversamente como los cuadrados de los tiempos periódicos, y los pesos sobre las superficies de los planetas o a otras distancias cualesquiera del centro son mayores o menores (por esta Proposición) en razón inversa del cuadrado de las distancias. De este modo, a partir de los tiempos periódicos de Venus en torno al Sol de 224 días y 16¾ horas, del satélite exterior de Júpiter alrededor de Júpiter de 16 días y $16^{8}/_{15}$ horas, del satélite de Huygens en torno a Saturno de 15 días y $22^{2}/_{3}$ horas y de la Luna en torno a la Tierra de 27 días 7 horas y 43 minutos, comparados con la distancia media de Venus al Sol y con las elongaciones heliocéntricas máximas del satélite exterior de Júpiter respecto al centro de éste de 8′, 16″, del satélite Huygens del centro de Saturno de 3′, 4″, y de la Luna del centro de la Tierra de 10′, 33″, haciendo un cálculo hallé que los pesos de cuerpos iguales a iguales distancias del centro del Sol, de Júpiter, de Saturno y de la Tierra eran hacia el Sol, hacia Júpiter, hacia Saturno y la Tierra como 1, $\frac{1}{1067}$, $\frac{1}{3021}$ y $\frac{1}{169282}$, respectivamente, y aumentando o disminuyendo las distancias, los pesos aumentaban o disminuían en razón del cuadrado: los pesos de cuerpos iguales hacia el Sol, Júpiter, Saturno y la Tierra a las distancias de 10 000, 997, 791, y 109 de sus centros, y por ende en sus superficies, serán como 10 000, 943,529, y 435 respectivamente. Cuánto pesan los cuerpos en la superficie de la Luna se verá después.

Corolario 2. También se sigue la cantidad de materia en cada planeta. Pues las cantidades de materia en los planetas son como sus fuerzas a iguales distancias de sus centros, esto es, en el Sol, Júpiter, Saturno y la Tierra como 1, $\frac{1}{1067}$, $\frac{1}{3021}$ y $\frac{1}{169282}$, respectivamente. Si la paralaje solar se estableciera mayor o menor que 10″ 30‴, la cantidad de materia en la Tierra deberá aumentarse o disminuirse en razón cúbica.

Corolario 3. También aparecen así las densidades de los planetas. Pues los pesos de cuerpos iguales y homogéneos hacia esferas homogéneas son en las superficies de éstas como los diámetros de las mismas, por la Proposición lxxii del Libro i, y por lo mismo las densidades de esferas heterogéneas son como aquellos pesos divididos por los diámetros de las esferas. Los diámetros verdaderos del Sol, de Júpiter, de Saturno y de la Tierra eran entre si como 10 000, 997, 791, y 109, y los pesos hacia ellos como 10 000, 943, 529, y 435 respectivamente, y por tanto las densidades son como 100, 94½, 67 y 400. La densidad de la Tierra resultante de este cálculo no depende de la paralaje solar, sino que se obtiene de la paralaje lunar, con lo cual aquí se establece con exactitud. Por consiguiente el Sol es algo más denso que Júpiter y éste que Saturno, mientras la Tierra es cuatro veces más densa que el Sol. Ya que éste, por su enorme calor, se enrarece. La Luna en cambio, es más densa que la Tierra, como se mostrará más tarde.

Corolario 4. Son, por tanto, más densos los planetas que son más pequeños, a igualdad de condiciones restantes. De este modo la fuerza de la gravedad se acerca más a la igualdad en sus superficies. Pero, a igualdad de condiciones, son más densos los planetas más próximos al Sol; Júpiter más que Saturno y la Tierra más que Júpiter. Efectivamente, había que situar a distintas distancias del Sol a los planetas para que cada uno según su densidad disfrutase de un mayor o menor grado de calor solar. Nuestra agua, si la Tierra estuviese situada en la órbita de Saturno se solidificaría y en la órbita de Mercurio se evaporaría al instante. Pues la luz del Sol, a la cual es proporcional el calor, es siete veces más densa en la órbita de Mercurio que aquí entre nosotros: y con el termómetro he hallado que con siete veces el calor estival del Sol al agua hierve. Y no hay duda de que la materia de Mercurio está acomodada al calor, y por lo mismo, sea más densa que la nuestra, toda vez que la materia más densa, exija mayor calor para realizar las operaciones naturales.

Proposición ix. Teorema ix

La gravedad, a partir de la superficie de los planetas hacia abajo, decrece en razón muy aproximada de la distancia al centro.

Si la materia del planeta es uniforme en cuanto a la densidad, esta Proposición se cumple exactamente, por la Proposición LXXIII del Libro I. Por lo tanto no habrá más error que el que se siga de la desigualdad de la densidad.

Proposición X. Teorema X

El movimiento de los planetas en los cielos puede conservarse durante mucho tiempo.

Se ha mostrado en el Escolio de la Proposición XL del Libro II que un globo de agua congelada, moviéndose libremente en nuestra atmósfera y describiendo la longitud de su semidiámetro, perdería por la resistencia del aire $^{1}/_{4586}$ parte de su movimiento. Pero la misma proporción resulta para cualquier globo muy aproximadamente de cualquier tamaño y a cualquier velocidad. Ahora bien, que el globo de nuestra Tierra es más denso que si fuera de agua lo infiero como sigue. Si este globo fuera todo de agua, las cosas todas que fuesen menos densas que el agua, por su menor gravedad específica, emergerían y flotarían. Por esa razón, un globo terroso cubierto de agua por todas partes, si fuese de mayor rareza que el agua, emergería por un lado, y toda el agua discurriendo desde allí se concentraría en la región opuesta. E igual es el caso de nuestra Tierra rodeada de mares por la mayor parte. Si ella no fuere más densa, emergería de los mares y con una parte de la misma de acuerdo con su grado de levedad sobresaldría del agua, concentrándose todos los mares hacia la región opuesta. Por la misma razón las manchas solares son más leves que la materia luminosa solar a la que se sobreponen. Y al formarse cualquiera de los planetas, cuando la masa era fluida, toda materia más grave tendía desde la masa líquida hacia el centro. Por lo cual, al ser la Tierra común periférica casi dos veces más grave que el agua, y un poco más abajo en las minas tres, cuatro y hasta cinco veces más densa, es probable que toda la materia de la Tierra junta sea cinco o seis veces más pesada que si fuera de agua; sobre todo porque, como se ha mostrado, la Tierra es casi cuatro veces más densa que Júpiter. Por ello si Júpiter fuera algo más denso que el agua, en el tiempo de treinta días en el que recorre la longitud de 459 semidiámetros suyos, perdería en un medio de la densidad de nuestro aire casi una décima parte de su movimiento. Pero como la resistencia de los medios disminuye en razón del peso y de la densidad, como en el caso del agua que al ser 13⅗ más leve que el mercurio es menos resistente en esa proporción; y el aire que es 860 veces más leve que el agua es también menos resistente en esa misma proporción, y si se considera en los cielos donde el peso del medio en el que se mueven los planetas

disminuye enormemente, la resistencia casi vendrá a desaparecer. Hemos mostrado, en efecto, en el Escolio de la Proposición XXII del Libro II, que si ascendemos hasta una altura de doscientas millas sobre la Tierra el aire sería allí más raro que sobre la superficie terrestre en una proporción de 30 a 0,000000000000 3998 y también como 75000000000000 a 1 aproximadamente. Por consiguiente, el planeta Júpiter girando en un medio de la densidad de ese aire más alto en el tiempo de 1 000 000 de años no perdería de su movimiento por la resistencia del medio ni una millonésima parte. En los espacios próximos a la Tierra nada hay que produzca resistencia salvo los vapores y exhalaciones del aire. Si éstos son extraídos con todo cuidado de un vaso de cristal cilíndrico cerrado, los graves caen dentro del vaso con toda libertad y sin resistencia sensible alguna; el mismo oro y una pluma muy leve soltados a la vez caen con igual velocidad, y describiendo en su caída una altura de cuatro, seis u ocho pies tocan el suelo a la vez, como se ha visto experimentalmente. Y por tanto, si nos trasladamos a los cielos, vacíos de aire y de exhalaciones, los planetas y los cometas sin resistencia sensible alguna se moverán por esos espacios durante larguísimo tiempo.

Hipótesis I

El centro del sistema del mundo está en reposo.

Esto lo conceden todos, si bien unos sostienen que la Tierra y otros que el Sol reposan en el centro del sistema. Veamos qué se sigue de aquí.

Proposición XI. Teorema XI

El centro común de gravedad de la Tierra, el Sol y todos los planetas está en reposo.

Pues (por el Corolario IV de las Leyes) dicho centro o está en reposo o se desplaza uniformemente en línea recta. Pero si tal centro se desplaza continuamente, el centro del mundo también se moverá, contra la Hipótesis.

Proposición XII. Teorema XII

El Sol es agitado por un movimiento permanente, pero jamás se aleja mucho del centro común de gravedad de todos los planetas.

Pues toda vez que (por el Corolario 2 de la Proposición VIII) la materia en el Sol es a la materia en Júpiter como 1067 a 1, y la distancia de Júpiter al Sol es al

semidiámetro solar en razón ligeramente mayor, el centro común de gravedad de ambos cae en un punto muy poco distante fuera de la superficie del Sol. Por la misma razón, puesto que la materia del Sol es a la de Saturno como 3021 a 1, y la distancia de Saturno al Sol es ligeramente menor respecto al semidiámetro solar, el punto común de gravedad de ambos caerá en un punto ligeramente interior a la superficie solar. Y repitiendo los elementos de este cálculo, si los planetas todos y la Tierra se situaran a un lado del Sol, el centro común de gravedad de todos ellos apenas se alejaría del centro del Sol el valor de un diámetro solar. En los demás casos la distancia entre esos centros siempre sería menor. Por consiguiente, dado que dicho centro de gravedad siempre está en reposo, el Sol a tenor de las diferentes ubicaciones de los planetas, se moverá hacia todas partes, pero jamás se apartará mucho de dicho centro.

COROLARIO. De aquí que el centro común de gravedad de la Tierra, el Sol y el resto de los planetas, haya de ser considerado como el centro del mundo. Puesto que la Tierra, el Sol y todos los demás planetas gravitan mutuamente entre sí y, por tanto, se mueven siempre según las leyes del movimiento en razón de sus fuerzas de gravedad, es evidente que sus centros móviles no se pueden tomar por el centro inmóvil del mundo. Si hubiera que situar en dicho centro al cuerpo hacia el que todos los otros gravitan en grado máximo (cual es la opinión corriente) tal privilegio habría que concedérselo al Sol. Mas como el Sol se mueve, hay que elegir un punto en reposo, respecto al cual el centro del Sol se separa muy poco, y del cual se alejaría aún menos si el Sol fuera más denso y más grande de suerte que se moviera menos.

PROPOSICIÓN XIII. TEOREMA XIII

Los planetas se mueven en elipses que tienen un foco en el centro del Sol, y con radios trazados a dicho centro describen áreas proporcionales a los tiempos.

Ya hemos considerado estos movimientos más arriba partiendo de los fenómenos. Ahora que se conocen los principios del movimiento, de ellos deducimos «*a priori*» los movimientos celestes. Puesto que los pesos de los planetas hacia el Sol son inversamente como los cuadrados de las distancias al centro del Sol, si el Sol reposase y los demás planetas no actuasen mutuamente entre ellos, sus órbitas serían elípticas, teniendo al Sol en el foco común, y describirían áreas proporcionales a los tiempos (por las Proposiciones I y XI y el Corolario 1 de la Proposición XIII del Libro I), pero las acciones de los planetas entre ellos son mínimas (de suerte que pueden despreciarse) y perturban menos los movimientos en elipses de los planetas en torno al Sol (por la Proposición LXVI del Libro I) que si tales movimientos se realizasen en torno a un Sol en reposo.

No se puede despreciar enteramente la acción de Júpiter sobre Saturno. Pues la gravedad hacia Júpiter es a la gravedad hacia el Sol (a iguales distancias) como 1 a 1067; por ello, en la conjunción de Júpiter y Saturno, al ser la distancia de Saturno a Júpiter respecto a la distancia de Saturno al Sol casi como 4 a 9, la gravedad de Saturno hacia Júpiter será a la gravedad de Saturno hacia el Sol como 81 a 16 x 1067, o como 1 a 211 aproximadamente. Y a esto se debe la perturbación de la órbita de Saturno en cada conjunción de este planeta con Júpiter, lo suficientemente sensible como para extrañar a los astrónomos. Según la diferente situación del planeta en estas conjunciones su excentricidad aumenta o disminuye y su afelio se adelanta o se retrasa notablemente, y el movimiento medio se acelera o se retarda alternativamente. Sin embargo el error total en su movimiento en torno al Sol debido a tamaña fuerza (además del movimiento medio) puede casi evitarse situando el foco inferior de su órbita en el centro común de gravedad de Júpiter y el Sol (por la Proposición LXVII del Libro I) y por tanto cuando resulta mayor no sobrepasa los dos minutos. Y el error máximo en el movimiento medio apenas supera los dos minutos por año. Por otra parte, en la conjunción de Júpiter y Saturno las gravedades aceleratrices del Sol hacia Saturno, de Júpiter hacia Saturno y de Júpiter hacia el Sol vienen a ser casi como 16, 81, y $\left(\frac{16 \text{ x } 81 \text{ x } 3021}{25}\right)$, o sea, 156 609, por lo cual la diferencia de las fuerzas de gravedad del Sol hacia Saturno y de Júpiter hacia Saturno es a la fuerza de gravedad de Júpiter hacia el Sol como 65 a 156 609 y también como 1 a 2409. Pero el poder máximo de Saturno para perturbar el movimiento de Júpiter es proporcional a esta diferencia, de suerte que la perturbación de la órbita de Júpiter es mucho menor que la de Saturno. Las perturbaciones de las demás órbitas son aún más pequeñas, salvo la de la Tierra sensiblemente perturbada por la Luna. El centro común de gravedad de la Tierra y la Luna discurre por una elipse en torno al Sol situado en el foco de la misma y, con un radio trazado al Sol, describe en ella áreas proporcionales a los tiempos, pero la Tierra gira con un movimiento mensual en torno a dicho centro común.

PROPOSICIÓN XIV. TEOREMA XIV

Los afelios y nodos de las órbitas están en reposo.

Los afelios están en reposo, por la Proposición XI del Libro I, al igual que los planos de las órbitas, por la Proposición I del mismo Libro, y al reposar los planos también reposan los nodos. No obstante, debido a las acciones mutuas de los planetas y cometas en revolución se originan algunas desigualdades, pero tan insignificantes que aquí pueden despreciarse.

Corolario 1. También están en reposo las estrellas fijas, toda vez que conservan las posiciones dadas respecto a los afelios y a los nodos.

Corolario 2. Y por tanto, al no percibirse paralaje alguna de las mismas como consecuencia del movimiento anual de la Tierra, sus fuerzas no producen efecto alguno sensible en nuestro sistema por la enorme distancia a que se hallan. Además de que las estrellas fijas, esparcidas igualmente por todas las partes del cielo, destruyen sus fuerzas mutuas con atracciones opuestas, por la Proposición lxx del Libro i.

Escolio[7]

Puesto que los planetas más próximos al Sol (es decir, Mercurio, Venus, Tierra y Marte) debido a la pequeñez de sus cuerpos influyen muy poco unos sobre otros, sus afelios y nodos están en reposo, salvo en la medida en que sean perturbados por las fuerzas de Júpiter, Saturno y los cuerpos superiores. Y de aquí se sigue, por la teoría de la gravedad, que sus afelios se mueven un poco hacia adelante respecto a las fijas y esto en una proporción de la potencia $\frac{3}{2}$ de sus distancias al Sol. De suerte que si el afelio de Marte en cien años se adelanta respecto a las fijas 33′ 20″, los afelios de la Tierra, de Venus y de Mercurio en cien años se adelantarán 17′ 40″; 10′ 53″ y 4′ 16″ respectivamente. Estos movimientos, por su insignificancia, se deprecian en esta Proposición.

Proposición xv. Problema i

Hallar los diámetros principales de las órbitas.

Han de tomarse en razón de la potencia $\frac{2}{3}$ de los tiempos periódicos, por la Proposición xv del Libro i, y después hay que aumentar cada uno en razón de la suma de las masas del Sol y cada planeta a la primera de las dos medias proporcionales entre dicha suma y el Sol, por la Proposición lx del Libro i.

Proposición xvi. Problema ii

Hallar las excentricidades y los afelios de las órbitas.

El problema se resuelve por la Proposición xviii del Libro i.

Proposición xvii. Teorema xv

Los movimientos diarios de los planetas son uniformes y la libración lunar se debe a su movimiento diario.

Es evidente por la Ley i del movimiento y por el Corolario 22 de la Proposición lxvi del Libro i. Ciertamente Júpiter gira respecto a las estrellas fijas en 9 h. 56′, Marte en 24 h. 39′, Venus en casi 23 h., la Tierra en 23 h. 56′, el Sol en 25½ días y la Luna en 27 d. 7 h. 43′. Está claro por los fenómenos que esto es así. Las manchas en el cuerpo solar vuelven al mismo sitio en el disco tras 27½ días aproximadamente respecto a la Tierra; y por tanto, respecto a las fijas el Sol gira aproximadamente en 25½ días. Y dado que el día lunar girando uniformemente sobre su eje es de un mes, la misma cara de ella mirará siempre hacia el otro foco de su órbita muy aproximadamente, y por consiguiente se desviará hacia acá o hacia allá de la Tierra según se halle el sitio de dicho foco. Esta es la libración longitudinal de la Luna; pues la latitudinal se debe a la latitud de la Luna y a la inclinación de su eje hacia el plano de la eclíptica. Esta teoría de la libración lunar fue expuesta por *N. Mercator* en su astronomía editada al principio de 1676 con detalle y gracias a las cartas que le envié. Con un movimiento semejante parece que gira alrededor de su eje el satélite exterior de Saturno, mirando siempre hacia éste con la misma cara. Pues al girar en torno a Saturno, cada vez que llega a la parte oriental de su órbita se ve escasamente y muchas veces deja de verse, cosa que puede deberse, como ha hecho notar *Cassini*, a algunas manchas en esa parte del cuerpo que entonces se orienta hacia la Tierra. Y también el satélite exterior de Júpiter parece girar en torno a su eje con un movimiento semejante, toda vez que en la parte del cuerpo opuesta a Júpiter tiene una mancha que aparece como si estuviera en Júpiter cada vez que pasa entre Júpiter y nuestros ojos.

Proposición xviii. Teorema xvi

Los ejes de los planetas son menores que los diámetros trazados perpendicularmente a dichos ejes.

Los planetas, si se suprime todo movimiento circular diario, deberían adoptar figura esférica debido a la igual gravedad de las partes por todos lados. Gracias a ese movimiento circular ocurre que las partes que se apartan del eje intentan ascender junto al ecuador. Por lo cual, si la materia fuese fluida aumentaría con su ascenso los diámetros ecuatoriales, mientras haría disminuir el eje con su descenso hacia los polos. Y así el diámetro de Júpiter (coincidiendo las observaciones de los astrónomos) se ve más corto entre los polos que de oriente a occidente. Y por la

misma razón, a no ser que nuestra Tierra fuera algo más alta en el ecuador que en los polos, los mares se hundirían hacia los polos, y elevándose junto al ecuador lo inundarían allí todo.

Proposición xix. Problema iii[8]

Hallar la proporción del eje de un planeta respecto a los diámetros perpendiculares al mismo.

Nuestro *Norwood*, midiendo hacia 1635 la distancia de 905 751 pies *londinenses* entre *Londres y York*, y observando una diferencia de latitud de 2.º 28′ dedujo que el valor de un grado era de 367 196 pies londinenses, es decir, de 57 300 toesas parisinas.

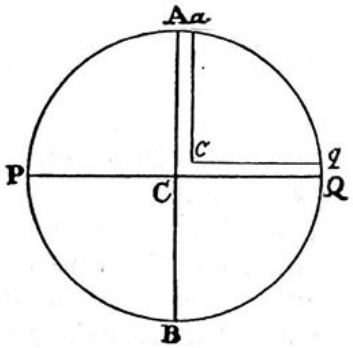
Picard, midiendo un arco de un grado y 22′ 55″ sobre el meridiano entre *Amiens* y *Malvoisine*, halló que el grado de arco equivalía a 57 060 toesas parisinas. *Cassini*, padre, midió la distancia sobre el meridiano desde *Collioure*, en el *Rosellón*, hasta el observatorio de París, y su hijo añadió la distancia desde el observatorio hasta la torre de la ciudad de *Dunquerque*. La distancia entera resultó ser de 486 156½ toesas, y la diferencia de latitudes entre *Collioure* y *Dunquerque* de 8 grados, 31′ 11⅝″. Por consiguiente, el arco de un grado resulta ser de 57 061 toesas parisinas. De esas medidas se sigue que el perímetro de la Tierra viene a ser de 123 249 600 pies parisinos y su semidiámetro de 19 615 800 pies parisinos, en la hipótesis de que la Tierra sea esférica.

En la latitud de *París* un grave en el tiempo de un segundo recorre al caer 15 pies, 1 pulgada y 1⁷⁄₉ líneas, como se dijo antes, es decir, 2173⁷⁄₉ líneas. El peso del cuerpo disminuye por el peso del aire ambiente. Y supongamos que esta disminución es del orden de 0,0011 del peso total, de suerte que dicho grave, cayendo en el vacío, recorrerá 2174 líneas en el tiempo de un segundo.

En cada día sideral de 23 h. 56′ 4″, un cuerpo que gire uniformemente en un círculo a la distancia de 19 615 800 pies del centro describirá en el tiempo de un segundo un arco de 1433, 46 pies, cuyo seno verso es de 0,0523656 pies, o de 7,54064 líneas. Por lo tanto, la fuerza con que los cuerpos descienden en la latitud de París es a la fuerza centrífuga de los cuerpos en el ecuador debida al movimiento diurno de la Tierra como 2174 a 7,54064.

La fuerza centrífuga de los cuerpos en el ecuador es a la fuerza centrífuga con que los cuerpos tienden a despegarse de la Tierra en la latitud de *París* de 48 grados, 50′, 10″, como el cuadrado del radio al seno del complemento de la susodicha latitud, esto es, como 7,54064 a 3, 267. Añádase esta fuerza a la fuerza con la que caen los graves a dicha latitud de *París* y el cuerpo en esa latitud cayendo con toda la fuerza de la gravedad en el tiempo de un segundo recorrerá 2177,267 líneas, o 15 pies parisinos, 1 pulgada y 5,267 líneas. Y toda la fuerza de la gravedad en dicha latitud será a la fuerza centrífuga de los cuerpos en el ecuador de la Tierra como 2177,267 a 7,54064 o como 289 a 1.

Por ello, si se representa la figura de la Tierra por APBQ, ahora ya no esférica, sino engendrada por la rotación de una elipse en torno al eje menor PQ, y ACQ*qca* es un canal lleno de agua que va desde el polo Q*q* al centro C*c* y después hasta el ecuador A*a*, el peso del agua en el brazo del canal AC*ca* tendrá que ser al peso del agua en el otro brazo del canal QC*cq* como 289 a 288, debido a que la fuerza centrífuga generada por el movimiento circular soportará una parte de las 289 del peso y la suprime mientras que el peso de las 288 partes del otro brazo sostiene a las restantes. Además, haciendo un cálculo (a partir del Corolario 2 de la Proposición xci del Libro i) hallo que si la Tierra constase de materia uniforme y quedase desprovista de todo movimiento, y su eje PQ fuese al diámetro AB como 100 a 101, la gravedad en el punto Q de la Tierra sería a la gravedad en el mismo punto Q de una esfera descrita con centro en C y radio PC o QC como 126 a 125. Y por la misma razón, la gravedad en el punto A hacia un esferoide descrito en torno al eje AB por rotación de la elipse APBQ es la gravedad en el mismo punto A de una esfera descrita con centro en C y radio AC, como 125 a 126. Pues la gravedad hacia la Tierra en el punto A es la media proporcional entre las gravedades hacia la esfera y hacia el susodicho esferoide; por cuanto que la esfera, disminuyendo el diámetro PQ en la razón de 101 a 100, tiende a la figura de la Tierra, al igual que esta figura, al disminuir en la misma razón un tercer diámetro que fuese perpendicular a los dos anteriores AB y PQ, se convierte en el susodicho esferoide; y la gravedad en el punto A, en uno y otro caso, disminuye en la misma proporción muy aproximadamente. Por consiguiente, la gravedad en el punto A hacia una esfera descrita con centro en C y radio AC es a la gravedad en el punto A hacia la Tierra como 126 a 125½, y la gravedad en el punto Q hacia una esfera descrita con centro en C y radio QC es a la gravedad en el punto A hacia la esfera descrita con centro en C y radio AC, en razón de los diámetros (por la Proposición lxxii del Libro i), es decir, como 100 a 101. Únanse ahora estas tres razones, 126 a 125, 126 a 125½, y 100 a 101; resultará que la gravedad en Q hacia la Tierra es a la gravedad en A hacia la Tierra como 126 x 126 x 100 a 125 x 125½ x 101, o como 501 a 500.



Ahora, dado que (por el Corolario 3 de la Proposición XCI del Libro I) la gravedad en cada brazo de uno y otro canal AC*ca* y QC*cq* es como la distancia de cada lugar al centro de la Tierra, si dichos brazos se separan mediante superficies transversales y equidistantes en partes proporcionales a los todos, los pesos de cada parte en el brazo AC*ca* serán a los pesos de cada otra parte en el otro brazo como las magnitudes y las gravedades aceleratrices conjuntamente; es decir, como 101 a 100 y 500 a 501, o sea, como 505 a 501. Y por consiguiente, si la fuerza centrífuga de una parte cualquiera en el brazo AC*ca* generada por el movimiento diurno fuese a su propio peso como 4 a 505, de suerte que restase cuatro partes del peso total de cada una dividida en 505, los pesos en uno y otro brazo permanecerían iguales y por tanto el fluido permanecería en equilibrio. Sin embargo la fuerza centrífuga de cada una de las partes es a su peso como 1 a 289, es decir, la fuerza centrífuga que debería ser $\frac{4}{505}$ partes del peso es solamente $\frac{1}{289}$ parte del mismo. Y por lo mismo digo, por la regla áurea, que si la fuerza centrífuga $\frac{4}{505}$ hace que la altura del agua en el brazo AC*ca* supere a la altura del agua en el brazo QC*cq* en $\frac{1}{100}$ parte de la altura total, la fuerza centrífuga $\frac{1}{289}$ hará que el exceso de altura en el brazo AC*ca* sea solamente $\frac{1}{289}$ parte de la altura del otro brazo QC*cq*. Por lo tanto, el diámetro de la Tierra en el ecuador es al diámetro de la misma en los polos como 230 a 229. De suerte que, siendo el semidiámetro medio terrestre, según la medición de *Picard*, de 19 615 800 pies parisinos, o de 3923,16 millas (suponiendo para la milla 5000 pies), la Tierra será más alta en el ecuador que en los polos en un exceso de 85 472 pies, o de $17\frac{1}{10}$ millas. Y su altura en el ecuador será casi de 19 658 600 pies, mientras en los polos

será de 19 573 000 pies.

Si un planeta fuera mayor o menor que la Tierra, manteniendo su misma densidad y su mismo tiempo periódico de revolución diaria, se mantendrá la proporción de la fuerza centrífuga a la gravedad y, por ello, se mantendrá también la proporción del diámetro entre polos respecto al diámetro en el ecuador. Pero si el movimiento diario se acelerase o retardase en una proporción cualquiera, la fuerza centrífuga aumentará o disminuirá en razón cuadrada de aquella proporción y, por lo mismo, la diferencia entre diámetros aumentará o disminuirá muy aproximadamente en razón del cuadrado de la susodicha proporción. Y si la densidad del planeta aumentase o disminuyese en una razón cualquiera, también la gravedad tendente hacia el mismo aumentará o disminuirá en la misma razón, y la diferencia entre diámetros, a la inversa, disminuirá en la razón del aumento de la gravedad o aumentará en la razón de la disminución de la gravedad. De donde, al girar la Tierra respecto a las fijas en 23 h. 56′, mientras Júpiter lo hace en 9 h. 56′ y ser los cuadrados de los tiempos como 29 a 5, mientras las densidades de los cuerpos en giro son como 400 a 94½, la diferencia de diámetros de Júpiter será a su diámetro menor como $\frac{29}{5}$ x $\frac{400}{94\frac{1}{2}}$ x $\frac{1}{229}$ a 1, o sea, como 1 a 9⅓, aproximadamente. De donde, al ser su diámetro mayor de 37″, su diámetro menor entre polos será de 33″ 25‴. Auméntese casi 3″ por la luz difusa, y los diámetros aparentes de este planeta resultan ser 40″ y 36″ 25‴; que resultan ser entre sí como 11⅙ a 10⅙, aproximadamente. Esto resulta así en la hipótesis de que el cuerpo de Júpiter sea uniformemente denso. Pero si su cuerpo fuese más denso hacia el plano ecuatorial que hacia los polos, sus diámetros podrían hallarse entre ellos en la relación de 12 a 11, o de 13 a 12, o quizá de 14 a 13. *Cassini* el año 1691 observó que el diámetro de Júpiter de este a oeste superaba en casi una decimoquinta parte de su longitud al otro diámetro. Y nuestro *Pound*, con un telescopio de 123 pies de largo y un buen micrómetro, midió los diámetros de Júpiter en 1719 con los resultados siguientes:

	Tiempo		**Diámetro máximo**	**Diámetro mínimo**	**Relación mutua**
	Día	***Hora***	***Partes***	***Partes***	
Enero	28	6	13,40	12,28	Como 12 a 11
Febrero	6	7	13,12	12,20	Como 13¾ a 12¾
Marzo	9	7	13,12	12,08	Como $12\frac{2}{3}$ a $11\frac{2}{3}$
Abril	9	9	12,32	11,48	Como 14½ a 13½

La teoría cuadra bastante bien con los fenómenos. Pues los planetas se calientan algo más por la luz del Sol hacia el ecuador, por lo que allí se endurecen algo más que hacia los polos.

Por otra parte, que la gravedad disminuye en el ecuador debido a la rotación diaria de nuestra Tierra, y por tanto que la Tierra asciende allí más que en los polos (siempre que sea uniformemente densa) se verá por los experimentos con péndulos de

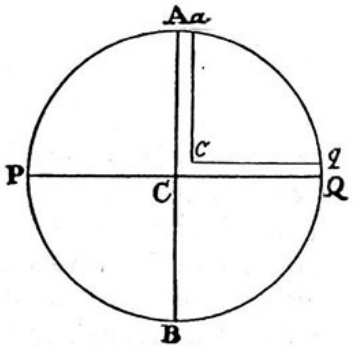
los que se da cuenta en la Proposición siguiente.

Proposición xx. Problema iv[9]

Hallar y comparar entre ellos los pesos de los cuerpos en las distintas regiones de esta Tierra.

Puesto que los pesos de los brazos desiguales del canal de agua ACQ*qca* son iguales, y los pesos de las partes, proporcionales a los brazos enteros y colocadas semejantemente en ellos, son entre sí como los pesos de los todos, y por lo mismo son iguales entre sí, los pesos de las partes iguales y situadas semejantemente en los brazos serán inversamente como los brazos, es decir, inversamente como 230 a 229. E igual es la razón entre cuerpos homogéneos e iguales cualesquiera situados semejantemente en los brazos del canal. Sus pesos son inversamente como los brazos, es decir, inversamente como la distancia de los cuerpos al centro de la Tierra. Por lo tanto, si los cuerpos se hallan en las partes superiores del canal, esto es, en la superficie de la Tierra, sus pesos serán entre sí como los inversos de sus distancias al centro de la Tierra. Y por la misma razón los pesos en otras regiones cualesquiera por toda la superficie de la Tierra son entre sí como los inversos de las distancias de esos lugares al centro, y por tanto, bajo la hipótesis de que la Tierra es un esferoide, están dados en proporción.

De donde se viene a concluir en el Teorema de que el incremento del peso, caminando desde el ecuador hacia los polos, es aproximadamente como el seno verso del doble de la latitud, o, lo que es lo mismo, como el cuadrado del seno recto de la latitud. Y casi en esa misma razón se halla el incremento del arco de los grados de latitud sobre el meridiano. Y por lo mismo, puesto que la latitud de *París* es de 48.º 50′, la de los lugares del ecuador de 00.º 00′ y la de los polos de 90.º y los duplos de los senos versos son 11 134, 00 000, y 20 000, siendo el radio 10 000, y la gravedad en los polos es a la gravedad en el ecuador como 230 a 229, mientras el exceso de la gravedad en el polo es a la gravedad en el ecuador como 1 a 229, el exceso de la gravedad en la latitud de *París* será a la gravedad en el ecuador como 1 x (11 334/20 000) a 229, o también como 667 a 2 290 000. Por consiguiente, las fuerzas totales de la gravedad en esos lugares serán entre ellas como 2 295 667 a 2 290 000. Por lo tanto, dado que la longitud de los péndulos que oscilan en tiempos iguales es como las gravedades, y en la latitud de *París* la longitud de un péndulo que oscile cada segundo es de 3 pies parisinos y 8½ líneas, o mejor, debido al peso del aire, de 8⅝ líneas, la longitud del péndulo en el ecuador se verá superada por la longitud del péndulo sincrónico de *París* con un exceso de 1,087 líneas. Y con cálculos similares se construye la tabla siguiente.



Latitud del lugar	Longitud del péndulo	Medida de un grado en el meridiano	Latitud del lugar	Longitud del péndulo	Medida de un grado en el meridiano
grados	*pies líneas*	*toesas*	*grados*	*pies líneas*	*toesas*
0	3 · 7,468	56637	46	3 · 8,461	57022
5	3 · 7,482	56642	7	3 · 8,494	57035
10	3 · 7,526	56659	8	3 · 8,528	57048
15	3 · 7,596	56687	9	3 · 8,561	57061
20	3 · 7,692	56724	50	3 · 8,594	57074
25	3 · 7,812	56769	55	3 · 8,756	57137
30	3 · 7,948	56823	60	3 · 8,907	57196
35	3 · 8,099	56882	65	3 · ?,044	57250
40	3 · 8,261	56945	70	3 · 9,162	57295
1	3 · 8,294	56958	75	3 · 9,258	57332
2	3 · 8,327	56971	80	3 · 9,329	57360
3	3 · 8,361	56984	85	3 · 9,372	57377
4	3 · 8,394	56997	90	3 · 9,387	57382
45	3 · 8,428	57010			

Por esta tabla se ve con claridad que la desigualdad de los grados es tan pequeña que en asuntos geográficos la figura de la Tierra puede considerarse esférica: sobre todo si la Tierra es algo más densa hacia el plano ecuatorial que hacia los polos.

Ahora bien, algunos astrónomos enviados a lejanas tierras para hacer observaciones astronómicas, han observado que los relojes de péndulo oscilaban más lentamente cerca del ecuador que en nuestras regiones. El primero fue el señor *Richer* que lo observó el año de 1672 en la isla de *Cayena*. Pues, mientras observaba el paso de las fijas por el meridiano en el mes de agosto, se dio cuenta de que su reloj se movía más lentamente que el movimiento medio del Sol con una diferencia de 2′ 28″ por día. Después preparó un péndulo simple que oscilaba al segundo, medido mediante un buen reloj y anotó la longitud del péndulo simple una vez por semana durante diez meses. Después, vuelto a *Francia*, comparó esa longitud del péndulo con la longitud del péndulo en *París* (que era de 3 pies parisinos y 8⅗ líneas) y halló que resultaba más corta en 1¼ líneas.

Más tarde nuestro *Halley* hacia 1677, navegando hacia la isla de *Sta. Elena*, observó que su reloj de péndulo se movía allí más lentamente que en Londres, pero no anotó la diferencia. Pero acortó la longitud del péndulo en más de una octava parte de pulgada, o sea, en una línea y media. Para hacer esto, como no alcanzaba el tornillo de la parte inferior del péndulo, interpuso un anillo de madera entre la tuerca del tornillo y el peso del péndulo.

Después, en el año 1682 los señores *Varin* y *Des Hayes* hallaron que la longitud del péndulo oscilante al segundo en el Real Observatorio de París era de 3 pies y $8\frac{5}{9}$ líneas. Con el mismo método hallaron que en la isla de *Gore* la longitud del péndulo isócrono era de 3 pies y $6\frac{5}{9}$ líneas, con una diferencia de 2 líneas respecto al anterior. Y ese mismo año navegando a las islas de *Guadalupe y Martinica* hallaron que la longitud del péndulo isócrono en esas islas era de 3 pies y 6½ líneas.

Posteriormente el señor *Couplet*, hijo, en junio de 1697, ajustó su reloj de péndulo al movimiento medio del Sol en el Real Observatorio de París de manera que concordase durante un tiempo largo con el movimiento del Sol. Después, navegando hasta *Lisboa* halló que en el mes de noviembre siguiente el reloj marchaba más lentamente que antes siendo la diferencia de 2′ 13″ cada 24 horas. Y al siguiente mes de marzo navegando hasta *Paraiba* observó que allí su reloj marchaba más despacio que en *París* siendo la diferencia de 4′ 12″ cada 24 horas. Afirma que el péndulo oscilando al segundo era en *Lisboa* 2½ líneas más corto y en *Paraiba* $3\frac{2}{3}$ más corto que en *París*. Más correctas hubieran sido las diferencias si las hubiera establecido en 1⅓ y $2\frac{5}{9}$. Ya que estas diferencias corresponden a las diferencias de tiempos de 2′ 13″ y 4′ 12″. No hay que fiarse de sus toscas observaciones.

En los años siguientes (1699 y 1700) *Des Hayes* viaja de nuevo a América y determinó que en las islas de *Cayena* y *Granada* la longitud del péndulo oscilando al segundo era algo menor de 3 pies y 6½ líneas, y que en la isla de *San Cristóbal* dicha longitud era de 3 pies y 6¾ líneas mientras que en la isla de *Sto. Domingo* era de 3 pies y 7 líneas.

Y en el año 1704, *Fueille* determinó en *Porto Bello* en América que la longitud del péndulo oscilante al segundo era de tres pies parisinos y sólo $5\frac{7}{12}$ líneas, es decir, casi tres líneas más corta que en *París*, pero la observación es errónea. Pues, pasando de allí a la isla *Martinica*, halla que la longitud del péndulo isócrono era solamente de tres pies parisinos y $5\frac{10}{12}$ líneas.

Ahora bien, la latitud de *Paraiba* es de 6.º 38′ sur, la de *Porto Bello*, 9.º 33″ norte y las latitudes de las islas de *Cayena*, *Gore*, *Guadalupe*, *Martinica*, *Granada*, *San Cristóbal* y *Santo Domingo*, son respectivamente, 4.º 55′, 14.º 40′, 14.º 00′, 14.º 44′, 12.º 6′, 17.º 19′ y 19.º 48′. Y el exceso de la longitud del péndulo en París sobre las longitudes de los péndulos isócronos observadas en esas latitudes es un poco mayor que el correspondiente a tenor de la tabla de longitudes del péndulo calculada poco más arriba. Y por tanto la Tierra es algo más alta en el ecuador que lo previsto en el cálculo anterior y algo más densa hacia el centro que en los pozos próximos a la

superficie, salvo que los calores de la zona tórrida aumentasen un tanto la longitud del péndulo.

Picard ha observado, efectivamente, que una varilla de hierro que en invierno medía un pie de larga con el frío de los hielos, calentada al fuego resultaba de un pie y un cuarto de línea. Después, *De la Hire* observó que una varilla de hierro que en tiempo igualmente invernal era de seis pies de longitud, cuando se exponía al Sol del verano resultaba de seis pies y dos tercios de línea. En el primer caso el calor era mayor que en el segundo pero en éste el calor fue mayor que el de las partes externas del cuerpo humano, pues lo metales con el calor del sol estival se calientan en gran medida. Pero la varilla del péndulo en el reloj oscilatorio nunca se expone al calor del sol estival, nunca recibe un calor igual al calor de la superficie externa del cuerpo humano. Por lo mismo la varilla del péndulo del reloj de tres pies de larga, será un poco más larga en verano que en invierno, pero el exceso escasamente superará ¼ de línea. Por consiguiente, la diferencia total de longitud de péndulos que sean isócronos en las diferentes latitudes, no puede atribuirse a las diferencias de calor. Tampoco a los errores de los astrónomos enviados por *Francia*. Pues, aunque sus observaciones no concuerdan exactamente entre ellas, sin embargo los errores son tan pequeños que se pueden despreciar. Y todos convienen en algo, en que los péndulos isócronos son más cortos en el ecuador que en el Real Observatorio de París, siendo la diferencia no menor de una línea y cuarto y no mayor de $2\frac{2}{3}$ líneas. Por las observaciones de *Richer* en *Cayena* la diferencia resultaba de una línea y cuarto. Por las observaciones de *Des Hayes* dicha diferencia corregida resultaba de una línea y media o de una línea y tres cuartos de línea. Por las observaciones menos exactas de otros dicha diferencia resultaba de casi dos líneas. Esta diferencia pudo deberse en parte a errores de observación, en parte a la desemejanza de las partes interiores de la Tierra y de la altura de los montes y en parte a los distintos calores del aire.

En mi opinión, una varilla de hierro de tres pies de larga es en Inglaterra en invierno un sexto de línea más corta que en verano. Por los calores ecuatoriales, réstese esta cantidad de la diferencia de una línea y cuarto observada por *Richer* y quedará $1\frac{1}{12}$ líneas; cosa que concuerda bastante bien con $1\frac{87}{1000}$ líneas establecidas antes por la teoría. *Richer* repitió cada semana durante diez meses las observaciones hechas en *Cayena* y comparó las longitudes de la varilla de hierro del péndulo anotadas allí con las anotadas de igual modo en *Francia*. Diligencia y cautela que parecen faltar en otros observadores. Si sus observaciones son de fiar, la Tierra será más alta en el ecuador que en los polos con un exceso de casi 17 millas, como la teoría había determinado más arriba.

Proposición xxi. Teorema xvii

Los puntos equinocciales retroceden y el eje de la Tierra en cada revolución anual se balancea inclinándose dos veces hacia la eclíptica y retornando dos veces a la posición inicial.

Se sigue del Corolario 20 de la Proposición LXXI del Libro I. Pero este movimiento de nutación debe ser muy pequeño, y apenas o, ni apenas, sensible.

Proposición XXII. Teorema XVIII[10]

Todos los movimientos lunares, y todas las desigualdades de los movimientos se siguen de los principios expuestos.

De la Proposición LXV del Libro I se sigue que los planetas mayores, mientras giran en torno al Sol, pueden arrastrar a otros planetas menores girando en torno a ellos mismos, y estos menores deben girar en elipses que tienen sus focos en los centros de los mayores. Pero por la acción del Sol sus movimientos resultarán perturbados de diversas formas, y sufrirán de las desigualdades que se observan en nuestra Luna. Esta, en efecto (por los Corolarios 2, 3, 4 y 5 de la Proposición LXVI del Libro I) se mueve rápidamente y con un radio trazado a la Tierra describe un área mayor que la correspondiente al tiempo y tiene una órbita menos curva y por ello se acerca más a la Tierra en las sicigias que en las cuadraturas, salvo en la medida en que la excentricidad lo impide. Ahora bien, la excentricidad es máxima (por el Corolario 9 de la Proposición LXVI) cuando el apogeo de la Luna ocurre en las sicigias, mientras que es mínima cuando acontece en las cuadraturas; y de aquí que la Luna es más veloz y más próxima a nosotros en el perigeo y más lenta y remota en el apogeo, en las sicigias que en las cuadraturas. Además, el apogeo avanza mientras que los nodos retroceden, pero con movimiento desigual. Pues el apogeo (por los Corolarios 7 y 8 de la Proposición LXVI) avanza más velozmente en las sicigias y regresa más lentamente en las cuadraturas, y el exceso del progreso sobre el regreso tiene una resultante anual hacia adelante. Los nodos, en cambio (por el Corolario 2 de la Proposición LXVI) reposan en las sicigias y retroceden muy rápidamente en las cuadraturas. Pero la latitud máxima de la Luna es mayor en sus cuadraturas (por el Corolario 10 de la Proposición LXVI) que en las sicigias, y el movimiento medio es más lento en el perihelio terrestre que en su afelio. Y éstas son las desigualdades más notables registradas por los astrónomos.

Pero hay también otras desigualdades no observadas por los anteriores astrónomos por medio de las cuales los movimientos lunares sufren perturbaciones tales que hasta la fecha no han podido ser reducidas a regla cierta alguna. Las velocidades o los movimientos horarios de los apogeos y nodos lunares y sus ecuaciones así como las excentricidades máximas en las sicigias y mínimas en las

cuadraturas, o la desigualdad llamada variación, aumentan y disminuyen anualmente (por el Corolario 14 de la Proposición LXVI) en razón del cubo del diámetro aparente del Sol. Y la variación, además, aumenta o disminuye en razón del cuadrado del tiempo entre cuadraturas muy aproximadamente (por los Corolarios 1 y 2 del Lema X y el Corolario 16 de la Proposición LXVI del Libro I), pero esta desigualdad, en los cálculos astronómicos, suele incluirse en la prostaféresis lunar y confundirse con ella.

Proposición XXIII. Problema V

Derivar los movimientos desiguales de los satélites de Júpiter y de Saturno a partir de los movimientos lunares.

Partiendo de los movimientos de nuestra Luna se derivan los movimientos análogos de las lunas o satélites de Júpiter del modo siguiente. El movimiento medio de los nodos del satélite exterior de Júpiter es al movimiento medio de los nodos de nuestra Luna en una razón compuesta de la razón cuadrada entre el tiempo periódico de la Tierra en torno al Sol y el tiempo periódico de Júpiter en torno al Sol, y de la razón simple entre el tiempo periódico del satélite en torno a Júpiter y el tiempo periódico de la Luna en torno a la Tierra (por el Corolario 16 de la Proposición LXVI del Libro I) y por tanto en cien años dicho nodo avanza 8.º 24′ hacia adelante. Los movimientos medios de los nodos de los satélites interiores son al movimiento del anterior como sus tiempos periódicos al tiempo periódico del exterior (por el mismo Corolario) y por tanto están dados. En cambio el movimiento del ápside de un satélite cualquiera hacia adelante es al movimiento regresivo de sus nodos como el movimiento del apogeo de nuestra Luna a su movimiento de los nodos (por el mismo corolario) y por tanto están dados. Pero este movimiento de los ápsides así hallado debe disminuirse en razón de 5 a 9 o de 1 a 2 aproximadamente, por razones que no hay espacio aquí de exponer. Las ecuaciones máximas de los nodos y del ápside de un satélite cualquiera casi son a las ecuaciones máximas de los nodos y el ápside de la Luna respectivamente como el movimiento de los nodos y del ápside de los satélites en el tiempo de una revolución de las ecuaciones primeras al tiempo de los nodos y del apogeo de la Luna en el tiempo de una revolución de las ecuaciones segundas. La variación de un satélite visto desde Júpiter es a la variación de la Luna como son entre sí todos los movimientos de los nodos en los tiempos en los que el satélite y la Luna regresan hacia el Sol, por el mismo Corolario; y por tanto, para el satélite exterior, no pasa de 5″ 12‴.

Proposición XXIV. Teorema XIX

El flujo y reflujo del mar proceden de las acciones del Sol y de la Luna.

Cada día, tanto lunar como solar, debe crecer y decrecer el mar dos veces como se desprende de los Corolarios 19 y 20 de la Proposición LXVI del Libro I, lo mismo que se desprende que la máxima altura del agua en los mares profundos y abiertos sigue en un espacio menor de seis horas al paso de los astros por el meridiano del lugar, como ocurre en todo el recorrido del mar *Atlántico* y *Etiópico* desde *Francia* al *Cabo de Buena Esperanza* al igual que en el mar *Pacífico* en la costa *chilena* y *peruana*; en todos los cuales litorales la marea ocurre a las dos, tres o cuatro horas, salvo cuando el movimiento se propaga desde el océano profundo a través de fondos de poco calado y se retarda hasta la hora quinta, sexta, séptima o más. Cuento las horas desde la llegada de ambos astros al meridiano del lugar, tanto bajo como sobre el horizonte, y entiendo por horas del día lunar las vigesimocuartas partes del tiempo en el cual la Luna con el movimiento aparente diurno regresa al meridiano del lugar. La fuerza del Sol o de la Luna para elevar el mar es máxima en el mismo paso del astro por el meridiano del lugar. Pero la fuerza impresa en el mar en ese momento permanece durante algún tiempo y aumenta por causa de la nueva fuerza impresa después, hasta que el mar alcanza la altura máxima, cosa que ocurre al cabo de una o dos horas aunque más frecuentemente sea al cabo de tres cuando alcance las costas, o incluso más, si el mar es poco profundo. Pero los dos movimientos que los dos astros provocan, no se ven separadamente, sino que constituyen un cierto movimiento mixto. En la conjunción de los astros o en la oposición también se juntan sus efectos, y se componen el flujo o reflujo máximos. En las cuadraturas el Sol aporta agua cuando la Luna la retrae, y la retrae cuando la Luna la aporta; y de la diferencia de los efectos surgirá la marea más pequeña de todas. Y puesto que, bajo el testimonio de la experiencia, es mayor el efecto de la Luna que el del Sol la altura máxima del agua ocurre aproximadamente en la tercera hora lunar. Fuera de las sicigias y de las cuadraturas, la marea máxima que por la sola fuerza lunar debería siempre ocurrir en la tercera hora lunar y con la sola fuerza solar en la tercera hora solar, con la composición de ambas fuerzas ocurre en un tiempo intermedio más próximo a la tercera hora lunar; por ello al pasar la Luna de las sicigias hacia las cuadraturas, cuando la tercera hora solar precede a la tercera lunar, la altitud máxima de las aguas precede también a la tercera hora lunar, y esto con intervalo máximo poco después de los octantes lunares; y con intervalos parecidos la marea máxima seguirá a la tercera hora lunar al pasar la Luna de las cuadraturas hacia las sicigias. Esto es así en mar abierto. Pues en las desembocaduras de los ríos las mareas máximas, en iguales circunstancias, alcanzan más tarde su «*ἀκμία*» (acmé).

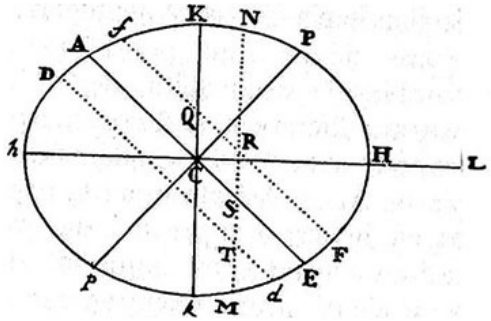
Pero los efectos de los astros dependen de sus distancias a la Tierra. Pues a menores distancias son mayores sus efectos, y menores a mayores distancias, y esto en razón del cubo de sus diámetros aparentes. Por tanto, el Sol, en invierno, hallándose en el perigeo, produce efectos mayores, y hace que las mareas en las

sicigias sean algo mayores y en las cuadraturas algo menores (en circunstancias iguales) que en verano; y la Luna en el perigeo de cada mes genera mareas más grandes que quince días antes o después, cuando se halla en el apogeo. De donde se sigue que dos mareas absolutamente máximas no se sigan una a otra en sicigias inmediatamente sucesivas.

También dependen los efectos de ambos astros de su declinación o distancia al ecuador. Pues si el astro se situase en el polo, atraería hacia allí a cada parte de agua de modo constante, sin aumento y disminución de la acción, con lo que no provocaría ninguna reciprocación de movimientos. Por consiguiente, los astros al apartarse del ecuador hacia los polos, pierden gradualmente sus efectos, y por ello generarán menores mareas en las sicigias solsticiales que en las equinocciales. En cambio, en las cuadraturas solsticiales provocarán mayores mareas que en las cuadraturas equinocciales, debido a que la Luna situada ahora en el ecuador supera en su efecto al del Sol más que nunca. Por tanto ocurren las mareas máximas en las sicigias y las mínimas en las cuadraturas de los astros, en torno a las fechas de uno y otro equinoccio. Y a una marea máxima en las sicigias acompaña siempre, como muestra la experiencia, una marea mínima en las cuadraturas. Mas, por la menor distancia del Sol a la Tierra en invierno que en verano, ocurre que las mareas máximas y mínimas precedan más frecuentemente que sigan al equinoccio de primavera y más frecuentemente le sigan que le precedan en el otoño.

También dependen los efectos de los astros de la latitud de los lugares. Represente A*p*EP a la Tierra cubierta de agua profunda por todas partes; C su centro; P, *p* los polos; AE el ecuador; F un punto cualquiera fuera del ecuador; F*f* el paralelo del punto; D*d* el paralelo correspondiente al anterior al otro lado del ecuador; L el lugar ocupado por la Luna tres horas antes; H el lugar de la Tierra perpendicularmente debajo; *h* el lugar opuesto al anterior; K, *k* los lugares distantes de los anteriores 90.º, CH, C*h* las alturas máximas del mar medidas desde el centro de la Tierra; y CK, C*k* las alturas mínimas: y si con los ejes H*h*, K*k*, se describe una elipse y con la revolución de dicha elipse se obtiene el esferoide HPK*hpk*, esta figura representará muy aproximadamente a la del mar, y CF, C*f*, CD, C*d*, serán las alturas del mar en los lugares F, *f*, D, *d*. Además, en la susodicha revolución de la elipse un punto cualquiera N, describirá un círculo NM secante de los paralelos F*f*, D*d* en unos lugares cualesquiera R, T, y del ecuador AE en S; CN será entonces la altura del mar en todos los lugares R, S, T, situados en este círculo. De aquí que, en la revolución diaria de cualquier punto F, el aflujo será máximo en F, en la tercera hora después del paso de la Luna por el meridiano sobre el horizonte; después, el reflujo máximo en Q tres horas después del ocaso de la Luna; después el aflujo máximo en *f* a la tercera hora siguiente al paso de la Luna por el meridiano bajo el horizonte; por fin el reflujo máximo en Q en la tercera hora después del orto lunar; y el aflujo posterior en *f* será menor que el aflujo anterior en F. Pues el mar entero se divide en dos flujos completamente hemisféricos, uno en el hemisferio KH*k* vertido hacia el lado boreal y

el otro en el hemisferio K*hk*, a los cuales se puede llamar flujo boreal y flujo austral. Estos flujos oponiéndose entre ellos y alternativamente acuden siempre a los meridianos de cada lugar, mediando entre ellos un tiempo de doce horas lunares. Y puesto que las regiones boreales participan en mayor grado del flujo boreal y las australes del austral, de ello se sigue que surjan mareas alternativamente mayores y menores en cada uno de los lugares, fuera del ecuador, por los que nacen o se ponen los astros. Pero la marea mayor, cuando la Luna declina hacia el vértice del lugar, ocurre casi en la hora tercera del paso de la Luna por el meridiano sobre el horizonte, y al cambiar la declinación de la Luna empieza a disminuir. Y la máxima diferencia de flujos ocurre en la época de los solsticios; sobre todo si el nodo ascendente de la Luna viene a ocurrir al principio de Aries. Así atestigua la experiencia que las mareas matutinas en invierno superan a las vespertinas y las vespertinas superan en verano a las matutinas, diferencia que en *Plymouth* alcanza un pie de altura, mientras que en *Bristol* llega a quince pulgadas, según las observaciones de *Colepress* y de *Sturmy*.



Pero los movimientos descritos hasta aquí cambian un tanto debido a aquella fuerza de reciprocación de las aguas, por la que la marea del mar, incluso habiendo cesado las acciones de los astros, podría continuar durante algún tiempo. Esta conservación del movimiento impreso disminuye la diferencia de mareas alternas; y hace mayores las mareas inmediatamente posteriores a las sicigias, mientras hace disminuir a las inmediatamente posteriores a las cuadraturas. De lo que resulta que las mareas alternas en *Plymouth* y *Bristol* no difieren entre sí más de un pie o quince pulgadas de altura; y que las mareas mayores de todas en esos mismos puertos no sean las primeras desde las sicigias, sino las terceras. También se retrasan todos los movimientos por su tránsito por los vados, hasta el punto de que las mareas mayores de todas en determinados estrechos y desembocaduras de ríos, sean las cuartas o incluso las quintas después de las sicigias.

Además puede ocurrir que la marea se propague desde el océano hacia un mismo puerto por estrechos diferentes, y llegue primero por un estrecho que por otro; en cuyo caso la misma marea, dividida en dos arribadas o más, podría componer movimientos nuevos de distintas clases. Imaginemos que dos mareas iguales llegan al

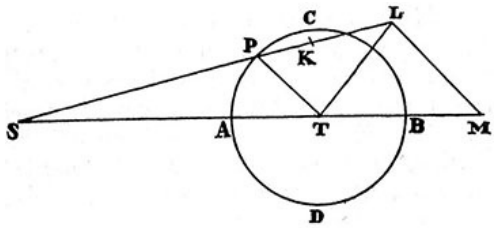
mismo puerto desde lugares distintos, de las cuales la primera precede a la segunda en seis horas, y acontezca en la tercera hora después del paso de la Luna por el meridiano del puerto. Si la Luna en este paso por el meridiano se hallare en el ecuador, vendrán cada seis horas aflujos iguales, los cuales incidiendo sobre los reflujos mutuos los igualarán a los aflujos, haciendo de este modo que en el espacio de ese día el agua repose tranquila. Si en ese momento la Luna declinase del ecuador, las mareas ocurrirán en el océano alternativamente mayores y menores, como se ha dicho; y de aquí se propagarán hacia el dicho puerto dos aflujos mayores y dos menores, en veces alternas. Pero los dos aflujos mayores compondrán una muy grande elevación de agua en el medio entre ambos, el aflujo mayor y menor harán que el agua ascienda a una altura media en el medio entre ambos, y entre los dos aflujos menores el agua ascenderá a la altura mínima. De este modo, en el espacio de veinticuatro horas, el agua alcanzará la máxima altura no dos veces como suele ocurrir, sino solamente una vez y una sola vez la mínima; y la altura máxima, si la Luna declina hacia el polo sobre el horizonte del lugar, ocurrirá en la hora sexta o en la treinta después del paso de la Luna por el meridiano, y al cambiar la declinación de la Luna se tornará en reflujo. Un ejemplo de todo lo cual ha ofrecido *Halley* a partir de las observaciones de los navegantes en el puerto de *Batshae* en el reino de *Tonquin* con una latitud boreal de 20.º 50′. Allí el agua al día siguiente del paso de la Luna por el ecuador se estanca, después, al declinar la Luna hacia el Norte empieza a fluir y refluir, no dos veces como en otros puertos, sino una sola vez cada día; y la marea ocurre al ocaso de la Luna mientras la bajamar máxima ocurre al orto lunar. Con la declinación de la Luna aumenta esta marea hasta el día séptimo u octavo, y después durante otros siete días decrece gradualmente como antes había crecido; y cesa con el cambio de declinación de la Luna, y al fin se torna en reflujo. A partir de ese momento el reflujo coincide con el ocaso de la Luna y el aflujo con el nacimiento, hasta que de nuevo cambie la declinación. Doble es la entrada a este puerto y a los estrechos próximos, una desde el océano chino entre el continente y la isla de *Leucoma* y la otra desde el mar Indico entre el continente y la isla de *Borneo*. Dejo para observaciones de las costas vecinas el determinar si la marea procedente del mar Indico en doce horas y la procedente del mar de China en seis a través de dichos estrechos, incidiendo así en las horas lunares tercera y novena, llegan a componer movimientos de este género, o si más bien es otra la condición de aquellos mares.

Hasta ahora he expuesto las causas de los movimientos de la Luna y de los mares. Ahora será conveniente añadir algo sobre la cantidad de los movimientos.

Proposición xxv. Problema vi

Hallar las fuerzas del Sol para perturbar los movimientos de la luna.

Represente S el Sol, T la Tierra, P la Luna, CABD la órbita de la Luna. Sobre SP tómese SK igual a ST; y sea SL a SK en razón cuadrada de SK a SP, y trácese LM paralela a PT; y si la gravedad aceleratriz de la Tierra hacia el Sol se representa por la distancia ST o SK, será SL la gravedad aceleratriz de la Luna hacia el Sol. Esta fuerza se compone de las partes SM y LM, de las cuales LM y la parte TM de SM perturban el movimiento de la Luna, como se expuso en la Proposición LXVI y sus Corolarios del Libro I. En tanto que la Tierra y la Luna giran en torno a un centro común de gravedad, también será perturbado el movimiento de la Tierra en torno a dicho centro por fuerzas similares; pero las sumas, tanto de las fuerzas como de los movimientos son susceptibles de ser referidos a la Luna, y representar las sumas de las fuerzas por medio de las líneas TM y ML análogas a ellas. La fuerza ML en su valor medio es a la fuerza centrípeta, con la que la Luna puede girar en torno a la Tierra en reposo a la distancia PT, como la razón cuadrada de los tiempos de la Luna en torno a Ja Tierra y de la Tierra en torno al Sol (por el Corolario 17 de la Proposición LXVI del Libro I), es decir, como la razón cuadrada de 27d. 7h. 43′ a 365 d. 6h. 9′, o sea, como 1000 a 178 725 o también como 1 a $178^{29}/_{40}$. Pero hemos hallado en la Proposición Cuarta que si la Tierra y la Luna giran; en torno a su centro común de gravedad, su distancia mutua media será de 60½ semidiámetros medios terrestres muy aproximadamente. Y la fuerza con la que la Luna puede girar en torno a la Tierra en reposo a la distancia PT de 60½ semidiámetros terrestres es a la fuerza con la que puede girar, en el mismo tiempo a la distancia de 60 semidiámetros, como 60½ a 60; y esta fuerza es a la fuerza de la gravedad entre nosotros como 1 a 60 x 60, muy aproximadamente. Por consiguiente, la fuerza media ML es a la fuerza de la gravedad en la superficie de la Tierra como 1 x 60½ a 60 x 60 x 60 x $178^{29}/_{40}$, o también como 1 a 638092,6. De suerte que por la proporción de las líneas TM, ML está dada también la fuerza TM; y éstas son las fuerzas del Sol por las que resultan perturbados los movimientos de la Luna. Q. E. I.

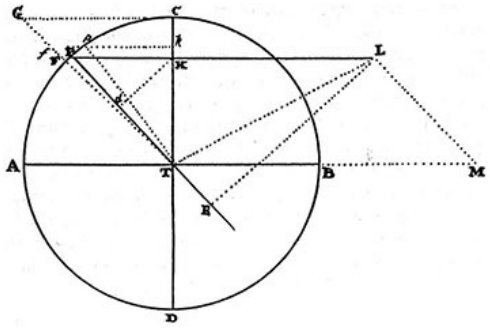


Proposición XXVI. Problema VII

Hallar el incremento horario del área que, con un radio trazado a la Tierra, describe la Luna en una órbita circular.

Hemos dicho que el área que describe la Luna con un radio trazado a la Tierra es proporcional al tiempo, salvo en la medida en que el movimiento lunar es perturbado por la acción del Sol. Nos proponemos investigar aquí la desigualdad del momento, o del incremento horario. Para que el cálculo resulte más sencillo, supongamos que la órbita de la Luna es circular, despreciemos todas las desigualdades, con la sola excepción de la que tratamos aquí. Pero por la enorme distancia del Sol, supongamos también que las líneas SP, ST son paralelas entre sí. De este modo la fuerza LM se reduce siempre a su valor medio TP, al igual que la fuerza TM se reduce a su valor medio 3PK. Estas fuerzas (por el Corolario II de las Leyes) componen la fuerza TL; y esta fuerza, si se hace descender la perpendicular LE sobre el radio TP, se descompone en las fuerzas TE, EL, de las cuales TE, actuando siempre según el radio TP ni acelera ni retrasa la descripción del área TPC realizada por el susodicho radio TP; mientras que EL actuando según la perpendicular, la acelera o retarda, en la medida en que acelera o retarda a la Luna. Esta aceleración de la Luna, a su paso de la cuadratura C a la conjunción A, realizada en cada momento de tiempo, es como la misma fuerza aceleratriz, es decir, como $\frac{3PK \times TK}{TP}$. Represéntese el tiempo por el movimiento medio lunar o (lo que casi es lo mismo) por el ángulo CTP, o también por el arco CP. Sobre CT elévese la normal CG igual a la propia CT. Y dividiendo el arco cuadrantal AC en innumerables partículas iguales P*p*, etc., por las que se representasen otras tantas partículas iguales de tiempo, y trazada *pk* perpendicular a CT, únanse TG con las prolongaciones de KP, *kp* cortándose en F y *f*; y FK será igual a TK, y K*k* será a PK como P*p* a T*p*, es decir, en una razón dada, y por consiguiente FK x K*k* o el área FK*kf* sera como $\frac{3PK \times TK}{TP}$, esto es, como EL; y, componiendo, el área total de GCKF como la suma de todas las fuerzas EL impresas en la Luna en el tiempo total CP, y por tanto también como la velocidad generada con esa suma, o sea, como la aceleración de la descripción del área CTP, o incremento del momento. La fuerza con la que la Luna podría girar en torno a la Tierra en reposo a la distancia TP, en su tiempo periódico CADB de 27d. 7h. 43′, haría que un cuerpo, cayendo en el tiempo CT, describiese la longitud ½CT, y adquiriría mientras tanto una velocidad igual a aquella con la que la Luna se mueve en su órbita. Esto es evidente por el Corolario 9 de la Proposición IV del Libro I. Pero, puesto que la perpendicular K*d* abatida sobre TP es la tercera parte de la propia EL, y la mitad de TP o de ML en los octantes, la fuerza EL en los octantes, cuando mayor resulta, superará a la fuerza ML en razón de 3 a 2, y por lo mismo, será a la fuerza con la cual la Luna puede girar en torno a la Tierra en reposo en su tiempo periódico, como 100 a $^2/_3$ x 17 872½, o sea, 11 915, y en el tiempo CT deberá generar una velocidad que fuese la $^{100}/_{11915}$ parte de la velocidad lunar, mientras que en el tiempo CPA generaría una velocidad mayor en

la proporción de CA a CT o TP. Represéntese la fuerza máxima en los octantes EL por medio del área FK x K*k*, igual al rectángulo ½TP x P*p*. Y la velocidad que la fuerza máxima puede generar en un tiempo cualquiera CP, será a la velocidad que toda fuerza menor EL genera en el mismo tiempo, como el rectángulo ½TP x CP al área KCGF: en cambio, en el tiempo entero CPA, las velocidades generadas serán entre sí como el rectángulo ½TP x CA y el triángulo TCG, o como el arco cuadrantal CA y el radio TP. Por lo tanto (por la Proposición IX del Libro V de los Elementos) la última velocidad, generada en todo el tiempo será $^{100}/_{11915}$ partes de la velocidad de la Luna. A esta velocidad de la Luna, que es análoga al momento medio del área, añádase o réstese la mitad de la otra velocidad; y si se representa el momento medio por el número 11 915, la suma 11 915 + 50, o sea, 11 965 representará el momento máximo del área en la sicigia A, mientras que la diferencia 11 915 - 50, o sea, 11 865 el momento mínimo de la misma en las cuadraturas. Por consiguiente, las áreas descritas en tiempos iguales en las sicigias y en las cuadraturas son entre sí como 11 965 a 11 865. Añádase al momento mínimo 11 865 un momento que sea a la diferencia 100 de los momentos como el trapecio FKCG al triángulo TCG (o lo que es lo mismo, como el seno cuadrado de PK al cuadrado del radio TP, esto es, como P*d* a TP) y la suma representará el momento del área cuando la Luna se halla en un lugar P intermedio.



Todo esto es así, en la hipótesis de que el Sol y la Tierra están en reposo, y la Luna gira en el tiempo sinódico de 27d. 7h. 43′. Pero como el período sinódico lunar sea en realidad de 29d. 12h. 44′, deben aumentarse los incrementos de los momentos en razón del tiempo, esto es, en razón de 1 080 853 a 1 000 000. De este modo el incremento total, que era $^{100}/_{11915}$ partes del momento medio será ahora de $^{100}/_{11023}$ partes del mismo. Por ello el momento del área en la cuadratura de la Luna será a su momento en la sicigia como 11 023 - 50 a 11 023 + 50, o sea, como 10 973 a 11 073; y a su momento, cuando la Luna se halla en otro lugar intermedio cualquiera P, como 10 973 a 10 973 + P*d*, siendo TP igual a 100.

Por lo tanto, el área que describe la Luna con un radio trazado a la Tierra en cada una de las partículas iguales de tiempo, es muy aproximadamente como la suma del

número 219,46 y el seno verso del doble de la distancia de la Luna a la cuadratura próxima, en un círculo cuyo radio es la unidad. Todo esto es así cuando la variación en los octantes es de valor medio. Pero si la variación es allí mayor o menor, dicho seno verso debe aumentarse o disminuirse en la misma razón.

Proposición XXVII. Problema VIII

Hallar la distancia de la Luna a la Tierra a partir del movimiento horario de la Luna.

El área que la Luna, con un radio trazado a la Tierra, describe en cada uno de los momentos de tiempo, es como el movimiento horario de la Luna y el cuadrado de la distancia de la Luna a la Tierra conjuntamente; y por tanto, la distancia de la Luna a la Tierra está en razón compuesta de la raíz cuadrada del área directamente y de la raíz cuadrada del movimiento horario inversamente. Q. E. I.

Corolario 1. De aquí que venga dado el diámetro aparente de la Luna, toda vez que es inversamente como su distancia a la Tierra. Intenten los astrónomos comprobar la medida en que esta regla concuerda con los fenómenos.

Corolario 2. De aquí también que la órbita lunar pueda definirse a partir de los fenómenos más exactamente que antes.

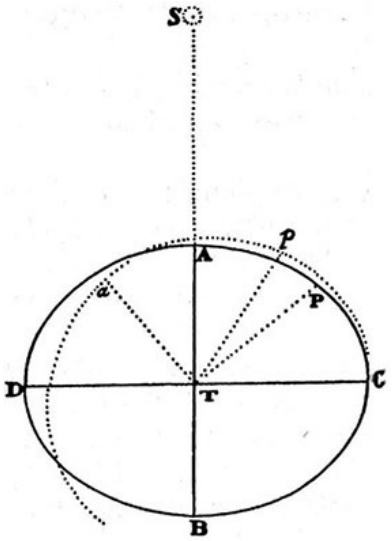
Proposición XXVIII. Problema IX

Hallar los diámetros de la órbita en que debe moverse la Luna sin excentricidad.

La curvatura de la trayectoria que describe un móvil, si es atraído según la perpendicular a dicha trayectoria, es directamente como la atracción e inversamente como el cuadrado de la velocidad. Supongo que las curvaturas de las líneas son entre sí como las razones últimas de los senos o de las tangentes de los ángulos de contacto pertenecientes a radios iguales, cuando tales radios disminuyen infinitamente. Pero la atracción de la Luna hacia la Tierra en las sicigias es el exceso de su gravedad hacia la Tierra sobre la fuerza del Sol 2PK (véase figura, Proposición XXV) por la que la gravedad aceleratriz de la Luna hacia el Sol supera a la gravedad aceleratriz de la Tierra hacia el Sol o es superada por ella. Pero en las cuadraturas dicha atracción es la suma de la gravedad de la Luna hacia la Tierra y de la fuerza solar KT, con la que la Luna es atraída hacia la Tierra. Si llamamos N a $\frac{AT + CT}{2}$ estas atracciones son como

$$\frac{178725}{AT^2} - \frac{2000}{CT \times N} \text{ y } \frac{178725}{CT^2} + \frac{1000}{AT \times N}$$

muy aproximadamente; o también como 178 725N x CT^2 - 2000AT^2 x CT y 178 725N x AT^2 + 1000CT^2 x AT. Pues, si la gravedad aceleratriz de la Luna hacia la Tierra se representa por el número 178 725, la fuerza media ML, que en las cuadraturas es PT o TK y atrae a la Luna hacia la Tierra, será 1000, mientras que la fuerza media TM en las sicigias será 3000; de la cual, si se resta la fuerza media ML, quedará 2000, fuerza con la que en las sicigias es atraída la Luna por la Tierra, y a la cual ya he denominado antes 2PK. Pero la velocidad de la Luna en las sicigias A y B es a su velocidad en las cuadraturas C y D, como CT a AT y el momento del área que describe la Luna con un radio trazado a la Tierra en las sicigias al momento de la misma área en las cuadraturas conjuntamente, esto es, como 11 073CT a 10 973AT.



Tómese inversamente el cuadrado de esta razón y directamente la primera razón y resultará la curvatura de la órbita lunar en las sicigias a la curvatura de la misma en las cuadraturas como 120 406 729 x 178 725AT^2 x CT^2 x N - 120 406 729 x 2000AT^4 x CT a 122 611 329 x 178 725AT^2 x CT^2 x N + 122 611 329 x 1000CT^4 x AT, esto es, como 2 151 969AT x CT x N - 24 081AT^3 a 2 191 371AT x CT x N + 12 261CT^3.

Puesto que la figura de la órbita lunar se desconoce, supongamos en su lugar la elipse DBCA, en cuyo centro T se halle la Tierra, y cuyo eje mayor DC yace entre las cuadraturas, mientras el menor AB yace entre las sicigias. Pero puesto que el plano de esta elipse se mueve con un movimiento angular en torno a la Tierra, y la trayectoria cuya curvatura consideramos debe ser descrita en un plano carente por

completo de movimiento angular, debemos considerar la figura que describe la Luna al girar en elipse sobre este plano, esto es, la figura C*pa*, cada uno de cuyos puntos *p* se halla tomando un punto cualquiera P en la elipse que represente el lugar de la Luna, y trazando T*p* igual a TP con la condición de que el ángulo PT*p* sea igual al movimiento aparente del Sol efectuado desde el momento de la cuadratura C; o (lo que viene a ser lo mismo) de modo que el ángulo CT*p* sea al ángulo CTP como el tiempo de la revolución sinódica al tiempo de la revolución periódica, o sea, como 29d. 12h. 44′ a 27d. 7h. 43′. Tómese, pues, el ángulo CT*a* en la misma razón al ángulo CTA, y sea la longitud T*a* igual a la longitud TA, y *a* será el ápside inferior y C el superior de esta órbita C*pa*. Estableciendo proporciones encuentro que la diferencia entre la curvatura de la órbita C*pa* en el vértice *a* y la curvatura del círculo con centro en T y radio TA es a la diferencia entre la curvatura de la elipse en el vértice A y la curvatura de ese mismo círculo como la razón cuadrada del ángulo CTP al ángulo CT*p*; y también que la curvatura de la elipse en A es a la curvatura del antedicho círculo como la razón cuadrada de TA a TC; y también que la curvatura del círculo antedicho es a la curvatura del círculo trazado con centro en T y radio TC como TC a TA; mientras que la curvatura de este último es a la curvatura de la elipse en C como la razón cuadrada de TA a TC; y también que la diferencia entre la curvatura de la elipse en el vértice C y la curvatura del segundo círculo mencionado es a la diferencia entre la curvatura de la figura T*pa* en el vértice C y la curvatura del mismo recién mencionado círculo como la razón cuadrada del ángulo CT*p* al ángulo CTP. Razones que, por lo demás, son fáciles de obtener a partir de los senos de los ángulos de contacto y de las diferencias de ángulos. Y comparándolas entre sí, resulta que la curvatura de la figura C*pa* en *a* es a su misma curvatura en C como AT^3 + 16 824/100 000 CT^2 x AT a CT^3 + 16 824/100 000 AT^2 x CT. Donde el número 16 824/100 000 representa la diferencia de los cuadrados de los ángulos CTP y CT*p* dividida por el cuadrado del ángulo menor CTP, o (lo que es lo mismo), la diferencia de los cuadrados de los tiempos 27d. 7h. 43′ y 29d. 12h. 44′ dividida por el cuadrado del tiempo 27d. 7h. 43′.

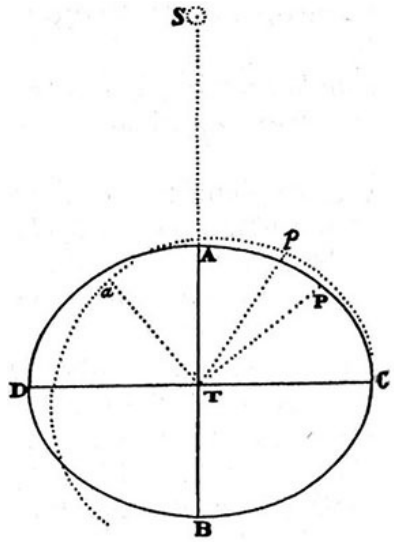
Por consiguiente, dado que *a* representa la sicigia de la Luna y C su cuadratura, la proporción hallada debe ser la misma que la proporción de la curvatura de la órbita de la Luna en las sicigias a su curvatura en las cuadraturas, proporción hallada más arriba. Por tanto, para hallar la proporción de CT a AT, multiplico entre sí medios y extremos. Y los términos resultantes divididos por AT x CT producen 2062,79CT^4 - 2 151 969N x CT^3 + 368 676N x AT x CT^2 + 36 342AT^2 x CT^2 - 362 047N x AT^2 x CT + 2 191 371N x AT^3 + 4051,4AT^4 = 0. Aquí, por la semisuma N de los términos AT y CT escribo 1, y por su semidiferencia escribo *x*, resulta CT = 1 + *x*, y AT = 1 - *x*; trasladando éstos a la ecuación y resolviendo la ecuación resultante se obtiene *x* igual a 0,00719, y por tanto el semidiámetro CT resulta 1,00719, y el semidiámetro AT 0,99281, números que son como $70\frac{1}{14}$ y $69\frac{1}{14}$ muy aproximadamente. Luego la distancia de la Luna a la Tierra en las sicigias es a su distancia en las cuadraturas

(dejando de lado la consideración de la excentricidad) como $69\frac{1}{14}$ a $70\frac{1}{14}$, o en números redondos como 69 a 70.

Proposición xxix. Problema x

Hallar la variación de la Luna.

Procede esta desigualdad en parte de la forma elíptica de la órbita lunar, y en parte de la desigualdad de los momentos del área que describe la Luna con un radio trazado a la Tierra. Si la Luna P se moviera en la elipse DBCA en torno a la Tierra en reposo en el centro de la elipse, y con el radio TP trazado a la Tierra describiera el área CTP proporcional al tiempo, y el semidiámetro mayor CT fuese al semidiámetro menor TA como 70 a 69, la tangente del ángulo CTP sería a la tangente del ángulo del movimiento medio calculado desde la cuadratura C como el semidiámetro TA de la elipse a su semidiámetro TC, o como 69 a 70. Pues la descripción del área CTP, al pasar la Luna de la cuadratura a la sicigia, debe acelerarse en una razón tal que el momento en la sicigia de la Luna sea a su momento en la cuadratura como 11 073 a 10 973, y de modo que el exceso del momento en un lugar intermedio cualquiera P sobre el momento cu la cuadratura sea como el cuadrado del seno del ángulo CTP. Cosa que ocurre con bastante exactitud si la tangente del ángulo CTP disminuye como la raíz cuadrada de la razón del número 10 973 al número 11 073, es decir, en la razón del número 68,6877 al número 69. De este modo, la tangente del ángulo CTP será ahora a la tangente del movimiento medio como 68,6877 a 70 y el ángulo CTP en los octantes, donde el movimiento medio es de 45.º, resultará de 44.º 27′ 28″, que restado del ángulo del movimiento medio de 45.º deja un resto de variación máxima de 32′ 32″. Esto sería así si la Luna, al pasar de la cuadratura a la sicigia, recorriera un ángulo de sólo 90.º. Pero por el movimiento de la Tierra, por el cual el Sol aparentemente se adelanta, la Luna, hasta que alcanza al Sol, describe un ángulo CT*a* mayor que un ángulo recto en la proporción del tiempo de la revolución sinódica lunar al tiempo de la revolución periódica, es decir, en la razón de 29d. 12h. 44′ a 27d. 7h. 43′. Y de esta manera todos los ángulos en torno al centro de T se dilatan en esa proporción, con lo que la variación máxima, que en otro caso sería de 32′ 32″, aumentada ahora en esa razón, se torna de 35′ 10″.



Este es su valor a la distancia media del Sol a la Tierra, ignorando las diferencias que puedan surgir de la curvatura del Orbe Máximo o de la mayor acción del Sol sobre la Luna menguante y nueva que creciente y llena. A otras distancias del Sol a la Tierra, la variación máxima está en razón compuesta de la razón cuadrada del tiempo de revolución sinódica lunar directamente (dado el tiempo del año) e inversamente de la razón cúbica de la distancia del Sol a la Tierra. Por consiguiente, en el apogeo del Sol, la variación máxima es de 33′ 14″, mientras que en su perigeo es de 37′ 11″, siempre que la excentricidad del Sol sea al semidiámetro transversal del Orbe Máximo como $16\frac{15}{16}$ a 1000.

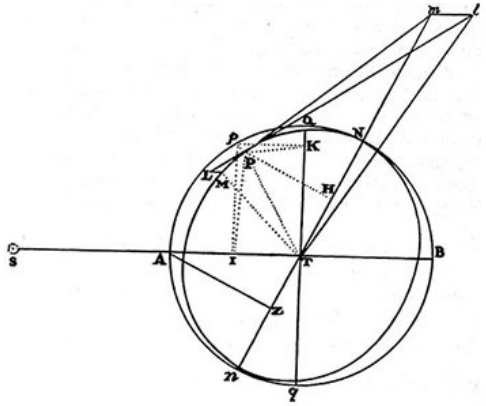
Hemos investigado hasta ahora la variación en una órbita no excéntrica, en la cual, por cierto, la Luna está siempre en sus octantes a la distancia media de la Tierra. Si la Luna, debido a su excentricidad, dista más o menos de la Tierra que si se hallase situada en dicha órbita, la variación puede ser entonces algo mayor o menor que lo establecido con la regla aquí dada; pero dejo para los astrónomos la tarea de determinar mediante los fenómenos el exceso o el defecto.

Proposición xxx. Problema xi

Hallar el movimiento horario de los nodos de la Luna en una órbita circular.

Representen S al Sol, T a la Tierra, P a la Luna, NP*n* la órbita de la Luna, N*pn* la proyección de la órbita sobre el plano de la eclíptica; N*n*, los nodos, *n*TN*m* la línea de los nodos prolongada infinitamente; PI, PK, perpendiculares trazadas sobre el plano

de la eclíptica; A, B, las sicigias de la Luna en el plano de la eclíptica; AZ, una perpendicular sobre la línea de los nodos N*n*; Q*q*, las cuadraturas de la Luna en el plano de la eclíptica, y *p*K una perpendicular sobre la línea Q*q* situada entre las curvaturas. La fuerza del Sol para perturbar el movimiento de la Luna (por la Prop. XXV) es doble, una proporcional a la línea LM del esquema de dicha Proposición, y la otra proporcional a la línea MT. Y la Luna es atraída hacia la Tierra con la primera de las fuerzas y hacia el Sol con la segunda según una línea recta paralela a la recta ST trazada de la Tierra al Sol. La primera fuerza LM actúa según el plano de la órbita lunar, y por ello en nada cambia la situación del plano. Por tanto, ésta debe ignorarse. La segunda fuerza MT, con la que se perturba el plano de la órbita lunar, es la misma que 3PK o que 3IT. Y esta fuerza (por la Propos. XXV) es a la fuerza con la cual la Luna podría girar uniformemente en su tiempo periódico y en un círculo en torno a la Tierra en reposo como 3IT al radio del círculo multiplicado por 178,725, o como IT al radio multiplicado por 59,575. Por lo demás, en este cálculo y en todo lo que sigue, considero a todas las líneas trazadas desde la Luna al Sol como paralelas a la línea que une la Tierra y el Sol, debido a que la inclinación disminuye los efectos en unos casos casi lo mismo que los aumenta en otros; y tratamos de hallar los movimientos medios de los nodos, ignorando estas menudencias, que harían al cálculo demasiado pesado.



Represente ahora PM al arco descrito por la Luna en un tiempo dado mínimo y ML un pequeño segmento cuya mitad podría recorrer la Luna en ese tiempo bajo la acción de la fuerza 3IT. Únanse PL, MP, y prolónguense hasta *m* y Z, donde cortan ál plano de la eclíptica; y sobre T*m* hágase descender la perpendicular PH. Y puesto que la recta ML es paralela al plano de la eclíptica, y por tanto no puede encontrarse con la recta *ml* que se halla en dicho plano, y sin embargo ambas rectas están en el plano común LMP*ml*, estas rectas serán paralelas, y por lo mismo los triángulos LMP y *lm*P serán semejantes. Pero como MP*m* está en el plano de la órbita por la cual se movía

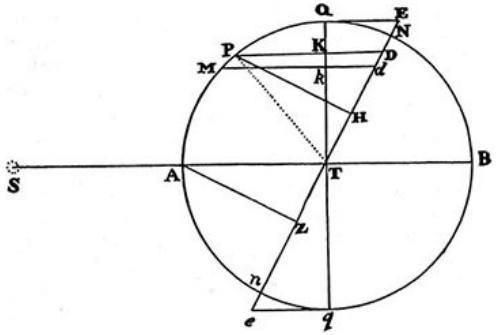
la Luna cuando estaba en P, el punto *m* caerá sobre la línea N*n* trazada por los nodos N*n*, de dicha órbita. Y puesto que la fuerza por la que se genera la mitad del pequeño segmento LM, si se imprimiese toda y a la vez en el punto P, generaría toda la dicha línea, y haría que la Luna se moviese en un arco cuya cuerda fuese LP, y, por tanto, trasladaría a la Luna del plano MP*m*T al plano LP*l*T, el movimiento angular de los nodos generado por dicha fuerza será igual al ángulo *m*T*l*. Pero *mi* es a *m*P como ML a MP, y, por tanto, como MP está dada por estar dado el tiempo, *ml* es como el rectángulo ML x *m*P, es decir, como el rectángulo IT x *m*P. Y el ángulo *m*T*l*, siempre que el ángulo T*ml* sea recto, es como $\frac{ml}{\mathrm{T}m}$ y, por tanto, como $\frac{\mathrm{IT} \times \mathrm{P}m}{\mathrm{T}m}$, es decir (por ser proporcionales *tm* y *m*P, TP y PH) como $\frac{\mathrm{IT} \times \mathrm{PH}}{\mathrm{TP}}$, y por tanto, por estar dado TP, como IT x HP. Pero si el ángulo T*ml*, o STN, fuera oblicuo el ángulo *m*T*l* será aún menor, en razón del seno del ángulo STN al radio, o también en razón de AZ a AT. Por consiguiente, la velocidad de los nodos es como IT x PH x AZ o también como el producto de los senos de los tres ángulos TPI, PTN, y STN.

Si estos ángulos fuesen rectos, estando los nodos en las cuadraturas y la Luna en las sicigias, la pequeña línea *mi* se irá al infinito, y el ángulo *m*T*l* resultará igual al ángulo *m*P*l*. En este caso, el ángulo *m*P*l* es al ángulo PTM, el cual describe la Luna en lomo a la Tierra en el mismo tiempo con su movimiento aparente, como 1 a 59,575. Pues el ángulo *m*P*l* es igual al ángulo LPM, es decir, al ángulo de deflexión de la Luna de la trayectoria recta, ángulo que la antedicha fuerza solar de 3IT podría generar en dicho tiempo dado, si entonces cesase la gravedad de la Luna; y el ángulo PTM es igual al ángulo de deflexión de la Luna de la trayectoria recta, ángulo que podría generar en el mismo tiempo la fuerza, con la que la Luna es retenida en su órbita, si cesare entonces la fuerza solar 3IT. Y estas fuerzas son entre ellas, como hemos dicho antes, como 1 a 59,575. Luego, como el movimiento medio horario de la Luna respecto a las fijas sea de 32′ 56″ 27‴ 12½⁗, el movimiento horario de los nodos en este caso será 33″ 10‴ 33⁗ 12⁗. Pero en otros casos este movimiento horario será a 33″ 10‴ 33⁗ 12⁗ como el producto de los senos de los tres ángulos TPI, PTN y STN (o de las distancias de la Luna a la cuadratura, al nodo y al Sol) al cubo del radio. Y cuantas veces el signo de un ángulo cualquiera pase de positivo a negativo, y de negativo a positivo, deberá pasar el movimiento regresivo a progresivo y el progresivo a regresivo. De donde se seguirá que los nodos progresarán siempre que la Luna se halle entre una de las cuadraturas y el nodo próximo a la cuadratura. En los demás casos regresan, y por el exceso del regreso sobre el progreso en cada mes van resultando retrógrados.

COROLARIO 1. De aquí que, si desde los extremos P y M de un arco muy pequeño dado y sobre la línea Q*q* que une las cuadraturas se hacen descender sendas perpendiculares PK, M*k* y se las prolonga hasta que corten la línea de los nodos N*n*

en D y *d*, el movimiento horario de los nodos será como el área MPD*d* y el cuadrado de la línea AZ conjuntamente. Pues sean PK, PH y AZ los tres senos antedichos. A saber, PK el seno de la distancia de la Luna a la cuadratura, PH el seno de la distancia de la Luna al nodo, y AZ el seno de la distancia del nodo al Sol; y la velocidad del nodo será como el producto PK x PH x AZ. Pues PT es a PK como PM a K*k*, y por tanto, por estar dadas PT y PM, K*k* es proporcional a PK. Y AT es a PD como AZ a PH, y por tanto PH, proporcional al rectángulo PD x AZ, y multiplicando las razones, PK x PH es como el producto K*k* x PD x AZ, y PK x PH x AZ como K*k* x PD x AZ^2, esto es, como el área PD*d*M y AZ^2 conjuntamente. Q. E. D.

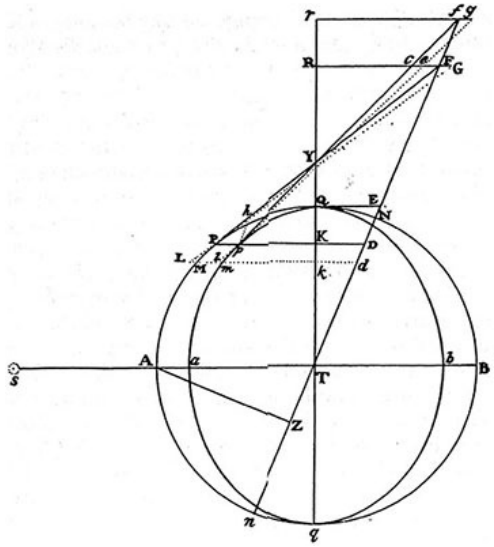
Corolario 2. En una posición dada cualquiera de los nodos, el movimiento horario medio es la mitad del movimiento horario en las sicigias de la Luna, y por tanto es a 16″ 35‴ 16⁗ 36⁗″ como el cuadrado del seno de la distancia de los nodos a las sicigias al cuadrado del radio, o también como AZ^2 a AT^2. Pues si la Luna, con movimiento uniforme, circulase por el semicírculo QA*q*, la suma de todas las áreas PD*d*M, en el tiempo en que la Luna pasa de Q a M, será el área QM*d*E, que termina en QE tangente del círculo; y en el tiempo en que la Luna alcanza el punto *n*, dicha suma será el área completa EQA*n* que describe la línea PD, después al pasar la Luna de *n* a *q*, la línea PD caerá fuera del círculo y describirá el área *nqe* limitada por *qe* tangente al círculo; la cual, puesto que antes los nodos regresaban, mientras ahora progresan, debe restarse del área primera, y puesto que es igual al área QEN dejará el semicírculo NQA*n*. Por tanto, la suma de todas las áreas PD*d*M, en el tiempo en que la Luna describe un semicírculo, es el área del semicírculo; y la suma de todas en el tiempo en que la Luna describe el círculo es el área de todo el círculo. Pero el área PD*d*M, cuando la Luna anda por las sicigias, es el rectángulo comprendido bajo el arco PM y el radio PT; y la suma de todas las áreas iguales a ésta, en el tiempo en que la Luna describe el círculo, es el rectángulo comprendido bajo toda la circunferencia y el radio del círculo; y este rectángulo, al ser igual a dos círculos, es el doble mayor que el rectángulo anterior. Por tanto, los nodos, con la misma velocidad continua que tienen en las sicigias lunares, describirían un espacio doblemente mayor del que de hecho describen; y por tanto, el movimiento medio, con el cual podrían describir, si se continuase uniformemente, un espacio desigual de por sí del realmente descrito, es la mitad del movimiento que tienen en las sicigias de la Luna. De donde, puesto que el movimiento horario máximo, si los nodos se hallan en las cuadraturas, es de 33″ 10‴ 33⁗″ 12⁗″, el movimiento horario medio, en este caso, será 16″, 35‴, 16⁗″, 36⁗″. Y, puesto que el movimiento horario de los nodos es siempre como AZ^2 y el área PD*d*M conjuntamente, y por lo mismo, el movimiento horario de los nodos en las sicigias de la Luna es como AZ^2 y el área PD*d*M conjuntamente, es decir (por estar dada PD*d*M descrita en las sicigias) como AZ^2, el movimiento medio será también como AZ^2, y por consiguiente este movimiento, cuando los nodos se encuentran fuera de las cuadraturas, será a 16″, 35‴, 16⁗″, 36⁗″ como AZ^2 a AT^2. Q. E. D.



Proposición xxxi. Problema xii

Hallar el movimiento horario de los nodos de la Luna en una órbita elíptica.

Represente Q*pmaq* una elipse trazada con el eje mayor Q*q* y el menor *ab*, QA*q*B un círculo circunscrito, T la Tierra en el centro común de ambos, S el Sol, *p* la Luna moviéndose en la elipse, y *pm* el arco que describe en una mínima partícula de tiempo dada, N y *n* los nodos unidos por la línea N*n*, *p*K y *mk* perpendiculares trazadas sobre el eje Q*q* y desde él prolongadas a ambos lados hasta que corten el círculo en P y M, y a la línea de los nodos en D y *d*. Y si la Luna, con radio trazado a la Tierra, describe un área proporcional al tiempo, el movimiento horario de un nodo en la elipse será como el área *p*D*dm* y AZ^2 conjuntamente.



Pues, si PF es tangente al círculo en P, y prolongada corta a TN en F, y *pf* es tangente a la elipse en *p* y prolongada corta a la misma TN en *f*, y si ambas tangentes concurren sobre el eje TQ en Y; y si ML representa el espado que la Luna, girando en el círculo y en el tiempo en que describe el arco PM, puede describir con movimiento transversal bajo la acción de la antedicha fuerza 3IT o 3PK, y *mi* representa el espacio que la Luna, girando en el mismo tiempo en la elipse, bajo la acción de la misma fuerza 3IT o 3PK, pudiera también describir y se prolongan LP y *lp* hasta que encuentren al plano de la eclíptica en G y *g*, y se unen FG y *fg*, de las cuales la prolongación de FG corta a *pf*, *pg* y TQ en *c*, *e* y R, respectivamente, y la prolongación de *fg* corta a TQ en *r*. Entonces, puesto que la fuerza 3IT o 3PK en el círculo es a la fuerza 3IT o 3PK en la elipse como PK a *p*K, o como AT a *a*T, el espacio ML generado por la primera fuerza será al espacio *ml* generado por la segunda como PK a *p*K, es decir, por la semejanza de figuras PYK*p* y FYR*c*, como FR a *c*R. Pero ML es a FG (por la semejanza de los triángulos PLM, PGF) como PL a PG, es decir (por ser paralelas L*k*, PK, GR) como *pl* a *pe*, es decir (por la semejanza de los triángulos *plm*, *cpe*) como *lm* a *ce*; e inversamente, como LM es a *lm*, o FR a *c*R, así es FG a *ce*. Y por lo mismo, si *fg* fuese a *ce* como *f*Y a *c*Y, es decir, como *fr* a *c*R (esto es, como *fr* a FR y FR a *c*R conjuntamente, es decir, como *f*T a FT y FG a *ce* conjuntamente) como al suprimir la razón de FG a ce en uno y otro quedan las razones de *fg* a FG y *f*T a FT, será *fg* a FG como *f*T a FT; y por consiguiente, los ángulos que subtenderían FG y *fg* a la Tierra T serían iguales entre ellos. Pero esos ángulos (por lo expuesto en la proporción anterior) son los movimientos de los nodos en el tiempo en que la Luna recorre en el círculo el arco PM y en la elipse el arco *pm*: y por tanto los movimientos de los nodos en el círculo y en la elipse serán iguales entre ellos. Todo esto será así siempre que *fg* sea a *ce* como *f*Y a *c*Y, esto es, si *fg* fuese igual a $\frac{ce \times fY}{cY}$. Pero por la semejanza de los triángulos *fgp*, *cep*, *fg* es a *ce* como *fp* es *cp*; y por tanto, *fg* es igual $\frac{ce \times fp}{cp}$; y por lo tanto, el ángulo que en realidad subtiende *fg* es al ángulo anterior que subtiende FG, esto es, el movimiento de los nodos en la elipse al movimiento de los nodos en el círculo como este *fg*, o $\frac{ce \times fp}{cp}$ al anterior *fg* o $\frac{ce \times fY}{cY}$, esto es, como *fp* x *c*Y a *f*Y x *cp*, o *fp* a *f*Y y *cY* a *cp*, es decir, si *ph* paralela a TN corta a FP en *h*, como F*h* a FY y FY a FP; esto es, como F*h* a FP o D*p* a DP, y por tanto, como el área D*pmd* al área DPMD. Y por consiguiente, puesto que (por el Corolario 1 de la Proposición XXX) el área segunda y AZ^2 conjuntamente son proporcionales al movimiento horario de los nodos en el círculo, el área primera y AZ^2 conjuntamente serán proporcionales al movimiento horario de los nodos en la elipse. Q. E. D.

COROLARIO. Por lo cual, como para una posición dada de los nodos, la suma de

todas las áreas *p*D*dm*, en el tiempo en que la Luna va desde una cuadratura hasta un lugar cualquiera m, viene a ser *mp*QE*d*, que está delimitada por la tangente a la elipse QE; y la suma de todas esas áreas en la revolución completa, venga a ser el área de la elipse entera, el movimiento medio de los nodos en la elipse será al movimiento medio de los nodos en el círculo, como la elipse al círculo; esto es, como Ta a TA, o como 69 a 70. Y por tanto, puesto que (por el Corolario 2 de la Proposición XXX) el movimiento horario medio de los nodos en el círculo es a 16″, 35‴, 16′′′′, 36′′′′′ como AZ^2 a AT^2, si se toma un ángulo de 16″, 21‴, 3′′′′, 30′′′′′ al ángulo de 16″, 35‴, 16′′′′, 36′′′′′ como 69 a 70, el movimiento horario medio de los nodos en la elipse será a 16″, 21‴, 3′′′′, 30′′′′′ como AZ^2 a AT^2; esto es, como el cuadrado del seno de la distancia del nodo al Sol al cuadrado del radio.

Pero por otra parte, la Luna con un radio trazado a la Tierra describe el área con mayor velocidad en las sicigias que en las cuadraturas, y por eso el tiempo en las sicigias se acorta mientras en las cuadraturas se alarga, y junto con el tiempo el movimiento de los nodos aumenta y disminuye. Pero el momento del área en las cuadraturas de la Luna era al momento del área en las sicigias como 10 973 a 11 073, y por ello, el momento medio en los octantes es al exceso en las sicigias y al defecto en las cuadraturas como la semisuma de ambas cantidades 11 023 a su semidiferencia 50. De donde, puesto que el tiempo de la Luna en cada una de las partículas iguales de la órbita es inversamente como su velocidad, el tiempo medio en los octantes será al exceso de tiempo en las cuadraturas y al defecto de tiempo en las sicigias, debidos a estas causas, como 11 023 a 50 muy aproximadamente. Pero marchando desde las cuadraturas hacia las sicigias, encuentro que el exceso de los momentos del área en cada lugar por encima del momento mínimo en las cuadraturas es como el cuadrado del seno de la distancia de la Luna hasta las cuadraturas muy aproximadamente; y por ello, la diferencia entre el momento en un lugar cualquiera y el momento medio en los octantes es como la diferencia entre el cuadrado del seno de la distancia de la Luna hasta las cuadraturas y el cuadrado del seno de 45 grados, o la mitad del cuadrado del radio; y el incremento del tiempo en cada lugar entre los octantes y las cuadraturas y su decremento entre los octantes y las sicigias se halla en la misma razón. Pero el movimiento de los nodos, en el tiempo en que la Luna recorre cada una de las partículas iguales de su órbita, se acelera o retarda en razón del cuadrado del tiempo. Pues este movimiento, mientras («caeteris paribus») la Luna recorre PM, es como ML, y ML está en razón del cuadrado del tiempo. Por lo cual, el movimiento de los nodos en las sicigias, realizado en el tiempo en que la Luna recorre partículas dadas de la órbita, disminuye como el cuadrado de la razón de 11 073 a 11 023; y el decremento es al resto del movimiento como 100 a 10 973, mientras que respecto al movimiento entero es como 100 a 11 073 muy aproximadamente. Pero el decremento en los lugares entre los octantes y las sicigias, y el incremento entre los lugares de los octantes y las cuadraturas es aproximadamente a este decremento como el movimiento total en dichos lugares al movimiento total en las sicigias y la diferencia

entre el cuadrado del seno de la distancia de la Luna hasta la cuadratura y la mitad del cuadrado del radio, a la mitad del cuadrado del radio conjuntamente. De donde, si los nodos se hallan en las cuadraturas y se toman dos lugares equidistantes a un lado y otro del octante, y otros dos a esa misma distancia uno de la sicigia y otro de la cuadratura, y de los decrementos de movimientos en los dos lugares entre la sicigia y el octante se restan los incrementos de movimientos en los otros dos lugares situados entre el octante y la cuadratura, el decremento restante será igual al decremento en la sicigia, como verá fácilmente quien haga el cálculo. Por consiguiente, el decremento medio, que debe restarse del movimiento medio de los nodos, es la cuarta parte del decremento en la sicigia. El movimiento horario total de los nodos en las sicigias, cuando la Luna con un radio trazado a la Tierra se suponía que describía un área proporcional al tiempo, era de 32″ 42‴ 7⁗. Y el decremento del movimiento de los nodos, en el tiempo en que la Luna, más veloz ahora, describe el mismo espacio, hemos dicho que era a este movimiento como 100 a 11 073; y por consiguiente, dicho decremento es de 17″ 43⁗ 11⁗⁗, cuya cuarta parte 4‴ 25⁗ 48⁗⁗ restada al movimiento horario medio hallado más arriba deja un resto de 16″ 16‴ 37⁗ 42⁗⁗ como movimiento horario medio corregido.

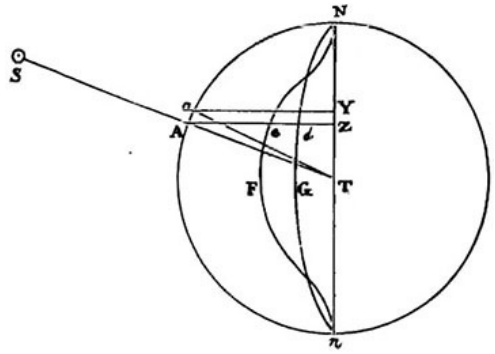
Si los nodos están fuera de las cuadraturas y se contemplan dos lugares equidistantes de las sicigias, uno a cada lado, la suma de los movimientos de los nodos, cuando la Luna se halla en dichos lugares, será a la suma de los movimientos cuando la Luna se halla en esos lugares y los nodos se hallan en las cuadraturas como AZ^2 a AT^2. Y los decrementos de los movimientos, debidos a las causas expuestas más arriba, serán entre sí como los movimientos mismos, y por tanto, los movimientos restantes serán entre sí como AZ^2 a AT^2, y los movimientos medios serán como los movimientos restantes. Por consiguiente, el movimiento horario medio corregido en cualquier situación dada de los nodos es a 16″ 16‴ 37⁗ 42⁗⁗ como AZ^2 a AT^2; es decir, como el cuadrado del seno de la distancia de los nodos a las sicigias al cuadrado del radio.

Proposición xxxii. Problema xiii

Hallar el movimiento medio de los nodos de la Luna.

El movimiento medio anual es la suma de todos los movimientos horarios medios del año. Imagínese que el nodo se halla en N y en cada una de las horas regresa a su lugar inicial, de suerte que, no obstante su movimiento propio, conserve siempre su lugar respecto a las estrellas fijas. Pero, mientras tanto, el Sol S, por el movimiento de la Tierra, se desplaza desde el nodo y completa uniformemente su curso anual aparente. Y sea A*a* arco mínimo dado que la recta TS trazada continuamente al Sol describe con su intersección con el círculo NA*n* en un tiempo mínimo dado: y el

movimiento horario medio (por lo que ya se ha mostrado) será como AZ^2, esto es (por ser proporcionales AZ, ZY), como el rectángulo AZ y ZY, es decir, como el área AZY*a*. Y la suma de todos los movimientos horarios medios desde el principio será como la suma de todas las áreas *a*YZA, es decir, como el área NAZ. Pero la máxima AZY*a* es igual al rectángulo bajo el arco A*a* y el radio del círculo, y por tanto, la suma de todos los rectángulos en el círculo entero es a la suma de otros tantos rectángulos máximos, como el área de todo el círculo al rectángulo bajo la circunferencia y el radio, es decir, como 1 a 2. Pero el movimiento correspondiente al rectángulo máximo era de 16″ 16‴ 37⁗ 42⁗⁗. Y este movimiento, en un año sidéreo completo de 365d. 6h. 9′ resulta de 39.º 38′ 7″ 50‴. Y por consiguiente, su mitad 19.º 49′ 3″ 55‴ es el movimiento medio de los nodos que corresponde al círculo entero. Y el movimiento de los nodos en el tiempo en que el Sol pasa de N hasta A es a 19.º 49′ 3″ 55‴ como el área NAZ al círculo entero.



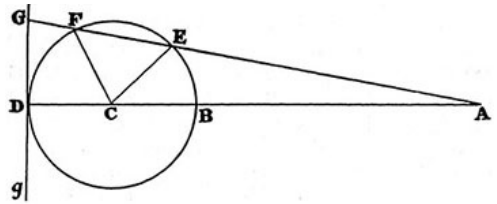
Esto es así en la hipótesis de que el nodo tras cada hora regresase al lugar inicial y de que el Sol tras el recorrido anual completo regresase al mismo nodo del que había partido. Pero por causa del movimiento del nodo ocurre que el Sol regresa antes al nodo, y ahora hay que calcular la abreviación del tiempo. Puesto que el Sol en el año entero completa 360 grados y el nodo con movimiento máximo completa en ese mismo tiempo 39.º 38′ 7″ 50‴, o 39,6355 grados, y el movimiento medio del nodo en un lugar cualquiera N es a su movimiento medio en sus cuadraturas como AZ^2 a AT^2; el movimiento del Sol al movimiento del nodo en N será como $360AT^2$ a $39{,}6355AZ^2$; es decir, como $9{,}0827646AT^2$ a AZ^2. De donde, si la circunferencia NA*n* del círculo entero se divide en partículas iguales A*a*, el tiempo en el que el Sol recorre una partícula A*a* si el círculo está en reposo, será al tiempo en que recorre esa misma partícula, si el círculo junto con los nodos gira en torno al centro T, inversamente como $9{,}0827646AT^2$ a $9{,}0827646AT^2 + AZ^2$. Pues el tiempo es inversamente como la velocidad con la que se recorre la partícula, y esta velocidad es la suma de las velocidades del Sol y del nodo. Por tanto, si el tiempo, en el que el Sol sin el movimiento del nodo recorrería el arco NA, se representa por el sector NTA, y

la partícula de tiempo en que recorrería el arco mínimo A*a*, se representa por la partícula AT*a* del sector; y (trazada la perpendicular *n*Y sobre N*n*) si sobre AZ se toma *d*Z de una longitud tal que el rectángulo *d*Z por ZY sea a la partícula AT*a* del sector como AT^2 a $9{,}0827646AT^2 + AZ^2$, es decir, de modo que *d*Z sea a ½AZ como AT^2 a $9{,}0827646AT^2 + AZ^2$, el rectángulo *d*Z por ZY representará el decremento de tiempo debido al movimiento de los nodos, durante todo el tiempo en que se recorre el arco A*a*. Y si el punto *d* es tangente a la curva N*d*G*n*, el área curvilínea N*d*Z será el decremento total en el tiempo en que se recorre el arco entero NA; y por ello el exceso del sector NAT sobre el área N*d*Z será dicho tiempo total. Y, puesto que el movimiento del nodo en menor tiempo es menor en razón del tiempo, también el área A*a*YZ deberá disminuir en esa proporción. Cosa que ocurrirá si sobre AZ se toma la longitud *e*Z tal que sea a la longitud AZ como AZ^2 a $9{,}0827646AT^2 + AZ^2$. Pues de este modo el rectángulo *e*Z por ZY será al área AZY*a* como el decremento del tiempo en que se recorre el arco A*a*, al tiempo total en que se recorrería si el nodo estuviese en reposo: y por lo tanto, dicho rectángulo corresponderá al decremento del movimiento del nodo. Y si el punto *e* fuera tangente a la curva N*e*F*n*, el área entera N*e*Z, que es la suma de todos los decrementos, corresponderá al decremento total en el tiempo en que es recorrido el arco AN; y el área restante NA*e* corresponderá al movimiento restante, que es el verdadero movimiento del nodo, en el tiempo en que el arco entero NA es recorrido por los movimientos conjuntos del Sol y del nodo. Ahora bien, el área del semicírculo es al área de la figura N*e*F*n*, hallada por el método de las series infinitas, como 793 a 60 muy aproximadamente. Pero el movimiento que correspondía al círculo entero era de 19.º 49′ 3″ 55‴ y por lo mismo, el movimiento correspondiente al doble de la figura N*e*F*n* es 1.º 29′ 58″ 2‴. El cual restado del movimiento anterior deja un resto de 18.º 19′ 5″ 53‴ movimiento total del nodo respecto a las fijas entre sus propias conjunciones con el Sol; y restado este movimiento de los 360 grados del movimiento anual del Sol, deja un resto de 341.º 40′ 54″ 7‴ movimiento del Sol entre esas mismas conjunciones. Pero este movimiento es al movimiento anual de 360 grados como el movimiento del nodo ya hallado de 18.º 19′ 5″ 53‴ a su movimiento anual, que será por tanto de 19.º 18′ 1″ 23‴. Este es el movimiento medio de los nodos en el año sidéreo. El mismo, por las tablas astronómicas es es de 19.º 21′ 21″ 50‴. La diferencia es menor que $\frac{1}{300}$ del movimiento total, y parece deberse a la excentricidad de la órbita lunar y a la inclinación hacia el plano de la eclíptica. Por causa de la excentricidad de la órbita el movimiento de los nodos se acelera demasiado, mientras que, por contra, debido a su inclinación se retarda un poco, reduciéndose así a su justa velocidad.

Proposición xxxiii. Problema xiv

Hallar el movimiento verdadero de los nodos de la Luna.

En un tiempo que sea como el área NTA - N*d*Z (de la figura precedente) este movimiento es como el área NA*e*, y por tanto está dado. Pero por la excesiva dificultad del cálculo es conveniente presentar la construcción del problema del modo siguiente. Con centro en C y una distancia cualquiera CD trácese el círculo BEFD. Prolónguese DC hasta A de modo que AB sea a AC como el movimiento medio a la mitad del movimiento medio verdadero, cuando los nodos están en las cuadraturas, es decir, como 19.º 18′ 1″ 23‴ a 19″ 49′ 3″ 55‴ y por tanto BC será a AC como la diferencia de movimiento 0.º 31′ 2″ 32‴ es al movimiento posterior 19″ 49′ 3″ 55‴, es decir, como 1 a $38\frac{3}{10}$; trácese después por el punto D la recta infinita G*g*, tangente al círculo en D; y si se toma el ángulo BCE o BCF igual al doble de la distancia entre el lugar del Sol y el nodo, hallado por medió del movimiento medio, y se traza AE o AF secante de la perpendicular DG en G, y se toma un ángulo que sea al movimiento total del nodo entré éste y las sicigias (es decir, a 9.º 11′ 3″) como la tangente DG a la circunferencia entera del círculo BED, y este ángulo (en cuyo lugar puede tomarse el ángulo DAG) se añade al movimiento medio de los nodos cuando éstos pasan de las cuadraturas a las sicigias o se sustrae del dicho movimiento medio cuando pasan de las sicigias a las cuadraturas, se tendrá su movimiento medio verdadero. Pues el movimiento verdadero así obtenido concuerda muy aproximadamente con el movimiento verdadero que resulta representando el tiempo por el área NTA - N*d*Z, y el movimiento del nodo por el área NA*e*; como comprobará quien se interese y haga los cálculos. Esta es la ecuación semestral del movimiento de los nodos. También hay una ecuación mensual, pero no es necesaria en absoluto para hallar la latitud de la Luna. Pues, dado que la variación de la inclinación de la órbita lunar hacia el plano de la eclíptica está sometida a una doble desigualdad, una semestral y otra mensual, su desigual mensual y la ecuación mensual de los nodos se compensan y corrigen entre ellas de tal modo que ambas pueden ser ignoradas a la hora de determinar la latitud de la Luna.



Corolario. De ésta y de la proposición precedente se desprende que los nodos en sus sicigias reposan, mientras en las cuadraturas retroceden con un movimiento horario de 16″ 19‴ 26‴′ y que la ecuación del movimiento de los nodos en los octantes es de 1.º 30′. Todo lo cual concuerda adecuadamente con los fenómenos celestes.

Escolio[11]

J. Machín, profesor de Astronomía en *Cresham y Henry Pemberton*, D. M. cada uno por su parte hallaron otra forma de determinar los movimientos de los nodos. Se ha hecho alguna mención de este método en otra parte. Y las cartas que he visto de ambos contenían dos proposiciones que concordaban entre ellas. Añadiré aquí la carta del *Sr. Machín*, primera que llegó a mis manos.

«SOBRE EL MOVIMIENTO DE LOS NODOS DE LA LUNA

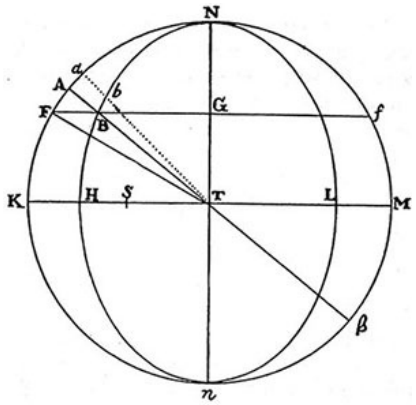
»Proposición I

»El movimiento medio del Sol desde el nodo se define por una media geométrica proporcional entre el movimiento medio del mismo Sol y aquel movimiento medio con el cual el Sol se separa velozmente del nodo en las cuadraturas.

»Sea T el lugar de la Tierra, N*n* la línea de los nodos de la Luna para un tiempo dado cualquiera, KTM una perpendicular a ella, TA una recta que gira alrededor del centro con la misma velocidad angular con que el Sol y el nodo se alejan uno de otro, de modo que el ángulo comprendido entre la recta en reposo N*n* y la recta en giro TA, siempre resulte igual a la distancia entre los lugares del Sol y del nodo. Ahora, si se divide una recta cualquiera TK en partes TS y SK que sean como el movimiento horario medio del Sol al movimiento horario medio del nodo en las cuadraturas, y se toma a la línea TH como media proporcional entre la parte TS y la línea entera TK, esta recta será, entre las demás, proporcional al movimiento medio del Sol desde el nodo.

»Pues, descríbase el círculo NK*n*M con centro en T y radio TK, y con el mismo centro y con las ejes TH y TN trácese la elipse NH*n*L, y en el tiempo en que el Sol se separa del nodo por el arco N*a*, si se traza la recta N*ba*, el área del sector NT*a* representará la suma de los movimientos del nodo y del Sol en ese mismo tiempo. Sea, pues, el arco mínimo A*a* un arco que describe la recta T*ba* girando uniformemente según la antedicha ley en una partícula dada de tiempo, y el sector mínimo TA*a* será como la suma de las velocidades con que el Sol y el nodo se desplazan en ese tiempo. Pero la velocidad del Sol es casi uniforme, por cuanto su pequeña desigualdad apenas produce variación en el movimiento medio de los nodos. La otra parte de esta suma, o sea, la velocidad del nodo en su valor medio, aumenta al apartarse de las sicigias como el cuadrado del seno de su distancia al Sol; por el Corolario de la Proposición XXXI del Libro III de los PRINCIPIA, y cuando es máxima en las cuadraturas con el Sol en K, alcanza la misma razón respecto a la velocidad del Sol que SK a TS, esto es, como (diferencia de los cuadrados de TK y TH o) el rectángulo KHM a TH al cuadrado. Pero la elipse NBH divide al sector

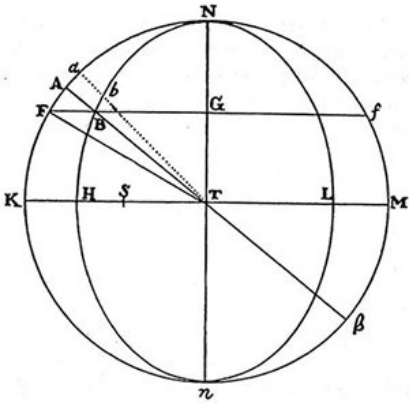
AT*a*, representación de la suma de estas dos velocidades, en dos partes AB*ba* y BT*b* proporcionales a las propias velocidades. Prolónguese, pues, BT hasta el círculo *β* y desde el punto B hágase descender sobre el eje mayor la perpendicular BG, que prolongada a ambos lados toque al círculo en los puntos F y *f* y puesto que el espacio AB*ba* es al sector TB*b* como el rectángulo AB*β* a BT cuadrado (pues dicho rectángulo es igual a la diferencia de cuadrados de TA y TB por estar la recta A*β* cortada igualmente en T y desigualmente en B). Por tanto esta razón, cuando el espacio AB*ba* es máximo en K, será la misma razón que la del rectángulo KHM a HT al cuadrado, pero la velocidad media máxima del nodo estaba a la velocidad del Sol en esta misma razón. Por consiguiente, en las cuadraturas, el sector AT*a* se divide en partes proporcionales a la velocidad. Y, puesto que el rectángulo KHM es a HT cuadrado como FB*f* a BG cuadrado, y el rectángulo AB*β* es igual al rectángulo FB*f*, el área pequeña AB*ba*, cuando es máxima, será al resto del sector TB *b* como el rectángulo AB*b* a BG cuadrado. Pero la razón de estas pequeñas áreas era siempre como el rectángulo AB*β* a BT cuadrado; y por lo tanto la pequeña área AB*ba* en el lugar A es menor que el área semejante en las cuadraturas, en razón cuadrada de BG a BT, esto es, en razón cuadrada del seno de la distancia del Sol al nodo. Y por tanto, la suma de todas las áreas pequeñas AB*ba*, o sea, el espacio ABN será como el movimiento del nodo en el tiempo en que el Sol se separa del nodo por el arco NA. Y el espacio restante, o sea, el sector elíptico NTB será como el movimiento medio del Sol en ese tiempo. Y por lo mismo, como el movimiento medio anual del nodo es aquel que ocurre en el tiempo en el que el sol completa su período, el movimiento medio del nodo desde el Sol será al movimiento medio del propio Sol como el área del círculo al área de la elipse, es decir, como la recta TK a la recta TH que es la media proporcional entre TK y TS; o lo que da lo mismo, como la media proporcional TH a la recta TS.



»Proposición II

»Dado el movimiento medio de los nodos de la Luna, hallar el movimiento verdadero.

»Sea el ángulo A la distancia del Sol al lugar medio del nodo, o el movimiento medio del Sol desde el nodo. Si se toma entonces un ángulo B cuya tangente sea a la tangente del ángulo A como TH a TK, esto es, como la raíz cuadrada de la razón del movimiento horario medio del Sol al movimiento medio horario del Sol desde el nodo cuando éste se halla en las cuadraturas, este ángulo B será la distancia del Sol al verdadero lugar del nodo. Pues únase FT y, por la demostración de la Proposición anterior, el ángulo FTN será la distancia del Sol del lugar medio del nodo, mientras que el ángulo ATN será la distancia del lugar verdadero, y las tangentes de estos dos ángulos son entre ellas como TK a TH.



»Corolario. De aquí que el ángulo FTA es la ecuación de los nodos de la Luna, y el seno de este ángulo cuando es máximo en los octantes, es al radio como KH a TK + TH. En cambio el seno de esta ecuación en otro lugar cualquiera A es al seno máximo como el seno de la suma de los ángulos FTN + ATN al radio: esto es, casi como el seno del doble de la distancia del Sol al lugar medio del nodo (o sea 2FTN) al radio.

»Escolio

»Si el movimiento horario medio de los nodos en las cuadraturas es de 16″ 16‴ 37⁗ 42⁗⁗, es decir, en el año sidéreo completo 39.º 38′ 7″ 50‴, TH será a TK como la raíz cuadrada de la razón del número 9,0827646 al número 10,0827646, esto es, como 18,6524761 a 19,6524761. Y por lo tanto, TH a HK como 18, 6524761 a 1, es

decir, como el movimiento del Sol en el año sidéreo al movimiento medio del nodo de 19.º 18′ 1″ 23⅔‴.

»Pero si el movimiento medio de los nodos de la Luna en 20 años Julianos es de 386.º 50′ 15″ como se deduce de las observaciones aportadas por la teoría de la Luna, el movimiento medio de los nodos en el año sidéreo será de 19.º 20′ 31″ 58‴. Y TH será a HK como 360.º es a 19.º 20′ 31″ 58‴, esto es, como 18,61214 a 1, de donde el movimiento horario medio de los nodos en las cuadraturas resulta de 16″ 18‴ 48′′′′. Y la ecuación máxima de los nodos en los octantes, de 1.º 29′ 57″».

Proposición xxxiv. Problema xv

Hallar la variación horaria de la inclinación de la órbita lunar hacia el plano de la eclíptica.

Representen A y a las sicigias; Q y *q* las cuadraturas; N y *n* los nodos; P el lugar de la Luna en su órbita; *p* la proyección de dicho lugar sobre el plano de la eclíptica, y *m*T*l* el movimiento momentáneo de los nodos como más arriba. Y si sobre la línea T*m* se hace descender la perpendicular PG, únase *p*G, y prolónguese hasta que corte a T*l* en *g* y únanse también P*g*; el ángulo PG*p* será la inclinación de la órbita lunar hacia el plano de la eclíptica cuando la Luna se halle en P; y el ángulo P*gp* será su inclinación al completar un momento de tiempo; y por lo tanto, el ángulo GP*g* será la variación momentánea de la inclinación. Pero este ángulo GP*g* es al ángulo GT*g* como TG a PG y PG y P*p* a PG conjuntamente. Y por consiguiente, si por un momento de tiempo se toma una hora, puesto que el ángulo GT*g* (por la Proposición xxx) es al ángulo de 33″ 10‴ 33′′′′ como IT x AZ a AT al cubo, el ángulo GP*g* (o la variación horaria de la inclinación) será al ángulo de 33″ 10‴ 33′′′′ como

$$\text{IT x AZ x TG x } \frac{\text{P}p}{\text{PG}} \text{ a AT}^3\text{. Q. E. I.}$$

Esto es así bajo la hipótesis de que la Luna girase en una órbita circular uniformemente. Pero si dicha órbita fuese elíptica, el movimiento medio de los nodos disminuye en la razón del eje menor al eje mayor; como se ha expuesto más arriba. Y en esa misma razón disminuye también la variación de la inclinación.

Corolario 1. Si sobre N*n* se eleva la perpendicular TF, y *p*M es el movimiento horario de la Luna en el plano de la eclíptica; y las perpendiculares *p*K, M*k* trazadas sobre QT y prolongadas ambas encuentran a TF en H y *h*: será IT a AT como K*k* a M*p*, y TG a H*p* como TZ a AT y por lo mismo IT x TG será igual a $\frac{\text{K}k \text{ x H}p \text{ x TZ}}{\text{M}p}$,

esto es, igual al área H*p*M*h* multiplicada por la razón $\frac{\text{TZ}}{\text{M}p}$: y en consecuencia, la variación horaria de la inclinación es a 33″ 10‴ 33′′′′ como H*p*M*h* multiplicada por AZ x $\frac{\text{TZ}}{\text{M}p}$ x $\frac{\text{P}p}{\text{PG}}$ es a AT^3.

Corolario 2. Y por tanto, si la Tierra y los nodos en cada hora completada se retirasen de sus nuevos lugares y regresaren en un instante siempre a sus lugares iniciales, de modo que su posición durante un mes periódico entero permaneciese dada; toda la variación de la inclinación durante dicho mes sería a 33″ 10‴ 33′′′′ como la reunión de todas las áreas H*p*M*h* generadas con la revolución del punto P unidas por los signos correspondientes + y -, multiplicada dicha reunión por AZ x TZ x $\frac{\text{P}p}{\text{PG}}$ a M*p* x AT^3, es decir, como el círculo entero QA*qa* multiplicado por AZ x TZ x $\frac{\text{P}p}{\text{PG}}$ a M*p* x AT^3, o sea, como la circunferencia DA*qa* multiplicada por AZ x TZ x $\frac{\text{P}p}{\text{PG}}$ a 2M*p* x AT^2.

Corolario 3. Por consiguiente, en una posición dada de los nodos, la variación horaria media que, continuada uniformemente durante un mes, podría generar aquella variación mensual, es a 33″ 10‴ 33′′′′ como AZ x TZ x $\frac{\text{P}p}{\text{PG}}$ a 2AT^2, o como P*p* x $\frac{\text{AZ x TZ}}{\text{½AT}}$ a PG x 4AT, es decir (puesto que P*p* es a PG como el seno de la inclinación susodicha al radio, y $\frac{\text{AZ x TZ}}{\text{½AT}}$ AT sea a 4AT como el seno del doble del ángulo AT*n* al cuádruplo del radio) como el seno de la misma inclinación multiplicado por el seno de la doble distancia entre los nodos y el Sol al cuádruplo del cuadrado del radio.

Corolario 4. Puesto que la variación horaria de la inclinación, cuando los nodos se hallan en las cuadraturas, es (por esta Proposición) al ángulo de 33″ 10‴ 33′′′′ como IT x AZ x TG $\frac{\text{P}p}{\text{PG}}$ a AT^3, es decir, como $\frac{\text{IT x TG}}{\text{½AT}}$ x $\frac{\text{P}p}{\text{PG}}$ a 2AT; esto es, como el seno de la doble distancia de la Luna a las cuadraturas multiplicado por $\frac{\text{P}p}{\text{PG}}$ al doble del radio: la suma de todas las variaciones horarias en el tiempo en que la Luna en esta

situación de los nodos, pasa de la cuadratura a la sicigia (esto es, en el espacio de 1776⅙ horas) será a la suma de otros tantos ángulos de 33″ 10‴ 33⁗ o sea, 5878″, como la suma de todos los senos de la doble distancia de la Luna a las cuadraturas multiplicada por $\frac{Pp}{PG}$ a la suma de otros tantos diámetros; esto es, como el diámetro multiplicado por $\frac{Pp}{PG}$ a la circunferencia; es decir, si la inclinación fuese de 5.º 1′, como 7 x $^{874}/_{10000}$ a 22, o 278 a 10 000. Y por lo mismo, la variación total procedente de la suma de todas las variaciones horarias es de 163″, o de 2′ 43″.

Proposición xxxv. Problema xvi

Hallar la inclinación de la órbita de la Luna hacia el plano de la eclíptica en un tiempo dado.

Sea AD el seno de la inclinación máxima, y sea AB el seno de la inclinación mínima. Bisecando a BD en C, y por centro en C y distancia BC trácese el círculo BGD. Sobre AC tómese CE en la misma razón respecto a EB en que está EB respecto a 2BA: y si para el tiempo dado se forma el ángulo AEG igual al doble de la distancia de los nodos a las cuadraturas, y sobre AD se hace descender la perpendicular GH: será AH el seno de la inclinación buscada.

Pues, GE^2 es igual a $GH^2 + HE^2 = BHD + HE^2 = HBD + HE^2 - BH^2 = HBD + BE^2 - 2BH \times BE = BE^2 + 2EC \times BH = 2EC \times AB + 2EC \times BH = 2EC \times AH$. Por consiguiente, al estar dado 2EC, GE^2 es como AH. Represente ahora *AEg* al doble de la distancia de los nodos a las cuadraturas después de completado un momento de tiempo dado, y el arco G*g*, por estar dado el ángulo GE*g*, será como la distancia GE. Pero H*h* es a G*g* como GH a GC, y por tanto H*h* es como el producto GH x G*g*, o GH x GE; es decir, como

$$\frac{GH}{GE} \times GE^2 \text{ o } \frac{GH}{GE} \times AH,$$

es decir, como AH y el seno del ángulo AEG conjuntamente. Por consiguiente, si AH en algún caso fuese el seno de la inclinación, ésta aumentará con incrementos iguales a los del seno de la inclinación, por el Corolario 3 de la Proposición anterior, y por lo mismo siempre permanecerá igual a dicho seno. Pero AH, cuando el punto G cae en uno de los puntos B o D, es igual a dicho seno, y por consiguiente siempre permanece igual a él. Q. E. D.

En esta demostración he supuesto que el ángulo BEG, que es el doble de la

distancia de los nodos a las cuadraturas, aumenta uniformemente. Pues no hay tiempo de examinar todas las minucias de las desigualdades. Imaginemos ahora que el ángulo BEG es recto, y en este caso G*g* sería el aumento horario de la doble distancia entre los nodos y el Sol; y la variación horaria de la inclinación en dicho caso (por el Corolario 3 de la anterior proposición) será a 33″ 10‴ 33⁗ como el producto del seno de la inclinación AH por el seno del ángulo recto BEG, que es el doble de la distancia de los nodos al Sol, al cuádruplo del cuadrado del radio; es decir, como el seno de la inclinación media AH al cuádruplo del radio; esto es (puesto que dicha inclinación media viene a ser de casi 5.º y 8½′) como su seno 896 al radio cuadruplicado 40 000, o como 224 a 10 000. Es por tanto la variación entera, correspondiente a la diferencia BD de los senos respecto a la susodicha variación horaria como el diámetro BD al arco G*g*; es decir, conjuntamente como el diámetro BD a la semicircunferencia BGD y el tiempo de $2079^{7}/_{10}$ horas en que el nodo pasa de las cuadraturas a las sicigias a una hora; o también, como 7 a 11 y $2079^{7}/_{10}$ a 1. Por lo cual, si se combinan todas estas razones, la variación total BD resultará ser respecto a 33″ 10‴ 33⁗ como 224 x 7 x $2079^{7}/_{10}$ a 110 000, es decir, como 29 645 a 1000, y de aquí que la susodicha variación BD resultará de 16′ 23½″.

Esta es la variación máxima de la inclinación en tanto no se considera el lugar de la Luna en su órbita. Pues la inclinación, si los nodos se hallan en las sicigias, no cambia en nada por los cambios de lugar de la Luna. Pero si los nodos se encuentran en las cuadraturas, la inclinación es menor cuando la Luna se halla en las sicigias que cuando se halla en las cuadraturas, con diferencia de 2′ 43″, como ya he indicado en el Corolario cuarto de la Proposición anterior. Y restando 1′ 21½, mitad de esta diferencia, de la variación media total B*d* en las cuadraturas lunares se tiene 15′ 2″, y sumada en las sicigias se tiene 17′ 45″. Así pues, si la Luna se hallase en las sicigias, la variación entera en el paso de los nodos desde las cuadraturas a las sicigias será de 17′ 45″: y por lo tanto, si la inclinación, cuando los nodos se hallan en las sicigias, fuese de 5.º 17′ 20″, la misma, cuando los nodos están en las cuadraturas y la Luna en las sicigias, será de 4.º 59′ 35″. Y que esto es así, se confirma por las observaciones.

Pero si ahora se deseare la inclinación de la órbita, cuando la Luna se halla en las sicigias y los nodos en un lugar cualquiera, sea AB a AD como el seno del ángulo de 4.º 59′ 35″ al seno del ángulo de 5.º 17′ 20″, y tómese el ángulo AEG igual al doble de la distancia de los nodos a las cuadraturas; y AH será el seno de la inclinación buscada. A la inclinación de esta órbita viene a ser igual su inclinación cuando la Luna dista 90.º de los nodos. Para otros lugares de la Luna, la desigualdad mensual que admite la variación de la inclinación se compensa en el cálculo de la latitud de la Luna, y hasta en cierto modo se anula, por la desigualdad mensual del movimiento de los nodos (como dije más arriba) y por tanto puede despreciarse al calcular dicha latitud.

Escolio[12]

Con estos cálculos de los movimientos lunares he querido mostrar que dichos movimientos pueden calcularse mediante la teoría de la gravedad a partir de sus causas. Mediante la misma teoría he hallado, además, que la ecuación anual del movimiento medio de la Luna se debía a la diferente dilatación de la órbita lunar por la fuerza del Sol, según el Corolario 6 de la Proposición LXVI del Libro I. Esta fuerza en el perigeo del Sol es mayor y dilata la órbita de la Luna; en su apogeo es menor y permite a dicha órbita contraerse. La Luna gira más lentamente en la órbita dilatada, y más rápidamente en la contraída; y la ecuación anual, por la que se compensa esta desigualdad, es nula en el apogeo y perigeo solares, mientras que en la distancia media de la Tierra y el Sol asciende casi a 11′ 50″, y en otros lugares es proporcional a la ecuación del centro del Sol; y se suma al movimiento medio de la Luna cuando la Tierra pasa de su afelio a su perihelio, mientras se resta en la parte opuesta de la órbita. Suponiendo el radio del Orbe Magno de 1000 y la excentricidad de la tierra 16⅞, esta ecuación, cuando es máxima, resulta de 11′ 49″ por la teoría de la gravedad. Pero la excentricidad de la Tierra parece ser algo mayor, y al aumentar la excentricidad esta ecuación debe aumentarse en la misma proporción. Sea la excentricidad de $16^{11}/_{12}$, y la ecuación máxima será de 11′ 51″.

También descubrí que en el perihelio de la Tierra, debido a la mayor fuerza del Sol, el apogeo y los nodos de la Luna se mueven más rápidamente que en su afelio, y ello en razón inversa del cubo de la distancia de la Tierra al Sol. Y de aquí surgen las ecuaciones anuales de estos movimientos proporcionales a la ecuación del centro solar. Pero el movimiento del Sol es inversamente proporcional al cuadrado de la distancia de la Tierra al Sol, y la máxima ecuación del centro generada por esta desigualdad es de 1.º 56′ 20″, que se corresponde con la excentricidad del Sol de $16^{11}/_{12}$ mencionada más arriba. Pero si el movimiento del Sol se hallase en razón inversa del cubo de la distancia, esta desigualdad generaría una ecuación máxima de 2.º 54′ 30″. Y por lo tanto, las ecuaciones máximas generadas por las desigualdades del apogeo y de los nodos de la Luna, son a 2.º 54′ 30″ como el movimiento medio diario del apogeo y el movimiento medio diario de los nodos de la Luna son al movimiento diario medio del Sol. De donde resulta la ecuación máxima del movimiento medio del apogeo de 19′ 43″ y la ecuación máxima del movimiento medio de los nodos de 9′ 24″. La primera ecuación se suma y la última se resta cuando la Tierra pasa del perihelio al afelio; y al contrario ocurre en la parte opuesta de la órbita.

También se ha establecido mediante la teoría de la gravedad que la acción del Sol sobre la Luna es algo mayor cuando el diámetro transversal de la órbita lunar pasa por el Sol que cuando se halla en ángulo recto con la línea que une a la Tierra y el Sol: y por ello la órbita lunar es algo mayor en el primer caso que en el segundo. Y de aquí surge otra ecuación del movimiento medio lunar, que depende del lugar del

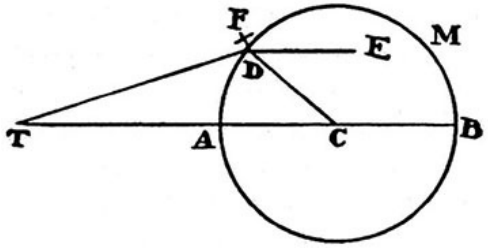
apogeo de la Luna respecto al Sol, la cual resulta máxima cuando coincide en el octante con el Sol; y es nula cuando el apogeo llega a las cuadraturas o a las sicigias: y se añade al movimiento medio al pasar el apogeo de la Luna desde la cuadratura del Sol a las sicigias mientras que se resta cuando el apogeo pasa de la sicigia a las cuadraturas. Esta ecuación, que llamaré semestral, en los octantes del apogeo, cuando es máxima, asciende a casi 3′ 45″, en la medida en que pude inferirlo de los fenómenos. Este es su valor a la distancia media del Sol a la Tierra. Pero aumenta o disminuye en razón inversa del cubo de la distancia del Sol, y por ende a la máxima distancia del Sol es de 3′ 34″ y a la mínima de 3′ 56″ aproximadamente: pero cuando el apogeo de la Luna se halla situado fuera de los octantes, resulta menor; y es a la ecuación máxima como el seno de la doble distancia del apogeo de la Luna hasta las inmediatas sicigias o la cuadratura al radio.

Por la misma teoría de la gravedad la acción del Sol sobre la Luna es algo mayor cuando la línea recta que pasa por los nodos de la Luna pasa también por el Sol que cuando dicha línea forma ángulo recto con la línea que une a la Tierra y el Sol. Y de aquí surge otra ecuación del movimiento medio lunar, que llamaré segunda semestral, y que es máxima cuando los nodos se encuentran en los octantes del Sol y que se desvanece cuando se hallan en las sicigias o en las cuadraturas, mientras en las demás posiciones de los nodos es proporcional al seno de la doble distancia de uno cualquiera de los nodos a la sicigia o cuadratura próxima; y se añade al movimiento medio de la Luna si el Sol está antes del nodo inmediato a la Luna y se resta si está delante, y en los octantes, cuando es máxima, asciende a 47″ a la distancia media del Sol y la Tierra, como infiero de la teoría de la gravedad. A otras distancias del Sol esta ecuación máxima en los octantes es inversamente proporcional al cubo de la distancia del Sol a la Tierra, y por ende en el perigeo del Sol asciende casi a 49″ y en su apogeo a casi 45″.

Por la misma teoría de la gravedad el apogeo de la Luna progresa máximamente cuando, o bien se halla en conjunción con el Sol, o bien en oposición a él, y retrocede cuando forma cuadratura con el Sol. Y la excentricidad se torna máxima en el primer caso y mínima en el segundo, por los Corolarios 7, 8 y 9 de la Proposición LXVI del Libro I. Y estas desigualdades, por los mismos Corolarios, son muy grandes y generan la ecuación principal del apogeo, a la que llamaré semestral. Y la ecuación máxima del semestre es de 12.º 18′ aproximadamente, en la medida en que pude inferirlo de las observaciones. Nuestro *Horrox* fue el primero en establecer que la Luna gira en una elipse en torno a la Tierra que está situada en su foco inferior. *Halley* situó el centro de la elipse en un epiciclo cuyo centro gira uniformemente en torno a la Tierra. Y del movimiento en el epiciclo surgen las desigualdades ya mencionadas en el progreso y regreso del apogeo y en el valor de la excentricidad. Imaginemos que la distancia media de la Luna a la Tierra se divide en 100 000 partes, y T represente a la Tierra, TC la excentricidad media de la Luna de 5505 partes. Prolónguese Te hasta B, de modo que CB sea el seno de la ecuación máxima

semestral de 12.º 18′ para el radio TC, y el círculo BDA trazado con centro en C y radio CB será el susodicho epiciclo en el que se halla situado el centro de la órbita lunar y gira según el orden de las letras BDA. Tómese el ángulo BCD igual al doble del argumento anual, o a la doble distancia del verdadero lugar del Sol al apogeo de la Luna igualado una vez, y CTD será la ecuación semestral del apogeo lunar y TD la excentricidad de la órbita de la Luna que se dirige hacia el apogeo igualado por segunda vez. Pues bien, una vez conocidos el movimiento medio de la Luna y el apogeo y la excentricidad, así como el eje mayor de la órbita de la Luna de 200 000 partes; a partir de esto se obtiene el verdadero lugar de la Luna en la órbita y su distancia a la Tierra, y esto por métodos muy conocidos.

En el perihelio de la Tierra, debido a la mayor fuerza del Sol, el centro de la órbita lunar se mueve más velozmente en torno al centro C que en el afelio, y esto en razón inversa del cubo de la distancia del Sol a la Tierra. Por estar la ecuación del centro del Sol comprendida en el argumento anual, el centro de la órbita de la Luna se mueve más velozmente en el epiciclo BDA en razón inversa del cuadrado de la distancia de la Tierra al Sol. Para que pueda moverse aún más deprisa en razón simple inversa de la distancia, trácese desde el centro D de la órbita hacia el apogeo de la Luna la recta DE, paralela a TC, y tómese el ángulo EDF igual al exceso del argumento anual antedicho sobre la distancia hacia adelante del apogeo de la Luna hasta el perigeo del Sol; o lo que es lo mismo, tómese el ángulo CDF igual al complemento de la verdadera anomalía del Sol respecto a 360 grados. Y sea DF a DC como el doble de la excentricidad del Orbe Magno respecto a la distancia media del Sol a la Tierra y el movimiento diario medio del Sol desde el apogeo de la Luna al movimiento diario medio del Sol desde su propio apogeo conjuntamente, es decir, como 33⅜ a 1000 y 52′ 27″ 16‴ a 59′ 8″ 10‴ conjuntamente, o como 3 a 100. E imaginemos que el centro de la órbita de la Luna se sitúa en el punto F y, mientras gira en el epiciclo de centro D y radio DF, el punto D progresa por la circunferencia del círculo DABD. Pues con esta proporción, la velocidad con la que el centro de la órbita lunar se moverá por una cierta línea curva descrita en torno al centro C, será inversamente como el cubo de la distancia del Sol a la Tierra muy aproximadamente, como es preciso.



Es difícil el cálculo de este movimiento, pero se torna más fácil mediante la

aproximación siguiente. Si la distancia media de la Luna a la Tierra es de 100 000 partes y la excentricidad es de 5505 partes, como antes, la recta CB o CD resultará de 1172¾ partes y DF de 35⅕ partes. Y esta recta, a la distancia TC, subtiende a la Tierra el ángulo generado por la traslación del centro de la órbita desde el punto D al punto F debido al movimiento de dicho centro: y el doble de esa línea en posición paralela y a la distancia del foco superior de la órbita de la Luna a la Tierra subtiende el mismo ángulo también generado en el movimiento del foco por dicha traslación y que por ello puede ser llamado segunda ecuación del centro. Y esta ecuación, a la distancia media de la Luna a la Tierra, es como el seno del ángulo comprendido entre la susodicha recta DF y la trazada desde el punto F hasta la Luna muy aproximadamente, y cuando es máxima resulta de 2′ 25″. Pues el ángulo comprendido entre la recta DF y la recta trazada desde F hasta la Luna se obtiene tanto restando el ángulo EDF de la anomalía media de la Luna, como añadiendo la distancia de la Luna al Sol a la distancia del apogeo de la Luna hasta el apogeo del Sol. Y como el radio es el seno del ángulo así hallado, así son también 2′ 25″ a la segunda ecuación, que ha de sumarse si aquella suma fuese menor que el semicírculo y restarse si fuese mayor. Así se obtendrá su longitud en las sicigias mismas de los astros.

Puesto que la atmósfera de la Tierra hasta una altura de 35 o de 40 millas refracta la luz del Sol, y al refractarla la dispersa por la sombra de la Tierra, y al dispersarla en los límites de la sombra dilata a la sombra, al diámetro de la sombra resultante de la paralaje, añado un minuto en los eclipses de la Luna, o acaso un minuto y un tercio.

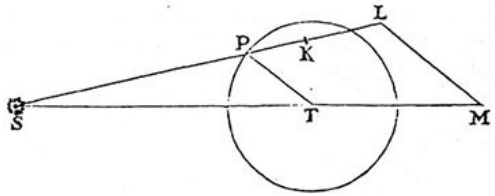
Pero la teoría de la Luna debe examinarse y establecerse mediante los fenómenos, primero en las sicigias, después en las cuadraturas y por último en los octantes. Y quien aborde esta tarea no sufrirá inconvenientes al utilizar los movimientos medios del Sol y de la Luna para el tiempo del Real Observatorio de *Greenwich* el último día de diciembre del año 1700, calendario antiguo, a saber: el movimiento medio del Sol ♑ 20.º 43′ 40″ y su apogeo ♋ 7.º 44′ 30″ y el movimiento medio de la Luna ♒ 15.º 21′ 00″, y su apogeo ♓ 8.º 20′ 00″, y del nodo ascendente ☊ 27.º 24′ 20″, y la diferencia de meridianos entre este observatorio y el Real de *París* es de 0h. 9m. 20sg., pero el movimiento medio de la Luna y de su apogeo no se tienen todavía con suficiente exactitud.

Proposición xxxvi. Problema xvii

Hallar la fuerza del Sol para mover el mar

La fuerza ML del Sol, o PT, en las cuadraturas lunares, a la hora de perturbar los movimientos lunares, era respecto a la Fuerza de la gravedad entre nosotros (por la Proposición xxv de este Libro) como 1 a 638092,6. Y la fuerza TM - LM, o 2PK en

las sicigias lunares era el doble mayor. Pero estas fuerzas, si descendiéramos a la superficie de la Tierra, disminuirán en razón de las distancias al centro de la Tierra, es decir, en razón de 60½ a 1; y por consiguiente, la fuerza anterior en la superficie de la Tierra es a la fuerza de la gravedad como 1 a 38 604 600. Con esta fuerza el mar se deprime en los lugares que distan 90 grados del Sol. Con la otra fuerza, que es el doble mayor, se eleva el mar tanto bajo el Sol como en la parte opuesta. La suma de las fuerzas es a la fuerza de la gravedad como 1 a 12 868 200. Y puesto que la misma fuerza induce el mismo movimiento, tanto si con ella deprime al agua en las regiones que distan 90 grados del Sol, como si la eleva en las regiones bajo el Sol u opuestas a él, esta suma será la fuerza total del Sol para agitar el mar; y tendrá el mismo efecto que si toda ella elevase el mar en las regiones bajo el Sol y opuestas al Sol, y en cambio no actuase en absoluto en las regiones que distan 90 grados del Sol.



Esta es la fuerza del Sol para mover el mar en un lugar dado cualquiera, cuando el Sol se halla, tanto en el vértice del lugar, como a su distancia media de la Tierra. En otras posiciones del Sol su fuerza para mover el mar es como el seno verso del doble de la altura del Sol sobre el horizonte del lugar, directamente, e inversamente como el cubo de la distancia del Sol a la Tierra.

COROLARIO. Puesto que la fuerza centrífuga de las partes de la Tierra originada por el movimiento terrestre diario, que es a la fuerza de la gravedad como 1 a 289, haga que la altura del agua en el ecuador supere la altitud de la misma sobre los polos en 85 472 pies *parisinos*, como se vio más arriba en la Proposición XIX, la fuerza solar de que hemos tratado, al ser a la fuerza de la gravedad como 1 a 12 868 200 y por tanto a la susodicha fuerza centrífuga como 289 a 12 868 200, o sea, como 1 a 44 527, hará que la altura del agua en las regiones bajo el Sol y sus opuestas supere a la altura de la misma en los lugares que distan 90 grados del Sol en solamente un pie *parisino* y once pulgadas con de pulgada. Pues ésta es la cantidad que, respecto a 85 472 pies, es como 1 a 44 527.

PROPOSICIÓN XXXVII. PROBLEMA XVIII[13]

Hallar la fuerza de la Luna para mover el mar.

La fuerza de la Luna para mover el mar hay que inferirla de su proporción

respecto a la fuerza del Sol, y esta proporción hay que inferirla de la proporción de los movimientos del mar generados por estas fuerzas. Antes de la desembocadura del río AVON, en la tercera milla más abajo de *Bristol*, en primavera y otoño el ascenso total del agua en la conjunción y en la oposición de los astros, según la observación de *Samuel Sturmy*, es de unos 45 pies, mientras que en las cuadraturas están sólo de 25. La altura primera es debida a la suma de las fuerzas, y la segunda a su diferencia. Por consiguiente, sean S y L las fuerzas del Sol y de la Luna situados sobre el ecuador y a distancias medias de la Tierra, y L + S será a L - S como 45 a 25, o como 9 a 5.

La subida del mar en el puerto ce *Plymouth* según observación de *Samuel Colepress* alcanza unos dieciséis pies de altura media, y en primavera y otoño la altura de la marea en las sicigias puede superar a su altura en las cuadraturas en siete u ocho pies más. Si la diferencia máxima de estas alturas fuese de nueve pies, L + S será a L - S como 20½ a 11½, o como 41 a 23. Proporción que coincide bastante bien con la anterior. Por la magnitud de las mareas en el puerto de *Bristol*, parecen más fiables las observaciones de *Sturmy*, y por ello nos atendremos a la proporción de 9 a 5 hasta que haya constancia de algo más seguro.

Por lo demás, debido a los movimientos de reciprocación de las aguas, las máximas mareas no coinciden con las sicigias de los astros sino que son las terceras *post*-sicigias, como se ha dicho, o son inmediatas al tercer paso de la Luna por el meridiano del lugar después de las sicigias, o mejor (como observa *Sturmy*), son las terceras después del día de novilunio o plenilunio, o también más o menos doce horas después del novilunio o plenilunio, y por ello vienen a incidir más o menos en la hora cuarenta y tres desde el novilunio o plenilunio. En el puerto susodicho vienen a ocurrir unas siete horas después del paso de la Luna por el meridiano del lugar; y por tanto siguen inmediatamente al paso de la Luna por el meridiano, cuando la Luna dista del Sol, o de la oposición del Sol, más o menos dieciocho o diecinueve grados hacia adelante. Verano e invierno son de gran importancia, no en los mismos solsticios, sino cuando el Sol dista de los solsticios una décima parte del curso entero, es decir, unos 36 ó 37 grados. E igualmente, la máxima subida del mar se origina por el paso de la Luna por el meridiano del lugar, cuando la Luna dista del Sol casi una décima parte del movimiento total de marea a marea. Sea esa distancia más o menos de 18½ grados. Y la fuerza del Sol a esta distancia de la Luna de las sicigias y de las cuadraturas será menor para aumentar o disminuir el movimiento del mar debido a la fuerza lunar que en las propias sicigias y cuadraturas en la razón del radio al seno del complemento del doble de esa distancia o sea del ángulo de 37 grados, es decir, en razón de 10 000 000 a 7 986 355. Y por lo mismo, en la proporción de más arriba hay que escribir 0,7986355S en lugar de S.

Pero también hay que disminuir la fuerza de la Luna en las cuadraturas, debido a la declinación lunar del ecuador. Pues en las cuadraturas, o más bien en el grado 18½ después de las cuadraturas, tiene una declinación aproximada de 23.º 13′. Y la fuerza

para mover el mar del astro en declinación ecuatorial disminuye en razón cuadrada del seno del complemento de la declinación muy aproximadamente. Y por tanto, la fuerza de la Lima en estas cuadraturas es tan solo de 0,8570327L. Por lo tanto, L + 0,7986355S es a 0,8570327L - 0,7986355S como 9 a 5.

Además los diámetros de la órbita en que debería moverse la Luna sin excentricidad son entre sí como 69 a 70; y por ello la distancia de la Luna a la Tierra en las sicigias es a su distancia en las cuadraturas como 69 a 70, en iguales restantes condiciones. Y sus distancias en el grado 18½ desde las sicigias cuando se genera la marea máxima, y en el grado 18½ desde las cuadraturas, cuando se genera la marea mínima, son a su distancia media como 69,098747 y 69,897345 a 69½. Pero las fuerzas de la Luna para mover el mar están en razón inversa del cubo de las distancias, y por ello las susodichas fuerzas en la máxima y mínima de estas distancias son a la fuerza en la distancia media como 0,9830427 y 1,017522 a 1. De donde resulta que 1,017522L + 0,7986355S es a 0,9830427 x 0,8570327L - 0,7986355S como 9 a 5. Y S a L como 1 a 4,4815. De suerte que siendo la fuerza del Sol a la de la gravedad como 1 a 12 868 200, la fuerza de la Luna será a dicha fuerza de la gravedad como 1 a 2 871 400.

COROLARIO 1. Dado que el agua bajo la acción de la fuerza del Sol sube a una altura de un pie, once pulgadas y $\frac{1}{30}$ de pulgada, la Luna con la propia fuerza hará que ascienda hasta una altura de ocho pies y $7\frac{5}{22}$ pulgadas, y con ambas fuerzas juntas hasta una altura de 10½ pies, y cuando la Luna está en el perigeo hasta 12½ pies y más, sobre todo cuando el viento colabora con su empuje. Semejante fuerza es bastante para provocar todos los movimientos del mar, y responde adecuadamente a la cantidad de movimiento. Pues en los mares que se extienden ampliamente de oriente a occidente, como en el *Pacífico* y en las zonas extratropicales de los mares *Atlántico* y *Etiópico*, el agua suele alcanzar alturas de seis, nueve, doce o quince pies. Pero en el mar *Pacífico*, más profundo y más extenso, se asegura que las mareas son mayores que en el *Atlántico* y en el *Etiópico*. Y es que para que la marea sea más alta, la anchura del mar de este a oeste no debe ser menor de noventa grados. En el mar *Etiópico* la subida del agua entre los trópicos es menor que en las zonas templadas debido a la estrechez del mar entre *África* y la *América* austral. En el medio del mar el agua no puede ascender, salvo que descienda a la vez en las dos orillas oriental y occidental, pese a que en nuestros estrechos mares haya de descender por veces alternas en dichas costas. Por esta razón el flujo y reflujo, en las islas muy distantes de las costas, suele ser muy pequeño. En los puertos en los que el agua se ve obligada a llenar y vaciar alternativamente las cuencas pasando con gran ímpetu por estrechos y poco profundos canales, los flujos y reflujos han de ser por fuerza mayores, como en *Plymouth* y en el puente de *Chepstow*, en *Inglaterra*; en el *Mont S. Michele* o en la ciudad de *Avranches* en *Normandía*, en *Cambaya y Pegu* en la *India* oriental. En dichos lugares el mar, entrando y saliendo con gran velocidad, tanto inunda los litorales como deja millas enteras de arenales. Y el ímpetu de las entradas y de las

salidas de agua no cesa hasta que el agua sube o desciende 30, 40, y hasta 50 pies, o más. Y lo mismo ocurre con los estrechos largos y poco profundos, como el de *Magallanes* o los que rodean *Inglaterra*. Las mareas, en semejantes puertos y estrechos por la fuerza de los accesos y recesos aumentan en gran medida. Mientras que en las costas con descenso rápido hacia un mar profundo y abierto, en las cuales el agua sin fuerza en la llegada y en el retroceso puede subir y bajar libremente, la magnitud de la marea responde a las fuerzas del Sol y de la Luna.

COROLARIO 2. Dado que la fuerza de la Luna para mover el mar es a la fuerza de la gravedad como 1 a 2 871 400, es evidente que dicha fuerza es mucho menor que la que es perceptible mediante experimentos con péndulos o cualesquiera otros estáticos o hidrostáticos. Únicamente dicha fuerza muestra un efecto sensible en las mareas.

COROLARIO 3. Toda vez que la fuerza de la Luna para mover el mar es a la correspondiente fuerza del Sol como 4,4815 a 1, y dichas fuerzas (por el Corolario 14 de la Proposición LXVI del Libro I) son como las densidades de los cuerpos de la Luna y del Sol y el cubo de los diámetros aparentes conjuntamente; la densidad de la Luna será a la densidad del Sol directamente como 4,4815 ale inversamente como el cubo del diámetro de la Luna al cubo del diámetro del Sol: esto es (siendo los diámetros medios aparentes de la Luna y del Sol de 31′, 16½ y de 32′, 12″), como 4891 a 1000. Pero la densidad del Sol era a la densidad de la Tierra como 1000 a 4000; y por tanto, la densidad de la Luna es a la densidad de la Tierra como 4891 a 4000, o como 11 a 9. Luego el cuerpo de la Luna es más denso y terroso que nuestra Tierra.

COROLARIO 4. Y puesto que el verdadero diámetro de la Luna obtenido de las observaciones astronómicas es al verdadero diámetro de la Tierra como 100 a 365, la masa de la Luna será a la masa de la Tierra como 1 a 39,788.

COROLARIO 5. Y la gravedad aceleratriz en la superficie de la Luna será casi un triplo menor que la gravedad aceleratriz en la superficie de la Tierra.

COROLARIO 6. Y la distancia del centro de la Luna al centro de la Tierra será a la distancia del centro de la Luna al centro común de gravedad de la Tierra y la Luna como 40,788 a 39,788.

COROLARIO 7. Y la distancia media del centro de la Luna al centro de la Tierra en los octantes de la Luna será aproximadamente de $60\frac{2}{5}$ semidiámetros máximos de la Tierra. Porque el semidiámetro máximo de la Tierra era de 19 658 600 pies parisinos, y la distancia media entre centros de Luna y Tierra consta de $60\frac{2}{5}$ de estos mismos semidiámetros y es igual a 1 187 379 440 pies. Esta distancia (por el Corolario anterior) es a la distancia del centro de la Luna al centro común de gravedad de la Tierra y la Luna como 40,788 a 39,788: por lo tanto, la distancia última es de 1 158 268 534 pies. Y, puesto que la Luna gira respecto a las fijas en 27 días, 7 horas y $43\frac{4}{9}$ minutos, el seno verso del ángulo descrito por la Luna en un minuto es de 12 752 341 para un radio de 1 000 000 000 000 000. Y el radio es a este seno verso como 1 158 268 534 pies son a 14,7706353 pies. Por consiguiente, la Luna, cayendo hacia la

Tierra con la fuerza con la que es retenida en órbita describirá en el tiempo de un minuto 14,7706353 pies. Y aumentando dicha fuerza en una razón como $178^{29}/_{40}$ a $177^{27}/_{40}$ se tendrá la fuerza total de gravedad en la órbita de la Luna, por el Corolario de la Proposición III. Y cayendo la Luna con esta fuerza describirá en el tiempo de un minuto 14,8538067 pies. Y a $1/_{60}$ de esa distancia de la Luna al centro de la Tierra, esto es, a la distancia de 197 896 573 pies del centro de la Tierra, un grave cayendo en el tiempo de un segundo describe también 14,8538067 pies. Por lo cual a la distancia de 19 615 800 pies, que son el semidiámetro medio de la Tierra, un grave cayendo describirá 15,11175 pies, o sea, 15 pies, 1 pulgada y $4^{1}/_{11}$ líneas. Este será el descenso de los cuerpos a la latitud de 45 grados. Y según la tabla precedente expuesta en la Proposición XX, el descenso será algo mayor en la latitud de *París*, con un exceso de casi $^{2}/_{3}$ de línea. Pues según este cálculo los graves en la latitud de *París* cayendo en el vacío describirán en un segundo casi 15 pies parisinos, 1 pulgada y $4^{4}/_{25}$ líneas. Y si la gravedad disminuyese, suprimiendo la fuerza centrífuga debida en esa latitud a la rotación diaria de la Tierra, los graves cayendo allí en el tiempo de un segundo describirán 15 pies, 1 pulgada y 1½ líneas. Y ya más arriba, en las Proposiciones IV y XIX, se ha mostrado que en la latitud de *París* los graves caen con esta velocidad.

COROLARIO 8. La distancia media entre los centros de la Tierra y la Luna en las sicigias de la Luna es de 60 semidiámetros máximos terrestres, restada una treintava parte del semidiámetro aproximadamente. Pero en las cuadraturas de la Luna las distancias medias entre los dichos centros es de $60^{5}/_{6}$ semidiámetros terrestres. Pues estas dos distancias son a las distancias medias de la Luna en los octantes como 69 y 70 a 69½, por la Proposición XXVIII.

COROLARIO 9. La distancia media del centro de la Tierra al de la Luna en las sicigias de la Luna es de 60,1 semidiámetros medios terrestres. Y en las cuadraturas de la Luna la distancia media entre esos mismos centros es de 61 semidiámetros terrestres menos un treintavo de semidiámetro.

COROLARIO 10. En las sicigias de la Luna su paralaje horizontal media a latitudes de 0, 30, 38, 45, 52, 60, 90 grados es de 57′ 20″; 57′ 16″; 57′ 14″; 57′ 12″; 57′ 10″; 57′ 8″; 57′ 4″, respectivamente.

No he tenido en cuenta en estos cálculos la atracción magnética de la Tierra, cuya cantidad es ciertamente muy pequeña y se desconoce. Pero si alguna vez se llegase a conocer esta atracción, y las medidas en grados sobre los meridianos, las longitudes de los péndulos isócronos en los distintos paralelos, las leyes del movimiento del mar, la paralaje lunar junto con los diámetros aparentes del Sol y de la Luna pudieran ser determinados con mayor exactitud sobre la base de los fenómenos, se podrían repetir estos cálculos con mayor precisión.

PROPOSICIÓN XXXVIII. PROBLEMA XIX

Hallar la figura del cuerpo lunar.

Si el cuerpo lunar fuese fluido como lo es nuestro mar, la fuerza de la Tierra para elevar dicho fluido tanto en las partes más cercanas como en las más lejanas sería a la fuerza de la Luna, por la cual nuestro mar es elevado en las partes que están bajo ella y en las opuestas, como la gravedad aceleratriz de la Luna hacia la Tierra es a la gravedad aceleratriz de la Tierra hacia la Luna, y el diámetro de la Luna al diámetro de la Tierra conjuntamente; es decir, como 39,788 a 1 y 100 a 365 conjuntamente, o como 1081 a 100. Por lo cual, dado que la fuerza lunar eleva nuestro mar hasta $8^{1}/_{2}$ pies, la fuerza terrestre elevaría el fluido lunar hasta 93 pies. En consecuencia, la figura lunar sería un esferoide, cuyo diámetro mayor prolongado pasaría por el centro de la Tierra, y superaría a los diámetros perpendiculares con un exceso de 186 pies. Por tanto, esa es la figura que posee la Luna, y debió adoptarla desde el principio. Q. E. I.

COROLARIO. Y esto hace que la misma cara de la Luna esté siempre vuelta hacia la Tierra. Y el cuerpo lunar no podría permanecer en otra posición, sino que oscilando retornaría siempre a esta posición. Sin embargo, las oscilaciones, debido a la pequeñez de las fuerzas que las provocan, serían muy lentas, hasta el punto de que la cara de la Luna que debería mirar siempre hacia la Tierra podría mirar hacia el otro foco de la órbita lunar (por lo dicho en la Proposición XVII), sin que se apartase al instante de dicha posición y se volviese hacia la Tierra.

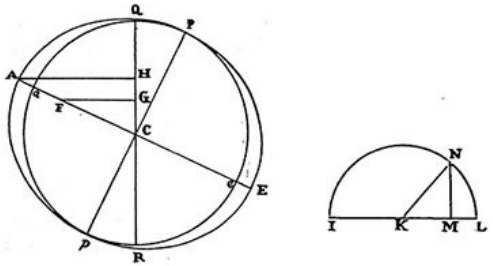
LEMA I[14]

Si APEp representa a la Tierra uniformemente densa, trazada con centro en C y polos P, p, y ecuador AE, y si con centro en C se imagina que trazamos una esfera Pape de radio CP; y sea QR un plano sobre el cual incide normalmente una recta trazada desde el centro del Sol al centro de la Tierra, mientras que cada partícula de toda la Tierra exterior PapAPepE que sobresale de la esfera descrita del modo dicho trata de alejarse hacia un lado y otro del plano QR, y los impulsos de cada una de las partículas es como su distancia al plano, digo, primero que la fuerza y eficacia totales de todas las partículas situadas en el círculo AE del ecuador y uniformemente dispuestas en anillo alrededor del globo en orden a hacer rotar a la Tierra en torno a su centro es a toda la fuerza y eficacia de otras tantas partículas situadas en el punto A sobre el ecuador, punto máximamente distante del plano QR, y en orden a producir un movimiento circular de la Tierra similar y en torno a su centro, como uno a dos. Y el susodicho movimiento circular tendrá lugar en torno a un eje situado en la sección común del ecuador y el plano QR.

Pues, con centro en K y semidiámetro IL trácese el semicírculo INLK. Imagínese que la semicircunferencia INL se divide en innumerables partes iguales, y desciendan sobre el diámetro IL desde cada parte N los senos NM. En tal caso la suma de los cuadrados de todos los senos NM será igual a la suma de los cuadrados de los senos KM, y una y otra suma juntas será igual a la suma de los cuadrados de otros tantos semidiámetros KN; y por consiguiente, la suma de todos los cuadrados NM será la mitad de la suma de los cuadrados de otros tantos semidiámetros KN.

Ahora divídase el perímetro del círculo AE en otras tantas partículas iguales y desde cada una de ellas F se hace descender sobre el plano QR la perpendicular FG, y lo mismo desde el punto A la perpendicular AH. Y la fuerza con que la partícula F se aleja del plano QR será como la susodicha perpendicular FG, por hipótesis, y dicha fuerza multiplicada por la distancia CG será la eficacia de la partícula F en orden a hacer girar a la Tierra en torno a su eje. Y por lo mismo, la eficacia de la partícula en el punto F será a la eficacia de la partícula en el punto A como FG x GC a AH x HC, es decir, como FC^2 a AC^2; y por tanto, la eficacia entera de todas las partículas en sus lugares F será a la eficacia de otras tantas partículas en el lugar A como la suma de todos los FC^2 a la suma de otros tantos AC^2, es decir (por lo ya demostrado) como uno a dos. Q. E. D.

Y toda vez que las partículas actúan apartándose perpendicularmente del plano QR, y esto igualmente a uno y otro lado del plano, harán girar a la circunferencia del círculo ecuatorial y a la Tierra anexa a él en torno a un eje situado tanto sobre dicho plano QR como sobre el plano del ecuador.

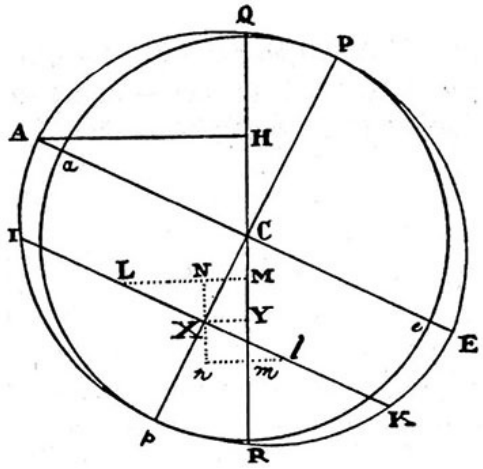


Lema II

Con los mismos supuestos, digo, en segundo lugar, que la fuerza y eficacia totales de todas las partículas situadas por todas partes fuera de la esfera, en orden a hacer girar a la Tierra en torno al mismo eje es a la fuerza total de las mismas partículas, uniformemente dispuestas alrededor del círculo ecuatorial en forma de anillo y en orden a mover a la Tierra con un movimiento circular semejante, como dos a cinco.

Pues sea IK un círculo menor cualquiera paralelo al ecuador AE, y sean L, *l* dos

partículas cualesquiera iguales situadas en dicho círculo fuera del globo P*ape*. Y si sobre el plano QR, que es perpendicular a un radio trazado al Sol, se trazan las perpendiculares LM, *lm*, las fuerzas totales con que tales partículas huyen del plano QR serán proporcionales a dichas perpendiculares LM, *lm*. Trácese entonces la línea L*l* paralela al plano P*ape* y biséquese en X, y por el punto X trácese N*n* paralela al plano QR y que encuentre a las perpendiculares LM, *lm* en N y *n*, y sobre el plano QR hágase caer la perpendicular XY. Y las fuerzas opuestas de las partículas L y *l* en orden a hacer girar a la Tierra en direcciones opuestas son como LM x MC y *lm* x *m*C, es decir, como LN x MC + NM x MC y I*n* x *m*C - *nm* x *m*C, o sea, LN x MC + NM x MC y LN x *m*C - NM x *m*C, y la diferencia entre ambas, LN x M*m* - NM x (MC + *m*C), es la fuerza de ambas partículas juntas para mover la Tierra en giro. La parte positiva de esta diferencia LN x M*m*, o sea, 2LN x NX es a la fuerza de dos partículas de igual magnitud, 2AH x HC, situadas en A, como LX^2 a AC^2. Y la parte negativa, NM x (MC + *m*C), o sea, 2XY x CY es a la fuerza de las mismas partículas situadas en A, 2AH x HC, como CX^2 a AC^2. Y por ello, la diferencia de las partes, es decir, la fuerza de las dos partículas L y *l* tomadas conjuntamente para hacer girar a la Tierra es a la fuerza de dos partículas iguales a ellas y situadas en el punto A en orden a hacer girar igualmente a la Tierra como LX^2 - CX^2 a AC^2. Pero si la circunferencia IK del círculo IK se divide en innumerables partículas iguales L, todos los LX^2 serán a otros tantos IX^2 como 1 a 2 (por el Lema I) y a otras tantas AC^2 como IX^2 a $2AC^2$; y todas las CX^2 a otras tantas AC^2 como $2CX^2$ a $2AC^2$. Por lo tanto, las fuerzas reunidas de todas las partículas alrededor del círculo IK son a las fuerzas reunidas de otras tantas partículas en el punto A como IX^2 - $2CX^2$ a $2AC^2$, y por lo mismo (por el Lema I) son respecto a las fuerzas reunidas de otras tantas partículas en torno al círculo AE como IX^2 - $2CX^2$ a AC^2.



Pero si ahora se divide el diámetro P*p* de la esfera en innumerables partes iguales sobre las que se sitúan otros tantos círculos IK, la materia en el perímetro de un círculo cualquiera IK será como IX^2: y por lo mismo, la fuerza de dicha materia para mover la Tierra en giro será como IX^2 por $IX^2 - CX^2$. Y la fuerza de esa misma materia si se hallase en el perímetro del círculo AE sería como $IX^2 \times AC^2$. Y por tanto, la fuerza de todas las partículas de toda la materia, situada fuera del globo en los perímetros de todos los círculos, es a la fuerza de otras tantas partículas situadas en el perímetro del círculo máximo AE, como todos los $IX^2 \times (IX^2 - 2CX^2)$ a otros tantos $IX^2 \times AC^2$, es decir, como todas las $AC^2 - CX^2$ por $AC^2 - 3CX^2$ a otras tantas AC^2, es decir, como todas las $AC^2 - CX^2$ por $AC^2 - 3CX^2$ a otras tantas $AC^2 - CX^2$ por AC^2, esto es, como todas las $AC^4 - 4AC^2 \times CX^2 + 3CX^4$ a otras tantas $AC^4 - AC^2 \times CX^2$, es decir, como toda la cantidad fluyente cuya fluxión es $AC^4 - 4AC^2 \times CX^2 + 3CX^4$, a toda la cantidad fluyente cuya fluxión es $AC^4 - AC^2 \times CX^2$; y en consecuencia, por el método de las fluxiones, como $AC^4 \times CX - \frac{4}{3}C^2 \times CX^3 + \frac{3}{5}X^5$ es a $AC^4 \times CX - \frac{1}{3}AC^2 \times CX^3$, o sea, si en lugar de CX se escribe todo C*p* o AC, como $\frac{4}{15}AC^5$ a $\frac{2}{3}AC^5$, lo que es como dos a cinco. Q. E. D.

Lema III

Con los mismos supuestos, digo, en tercer lugar, que el movimiento de toda la Tierra en torno al eje antedicho, resultante de los movimientos de todas las partículas, será al movimiento del anillo susodicho en torno al mismo eje en una razón compuesta de la razón de la materia en la Tierra a la materia en el anillo y de la razón de tres cuadrados del cuadrante de cualquier círculo a dos cuadrados de su diámetro; es decir, en razón de la materia a la materia y del número 925 275 al número 1 000 000.

Pues el movimiento de un cilindro que gira en torno a su eje inmóvil es al movimiento de la esfera inscrita que gira a la vez, como cuatro cuadrados iguales cualesquiera son a tres de los círculos inscritos en ellos: y el movimiento del cilindro es al movimiento de un anillo muy delgado que rodease el lugar común de contacto, como el doble de la materia en el cilindro al triple de la materia en el anillo; y este movimiento del anillo, continuado uniformemente en torno al eje del cilindro, es al movimiento uniforme de éste en torno a su propio diámetro, realizado en el mismo tiempo periódico, como la circunferencia del círculo al doble del diámetro.

Hipótesis II

Si, suprimiendo todo el resto de la Tierra, el anillo predicho, sólo en derredor de la

Tierra, girase en torno al Sol con movimiento anual, mientras que en torno a su propio eje, inclinado éste sobre el plano de la eclíptica con un ángulo de 23½ grados, girase también con movimiento diario: el movimiento de los puntos equinocciales sería el mismo, tanto si el susodicho anillo fuera fluido, como si estuviese constituido de materia firme y rígida.

Proposición xxxix. Problema xx[15]

Hallar la precesión de los equinoccios

El movimiento medio horario de los nodos de la Luna en una órbita circular, cuando los nodos están en cuadraturas, era de 16″ 35‴ 16⁗ 36⁗′, y su mitad 8″ 17″ 38⁗′ 18⁗′, es (por las razones dadas antes) el movimiento medio horario de los nodos en dicha órbita. Y el año sidéreo entero sea de 20.º 11′ 46″. Puesto que en semejante órbita los nodos de la Luna completarían cada año los 20.º 11′ 46″ hacia adelante, y si hubiese varias lunas, los movimientos de los nodos de cada una (por el Corolario 16 de la Proposición lxvi del Libro i) serían como los tiempos periódicos, si la Luna en el tiempo de un día sidéreo girase junto a la superficie de la Tierra, el movimiento anual de los nodos sería a 20.º 11′ 46″ como el día sidéreo de 23h. 56′ al tiempo periódico de la Luna de 27d. 7h. 43′; es decir, como 1436 a 39 343. E igual es la razón de los nodos de un anillo de lunas rodeando a la Tierra, tanto si dichas lunas no se tocan, como si fueran líquidas y formasen un anillo continuo, o si ese anillo fuese rígido y se hiciese inflexible.

Supongamos, pues, que dicho anillo es igual, en cantidad de materia, a toda la Tierra P*ap*AP*ep*E que sobresale de la esfera P*ape*, (vde. fig. del Lema ii); y puesto que dicha esfera es a la dicha Tierra que sobresale como aC^2 a $AC^2 - aC^2$, esto es (al ser PC semidiámetro menor de la Tierra o también *a*C al semidiámetro mayor AC como 229 a 230) como 52 441 a 459; si este anillo ciñese a la Tierra por el ecuador y ambos girasen a la vez en torno al diámetro del anillo, el movimiento del anillo sería al movimiento de la esfera interior (por el Lema iii de éste) como 459 a 52 441 y 1 000 000 a 925 275 conjuntamente, es decir, como 4590 a 485 223; y por tanto el movimiento del anillo sería a la suma del movimiento del anillo y del movimiento del globo como 4590 a 489 813. De donde, si el anillo se adhiriese al globo y comunicase a éste el movimiento con el cual regresan sus nodos o puntos equinocciales, el movimiento restante en el anillo será a su movimiento anterior como 4590 a 489 813; y por lo mismo, el movimiento de los puntos equinocciales disminuirá en la misma razón. Por consiguiente, el movimiento anual de los puntos equinocciales del cuerpo compuesto del anillo y del globo será al movimiento de 20.º 11′ 46″ como 1436 a 39 343 y 4590 a 489 813 conjuntamente, es decir, como 100 a 292 369. Pero las fuerzas por las cuales retroceden los nodos de las lunas (como expuse antes) y por tanto

también retroceden los puntos equinocciales (esto es, las fuerzas 3IT, en la figura de la Proposición XXX) son en cada partícula como las distancias de cada una de esas partículas al plano QR, y con esas fuerzas huyen las partículas del plano; y por ello (por el Lema II), si la materia del anillo se distribuyera por toda la superficie del globo al modo de la figura P*ap*AP*ep*E para dar lugar a la parte sobresaliente de la Tierra, toda la fuerza y eficacia de todas las partículas para hacer girar a la Tierra en torno de un diámetro ecuatorial, y por tanto para mover los puntos equinocciales, resultará menor que antes, y ello en razón de 2 a 5. Y por eso el regreso anual de los equinoccios sería ahora respecto a 20.º 11′ 46″, como 10 a 73 092; y por tanto resultaría de 9″ 56‴ 50‴.

Además, este movimiento, debido a la inclinación del plano ecuatorial hacia el plano de la eclíptica, debe ser disminuido, y esto en la razón del seno de 91 706 (que es el seno del complemento de 23½ grados) al radio 100 000. Con lo cual el movimiento resultará ahora de 9″ 7‴ 20⁗. Esta es la precesión anual de los equinoccios debida a la fuerza del Sol.

Pero la fuerza de la Luna para mover el mar era a la del Sol como 4,4815 a 1 aproximadamente. Y la fuerza de la Luna para mover los equinoccios es a la del Sol en la misma proporción. Y de aquí resulta que la precesión anual de los equinoccios debida a la fuerza de la Luna sea de 40″ 52‴ 52⁗, y la precesión de los equinoccios total anual debida a ambas fuerzas es de 50″ 00‴ 12⁗. Este movimiento concuerda con los fenómenos. Pues la precesión de los equinoccios según las observaciones astronómicas es de cincuenta segundos al año más o menos.

Si la altura de la Tierra en el ecuador superase a la altura en los polos en más de 17⅙ millas, la materia terrestre sería más rara en la circunferencia que en el centro, y la precesión de los equinoccios debería aumentar debido a dicha altura, mientras debería disminuir debido a la rareza.

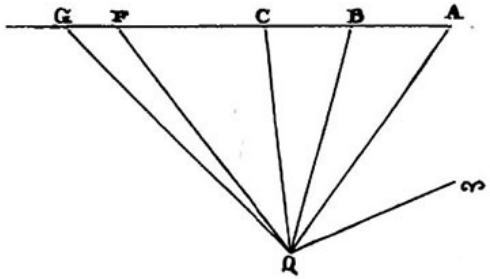
Ya hemos descrito el sistema del Sol, la Tierra, la Luna y los planetas; resta sólo añadir algunas cosas sobre los cometas.

Lema IV

Los cometas están más arriba de la Luna y circulan por las regiones planetarias.

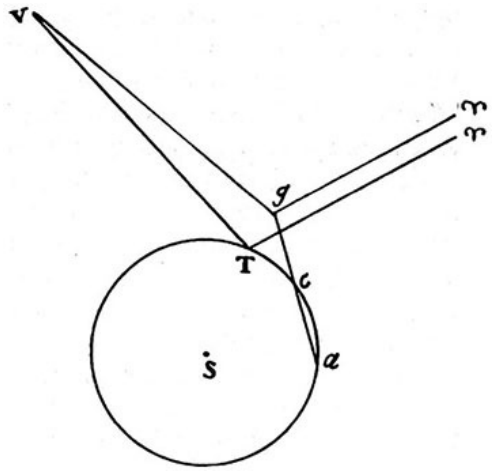
Al igual que la falta de paralaje diurna alejó a los cometas fuera y más allá de las regiones sublunares, de igual modo por la paralaje anual se llega a la convicción de su descenso a las regiones planetarias. Pues los cometas, que avanzan según el orden de los signos zodiacales, son todos al final de su aparición o muy lentos o retrógrados cuando la Tierra se halla entre ellos y el Sol; y mucho más veloces si la Tierra se halla en oposición. Y por el contrario, los que caminan contra el orden de los signos son más veloces al final de su aparición si la Tierra se halla entre ellos y El Sol; pero

son mucho más lentos y hasta retrógrados si la Tierra se halla en las partes opuestas. Esto sucede sobre todo por el diferente movimiento de la Tierra según su sitio, como ocurre con los planetas, que según el movimiento de la Tierra vaya en la misma o contraria dirección, ahora parecen retrógrados, después muy lentos y otras veces muy rápidos. Si la Tierra marcha en la misma dirección del cometa y es transportado por un movimiento angular en torno al Sol tan veloz que hace que una recta que pase por la Tierra y por el cometa siempre converja en las regiones ulteriores al cometa, éste, visto desde la Tierra, aparecerá siempre retrógrado por su movimiento más lento; pero si la Tierra es transportada con movimiento más lento el movimiento del cometa (descontado el movimiento de la Tierra) resulta por lo menos más lento. Pero cuando la Tierra camina en dirección contraria, el cometa parece por ello más veloz. La distancia del cometa, en cambio, se obtiene a partir de la aceleración, de la retardación o del movimiento retrógrado del modo siguiente.



Sean ♈ QA, ♈ QB, ♈ QC, tres longitudes del cometa observadas al principio del movimiento y sea ♈ QF la última longitud observada, cuando el cometa deja de verse. Trácese la recta ABC, cuyas partes AB, BC están situadas entre las rectas QA y QB, QB y QC, y son además entre sí como los tiempos comprendidos entre las tres primeras observaciones. Prolónguese AC hasta G, de modo que AG sea a AB como el tiempo comprendido entre la primera y la última observación e$ al tiempo entre la primera y la segunda, y únanse QG. Y si el cometa se moviese en línea recta y la Tierra o estuviese en reposo o avanzase con movimiento uniforme también en línea recta, el ángulo ♈ QC sería la longitud del cometa en el momento de la última observación. Pues el ángulo FQG, que es la diferencia de longitudes, se debe a la desigualdad de los movimientos del cometa y de la Tierra. Pero este ángulo, si la Tierra y el cometa se mueven en sentidos contrarios, se suma al ángulo ♈ QG, y de este modo convierte al movimiento aparente del cometa en más rápido; pero si el cometa marcha hacia la misma dirección que la Tierra, se resta de dicho ángulo y convierte con ello al movimiento del cometa o en más lento o, incluso, en retrógrado, como ya he explicado. Surge, pues, este ángulo principalmente del movimiento de la Tierra, y por ello ha de considerarse con toda razón como la paralaje del cometa, dejando de lado tal vez algún incremento o decremento del mismo que pueda deberse

a algún movimiento desigual del cometa en su propia órbita. Mas la distancia del cometa se infiere de dicha paralaje como sigue. Represente S al Sol, *ac*T a la eclíptica, *a* el lugar de la Tierra en la primera observación, *c* el lugar de la Tierra en la observación tercera, T el lugar de la Tierra en la observación última, y T♈ una línea recta trazada al principio de Aries. Tómese al ángulo ♈ TV igual al ángulo ♈ QF, es decir, igual a la longitud del cometa cuando la Tierra se halla en T. Únanse *ac* y prolónguese hasta *g* de modo que *ag* sea a *ac* como AG a AC, y *g* será el lugar que ocuparía la Tierra en el momento de la última observación con un movimiento uniformemente continuo por la recta *ac*. Por ello, si se traza *g*♈ paralela a T♈ y se toma al ángulo ♈*g*V igual al ángulo ♈ QG, dicho ángulo ♈ *g*V será igual a la longitud del cometa visto desde el punto *g*; y el ángulo TV*g* será la paralaje debida a la traslación de la Tierra desde el punto *g* hasta el punto T: y por consiguiente, V será el lugar del cometa en el plano de la eclíptica. Pero este punto V suele hallarse más bajo que la órbita de *Júpiter*.



Y lo mismo se infiere de la curvatura de la trayectoria de los cometas. Estos cuerpos se desplazan por círculos máximos prácticamente, cuando su movimiento es más rápido; pero al final de su curso, cuando la parte de movimiento aparente debida a la paralaje está en mayor proporción al movimiento aparente, se separan de dichos círculos, y siempre que la Tierra se mueve hacia una dirección, ellos se desvían hacia otra. Semejante deflexión se debe sobre todo a la paralaje, en cuanto que responde al movimiento de la Tierra; y su notable magnitud, según mis cálculos, sitúa a los cometas en el momento de desaparecer bastante por debajo de Júpiter. De donde se sigue que en los perigeos y perihelios, cuando están más cerca, descienden muchas veces más abajo de la órbita de Marte y de los planetas inferiores.

También se confirma la proximidad de los cometas por la luz de sus cabezas. Pues el brillo de un cuerpo celeste iluminado por el Sol y que se va alejando hacia lo

remoto del cielo disminuye en razón de la cuarta potencia de la distancia: a saber, en razón cuadrada por el aumento de la distancia del cuerpo al Sol y en otra razón cuadrada por la disminución del diámetro aparente. De donde, si se da la cantidad de la luz y el diámetro aparente de un cometa se tendrá dada también la distancia, diciendo que su distancia será a la distancia de un planeta, directamente como la razón de un diámetro a otro e inversamente como la razón de la raíz cuadrada de la luz de uno a la raíz cuadrada de la luz del otro. Así, el diámetro mínimo de la cabellera del cometa de 1682, observado por *Flamsteed* con un telescopio de 16 pies y medido con micrómetro era de 2′ 0″; pero el núcleo o estrella central de la cabeza apenas ocupaba una décima parte de esa latitud, de suerte que su diámetro sólo era de 11″ ó 12″. Pero en luz y claridad de su cabeza superaba a la cabeza del cometa de 1680, y competía con las estrellas de primera y segunda magnitud. Supongamos que Saturno, con su anillo, fuese cuatro veces más brillante: y puesto que la luz del anillo casi iguala a la luz del globo interior y el diámetro aparente del globo viene a ser de 21″ aproximadamente, la luz, por tanto, del globo y del anillo juntas igualarán a la luz de un globo que fuese de 30″ de diámetro: la distancia del cometa será a la distancia de Saturno como 1 a √4 inversamente y como 12″ a 30″ directamente, es decir, como 24 a 30 o como 4 a 5. Además, el cometa del año 1665, como lo atestigua *Hewelcke*, superaba en brillo a casi todas las estrellas fijas, incluso al mismo Saturno, quizás porque su color era mucho más vivo. Resultaba este cometa mucho más luminoso que el aparecido a finales del año anterior y que era comparado con las estrellas de primera magnitud. La anchura de la cabeza era casi de 6′, pero el núcleo, comparado con los planetas mediante el telescopio era claramente menor que Júpiter y ora se juzgaba menor que el cuerpo interior de Saturno, ora igual a dicho cuerpo. Por consiguiente, al ser el diámetro de la cabeza de los cometas muy raramente superior a 8′ ó 12′ y el diámetro del núcleo o estrella central ser sólo una décima o decimoquinta parte del diámetro de la cabeza, es evidente que estas estrellas son a lo sumo de la misma magnitud aparente que los planetas. Por lo cual, dado que su luz se pueda comparar a veces con la de Saturno, y hasta alguna vez superarla, está claro que todos los cometas en los perihelios se han de situar más abajo de Saturno o no muy por encima. Y se equivocan por completo quienes los alejan hasta las regiones de las fijas: pues con semejante solución no deberían ser iluminados por nuestro Sol más de lo que nuestros planetas lo son por las estrellas fijas.

Hemos tratado de estas cosas sin tener en cuenta el oscurecimiento de los cometas debido a ese humo abundante y espeso que rodea a la cabeza, apareciendo ésta siempre como velada por una nube. Pues cuanto más oscuro se torna el cuerpo debido a dicha nube, tanto más necesario es que esté más cerca del Sol para que la cantidad de luz que refleja emule a los planetas. Por ello es muy verosímil que los cometas desciendan bastante por debajo de la órbita de Saturno, como hemos observado por la paralaje. Y esto mismo se confirma magníficamente por las colas. Estas, o se originan por la reflexión producida por el humo esparcido por el éter, o se originan de la luz de

la cabeza. En el primer caso hay que disminuir la distancia de los cometas, para que el humo que sale continuamente de la cabeza no se halle propagado por espacios demasiado amplios y con una velocidad increíble. En el segundo caso hay que atribuir toda la luz, tanto de la cola como de la cabellera, al núcleo de la cabeza. Ahora bien, si imaginamos toda esa luz reunida y encerrada dentro del disco del núcleo con certeza, siempre que emita una cola máxima y luminosa, superará con mucho al brillo de Júpiter. Pues, con un diámetro aparente menor emite más luz, y por ende será mucho más iluminado por el Sol, con lo que estará también mucho más cerca del mismo Sol. E incluso, por este argumento, deben situarse bajo la órbita de Venus las cabezas que se ocultan bajo el Sol y que, a veces, emiten colas enormes y muy brillantes a modo de tizones encendidos. Pues toda esa luz, si se supone reunida en la estrella, superaría a la propia Venus, por no decir a muchas Venus.

Y lo mismo, finalmente, se sigue de la luz de las cabezas que crece al alejarse los cometas desde la Tierra hacia el Sol y decrece al apartarse del Sol hacia la Tierra. De modo que el último cometa del año 1665 (según las observaciones de *Hewelcke*) desde el momento de su aparición perdía continuamente movimiento aparente y por ello ya había pasado de su perigeo; pero el brillo de su cabeza crecía no menos de día en día, hasta que el cometa dejó de verse cubierto por los rayos solares. El cometa de 1683 (observado por el mismo *Hewelcke*) a fines del mes de julio, al aparecer por primera vez, se movía muy lentamente, recorriendo en su órbita unos 40 ó 45 minutos al día. Desde esas fechas su movimiento diario aumentó continuamente hasta el 4 de septiembre, en que alcanzó casi 5 grados. Luego durante todo este tiempo el cometa se acercaba a la Tierra. Cosa que también se infiere del diámetro de la cabeza medido con micrómetro: toda vez que *Hewelcke* lo fijó el 6 de agosto en tan sólo 6′ 5″, incluida la cabellera, mientras que en 2 de septiembre era de 9′ 7″. Por consiguiente, la cabeza al principio aparecía mucho menor que al final del movimiento, mientras que sin embargo al principio, en las proximidades del Sol, era mucho más luminosa que hacia el final, como dice el mismo *Hewelcke*. Por lo tanto, durante todo este tiempo, debido a su alejamiento del Sol, iba decreciendo en luminosidad, pese a su acercamiento hacia la Tierra. El cometa de 1618, hacia el mes de diciembre, y éste de 1680, hacia finales del mismo mes, se movían muy rápidamente, y por lo mismo estaban entonces en los perigeos. Pero el máximo resplandor de sus cabezas aconteció casi dos semanas antes, cuando acababan de salir de los rayos solares; mientras que el máximo resplandor de las colas un poco antes, en la mayor proximidad al Sol. La cabeza del cometa primero, según las observaciones de *Cysat* de 1 de diciembre (ahora ya en el perigeo) con poca menos magnitud, había decrecido mucho en brillo o nitidez luminosa. El 7 de enero, *Kepler*, inseguro ya de la cabeza del cometa, puso fin a la observación. El 12 de diciembre vio y observó *Flamsteed* la cabeza del segundo cometa a una distancia de nueve grados del Sol; cosa apenas concebible para una estrella de tercera magnitud. El 15 y 17 se presentó como una estrella de tercera magnitud, aunque disminuido su brillo por las nubes que brillaban

junto al Sol poniente. El 26 de diciembre, moviéndose muy rápidamente, y casi hallándose en el perigeo, brillaba menos que la estrella de la boca de Pegaso, de tercera magnitud. El 3 de enero parecía una estrella de cuarta magnitud, el 9 de enero, de quinta magnitud, el 13 de enero desapareció por el resplandor de la Luna creciente. El 25 de enero apenas igualaba a las estrellas de séptima magnitud. Si a un lado y otro del perigeo se toman tiempos iguales, las cabezas, que en dichos tiempos se hallan en regiones distantes, deberían brillar igual por sus distancias iguales de la Tierra, y brillaban sin embargo máximamente hacia el lado del Sol mientras que hacia la otra parte del perigeo se desvanecían. Por tanto, por la gran diferencia de luz en uno y otro lugar, se infiere la gran proximidad del Sol y del cometa en el primer lugar. Pues la luz de los cometas suele ser regular, y parece máxima cuando las cabezas se mueven muy velozmente y se hallan, por tanto, en los perigeos; salvo en la medida en que es mayor en las cercanías del Sol.

COROLARIO 1. Por consiguiente, los cometas brillan con la luz que reflejan del Sol.

COROLARIO 2. También de lo dicho se sigue el motivo por el que los cometas son tan frecuentes en la región del Sol. Si se viesen en las regiones de más allá de Saturno, deberían aparecer frecuentemente en las regiones opuestas al Sol. Pues los que anduviesen por estos lugares serían más próximas a la Tierra y, en cambio, el Sol ocultaría a los demás interponiéndose ante ellos. Pero al repasar la historia de los cometas encuentro que se han visto cuatro o cinco veces más hacia el lado del hemisferio solar que hacia el opuesto, además de otros, sin duda no pocos, obscurecidos por la luz del Sol. Pero además, al descender hacia nuestras regiones ni emiten colas, ni son iluminados por el Sol lo bastante para que puedan ser observados a simple vista antes de que se hallen más cerca que el mismo Júpiter. Pero la mayor parte del espacio descrito con tan corto radio en torno al Sol cae del lado de la Tierra que mira hacia el Sol, y los cometas, en dicha mayor parte y como mucho más próximos al Sol, suelen ser iluminados en mayor grado.

COROLARIO 3. Con ello también es evidente que los cielos se hallan desprovistos de resistencia. Pues los cometas, que siguen trayectorias oblicuas y a veces contrarias al curso de los planetas, se mueven libremente en todas direcciones, y conservan sus movimientos por tiempo muy largo, incluso contra el curso de los planetas. Si no me engaño, son una especie de planetas y retornan en órbita con un movimiento perpetuo. Pues la afirmación de algunos autores de que se trata de meteoros, basando su argumento en los cambios continuos de las cabezas, parece carecer de fundamento. Las cabezas de los cometas están rodeadas de enormes atmósferas, y las atmósferas más bajas deben ser más densas. De donde, las nubes no son los mismos cuerpos de los cometas en los que se observan dichos cambios. Así también la Tierra, vista desde los planetas, brillará sin duda con la luz de sus nubes, mientras que el cuerpo sólido apenas será visible bajo las nubes. También los cinturones de Júpiter se forman en las nubes de aquel planeta y cambian entre sí de lugar, dejando apenas verse el cuerpo

sólido de Júpiter a través de ellas. Mucho más los cuerpos de los cometas habrán de esconderse bajo sus atmósferas más grandes y más densas.

Proposición xl. Teorema xx[16]

Los cometas se mueven en secciones cónicas que tienen los focos en el centro del Sol, y con radios trazados al Sol describen áreas proporcionales a los tiempos.

Es evidente por el Corolario 1 de la Proposición xiii del Libro i comparado con las Proposiciones viii, xii y xiii del Libro iii.

Corolario 1. De aquí que, si los cometas giran en órbita, las órbitas serán elipses, y los tiempos periódicos serán a los tiempos periódicos de los planetas en razón de la potencia ³/₂ de los ejes principales. Y por ello los cometas que hacen la mayor parte de su recorrido en regiones ultraplanetarias y describen por ello órbitas con ejes mayores giran con mayor lentitud. De suerte que si el eje de la órbita de un cometa es cuatro veces mayor que el eje de la órbita de Saturno, el tiempo de revolución del cometa será al tiempo de revolución de Saturno, es decir, a 30 años, como 4 √4 (o sea, 8) a 1, y por tanto será de 240 años.

Corolario 2. Pero las órbitas serán tan próximas a parábolas que en su lugar pueden utilizarse parábolas sin errores sensibles.

Corolario 3. Y, por lo mismo (por el Corolario 7 de la Proposición xvi del Libro i), la velocidad de todo cometa será a la velocidad de un planeta cualquiera que gire en círculo en torno al Sol como la raíz cuadrada de la doble distancia del planeta al centro del Sol a la distancia del cometa al centro del Sol muy aproximadamente. Supongamos que el radio del Orbe Máximo, o sea, que el semidiámetro máximo de la elipse en que gira la Tierra tiene 100 000 000 partes: la Tierra con su movimiento diario medio describirá 1 720 212 de estas partes, y con su movimiento horario 71 675½ partes. Por tanto, un cometa a la misma distancia que la Tierra del Sol, con una velocidad que fuese a la velocidad de la Tierra como √2 a 1, describirá con su movimiento diario 2 432 747 partes, y con su movimiento horario 101 364¼ partes. Pero a distancias mayores o menores, tanto el movimiento diario como el horario será respecto a este movimiento diario y horario inversamente como la raíz cuadrada de las distancias, y por tanto está dado.

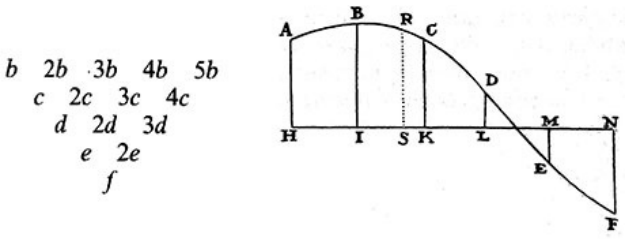
Corolario 4. Por consiguiente, si el «latus rectum» de la parábola es cuatro veces mayor que el radio del Orbe Máximo, y el cuadrado de dicho radio se supone que es de 100 000 000 partes, el área que el cometa describe cada día con un radio trazado al Sol será de 1 216 373½ partes, y en cada hora dicha área será de 5068½ partes. Pero si el «latus rectum» fuera mayor o menor en una razón cualquiera, el área diaria y horaria será mayor o menor en razón de la raíz cuadrada de esa razón.

Lema v

Hallar una línea curva de especie parabólica que pase por cuantos puntos dados se quiera.

Sean tales puntos A, B, C, D, E, F, etc., y desde ellos hasta una recta cualquiera de posición dada HN háganse descender otras tantas perpendiculares, AH, BI, CK, DL, EM, FN.

Caso 1. Si los intervalos HI, IK, KL, etc., entre los puntos H, I, K, L, M, N, son iguales, tomemos las primeras diferencias de las perpendiculares AH, BI, CK, etc., como *b*, 2*fe*, 3*b*, 4*b*, 5*b*, etc., las segundas como *c*, 2*c*, 3*c*, 4*c*, etc., las terceras como *d*, 2*d*, 3*d*, etc., esto es, de modo que AH - BI = b, BI - CK = 2*b*, CK - DL = 3*b*, DL + EM = 4*b*, -EM + FN = 5*b*, etc., entonces, *b* - 2*b* = *c*, etc., hasta llegar así a la última diferencia que aquí es *f*. Elevando después una perpendicular cualquiera RS que sería una ordenada aplicada a la curva que se busca: con el fin de hallar su longitud supongamos que los intervalos HI, IK, KL, LM, etc., son unidades, y escribe AH = a, - HS = *p*, ½*p* x - IS = *q*, ⅓*q* x +SK = *r*, ¼*r* x +SL = *s*, ⅕*s* x + SM = *t*, siguiendo así hasta la penúltima perpendicular que es ME y anteponiendo los signos negativos a los términos HS, IS, etc., situados hacia el lado A del punto S, y los signos positivos a los términos SK, SL, etc., que están situados hacia el otro lado del punto S. Y si los signos están bien respetados, será RS = *a* + *bp* + *cq* + *dr* + *es* + *ft*, etcétera.



Caso 2. Pero si los intervalos HI, IK, etc., entre los puntos H, I, K, L, etc., fuesen desiguales, toma *a b*, 2*b*, 3*b*, 4*b*, 5*b*, como primeras diferencias de las perpendiculares AH, BI, CK, etc., separadas por los intervalos entre perpendiculares; y *a c*, 2*c*, 3*c*, 4*c*, etc., como segundas diferencias separadas por cada dos intervalos, y *a d*, 2*d*, 3*d*, etc., como terceras diferencias separadas por intervalos de tres en tres, *a e*, 2*e*, etc., como cuartas diferencias divididas por los intervalos de cuatro en cuatro, y así sucesivamente; es decir, de modo que $b = \frac{AH - BI}{HI}$, $2b = \frac{BI - CK}{IK}$ $3b = \frac{CK - DL}{KL}$, etc., después $c = \frac{b - 2b}{HK}$, $2c = \frac{2b - 3b}{IL}$ 3c = $c = \frac{3b - 4b}{KM}$, etc., y luego d= $\frac{c - 2c}{HL}$, $2d =$

$\frac{2c - 3c}{IM}$ etc. Halladas las diferencias, escribe AH = *a*, -HS - *p*, *p* x -IS - *q*, *q* x +SK = *r*, *r* x +SL = *s*, *s* x +SM = *t*; siguiendo así hasta la penúltima perpendicular ME, la ordenada RS será igual a la suma $a + bp + cq + dr + es + ft$, etc.

COROLARIO. De aquí que las áreas de todas las curvas puedan calcularse muy aproximadamente. Pues si se hallan unos puntos de la curva a cuadrar, y se imagina que por esos puntos pasa una parábola, el área de esa parábola será la misma que la de la curva a cuadrar muy aproximadamente. Y una parábola se puede cuadrar siempre geométricamente por métodos bien conocidos.

LEMA VI

A partir de algunos lugares observados de un cometa, hallar su lugar en un momento cualquiera intermedio dado.

Sean HI, IK, KL, LM los tiempos entre observaciones (en la figura anterior), sean HA, IB, KC, LD, ME cinco longitudes observadas del cometa, HS el tiempo dado entre la observación primera y la longitud buscada. Y si por los puntos A, B, C, D, E, se supone trazada la curva regular ABCDE, y por el Lema anterior se halla su ordenada RS, será RS la longitud buscada.

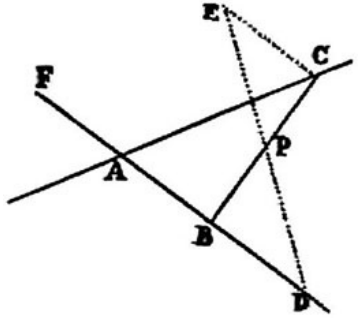
Con el mismo método sobre la base de cinco latitudes, observadas se halla la latitud en un momento dado.

Si las diferencias de longitudes observadas son pequeñas, tales como 4 ó 5 grados solamente, bastarán tres o cuatro observaciones para hallar las nuevas longitudes o latitudes. Pero si las diferencias son mayores, por ejemplo de 10 ó 20 grados, habrá que disponer de cinco observaciones.

LEMA VII

Por un punto dado P trazar una recta BC cuyas partes PB, PC, cortadas por dos rectas AB, AC de posición dada, se hallen entre sí en una razón dada.

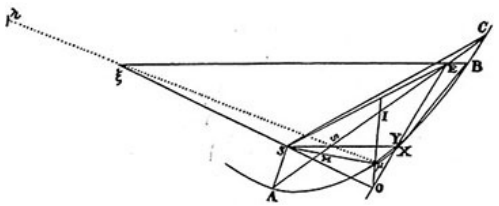
Desde dicho punto P hasta AB una cualquiera de las dos rectas trácese otra recta cualquiera PD, y prolónguese ésta hacia la otra recta AC hasta E, de modo que PE sea a PD en la dicha razón dada. Sea EC paralela a la recta AD; y si se traza CPB, entonces PC será a PB como PE a PD. Q. E. F.



LEMA VIII[17]

Sea ABC una parábola con foco en S. Con la cuerda AC bisecada en I sepárese el segmento ABCI, cuyo diámetro es Iμ y cuyo vértice es μ. Sobre la prolongación de Iμ tómese μO igual a la mitad de la propia Iμ. Únase OS y prolónguese ésta hasta ξ de modo que Sξ sea igual a 2SO. Y si el cometa B se mueve por el arco CBA, y se traza ξB secante de AC en E: digo que el punto E separará de la cuerda AC el segmento AE muy aproximadamente proporcional al tiempo.

Pues únase EO cortando al arco parabólico ABC en Y, y trácese μX tangente a dicho arco en el vértice μ y que se encuentre con la línea EO en X; y el área curvilínea AEXμA será al área curvilínea ACYμA como AE a AC. Por lo tanto, dado que el triángulo ASE es al triángulo ASC en esa misma razón, el área total ASEXμA será al área total ASCYμA como AE a AC. Pero como ξO es a SO como 3 a 1, y EO a XO en la misma razón, SX será paralela a EB, y por tanto, si se une BX, el triángulo SEB será igual al triángulo XEB. Por lo cual, si al área ASEXμA se le añade el triángulo EXB, y de la suma se resta el triángulo SEB, el área ASBXμA permanecerá igual al área ASEXμA, y por lo mismo seguirá siendo al área ASCYμA como AE a AC. Pero el área ASBYμA es muy aproximadamente igual al área ASBXμA, y el área ASBYμA es al área ASCYμA como el tiempo del arco descrito AB al tiempo del arco total descrito AC. Y por tanto, AE es a AC en la razón de los tiempos muy aproximadamente. Q. E. D.



COROLARIO. Cuando el punto B cae en el vértice *p* de la parábola, AE está a AC exactamente en la razón de los tiempos.

ESCOLIO

Si unimos $\mu\xi$ cortando a AC en δ, y sobre ella se toma ξn de tal modo que sea a μB como 27MI a 16Mμ, trazando B*n*, ésta cortará a la cuerda AC en razón de los tiempos más exactamente que antes. Pero el punto *n* debe estar más allá del punto ξ, si el punto B dista más del vértice principal de la parábola que el punto μ; y debe estar más acá si dista menos de dicho vértice.

LEMA IX

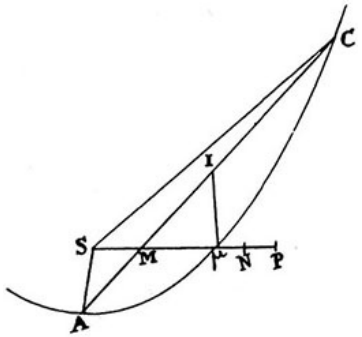
Las rectas Iμ y μM y la longitud $\frac{AIC}{4S\mu}$ *son iguales entre sí.*

Pues 4Sμ es el «latus rectum» de la parábola correspondiente al vértice μ.

LEMA X

Si Sμ se prolonga hasta N y P de modo que μN sea la tercera parte de μI, y SP sea a SN como SN a Sμ, un cometa, en el tiempo en que describe el arco AμC y desplazándose con una velocidad que sea siempre igual a la que posee a la altura SP, describirá una longitud igual a la cuerda AC.

Pues si un cometa, en el mencionado tiempo y con la velocidad que tiene en μ, avanza uniformemente sobre una recta tangente a la parábola en μ, el área que describiría con un radio trazado al punto S sería igual al área parabólica ASCμ. Por consiguiente, el contenido bajo la longitud descrita en la tangente y bajo la longitud Sμ sería al contenido bajo las longitudes AC y SM, como el área ASCμ al triángulo ASC, es decir, como SN a SM. Por lo cual, AC es a la longitud descrita en la tangente como Sμ a SN. Pero al ser la velocidad del cometa a la altura SP (por el Corolario 6 de la Proposición XVI del Libro I) respecto a su velocidad a la altura Sμ inversamente como la raíz cuadrada de SP a Sμ, esto es, como la razón de Sμ a SN, la longitud descrita con esa velocidad y en los mismos tiempos será a la longitud descrita en la tangente como Sμ a SN. Por consiguiente, dado que AC y la longitud descrita con esta nueva velocidad se hallan en la misma razón respecto a la longitud descrita en la tangente, son iguales entre sí. Q. E. D.

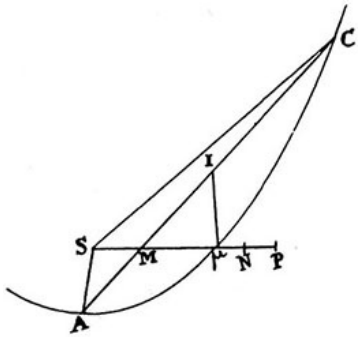


COROLARIO. Luego un cometa con la velocidad que tiene a la altura Sμ + $\frac{2}{3}$Iμ recorrerá en el mismo tiempo la cuerda AC, muy aproximadamente.

LEMA XI

Si un cometa privado de todo movimiento se dejare caer hacia el Sol desde la altura SN, o Sμ + $\frac{1}{3}$Iμ, y fuese urgido hacia el Sol siempre uniformemente por la misma fuerza que le urgía al principio; el tal cometa, en la mitad de tiempo en el que describía en su órbita el arco AC, describirá ahora en su descenso un espacio igual a la longitud Iμ.

Pues el cometa, en el tiempo en que describa el arco parabólico AC, en ese mismo tiempo y con la misma velocidad que tiene a la altura SP (por el Lema anterior) describirá la cuerda AC, y por tanto (por el Corolario 7 de la Proposición XVI del Libro I) en el mismo tiempo, girando con la fuerza de su gravedad en un círculo, cuyo diámetro fuese SP, describiría un arco cuya longitud sería a la cuerda AC del arco parabólico, como la raíz cuadrada de uno a la de dos. Y por consiguiente, con el peso que tiene hacia el Sol a la altura SO y cayendo hacia el Sol desde esa altura, describirá en la mitad de dicho tiempo (por el Corolario 9 de la Proposición IV del Libro I) un espacio igual al cuadrado de la semicuerda dividido por el cuádruplo de la altura SP, es decir, el espacio $\frac{AI^2}{4SP}$. Ahora bien, como el peso del cometa hacia el Sol a la altura SN es al peso del mismo hacia el Sol a la altura SP como SP a Sμ, el cometa con el peso que tiene a la altura SN, en el mismo tiempo, cayendo hacia el Sol, describirá el espacio $\frac{AI^2}{4S\mu}$, es decir, un espacio igual a la longitud Iμ o Mμ. Q. E. D.



Proposición xli. Problema xxi

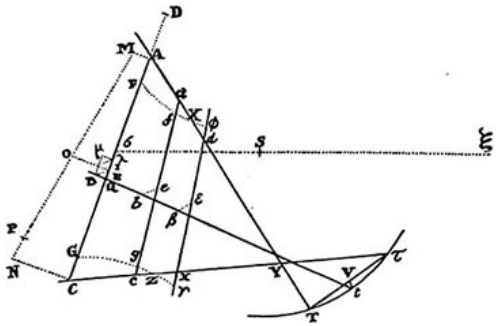
Determinar la trayectoria de un cometa que se mueve en una parábola a partir de tres observaciones dadas.

He abordado este problema verdaderamente difícil de muchas formas y propuse en el Libro Primero algunos problemas relativos a su resolución. Más tarde se me ocurrió la siguiente solución un poco más sencilla.

Elíjanse tres observaciones separadas entre sí por espacios de tiempo aproximadamente iguales. Pero de modo que el intervalo de tiempo en que el cometa se mueve más lentamente sea algo mayor que el otro, quizá de modo que la diferencia entre los tiempos sea a la suma de los mismos como la suma de los dichos tiempos a seiscientos días más o menos; o también como si el punto E (en la Figura del Lema viii) cayese sobre el punto M, y desde allí se desplazase hacia I más bien que hacia A. Si tales observaciones no están disponibles, hay que hallar un nuevo lugar del cometa por medio del Lema vi.

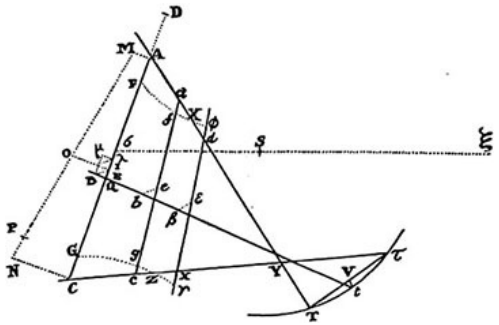
Representen S al Sol, T, *t*, *τ* tres lugares de la Tierra sobre el orbe máximo, TA, *t*B, *τ*C, tres longitudes observadas del cometa, V el tiempo entre la primera y la segunda observación, W el tiempo entre la segunda y la tercera observación, X la longitud que el cometa puede alcanzar a describir en todo el tiempo dicho con la velocidad que tiene a la distancia media de la Tierra al Sol, y que tendrá que hallarse (por el Corolario 3 de la Proposición xl del Libro iii), y *t*V una perpendicular sobre la cuerda T*τ*. Sobre la longitud media observada *t*B tómese un lugar cualquiera B como lugar del cometa en el plano de la eclíptica y desde él trácese hasta el Sol S la línea BE que sea a la sagita *t*V como el producto de SB por St^2 es al cubo de la hipotenusa del triángulo rectángulo cuyos lados son SB y la tangente de la latitud del cometa en la segunda observación a un radio *t*B. Trácese por el punto E (por el Lema vii) la recta AEC, cuyas partes AE y EC terminan en las rectas TA y *τ*C y deben ser entre sí como los tiempos V y W: y entonces A y C serán los lugares del cometa en la primera

y tercera observación sobre el plano de la eclíptica muy aproximadamente, siempre que B sea su lugar de la segunda observación asumido correctamente.



Sobre AC, bisecada en I, elévese la perpendicular I*i*. Trácese por el punto B la línea invisible Bz paralela a la propia AC. Únase la línea invisible Sz secante de AC en *λ* y complétese el paralelogramo *i*Iλ*μ*. Tómese I*σ* igual a 3Iλ, y por el Sol S trácese la línea también oculta *σξ* igual a 3S*σ* + 3*iλ* Y ahora, borrando las letras A, E, C, I, tracemos desde el punto B hacia el punto *ξ* una nueva línea invisible BE, que sea a la primera BE en razón cuadrada de la distancia BS a la cantidad S*μ* + ⅓*iλ*. De nuevo trácese por el punto E la recta AEC en la misma proporción que antes, o sea, de modo que sus partes AE y EC estén entre sí como los tiempos V y W entre observaciones. Y A y C serán los lugares del cometa más exactamente.

Sobre AC, bisecada en I, elévense las perpendiculares AM, CN, IO; de las cuales AM y CN sean tangentes de las latitudes en las observaciones primera y tercera con radios TA y *τ*C. Únase MN cortando a IO en O. Constrúyase el rectángulo *i*Iλ*μ*, como antes. Sobre la prolongación de IA tómese ID igual a S*μ* + ⅔ *iλ*. Después, sobre MN hacia N tómese MP, que sea respecto a la longitud X hallada antes, como la raíz cuadrada de la distancia media de la Tierra al Sol (o semidiámetro del Orbe Magno) a la distancia OD. Si el punto P cae en el punto N, serán A, B, C, los tres lugares del cometa por los que debe circular su órbita en el plano de la eclíptica. Pero si el punto P no cae en el punto N, sobre la recta AC tómese CG igual a NP de modo que los puntos G y P caigan hacia el mismo lado de la recta NC.



Con el mismo método con el que se hallaron los puntos E, A, C, G, partiendo de un punto B supuesto, hállense, partiendo de otros puntos supuestos cualesquiera *b* y *β*, unos nuevos puntos *e*, *a*, *c*, *g* y *ε*, *α*, *χ*, *γ*. Después, si por G, *g*, *γ*, se traza la circunferencia del círculo C*gγ*, secante de la recta τC en Z, será Z el lugar del cometa en el plano de la eclíptica. Y si sobre AC, *ac*, *αχ* se toman AF, *af*, *αϕ*, respectivamente iguales a CG, *cg*, *χγ* y por los puntos F, *f*, *ϕ*, se traza la circunferencia del círculo F*fϕ* secante en X de la recta AT, será el punto X otro lugar del cometa en el plano de la eclíptica. En los puntos X y Z elévense tangentes de las latitudes del cometa con radios TX y τZ; y se tendrán dos lugares del cometa en su propia órbita. Finalmente (por la Proposición XIX del Libro I), con foco en S trácese por dichos dos lugares una parábola y ésta será la trayectoria del cometa. Q. E. I.

La demostración de esta construcción se sigue de los Lemas: toda vez que la recta AC es cortada en E en razón de los tiempos, por el Lema VII, como se exige por el Lema VIII: y BE, por el Lema XI, sea la parte de la recta BS o B*ξ* que yace en el plano de la eclíptica entre el arco ABC y la cuerda AEC; y MP (por el Corolario del Lema X) es la longitud de la cuerda del arco que debe describir el cometa en su órbita propia entre la primera y la tercera observación, y que debería ser igual a MN, siempre que B fuere el verdadero lugar del cometa en el plano de la eclíptica.

Por lo demás, no conviene elegir para B, *b*, *β*, unos puntos cualesquiera, sino más bien unos puntos cercanos al verdadero. Si el ángulo AQ*t*, en el cual corta a la recta *t*B la proyección de la órbita sobre el plano de la eclíptica, se conociese aproximadamente, en dicho ángulo hay que trazar la recta invisible AC, la cual ha de ser a ⅘Tτ como la raíz cuadrada de SQ a la de ST. Y trazando la recta SEB, cuya parte EB sea igual a la longitud V*t*, se hallará el punto B que habríamos de manejar la primera vez. Después, borrando la recta AC y trazándola de nuevo según la precedente construcción y además hallada la longitud MP, tómese sobre *t*B el punto *b*, de modo tal que, si TA y τC son secantes entre sí en Y, la distancia Y*b* sea a la distancia YB en una razón compuesta de la razón de MP a MN y de la razón de la raíz cuadrada de SB a la de S*b*. Y con el mismo método habrá de hallarse el tercer punto *β*, y, si es preciso, repetir por tercera vez la operación. Pero con este método,

bastarán dos operaciones como mucho. Pues si la distancia B*b* resultase muy pequeña, una vez hallados los puntos F, *f*, y G, *g*, trazando las rectas F*f* y G*g* cortarán a TA y τC en los puntos buscados X y Z.

Ejemplo[18]

Propóngase el cometa del año 1680. Su movimiento, observado por *Flamsteed* y calculado a partir de las observaciones y corregido por *Halley* a partir de las mismas observaciones, ofrece la siguiente tabla.

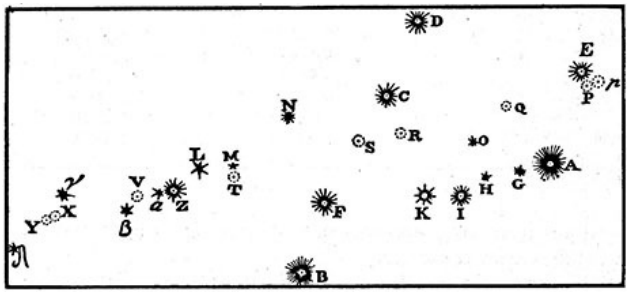
	Tiempo		*Longitud del Sol*	*El cometa*	
	Aparente	*Verdadero*		*Longitud*	*Latitud norte*
	h m	h m s	° ′ ″	° ′ ″	° ′ ″
1680, *dic*. 12	4,46	4,46,00	♑ 1,51,23	♑ 6,32,30	8,28,00
21	6,32½	6,36,59	11,06,44	♒ 5,08,12	21,42,13
24	6,12	6,17,52	14,09,26	18,49,23	25,23, 5
26	5,14	5,20,44	16,09,22	28,24,13	27,00,52
29	7,55	8,03,02	19,19,43	♓ 13,10,41	28,09,58
30	8,02	8,10,26	20,21,09	17,38,20	28,11,53
1681, *ene*. 5	5,51	6,01,38	26,22,18	♈ 8,48,53	26,15, 7
9	6,49	7,00,53	♒ 0,29,02	18,44,04	24,11,56
10	5,54	6,06,10	1,27,43	20,40,50	23,43,52
13	6,56	7,08,55	4,33,20	25,59,48	22,17,28
25	7,44	7,58,42	16,45,36	♉ 9,35,00	17,56,30
30	8,07	8,21,53	21,49,58	13,19,51	16,42,18
feb. 2	6,20	6,34,51	24,46,59	15,13,53	16,04, 1
5	6,50	7,04,41	27,49,51	16,59,51	15,27, 3

A estas observaciones añádanse algunas nuestras.

	Tiempo aparente	*El cometa*	
		Longitud	*Latitud norte*
	h m	° ′ ″	° ′ ″
1681, *feb*. 25	8,30	♉ 26,18,35	12,46,46
27	8,15	27,04,30	12,36,12
mar. 1	11,0	27,52,42	12,23,40
2	8,0	28,12,48	12,19,38
5	11,30	29,18, 0	12,03,16
7	9,30	♊ 0,04,	11,57, 0
9	8,30	0,43, 4	11,45,52

Estas observaciones fueron hechas con un telescopio de siete pies y con micrómetro e hilos colocados en el foco del telescopio; mediante dichos instrumentos determinamos tanto las posiciones de las fijas entre ellas como las posiciones del cometa respecto a las fijas. Represente A la estrella de cuarta magnitud del talón izquierdo de Perseo (*Bayer, o*), B la estrella de tercera magnitud siguiente en el pie izquierdo (*Bayer, ζ*), y C la estrella de sexta magnitud (*Bayer, η*) en el talón del

mismo pie, E, F, G, H, I, K, L, M, N, O, Z, *α*, *β*, *γ*, *δ*, las otras estrellas menores del mismo pie. Y sean *p*, P, Q, R, S, T, V, X, los lugares del cometa en las observaciones descritas más arriba; y siendo la distancia AB de 80$^{7}/_{12}$ partes, AC era de 52¼, BC de 58⅚, AD de 57$^{5}/_{12}$, BD de 82$^{6}/_{11}$, CD de 23⅔ AE de 29$^{4}/_{7}$, CE de 57½ DE de 49$^{11}/_{12}$, AI de 27$^{7}/_{12}$, BI de 52⅙, CI de 36$^{7}/_{12}$, DI de 53$^{5}/_{11}$, AK de 38⅔ BK de 43, CK de 31$^{5}/_{9}$ FK de 29, FB de 23, FC de 36¼, AH de 18$^{6}/_{7}$, DH de 50⅞, BN de 46$^{5}/_{12}$, CN de 31⅓, BL de 45$^{5}/_{12}$, NL de 31$^{5}/_{7}$. HO era a HI como 7 a 6 y prolongada pasaba por entre las estrellas D y E, de suerte que la distancia de la estrella D desde esa recta resultará de ⅙CD. LM era a LN como 2 a 9 y, prolongada, pasaba por la estrella H. Con éstas se determinaban las posiciones de las fijas entre ellas.



Más tarde *Pound*, de entre los nuestros, ha observado de nuevo las posiciones de estas fijas entre ellas y consignó en la siguiente tabla sus longitudes y latitudes.

Estrellas fijas	*Longitudes*	*Latitud norte*
	° ′ ″	° ′ ″
A	26,41,50	12, 8,36
B	28,40,23	12,17,54
C	27,58,30	12,40,25
E	26,27,17	12,52, 7
F	28,28,37	11,52,22
G	26,56,08	12, 4,58
H	27,11,45	12, 2, 1
I	27,25, 2	11,53,26
K	27,42, 7	11,53,26
L	29,33,34	12, 7,48
M	29,18,54	12, 7,20
N	28,48,29	12,31, 9
Z	29,44,48	11,57,13
α	29,52, 3	11,55,48
β	0, 8,23	11,48,56
γ	0,40,10	11,55,18
δ	1, 3,20	11,30,42

Pues bien, las posiciones que observé del cometa respecto a estas fijas eran como sigue.

El viernes 25 de febrero, calendario antiguo, a las 8½ p. m., el cometa hallándose en *p* distaba de la estrella E menos de $^{3}/_{13}$AE y más de ⅕AE, por lo cual, aproximadamente distaba $^{3}/_{14}$AE; y el ángulo A*p*E era un poco obtuso, pero casi recto. Pues si se abatía una perpendicular sobre *p*E desde A, la distancia del cometa de dicha perpendicular era de ⅕*p*E.

Esa misma noche, a las 9½ h. la distancia del cometa situado en P a la estrella E era mayor que $\frac{1}{4½}$AE, y menor que $\frac{1}{5¼}$AE y por tanto igual a $\frac{1}{4⅞}$AE, ó $\frac{8}{39}$AE aproximadamente. Pues la distancia del cometa desde la perpendicular trazada desde la estrella A sobre la recta PE era de ⅘PE.

El domingo 27 de febrero a las 8¼ p. m. estando el cometa en Q distaba de la estrella O una distancia igual a la de las estrellas O y H, y la recta QO, prolongada, pasaba entre las estrellas K y B. No pude, debido a las nubes, determinar más exactamente la posición de dicha recta.

El martes 1 de marzo a las 11 p. m., con el cometa en R, se hallaba situado exactamente entre las estrellas K y C, y la parte CR de la recta CRK era algo mayor que ⅓CK, y un poco menor que ⅓CK + $^{1}/_{16}$CR, y por tanto igual a ⅓CK + $^{1}/_{16}$CR, o también $^{16}/_{45}$CK.

El miércoles 2 de marzo, a las 8 p. m., con el cometa en S la distancia a la estrella C era de $^{4}/_{9}$FC aproximadamente. La distancia de la estrella F a la prolongación de la recta CS era de ½4FC; y la distancia de la estrella B de dicha recta era cinco veces mayor que la distancia de la estrella F. Asimismo, la prolongación de la recta NS pasaba entre las estrellas H y I, pasando cinco o seis veces más cerca de H que de I.

El sábado 5 de marzo, a las 11½ p. m., con el cometa en T, la recta MT era igual a ½ML, y la prolongación de la recta LT pasaba entre B y F, cuatro o cinco veces más cerca de F que de B, separando de BF una quinta o sexta parte de la misma hacia F. Y la prolongación de MT pasaba por fuera del espacio BF hacia el lado de la estrella B, cuatro veces más cerca de B que de F. M era una estrella muy pequeña que apenas se divisaba por el telescopio, y L era una estrella mayor, como de octava magnitud.

El lunes 7 de marzo, a las 9½ p. m., con el cometa en V, la prolongación de la recta V*α* pasaba entre B y F, segregando hacia F½ de BF, y era respecto a la recta V*β* como 5 a 4. Y la distancia del cometa de la recta *αβ* era ½V*β*.

El miércoles 9 de marzo, a las 8½ p. m., con el cometa en X, la recta *γ*X era igual a ¼*γδ*, y la perpendicular trazada desde la estrella *δ* a la recta *γ*X era ⅖*γδ*.

Esa misma noche, a las 12, con el cometa en Y, la recta *γ*Y era igual a ⅓*γδ*, o un poco menor, quizá $^{5}/_{16}$*γδ*, y la perpendicular trazada desde la estrella *δ* hasta la recta *γ*Y era igual a ⅙*γδ* ó $^{1}/_{7}$*γδ* muy aproximadamente. Pero por la cercanía al horizonte apenas se podía ver el cometa, ni tampoco su lugar se podía determinar con tanta precisión como en las observaciones anteriores.

Con base en estas observaciones mediante la construcción de figuras y mediante cálculos obtuve las longitudes y latitudes del cometa, y nuestro *Pound* corrigió los lugares del cometa a partir de las correcciones de los lugares de las fijas, y más arriba hemos dado los lugares corregidos. Me serví de un micrómetro construido con no mucha delicadeza, pero los errores de longitud y latitud (en la medida en que se deban a nuestras observaciones) escasamente superan un minuto. Por otra parte, el cometa (según nuestras observaciones) al final de su movimiento empezó a desviarse notablemente hacia el norte desde el paralelo en que había estado desde finales del mes de febrero.

Pero ahora vayamos a la determinación de la órbita del cometa; de las observaciones descritas hasta aquí seleccioné tres, realizadas por *Flamsteed* en diciembre 21, enero 5, y enero 25. A partir de ellas hallé que Sí era 9842,1 partes y V*t* 455 partes de 10 000 en que estuviera dividido el semidiámetro del Orbe Magno. Entonces, suponiendo para la primera operación a *t*B de 5657 partes, hallé para SB 9747, para BE la primera vez 412, para S*μ* 9503, para *iλ* 413: para BE la segunda vez 421, para OD 10 186, para X 8528,4, MP 8450, MN 8475, NP 25. De donde, obtuve para la segunda operación *tb* 5640. Y mediante esta operación hallé por fin las distancias TX 4775 y *τ*Z 11 322. Definiendo la órbita con estos datos, hallé sus nodos, el descendiente en ☋ y el ascendente en ♑ 1.º 53′; la inclinación de su plano respecto al de la eclíptica de 61.º 20⅓; su vértice (o perihelio del cometa) distaba del nodo 8.º 38′, y se hallaba en ♐ 27.º 43′ latitud sur; mientras su «latus rectum» era de 236,8 y el área descrita cada día con un radio trazado al Sol era de 93 585, suponiendo el cuadrado del semidiámetro del Orbe Magno de 100 000 000; y el cometa caminó por esta órbita siguiendo el orden de los signos y se halló en el vértice de la órbita o perihelio el 8 de diciembre a las Oh. 4′ p. m. Todo esto lo determiné mediante escala de partes iguales y cuerdas de ángulos obtenidas de la tabla de senos naturales gráficamente, construyendo un dibujo lo bastante grande como para que el semidiámetro del Orbe máximo (de 10 000 partes) fuese casi igual a 16⅓ pulgadas de pie inglés.

Finalmente, para cerciorarme de si el cometa se movía por una órbita así hallada, deduje mediante operaciones, en parte aritméticas y en parte gráficas, los lugares del cometa en dicha órbita en los momentos de algunas observaciones, como se pueden ver en la tabla siguiente.

El cometa

		Dist. al Sol	*Long. calc.*	*Latit. calc.*	*Long. observ.*	*Latit. observ.*	*Difer. de long.*	*Difer. de latit.*
Dic.	12	2792	6.º 32′	8.º 18½	6.º 31½	8.º 26	+1	-7½
	29	8403	13,13 2/3	28,00	13,11¾	28,10 1/12	+2	-10 1/12
Feb.	5	16 669	17,00	15,29 2/3	16,59⅞	15,27⅖	+0	+2¼
Mar.	5	21 737	29,19¾	12,4	29,20 6/7	12,03½	-1	+½

Pero, más tarde, nuestro *Halley* determinó mediante cálculo aritmético la órbita con mayor exactitud que la permitida por las descripciones de líneas; y en verdad vino a mantener el lugar de los nodos en ☋ y ♑ 1.º 53′, la inclinación del plano de la órbita respecto a la eclíptica de 61.º 20⅓ y el momento del perihelio del cometa en *diciembre*, 8d. 0h. 4′: por otra parte, la distancia del perihelio al nodo ascendente al medirla halló que era de 9.º 20′, y el «latus rectum» de la parábola de 2430 partes, siendo la distancia media de la Tierra al Sol de 100 000 partes. A partir de estos datos, haciendo un cálculo aritmético muy exacto, calculó los lugares del cometa en los momentos de la observación como sigue.

Tiempo verdadero	*El cometa*			*Errores de Longitud · Latitud*	
	Distancia al Sol	*Longitud calculada*	*Latitud calculada*		
d, h, m		° ′ ″	° ′ ″	° ′ ″	′ ″
Dic. 12,04,46	28 028	♑ 6,29,25	8,26,00	-3,05	-2,00
21,06,37	61 076	♒ 5,06,30	21,43,20	-1,42	+1,07
24,06,18	70 008	18,48,20	25,22,40	-1,03	-0,25
26,05,20	75 576	28,22,45	27,01,36	-1,28	+0,44
29,08,03	84 021	♓ 13,12,40	28,10,10	+1,59	+0,12
30,08,10	86 661	17,40,05	28,11,20	+1,45	-0,33
Ene. 5,06,1½	101 440	♈ 8,49,49	26,15,15	+0,56	+0,08
9,07,00	110 959	18,44,36	24,12,54	+0,32	+0,58
20,06,06	113 162	20,41,00	23,44,10	+0,10	+0,18
13,07,09	120 000	26,00,21	22,17,30	+0,33	+0,02
25,07,59	145 370	♉ 9,33,40	17,57,55	-1,20	+1,25
30,08,22	155 303	13,17,41	16,42,07	-2,10	-0,11
Feb. 2,06,35	160 951	15,11,11	16,04,15	-2,42	+0,14
5,07,4½	166 686	16,58,55	15,29,13	-0,41	+2,00
25,08,41	202 570	26,15,46	12,48,00	-2,49	+1,10
Mar. 5,11,39	216 205	29,18,35	12,05,40	+0,35	+2,14

También apareció este cometa en el mes de *noviembre* precedente y fue observado en *Coburgo* de *Sajonia* por *Gottfried Kirch* los días cuatro, seis, y once de dicho mes, calendario antiguo; y de las posiciones del mismo respecto a las estrellas fijas cercanas, posiciones observadas con bastante exactitud gracias a un telescopio a veces de dos pies y a veces de diez, junto con la diferencia de diez grados entre *Coburgo* y *Londres* y los lugares de las fijas observados por nuestro *Pound*, determinó nuestro *Halley* los lugares del cometa como sigue.

Nov. 3d. 17h. 2′, en el tiempo aparente de *Londres*, el cometa estaba en ☊ 29.º 5T con latitud boreal 1.º 17′ 45″.

Nov. 5d. 15h. 58′, el cometa estaba en ♍ 3.º 23′ con lat. boreal 1.º 6′.

Nov. 10d. 16h. 3T, el cometa distaba igual de las estrellas de Leo *σ* y *τ* de *Bayer*, pero aún no alcanzaba la línea que une a ambas estrellas, aunque distaba poco de ella. En el catálogo de estrellas de *Flamsteed σ* estaba entonces ♍ 14.º 15′, con latitud boreal de 1.º 41′, casi, mientras *τ* estaba ♍ 17.º 3½, con una latitud austral de 0.º 34′.

El punto medio entre estas estrellas ♍ 15.º 39¼ con latitud boreal de 0.º 33½. Sea la distancia del cometa a la susodicha línea de 10′ ó 12′ aproximadamente, y la diferencia de las longitudes del cometa y de dicho punto medio será de 7′, y las diferencias de latitudes de casi 7½. Y por ende el cometa estaba en ♍ 15.º 32′, con una latitud boreal de casi 26′.

La observación primera de la situación del cometa hasta algunas fijas menores fue más que suficientemente exacta. También la segunda fue suficientemente exacta. En la tercera, que fue menos exacta, pudo ser el error de menos de seis o siete minutos, y escasamente mayor. La longitud, en la primera observación, que fue la más exacta de todas, estaba calculada en la susodicha órbita parabólica en ♌ 29.º 30′ 22″, y latitud boreal de 1.º 25′ 7″, y su distancia al Sol de 115 546.

Mas he aquí que *Halley*, dándose cuenta de que un gran cometa había aparecido cuatro veces con intervalo de 575 años, a saber, el mes de septiembre posterior a la muerte de *Julio César*, el año de *Jesucristo*, 531, siendo *Cónsules Lampadio* y *Oreste*, el año de *Cristo* 1106 en el mes *de febrero y a fines del año* 1680, *y ello con una cola larga y brillante* (salvo el de la muerte de *César* en el que por la mala posición de la Tierra la cola brilló menos), se propuso hallar una órbita elíptica cuyo eje mayor fuera de 1 382 957 partes, siendo la distancia media de la Tierra al Sol de 10 000 partes, órbita en la cual hubiera podido girar de hecho el cometa en 575 años. Y situando el nodo ascendente en ♋ 2.º 2′, la inclinación del plano de la órbita respecto al de la eclíptica en 61.º 6′ 48″, el perihelio del cometa en dicho plano en ♐ 22.º 44′ 25″, el momento exacto del perihelio en diciembre 7d. 23h. 9′, la distancia del perihelio al nodo ascendente en el plano de la eclíptica en 9.º 17′ 35″, y el eje conjugado en 1848,2, calculó el movimiento del cometa en esta órbita elíptica. Sus lugares, tanto los derivados de las observaciones como los resultantes de los cálculos aparecen en la tabla siguiente.

Las observaciones de este cometa desde el principio al final concuerdan con el movimiento del cometa en la órbita recién descrita no menos que suelen concordar los movimientos de los planetas con sus respectivas teorías, y al concordar vienen a probar que se trata de uno y el mismo cometa que apareció en todo ese tiempo y cuya órbita fue correctamente establecida aquí.

En la tabla precedente omitimos las observaciones de los días 16, 18, 20 y 23 de noviembre por ser menos precisas. Pero también en esos días el cometa fue observado. Efectivamente, *Ponthio* y sus compañeros, el 17 de nov. (calendario antiguo) en *Roma* a las 6h. de la mañana, es decir, a las 5h. 10′ de *Londres*, con hilos tirados a las estrellas fijas, observaron el cometa en ♎ 8.º 30′ con latitud sur de 0.º 40′. Aparecen sus observaciones en un tratado que publicó *Ponthio* sobre este cometa. *Celtio*, que estaba presente y comunicó a *Cassini* por carta sus observaciones, vio el cometa a esa hora en ♎ 8.º 30′ con latitud austral de 0.º 30′. A la misma hora *Gallet* en *Avignon* (es decir, a las 5h. 42′ de la mañana de *Londres*) vio el cometa en ♎ 8 grados sin latitud. Aunque el cometa, por la teoría, estaba ahora en

♎ 8.º 16′ 45″ con latitud austral de 0.º 53′ 7″.

El 18 de noviembre a las 6h. 30′ de la mañana en *Roma* (es decir, a las 5h. 40′ de *Londres) Ponthio* vio el cometa en ♎ 13.º 30′ con latitud austral de 1.º 20′. *Cellio* en ♎ 13.º 30′, con latitud austral de 1.º 00′. *Gallet*, por su parte, en *Avignon* a las 5,30′ vio al cometa en ♎ 13.º 00′ con latitud austral de 1.º 00′. Y el R. P. *Ango*, en el Colegio de *La Fleche*, en *Francia*, a las 5 de la mañana (es decir, a las 5,9′ en Londres) vio el cometa en medio de dos pequeñas estrellas, una de las cuales es la central de tres en línea recta de la mano austral de Virgo, *Bayer* ♆, y la otra es la última del ala, *Bayer θ*. Por lo cual el cometa estaba entonces en ♎ 12.º 46′ con latitud austral de 50′. El mismo día, en *Boston* en *Nueva Inglaterra*, a 42½ grados de latitud a las cinco de la mañana (es decir, a las 9h. 44′ de la mañana de Londres) fue visto el cometa cerca de ♎ 14 grados con latitud austral de 1.º 30′, como me ha hecho saber el Ilustre *Halley*.

El 19 de nov., a las 4½ de la mañana en *Cambridge*, el cometa (observado por cierto joven) distaba de Spica ♍ casi 2 grados hacia el noroeste. Pero Spica se encontraba en ♎ 19.º 23′ 47″ con latitud austral de 2.º 1′ 59″. El mismo día en *Boston*, *Nueva Inglaterra*, a las 5h. a. m. el cometa distaba de Spica ♍ 1 grado, siendo la diferencia de latitudes 40′. El mismo día en *Jamaica*, el cometa distaba de Spica casi 1 grado. El mismo día el señor *Arthur Storer* en el río *Patuxent*, junto a *Hunting-Creek*, en Maryland en los confines de *Virginia*, a 38½ grados de latitud y a las cinco de la mañana (esto es, a las lOh. de *Londres*) vio el cometa sobre Spica ♏ y casi junto con ella, siendo la distancia entre ellos de casi ¾ de grado. Y comparando estas observaciones entre sí infiero que a las 9h. 44′ de *Londres* el cometa estaba en ♎ 18.º 50′ con latitud austral de casi 1.º 25′. Y el cometa según la teoría estaba ahora en ♎ 18.º 52′ 15″, con latitud austral de 1.º 26′ 54″.

El 20 de nov., *Montenari*, profesor de astronomía en *Padua*, a las seis de la mañana en *Venecia* (es decir, a las 5h. 10′ en *Londres*) vio al cometa en ♎ 23 grados con latitud austral de 1.º 30′. El mismo día en *Boston*, el cometa distaba de Spica ♍ 4 grados de longitud a oriente y por tanto estaba en ♎ 23.º 24′ aproximadamente.

El 21 de nov., *Ponthio* y sus colegas a las 7¼ de la mañana observaron el cometa en ♎ 27.º 50′ con latitud austral 1.º 16′. *Cellio* en ♎ 28 grados. *Ango* a las cinco de la mañana en ♎ 27.º 45′. *Montenari* en ♎ 27.º 51′. El mismo día en la isla de *Jamaica* se vio cerca del principio de Escorpio, y tuvo casi la misma latitud que Spica de Virgo, es decir, 2.º 2′. El mismo día a las cinco de la mañana en *Ballasore* en la *India Oriental* (es decir, a las 11h. 20′ de la noche precedente en *Londres*) se tomó una distancia del cometa de Spica ♍ de 7.º 33′ hacia oriente. Se hallaba en línea recta entre Spica y Libra, y por ende se hallaba en ♎ 26.º 58′, con latitud austral de casi 1.º 11′; y después de las 5h. 40′ (hacia las cinco de la mañana de *Londres*) estaba en ♎ 28.º 12′ con latitud sur de 1.º 16′. Según la teoría el cometa se hallaba en ese momento en ♎ 28.º 10′ 36″, con latitud austral de 1.º 53′ 35″.

El 22 de noviembre el cometa fue visto por *Montenari* en ♍ 2.º 33′. En *Boston*,

en *Nueva Inglaterra*, apareció en ♏ 3 grados casi y con la misma latitud prácticamente que antes, esto es, 1.º 30′. El mismo día a las cinco de la mañana en *Ballasore* se observaba el cometa en ♏ 1.º 50′; y por tanto, a las cinco de la mañana en *Londres* el cometa estaba en ♏ 3.º 5′ aproximadamente. El mismo día en *Londres*, a las 6½ de la mañana, nuestro *Hooke* vio el cometa en ♏ 3.º 30′ aproximadamente, y esto en una línea recta que pasa por Spica de Virgo y Corazón de Leo, si bien no exactamente, sino separándose un poquito hacia el norte. *Montenari* también observó que la línea trazada desde el cometa a Spica ese día y los siguientes pasaba por el lado sur de Corazón de Leo, mediando un muy pequeño espacio entre Corazón de Leo y esa línea. La línea recta que pasa por Corazón de Leo y Spica de Virgo cortó a la eclíptica en ♏ 3.º 46′, con un ángulo de 2.º 51′. Y si el cometa hubiese estado en esa línea en ♏ 3 grados, su latitud hubiera sido de 2.º 26′. Pero como el cometa, de acuerdo con *Hooke* y *Montenari*, distaba un poco de esa línea hacia el norte, su latitud fue un poco menor. El día 20, por la observación de *Montenari*, su latitud era sensiblemente igual a la de Spica ♍ y era de casi 1.º 30′, y de acuerdo con *Hooke*, *Montenari* y *Ango*, aumentaba continuamente, y por tanto ahora era sensiblemente mayor que 1.º 30′. Entre los límites establecidos de 2.º 26′ y 1.º 30′, la latitud con un valor medio estaría aproximadamente en 1.º 58′. La cola del cometa, de acuerdo con *Hooke y Montenari* se dirigía hacia Spica ♍, declinando un tanto de dicha estrella, según *Hooke* hacia el sur, según *Montenari* hacia el norte; por ello la declinación apenas fue sensible, y siendo la cola casi paralela al ecuador, se apartaba un tanto de la oposición del Sol hacia el norte.

El 23 de nov., calendario antiguo, a las cinco de la mañana en *Nuremberg* (es decir, a las 4½ en *Londres*) el señor *Zimmerman* vio el cometa en ♏ 8.º 8′, con latitud austral 2.º 31′, tomando sus distancias de las estrellas fijas.

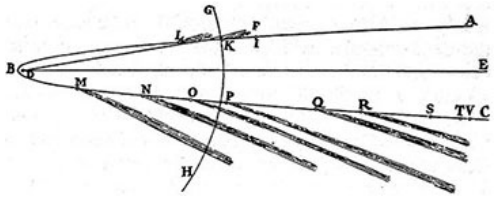
El 24 de nov., antes de la salida del Sol *Montenari* vio el cometa en ♏ 12.º 52′ al lado norte de la recta que pasa por Corazón de Leo y Spica de Virgo, y por tanto, la latitud era algo menor que 2.º 38′. Como hemos dicho esta latitud, de acuerdo con las observaciones de *Montenari*, *Ango* y *Hooke*, aumentaba continuamente; por tanto ahora era algo mayor que 1.º 58′; y por tanto puede establecerse sin mayor error en un valor medio de 2.º 18′. *Ponthio* y *Gallet*, pretenden que a estas alturas ya había decrecido la latitud, mientras *Cellio* y el observador de *Nueva Inglaterra* retienen casi el mismo valor, a saber un grado o un grado y medio. Más inexactas son las observaciones de *Ponthio y Gallet*, sobre todo las que realizaban tomando acimuts y alturas, como las de *Gallet*; mejores resultan las realizadas por las posiciones del cometa respecto a las fijas por *Montenari*, *Hooke*, *Ango* y el observador de *Nueva Inglaterra* y a veces por *Ponthio y Cellio*. El mismo día a las cinco de la mañana en *Ballasore* el cometa se veía en ♏ 11.º 45′; y por tanto a las cinco de la mañana en *Londres* estaba casi en ♏ 13 grados. Pero por la teoría el cometa estaba ahora en ♏ 13.º 22′ 42″.

El 25 de nov., antes de la salida del Sol, *Montenari* vio al cometa casi en ♏ 17¾º.

Y *Cellio* observó en ese momento que el cometa estaba en línea recta entre la estrella brillante del fémur derecho de Virgo y el plato austral de Libra, y esta recta corta la trayectoria del cometa en ♏ 18.º 36′. En cambio, por la teoría el cometa ahora estaba en casi ♏ 183⅓.º.

Por consiguiente, estas observaciones concuerdan con la teoría, en la medida en que concuerdan entre ellas, y, al concordar, prueban que fue uno y el mismo el cometa que se vio durante todo el tiempo desde el día cuatro de noviembre hasta el día nueve de marzo. La trayectoria de este cometa cortó dos veces el plano de la eclíptica, y por lo tanto no fue rectilínea. Cortó a la eclíptica no en partes opuestas del cielo, sino al final de Virgo y al principio de Capricornio, con un intervalo de casi 98 grados; y por lo mismo el curso del cometa se desviaba mucho del Círculo Máximo. Pues en el mes de noviembre su curso declinaba de la eclíptica al menos tres grados al sur y, después, en el mes de diciembre se apartaba de la eclíptica hacia el norte en 29 grados, declinando, como observó *Montenari*, las dos partes de la órbita con las cuales el cometa tendía hacia el Sol y se apartaba del mismo, con un ángulo aparente de más de treinta grados. Este cometa caminaba a lo largo de nueve signos, a saber, desde el último grado de Leo al Principio de Géminis, a partir del signo de Leo por el que pasaba antes de empezar a ser visto; y no existe ninguna otra teoría por la cual un cometa recorra tan gran porción del cielo con un movimiento regular. Su movimiento fue muy desigual. Pues en torno al día 20 de noviembre recorrió casi 5 grados cada día; después, con un movimiento retardado, entre el 26 de nov. y el 12 de diciembre, en un tiempo de 15½, sólo recorrió 40 grados: después con movimiento de nuevo acelerado, recorrió casi 15 grados cada día, antes de que el movimiento empezase de nuevo a retardarse. Y la teoría que mejor responde a un movimiento tan desigual a través de la mayor parte del cielo y que obedece las mismas leyes que la teoría de los planetas y concuerda exactamente con las observaciones astronómicas más exactas, no puede no ser verdadera.

Por lo demás, me ha parecido oportuno representar en el dibujo anexo, dibujadas sobre el plano de la trayectoria, tanto la trayectoria descrita por el cometa como la cola verdadera que proyectó en cada uno de los lugares: en él, ABC representa la trayectoria del cometa, D el Sol, DE el eje de la trayectoria, DF la línea de los nodos, GH la intersección de la trayectoria y de la esfera del Orbe Máximo, I el lugar del cometa el 4 de nov. de 1680, K su lugar el 11 de nov., L el lugar del 19 de nov., M el lugar del 12 de dic., N el lugar del 21 de dic., O el lugar del 29 de dic., P el lugar del 5 de enero siguiente, Q el lugar del 25 de enero. R el lugar del 5 de febrero, S el lugar del 25 de febrero, T el lugar del 5 de marzo, y V el lugar del 9 de marzo. Para definir la cola hice las observaciones siguientes.



El 4 y 6 de nov. la cola aún no apareció. El 11 de nov. empezándose a ver ya la cola, no parecía mayor de medio grado de larga con un telescopio de diez pies. El 17 de nov. *Ponthio* observó una cola de más de 15 grados de largo. El 18 de nov. se veía en *Nueva Inglaterra* una cola de más de 30 grados de largo, directamente opuesta al Sol y se alargaba hasta la estrella ♂, que se hallaba entonces en ♍ 9.º 54′. El 19 de nov. en *Maryland* se vio una cola de 15 ó 20 grados de largo. El 10 de diciembre (observador, *Flamsteed*) la cola pasaba por medio de la distancia entre la cola de la serpiente de Ofíuco y la estrella *δ* del ala sur de Águila, y terminaba cerca de las estrellas A, *ω*, *b*, de las tablas de *Bayer*. Por consiguiente, el final estaba ♑ 19½º con una latitud boreal de casi 34¼º. El 11 de diciembre la cola subía hasta casi la cabeza de Sagitario (*Bayer α*, *β*) acabando en ♑ 26.º 43′, con una latitud boreal de 38.º 34′. El 12 de diciembre la cola pasaba por medio de Sagitario, no sobrepasando mucho, acabando en ♒ 4 grados con una latitud boreal de casi 42½º. Entiéndase esto de la longitud de la cola más brillante. Pues con luz más oscura, quizá en un cielo más transparente, la cola fue observada en *Roma* el 12 de diciembre a las 5h. 40′ (observador *Ponthio*) extendiéndose hasta 10 grados por encima de la cola del *Cisne*; y partiendo desde esta estrella hacia el Noroeste su borde sobresalía 45′. En estos días, por tanto, la cola era de ancha 3 grados, en el lado de arriba, y por consiguiente, su centro distaba de dicha estrella 2.º 15′ hacia el sur, y el extremo superior se hallaba en ♓ 22 grados con una latitud boreal de 61 grados. Y de aquí que la cola era de casi 70 grados de largo, el 21 de diciembre. Subía casi hasta la silla de *Casiopea*, distando igualmente de *β* y de *Schedir*, manteniendo una distancia igual de una y otra a la distancia mutua entre ellas, acabando, por tanto, en ♈ 24 grados con una latitud de 47½º. El 29 de diciembre la cola tocaba a *Scheat* situada a la izquierda, y el intervalo de las dos estrellas del pie boreal de *Andrómeda* lo cubría exactamente, y era de 54 grados de larga; por lo tanto acababa en ♉ 19 grados, con una latitud de 35 grados. El 5 de enero, la cola tocó la estrella *π* del pecho de Andrómeda en el lado derecho, y a la estrella *μ* del cinturón en el lado izquierdo; y (según mis observaciones) era de 40 grados de larga; estaba curvada y su lado convexo miraba al sur. Formaba con un círculo que pasase por el Sol y la cabeza del cometa un ángulo de 4 grados junto a la cabeza del cometa; pero hacia el otro extremo se inclinaba hacia dicho círculo en un ángulo de 10 u 11 grados y la cuerda de la cola con dicho círculo contenían un ángulo de 8 grados. El 13 de enero, la cola, con una luz bastante perceptible, terminaba entre *Alamech* y *Algol* y con luz muy débil se acababa en la zona de la estrella *κ* del costado de Perseo. La distancia del final de la cola hasta el círculo que pasaba por el

Sol y el cometa era de 3.º 50′, y la inclinación de la cuerda de la cola hacia dicho círculo era de 8½º. En 25 y 26 de enero, la cola con luz muy tenue se alargaba hasta una longitud de 6 ó 7 grados; y las dos noches siguientes con un cielo muy sereno, con luz muy tenue y apenas perceptible alcanzaba una longitud de 12 grados y poco más. Su eje se dirigía a la estrella brillante del hombro oriental del Auriga, con toda exactitud, y por ello declinaba de la oposición al Sol hacia el norte en un ángulo de 10 grados. Finalmente, el 10 de febrero, con instrumentos oculares observé una cola de 2 grados de longitud. Pues la antedicha luz tenue con gafas no se veía. Pero *Ponthio*, el 7 de febrero, escribe que vio la cola de 12 grados de longitud. El 25 de febrero y siguientes el cometa se vio sin cola.

A quien considere la órbita recién descrita y medite en los fenómenos restantes de este cometa no le costará mucho concluir que los cuerpos de los cometas son sólidos, compactos, fijos y duraderos, al modo de los cuerpos de los planetas. Pues si nada más fuesen que vapores o exhalaciones de la Tierra, del Sol o los planetas, este cometa, a su paso por las inmediaciones del Sol, debió disiparse al instante. Pues el calor del Sol es como la densidad de los rayos, es decir, como el inverso del cuadrado de la distancia del lugar hasta el Sol. Y por lo mismo, como la distancia del cometa hasta el centro del Sol el día 8 de diciembre, cuando se encontraba en el perihelio, era a la distancia de la Tierra al centro del Sol casi como 6 a 1000, el calor del Sol junto al cometa en ese momento era al calor del Sol estival entre nosotros como 1 000 000 a 36, o como 28 000 a 1. Pero el calor del agua en ebullición es casi tres veces mayor que el calor que alcanza la tierra seca al calor del verano, como he comprobado: y el calor del hierro candente (si no es mala mi conjetura) es tres veces mayor o cuatro que el del agua en ebullición; y por eso el calor que la arena seca en la superficie del cometa al pasar éste por el perihelio recibió de los rayos solares sería unas 2000 veces mayor que el calor del hierro candente. Con semejante calor los vapores y las exhalaciones, así como toda materia volátil, deberían consumirse y disiparse al instante.

Por consiguiente, el cometa en su perihelio recibió del Sol un enorme calor, y puede conservar dicho calor por mucho tiempo. Pues un globo de hierro candente de una pulgada de diámetro apenas pierde todo su calor en el espacio de una hora expuesto al aire. Pero un globo mayor conservará el calor más tiempo en razón del diámetro, toda vez que la superficie (en cuya proporción se refrigera por contacto con el aire ambiente) es menor en dicha razón respecto a la cantidad de materia caliente contenida. Por lo cual, un globo de hierro candente igual a esta Tierra, es decir, de 40 000 000 de pies de ancho, más o menos, tardaría en enfriarse otros tantos días, o unos 50 000 años. Pero sospecho que la duración del calor, por causas desconocidas, aumenta en razón menor que la del diámetro: y quisiera que se investigase la verdadera razón mediante experimentos.

Por lo demás, hay que notar que el cometa en el mes de diciembre, cuando se había calentado al Sol hasta ese punto, emitía una cola mucho mayor y más brillante

que antes, en noviembre, cuando aún no había alcanzado el perihelio. Y generalmente todas las colas máximas y resplandecientes salen de los cometas inmediatamente después de su paso por la región solar. El calentamiento del cometa comporta, por tanto, el crecimiento de la cola. Y de ello me parece inferir que la cola no es otra cosa que un vapor sumamente tenue que emite la cabeza o núcleo del cometa debido a su calor.

Por otra parte, sobre las colas de los cometas tres son las opiniones; o bien que son resplandor del Sol que se propaga a través de la cabeza transparente de los cometas, o bien que se originan de la refracción de la luz en su camino desde la cabeza del cometa hasta la Tierra, o bien, finalmente, que son una nube o vapor que surge constantemente de la cabeza del cometa y se aleja hacia las partes opuestas al Sol. La primera opinión es propia de quienes aún no se han adentrado en la ciencia de la óptica. Pues el resplandor del Sol en una habitación oscura no se ve más que en la medida en que la luz es reflejada por partículas de polvo y de humo suspendidas siempre en el aire: y por lo mismo es más brillante en un aire más lleno de humos gruesos, y hiere con más fuerza a la vista; en un aire más enrarecido es mucho más tenue y se percibe con mayor dificultad: de suerte que en los cielos sin materia reflectante ninguna reflexión es posible. La luz no se ve en tanto en que está en el resplandor, sino en tanto en que de allí es reflejada hacia nuestros ojos. Pues la visión no ocurre más que por los rayos que inciden sobre los ojos. Por lo tanto se requiere alguna materia en la región de la cola, a no ser que el cielo entero brille uniformemente iluminado por la luz del Sol. La segunda opinión se ve acosada por múltiples dificultades. Las colas jamás tienen varios colores: cosa que suele acompañar inseparablemente a las refracciones. La luz de las fijas y de los planetas transmitida hasta nosotros con perfecta distinción demuestra que el medio celeste no goza de ningún poder de refracción. Pues, que los *Egipcios* vieran algunas veces, como se dice, estrellas fijas con cola, esto, además de ser muy raro, debe achacarse a una refracción fortuita de las nubes. Pues la radiación y los destellos de las fijas deben achacarse a refracciones bien de los ojos, bien del aire en movimiento, toda vez que desaparecen tan pronto como se arriman los ojos al telescopio. El temblor del aire y de los vapores que ascienden hace que los rayos se descoloquen de vez en cuando del estrecho espacio de la pupila, mientras no ocurre en el espacio más ancho del objetivo del telescopio. De aquí, que el centelleo tenga lugar en el primer caso y no en el segundo, y que cese en este segundo caso muestra la regular trasmisión de la luz por los cielos sin ninguna refracción sensible. Si alguien sostiene que las colas de los cometas no suelen verse cuando su luz no es suficientemente fuerte porque entonces los rayos secundarios no tienen fuerzas bastantes para herir a los ojos, y por esa razón no se ven las colas de las fijas, se ha de tener presente que la luz de las fijas puede aumentarse con los telescopios más de cien veces, y ni entonces se ven colas. Y la luz de los planetas es más abundante, pero colas ninguna; mientras los cometas frecuentemente tienen larguísimas colas, cuando las cabezas tienen luz muy tenue y

velada. Y así, el cometa de 1680, en diciembre, cuando la luz de su cabeza apenas igualaba a la de las estrellas de segunda magnitud, emitía una cola de notable resplandor y una longitud de 40, 50, 60 ó 70 y más grados. Más tarde, en enero 27 y 28, la cabeza parecía una estrella de séptima magnitud, mientras la cola, con una luz bastante tenue, aunque suficientemente sensible, alcanzaba una longitud de 6 ó 7 grados, y con una luz muy difusa que apenas sí se podía ver, alcanzaba hasta 12 grados o más: como se ha dicho antes. Pero el 9 y 10 de febrero, cuando la cabeza había dejado de verse a simple vista, he visto la cola de dos grados de largo con el telescopio. Además, si la cola se originase de la refracción de materia celeste y se inclinase de la oposición al Sol de acuerdo con la figura de los cielos, tal inclinación, en las mismas regiones del cielo, debería ocurrir siempre hacia la misma dirección. Pero el cometa de 1680, el 28 de diciembre a las 8½ p. m. en *Londres*, se hallaba en ♓ 8.º 41′ con una latitud boreal de 28.º 6′, estando el Sol en ♑ 18.º 26′. Y el cometa de 1577 el 29 de diciembre se hallaba en ♓ 8.º 41′ con una latitud boreal de 28.º 40′, estando también el Sol en ♑ 18.º 26′ aproximadamente. En ambos casos la Tierra se hallaba en el mismo lugar y el cometa aparecía en el mismo sitio del cielo, pero en el primer caso la cola del cometa (según observaciones mías y de otros) declinaba de la oposición del Sol con un ángulo de 4½ grados hacia el norte; en el segundo caso, en cambio (según las observaciones de *Tychó*) la declinación era de 21 grados hacia el sur. Descartada, por tanto, la refracción de los cielos, sólo queda derivar los fenómenos de las colas de alguna materia que refleja la luz.

Que las colas se originan de las cabezas y ascienden hacia las regiones opuestas al Sol se confirma por las leyes que observan. Como, p. e., que situadas en los planos de las órbitas de los cometas que pasan por el Sol, se desvían de la oposición del Sol siempre hacia aquellas partes que, al avanzar, van dejando atrás las cabezas en dichas órbitas. Que para un espectador situado en esos planos aparecen en las partes directamente opuestas al Sol; pero al separarse de esos planos el espectador nota enseguida la desviación, y poco a poco aparece mayor. Que la desviación, con las demás condiciones iguales, es menor cuando la cola es más oblicua respecto a la órbita del cometa, lo mismo que cuando la cabeza del cometa se acerca más al Sol; sobre todo si el ángulo de la desviación se toma junto a la cabeza del cometa. Además de que las colas que no se desvían aparecen rectas, mientras que las que se desvían aparecen curvadas. Que la curvatura es mayor cuando la desviación es mayor, y más sensible, en igualdad de circunstancias, cuando la cola es más larga: pues, en las muy cortas, la curvatura difícilmente se distingue. Que el ángulo de desvío es menor junto a la cabeza del cometa, y mayor junto al extremo opuesto de la cola, y por tanto que la cola con su lado convexo mira a las partes de las que se desvía, las cuales están en la recta infinita, trazada desde la cabeza del cometa al Sol. Y que las colas que son más largas y extensas y brillan con luz más viva, son algo más resplandecientes por el lado convexo y perfiladas algo menos imprecisamente que por el lado cóncavo. Por consiguiente, los fenómenos de las eolias dependen del movimiento de la cabeza y no

de la región del cielo en que se ve la cabeza; y por lo mismo, no ocurren por la refracción de los cielos, sino que se originan de la cabeza que proporciona materia. Pues, al igual que en nuestro aire el humo de un cuerpo cualquiera ardiendo se dirige hacia arriba, y esto perpendicularmente si el cuerpo está en reposo o bien oblicuamente si el cuerpo se mueve hacia un lado: del mismo modo en los cielos, donde los cuerpos gravitan hacia el Sol, los humos y vapores deben ascender desde el Sol (como ya se ha dicho) y dirigirse hacia arriba o en dirección recta si los cuerpos humeantes están en reposo; o bien oblicuamente, si el cuerpo avanzando continuamente va dejando siempre atrás los lugares hacia los cuales habían ascendido las partes superiores del vapor. Y la oblicuidad esa siempre será menor cuando más veloz sea el ascenso del vapor: o sea, en las proximidades del Sol y junto al cuerpo fumante. Pero por la diversidad de oblicuidad se curvará la columna de vapor: y debido a que en el lado anterior de la columna el vapor es algo más reciente, por eso mismo es allí también algo más denso y por ello refleja una luz más abundante a la vez que se perfila con un límite menos impreciso. Sobre los movimientos súbitos e inciertos de las colas o sobre sus figuras irregulares, descritos a veces por algunos, nada añado aquí, por cuanto que pudieran deberse a movimientos y cambios de nuestro aire o a movimientos de las nubes que oscurecen en parte las colas, o quizá a partes de la vía láctea que pudieran confundirse con las colas en tránsito por ella y contemplarse como partes de las mismas.

A partir de la rareza de nuestro aire se comprende que de las atmósferas de los cometas puedan surgir vapores suficientes para llenar tan enormes espacios. Pues el aire junto a la superficie de la Tierra ocupa un espacio casi 850 veces mayor que el mismo peso de agua, y por tanto, una columna cilíndrica de aire de 850 pies de alto pesa igual que una columna de agua de un pie y de la misma anchura. Y una columna de aire que llegue hasta la cima de la atmósfera iguala con su peso a una columna de agua que tenga 33 pies de alto aproximadamente; y por ello, si se quitase de toda la columna aérea la parte inferior de 850 pies de alto, la parte restante superior igualaría con su peso a una columna de agua de 32 pies de alto. Y a partir de ello (por una regla confirmada mediante muchos experimentos, de que la compresión del aire es como el peso del aire sobrepuesto, y de que la gravedad es inversamente como el cuadrado de la distancia de los lugares al centro de la Tierra), calculando de acuerdo con el Corolario de la Proposición XXII del Libro II, he hallado que el aire, si se sube hasta una altura desde la superficie terrestre de un semidiámetro de la Tierra, tendrá una rareza mayor con mucho que entre nosotros, en una razón mayor que la que existe entre todo el espacio bajo la órbita de Saturno y un globo descrito con un diámetro de una pulgada. Y por lo mismo, un globo de nuestro aire de una pulgada de ancho, con la rareza que tendría a la altura de un semidiámetro terrestre, llenaría todas las regiones de los planetas hasta la esfera de Saturno y más allá. Por consiguiente, puesto que el aire a más altura todavía se enrarece mucho más, y la cabellera o atmósfera de un cometa surgiendo desde su centro, sea casi diez veces

más alta que la superficie del núcleo, y después la cola todavía ascienda mucho más arriba, la cola deberá ser sumamente enrarecida. Y aunque, por ser mucho más gruesa la atmósfera de los cometas y por la enorme gravitación de los cuerpos hacia el Sol y de las partículas de aire y vapor entre ellas, pudiera ocurrir que el aire en los espacios celestes y en las colas de los cometas no se enrarezca tanto, no obstante, una muy pequeña cantidad de aire y de vapor basta y sobra para todos estos fenómenos de las colas, como se ve por los dichos cálculos. Pues, hasta por los astros que se traslucen a través de ellas se puede colegir la gran rarefacción de las colas. La atmósfera terrestre, con un grosor de unas pocas millas, iluminada por la luz del Sol oscurece a todos los astros y hasta casi extingue a la propia Luna; mientras tanto, a través del inmenso grosor de las colas iluminadas igualmente por la luz del Sol se perciben hasta los menores astros sin detrimento de la claridad. Y no suele ser mayor el brillo de muchas colas que el de nuestro aire reflejando la luz del Sol en un rayo, de una o dos pulgadas de ancho dentro de un cuarto oscuro.

En cuánto tiempo asciende el vapor desde la cabeza hasta el extremo de la cola casi puede conocerse trazando una recta desde el extremo de la cola hasta el Sol y anotando el punto en que dicha recta corta a la trayectoria. Pues el vapor al final de la cola, si sube en línea recta desde el Sol, comenzó a ascender en el momento en que la cabeza se hallaba en el punto de la intersección. Pero el vapor no asciende en línea recta desde el Sol, sino que, reteniendo el movimiento que tenía el cometa antes de su ascenso, y componiéndolo con el movimiento de su propio ascenso, asciende oblicuamente. Por lo cual sería mejor solución del problema que dicha recta secante de la trayectoria fuera paralela a la longitud de la cola, o mejor (por el movimiento curvilíneo del cometa) que sea divergente de la línea de la cola. De este modo hallé que el vapor que se hallaba en el extremo de la cola el 25 de enero había comenzado a ascender de la cabeza el día 11 de diciembre anterior, y por lo mismo había consumido en su ascenso completo más de 45 días. Pero la cola entera que apareció el día 10 de diciembre, había ascendido en los dos días que habían sido consumidos por el perihelio del cometa. Por lo tanto, en las proximidades del Sol, el vapor al principio ascendía muy rápidamente, y después con un movimiento continuamente retardado por su gravedad continuaba ascendiendo, y al ascender aumentaba la longitud de la cola, mientras la cola, cuando apareció, constaba casi de todo el vapor que había ascendido durante el perihelio; y el primer vapor que ascendió y dio lugar al extremo de la cola no se disipó hasta que, por su exagerada distancia tanto del Sol iluminante como de nuestros ojos, dejó de verse. Por eso también, las colas de otros cometas, que son más cortas, no ascienden con un movimiento rápido y continuo desde las cabezas y desaparecen enseguida, sino que son columnas permanentes de vapores y exhalaciones propagadas desde la cabeza con un movimiento muy lento de muchos días, las cuales, al participar de aquellos movimientos de la cabeza que recibieron en el inicio continúan moviéndose por los cielos a la vez que las cabezas. Y esto de nuevo permite inferir que los espacios están desprovistos de fuerza de

resistencia, por cuanto que en ellos no sólo los cuerpos sólidos de los planetas y de los cometas, sino también los vapores sumamente enrarecidos de las colas llevan a cabo con toda libertad sus rapidísimos movimientos y los conservan durante largo tiempo.

Atribuye *Kepler* el ascenso de las colas desde las atmósferas de las cabezas así como su dirección hacia las partes opuestas al Sol a la acción de los rayos de luz que arrastran con ellos a la materia de las colas. Y no parece lejos de la razón el que un aura tan sumamente tenue y en unos espacios completamente libres cediera a la acción de los rayos, pese a que en nuestras regiones sumamente entorpecidas las materias gruesas no puedan ser desplazadas por los rayos del Sol. Piensa otro que pueden darse partículas tanto leves como graves, y que la materia de las colas levita y por su levitación asciende desde el Sol. Pero, puesto que la gravedad de los cuerpos terrestres es como la cantidad de materia en ellos, y por tanto, manteniendo la cantidad de materia, no pueda aumentar ni disminuir, sospecho que el ascenso sea debido más bien a la rarefacción de la materia de las colas. El humo asciende por la chimenea gracias al impulso del aire en el que flota. Dicho aire, enrarecido por el calor asciende al disminuir su gravedad específica y arrastra consigo al humo contenido en él. ¿Por qué no podrían ascender de igual modo desde el Sol las colas de los cometas? Pues los rayos del Sol no agitan los medios que atraviesan a no ser en la reflexión y la refracción. Las partículas reflectantes calentadas por esa acción calientan el aura etérea en que se hallan inmersas. Esta se enrarece por el calor que le ha sido comunicado, y por la disminución, debida al enrarecimiento, de su gravedad específica con la cual tendía primero hacia el Sol, ascenderá y arrastrará consigo a las partículas reflectantes de las que se compone la cola; también conduce al ascenso de los vapores el hecho de que éstos giran en torno al Sol y con esta acción tratan de apartarse del Sol, mientras la atmósfera del Sol o reposa, al igual que la materia celeste, de modo total o bien gira más lentamente con el solo movimiento que hubiese recibido de la rotación solar. Estas son las causas del ascenso de las colas en las proximidades del Sol, cuando las órbitas son más curvas y los cometas se hallan dentro de la atmósfera más densa, y por ello más grave, del Sol, y cuando emiten colas de mayor longitud. Pues las colas que nacen entonces, conservando su movimiento y gravitando entre tanto hacia el Sol, se moverán en elipses en torno al Sol al igual que las cabezas, y con ese movimiento siempre acompañarán a las cabezas adhiriéndose a ellas con toda libertad. Pues la gravedad de los vapores hacia el Sol no tendrá mayor eficacia para que las colas se aparten después de las cabezas hacia el Sol, que la gravedad de las cabezas para hacer que éstas se separen de las colas. Con la gravedad común o bien caen a la vez hacia el Sol, o bien son retardadas a la vez en su ascenso; por lo cual dicha gravedad no impide que colas y cabezas alcancen una posición mutua cualquiera, debida a las causas expuestas o a otras cualesquiera y, después, la conserven con toda libertad.

Pero las colas que nacen en los perihelios de los cometas se alejarán a muy

distantes regiones junto con sus cabezas, y allí o regresan tras una larga serie de años hasta nosotros junto con ellas o, más bien, enrarecidas se desvanecen allí poco a poco. Pues, luego, cuando las cabezas descienden de nuevo al Sol, unas colas nuevas y minúsculas habrán de propagarse desde las cabezas, y más tarde, en los perihelios de dichos cometas, que llegan a descender hasta la atmósfera del Sol, aumentan enormemente. Pues el vapor, en aquellos espacios completamente libres, se enrarece y dilata continuamente. Por esa razón la cola entera en el extremo superior es mucho más ancha que junto a la cabeza del cometa. Y es razonable pensar que el vapor continuamente dilatado por esa rarificación, acaba por difundirse y dispersarse por todos los cielos, para desde allí poco a poco ser atraído por su gravedad hacia los planetas y mezclarse con las atmósferas de ellos. Pues de igual modo que necesariamente se requieren los mares para la constitución de esta Tierra, y esto de suerte que de ellos pueda excitar el calor del Sol abundantes y suficientes vapores que, o acumulados en nubes caigan como lluvia y rieguen y nutran a toda la tierra para la germinación de los vegetales, o bien condensados en las frías cumbres de los montes (como algunos razonablemente sostienen) discurran en fuentes y ríos: así también, para la conservación de los mares y de los humores en los planetas parecen requerirse los cometas, merced a cuyas exhalaciones y a cuyos vapores pueda suplirse y restaurarse cuanto licor se consume por la vegetación y la putrefacción y se convierte en tierra árida. Pues, todo lo vegetal crece absolutamente de licores, y después, en su mayor parte, se convierte por putrefacción en tierra árida, y el limo resulta siempre de los licores putrefactos. De aquí que la materia árida terrosa aumente cada día, mientras que los licores, salvo que aumenten por aportes exteriores, deberán decrecer continuamente, hasta que por fin desaparezcan. Por lo demás, sospecho que esa parte de nuestro aire, mínima pero sutilísima y óptima además de necesaria para la vida de todas las cosas, proviene principalmente de los cometas.

Las atmósfera de los cometas, en el descenso de éstos hacia el Sol, desparramándose hacia las colas disminuyen, y (ciertamente hacia el lado que mira al Sol) se hacen más estrechas; y viceversa, en su regreso desde el Sol, cuando ya se desparraman menos hacia las colas, se ensanchan; si es que *Hewelcke* hizo bien la anotación de sus apariencias. Pues aparecen mínimas cuando las cabezas calentadas por el Sol en tal modo han dado lugar a colas máximas y muy brillantes, y los núcleos están rodeados de humo quizá más denso y negro en las partes inferiores de las atmósferas. Pues todo humo suscitado por un calor muy grande suele ser más grueso y negro. Así la cabeza del cometa al que nos hemos referido, a iguales distancias del Sol y de la Tierra aparecía más oscuro después del perihelio que antes. Pues en el mes de diciembre, solía compararse con estrellas de tercera magnitud, mientras en noviembre lo era con estrellas de primera y segunda. Y quienes vieron ambos, describen como más grande al primero. Pues para un cierto joven de *Cambridge*, el 19 de noviembre, este cometa con su luz un tanto plomiza y opaca igualaba a Spica

de Virgo, y brillaba más claramente que después. Y el 20 de noviembre del calendario antiguo, el cometa parecía a *Montenari* mayor que las estrellas de primera magnitud, siendo la cola de dos grados de longitud. Y el señor *Storer*, en carta que se halla en mis manos, escribe que la cabeza del cometa en el mes de diciembre cuando emitía la cola máxima y más brillante, era pequeño y en su magnitud visible mucho menor que el cometa que había aparecido en el mes de noviembre antes de la salida del Sol. Conjeturaba que la razón de ello era que la materia de la cabeza al principio era más abundante y que poco a poco se habría consumido.

De igual modo parece que hay que contemplar el hecho de que las cabezas de otros cometas que emitieron colas muy grandes y luminosas, aparecieran muy oscuras y pequeñas. Pues el año 1668, el 5 de marzo, calendario nuevo, a las siete de la tarde el R. P. *Valentín Estancel*, estando en *Brasil*, vio un cometa próximo al horizonte hacia el suroeste, con una cabeza mínima y apenas distinguible, pero con una cola sumamente brillante, hasta el punto de que los que se hallaban en el litoral veían con toda facilidad su imagen reflejada en el mar. Ya que era la imagen de una estela brillante de 23 grados de longitud, inclinada desde occidente hacia el sur y casi paralela al horizonte. Pero semejante resplandor solamente duró tres días, decreciendo después notablemente; y mientras decrecía el resplandor aumentó la magnitud de la cola. Y en *Portugal* también se dice que llegó a ocupar casi la cuarta parte del cielo (es decir, 45 grados) extendiéndose de occidente a oriente con un gran resplandor; y ni siquiera apareció entera, al hallarse siempre la cabeza circulando bajo el horizonte de estas regiones. Es evidente por el incremento de la cola y la merma del resplandor que la cabeza regresaba desde el Sol, y que estuvo próxima a él al principio, al igual que el cometa de 1680. Y en la *Crónica Sajona* se lee también sobre un cometa semejante del año 1106 «cuya estrella era pequeña y oscura» (como el del año 1680) «pero el resplandor que salió de ella era muy luminoso y como una estela ardiente que se dirigía hacia oriente y norte», como refiere también *Hewelcke* citando al monje *Simeón* de *Durham*. Apareció a principios del mes de febrero, al atardecer hacia el suroeste. De aquí y de la posición de la cola se infiere que la cabeza estaba próxima al Sol. Dice *Mateo de París* que «*distaba del Sol casi un codo, emitiendo de sí un gran rayo desde la hora tercia*» (más exacto sería la sexta) «*hasta la hora nona*». Así era también aquel ardiente cometa descrito por *Aristóteles* en *Meteoros*, Libro I, 6, «*cuya cabeza no se vio el primer día debido a que se hallaba ante el Sol o al menos bajo los rayos solares, pero al día siguiente se vio cuanto fue posible. Pues se apartó una mínima distancia del Sol, y enseguida se ocultó. Por causa del excesivo brillo*» (de la cola) «*todavía no aparecía el fuego difuso de la cabeza, pero pasado un tiempo*» (dice Aristóteles), «*como* (la cola) *brillase ya menos, se le restituyó su cara* (a la cabeza del) *al cometa. Extendió su resplandor hasta la tercera parte del cielo* (esto es, a 60 grados). *Y apareció en invierno* (el año 4 de la Olimpíada 101) y *ascendiendo hasta el cinturón de Orión allí desapareció*». El cometa del año 1618, que emergió de entre los rayos solares con una enorme cola,

parecía igualar y hasta superar a las estrellas de primera magnitud, pero han aparecido no pocos cometas mayores que tenían colas más breves. De ellos algunos dicen que igualaron a Júpiter, otros a Venus o incluso a la Luna.

Hemos dicho que los cometas son una clase de planetas que giran en órbitas muy excéntricas en torno al Sol. Y, al igual que de los planetas sin cola suelen ser menores los que giran en órbitas menores y más cercanas al Sol, así también los cometas, que en sus perihelios se acercan más al Sol, parece razonable que sean mucho menores, para que no lleguen a perturbar demasiado con su atracción al Sol. Omito, pues, la determinación de los diámetros transversales de las órbitas y los tiempos periódicos a partir de la comparación de los cometas que regresan por las mismas órbitas tras largos intervalos de tiempo. Pero mientras tanto, la siguiente proposición puede dar alguna luz en este asunto.

Proposición XLII. Problema XXII[19]

Corregir la trayectoria hallada de un cometa.

Operación 1. Tómese la posición del plano de la trayectoria tal como fue hallada por medio de la Proposición anterior; y selecciónense tres lugares del cometa determinados por observaciones cuanto más exactas sea posible, y lo más distantes entre sí; y sea A el tiempo entre la primera y la segunda observación, y B el tiempo entre la segunda y la tercera. Pero conviene que el cometa se halle en alguno de ellos en el perigeo, o por lo menos que no ande muy lejos del perigeo. Hállese mediante operaciones trigonométricas, a partir de esos lugares aparentes, los tres lugares verdaderos del cometa en el plano aquel de la trayectoria que se ha tomado. Después, por los lugares hallados, en torno al Sol o foco, mediante operaciones aritméticas realizadas con ayuda de la Proposición XXI del Libro I, trácese una sección cónica, y sean D y E sus áreas, delimitadas por los radios trazados al Sol desde los lugares hallados; a saber D el área entre la observación primera y segunda, y E el área entre la segunda y la tercera. Y sea T el tiempo total en el que el área entera D + E debe ser descrita con la velocidad del cometa hallada por la Proposición XVI del Libro I.

Operación 2. Auméntese la longitud de los nodos del plano de la trayectoria, añadiendo a aquella longitud 20′ ó 30′, que llamaremos P, pero conservando la anterior inclinación hacia el plano de la eclíptica. Después, a partir de los tres antedichos lugares observados del cometa, encuéntrense en este nuevo plano tres lugares verdaderos, como antes: después, trácese también una órbita que pase por esos tres lugares, y las áreas de la misma descritas entre observaciones, que sean ahora *d* y *e*, al igual que el tiempo entero *t*, en el que debiera describirse el área entera *d* + *e*.

Operación 3. Manténgase la longitud de los nodos de la primera operación y

auméntese la inclinación del plano de la trayectoria hacia el plano de la eclíptica, añadiendo a su inclinación inicial 20′ ó 30′, que denominaremos Q. Después, a partir de los antedichos tres lugares aparentes observados del cometa, encuéntrense tres lugares verdaderos en este plano, y trácense la órbita que pase por esos lugares y las dos áreas comprendidas entre las observaciones y que sean éstas δ y ε, y el tiempo total τ en el cual debiera describirse el área entera $\delta + \varepsilon$.

Ahora sea, C a 1 como A a B, y G a 1 como D a E, y g a 1 como d a e, y γ a 1 como δ a ε; y sea S el tiempo verdadero entre la primera y tercera observación; y observando con cuidado las reglas de los signos + y - hállense los números m y n, de modo que ocurra 2G - 2C = mG - mg + nG - $n\gamma$, y 2T - 2S = mT - mt + nT - $n\tau$. Y si en la primera operación I representa la inclinación del plano de la trayectoria hacia el plano de la eclíptica, y K representa la longitud de uno cualquiera de los nodos, I + nQ será la verdadera inclinación del plano de la trayectoria hacia el plano de la eclíptica, mientras K + mP será la verdadera longitud del nodo. Y finalmente, si en la primera operación, en la segunda y en la tercera, las cantidades R, r, y p, representan los «latera recta» de la trayectoria, mientras las cantidades $\frac{1}{\mathrm{L}}$, $\frac{1}{l}$, $\frac{1}{\lambda}$, representan sus lados transversales, respectivamente: el verdadero «latus rectum» será R + mr - mR + np - nR, mientras que el verdadero lado transverso de la trayectoria que describe el cometa será $\frac{1}{\mathrm{L} + ml - m\mathrm{L} + n\lambda - n\mathrm{L}}$. Pero una vez dado el lado transverso también está dado el tiempo periódico del cometa. Q. E. F.

Por lo demás, los tiempos periódicos de los cometas que giran y los lados transversos de las órbitas no se podrán determinar con bastante exactitud, a no ser por la mutua comparación de los cometas que aparecen en tiempos distintos. Si muchos cometas, después de iguales intervalos de tiempo, se encuentra que han descrito la misma órbita habrá que concluir que todos éstos eran uno y el mismo cometa girando en la misma órbita. Y entonces, por los tiempos de las revoluciones se tendrán dados los lados transversales de las órbitas y a partir de estos lados se determinarán las órbitas elípticas.

A este fin habrá, pues, que calcular las trayectorias de muchos cometas, bajo la hipótesis de que son parabólicas. Pues esta clase de trayectorias concuerdan muy aproximadamente siempre con los fenómenos. Esto se ve claramente, no sólo por la trayectoria parabólica del cometa del año 1680, que antes he comparado con las observaciones, sino también por la de aquel célebre cometa que apareció los años 1664 y 1665 y fue observado por *Hewelcke*. Este calculó, aunque con poca exactitud, sus longitudes y latitudes sobre la base de sus observaciones. Con esas mismas observaciones, nuestro *Halley* calculó de nuevo los lugares de ese cometa, y así al fin, por los lugares hallados de ese modo, determinó la trayectoria del cometa. Y halló

que su nodo ascendente estuvo en ♊ 21.º 13′ 55″, la inclinación de la órbita hacia el plano de la eclíptica de 21.º 18′ 40″, la distancia del perihelio al nodo en la órbita 49.º 27′ 30″. El perihelio en ♌ 8.º 40′ 30″, con latitud heliocéntrica austral de 16.º 1′ 45″ en tiempo igualado de *Londres* o a las 13h. 8′ p. m. de *Dantzig*, calendario antiguo, y el «latus rectum» de la parábola fue de 410 286, para una distancia media de la Tierra al Sol de 100 000. La medida en que los lugares del cometa calculados en esta órbita concuerdan con las observaciones se verá claramente por la siguiente tabla calculada por *Halley*.

En el mes de febrero de principios del año 1665, la primera estrella de Aries, a la que llamaré en lo sucesivo γ se hallaba en ♈ 28.º 30′ 15″, con una latitud boreal de 7.º 8′ 58″. La segunda de Aries estaba en ♈ 29.º 17′ 18″, con una latitud boreal de 8.º 28′ 16″. Y una estrella de séptima magnitud, que llamaré A, estaba en ♈ 28.º 24′ 45″, con latitud boreal de 8.º 28′ 33″. Y el cometa, el 7 de febrero a las 0h. 7′ 30″ de *París* (es decir, el día 7 de febrero a las 0h. 8′ 37″ de *Dantzig*) calendario antiguo, formaba con dichas estrellas γ y A un triángulo rectángulo en γ. Y la distancia del cometa a la estrella γ era igual a la distancia entre las estrellas γ y A, esto es, de 1.º 20′ 26″ en el paralelo de la latitud de la estrella γ. Por lo cual, si de la longitud de la estrella γ se resta 1.º 20′ 26″, quedará la longitud del cometa ♈ 27.º 9′ 49″. *Auzout*, basándose en esta observación suya, situó al cometa en ♈ 27.º 0′ aproximadamente. Y según el dibujo, con el que *Hooke* representó su movimiento, el cometa estaba ahora en ♈ 26.º 59′ 24″. Considerando la media entre ambos lo he puesto en ♈ 27.º 4′ 46″. *Auzout* por la misma observación situó la latitud del cometa en 7.º y 4′ ó 5′ norte. Más exacto hubiera sido en 7.º 3′ 29″, puesto que la diferencia de latitudes del cometa y de la estrella γ era igual a la diferencia de longitudes de las estrellas γ y A.

Tiempo aparente en Dantzig (s.a.)	Distancias del cometa observadas desde		Lugares observados		Lugares calculados en la órbita
Diciembre d h m		° ′ ″		° ′ ″	° ′ ″
3. 18. 29½	El corazón de León	46,24,20	Long. ♎	7,01,00	♎ 7, 1,29
	La punta de Virgo	22,52,10	Lat. S.	21,39, 0	21,38,50
4. 18. 1½	El corazón de León	46, 2,45	Long. ♎	6,15, 0	♎ 6,16, 5
	La punta de Virgo	23,52,40	Lat. S.	22,24, 0	22,24, 0
7. 17. 48	El corazón de León	44,48, 0	Long. ♎	3, 6, 0	♎ 3, 7,33
	La punta de Virgo	27,56,40	Lat. S.	25,22, 0	25,21,40
17. 14. 13	El corazón de León	53,15,15	Long. ♌	2,56, 0	♌ 2,56, 0
	El hombro derecho de Orión	45,43,30	Lat. S.	49,25, 0	49,25, 0
19. 9. 25	Proción	35,13,50	Long. ♊	28,40,30	♊ 28,43, 0
	La estrella brillante de la mandíbula de Ballena	52,56, 0	Lat. S.	45,48, 0	45,46, 0
20. 9. 53½	Proción	40,49, 0	Long. ♊	13,03, 0	♊ 13, 5, 0
	La estrella brillante de la mandíbula de Ballena	40,04, 0	Lat. S	39,54, 0	39,53, 0
21. 9. 9½	El hombro derecho de Orión	26,21,25	Long. ♊	2,16, 0	♊ 2,18,30
	La estrella brillante de la mandíbula de Ballena	29,28, 0	Lat. S.	33,41, 0	33,39,40
22. 9. 0	El hombro derecho de Orión	29,47, 0	Long. ♉	24,24, 0	♉ 24,27, 0
	La estrella brillante de la mandíbula de Ballena	20,29,30	Lat. S.	27,45, 0	27,46, 0
26. 7. 58	La estrella brillante de Aries	23,20, 0	Long. ♉	9, 0, 0	♉ 9, 2,28
	Aldebarán	26,44, 0	Lat. S.	12,36, 0	12,34,13
27. 6. 45	La estrella brillante de Aries	20,45, 0	Long. ♉	7, 5,40	♉ 7, 8,45
	Aldebarán	28,10, 0	Lat. S.	10,23, 0	10,23,13
28. 7. 39	La estrella brillante de Aries	18,29, 0	Long. ♉	5,24,45	♉ 5,27,52
	Palilicium	29,37, 0	Lat. S.	8,22,50	8,23,37
31. 6. 45	El cinturón de Andrómeda	30,48,10	Long. ♉	2, 7,40	♉ 2, 8,20
	Palilicium	32,53,30	Lat. S.	4,13, 0	4,16,25
Enero 1665 7. 7. 37½	El cinturón de Andrómeda	25,11, 0	Long. ♈	28,24,47	♈ 28,24, 0
	Palilicium	37,12,25	Lat. N.	0,54, 0	0,53, 0
13. 7. 0	La cabeza de Andrómeda	28, 7,10	Long. ♈	27, 6,54	♈ 27, 6,39
	Palilicium	38,55,20	Lat. N.	3, 6,50	3, 7,40
24. 7. 29	El cinturón de Andrómeda	20,32,15	Long. ♈	26,29,15	♈ 26,28,50
	Palilicium	40, 5, 0	Lat. N.	5,25,50	5,26, 0
Febrero 7. 8. 37			Long. ♈	27, 4,46	♈ 27,24,55
			Lat. N.	7, 3,29	7, 3,15
22. 8. 46			Long. ♈	28,29,46	♈ 28,29,58
			Lat. N.	8,12,36	8,10,25
Marzo 1. 8. 16			Long. ♈	29,18,15	♈ 29,18,20
			Lat. N.	8,36,26	8,36,12
7. 8. 37			Long. ♉	0, 2,48	♉ 0, 2,42
			Lat. N.	8,56,30	8,56,56

(Tabla de Halley)

El 22 de febrero a las 7h. 30′ de *Londres*, es decir, el 22 de febrero a las 8h. 46′ en *Dantzig*, la distancia del cometa a la estrella A, según las observaciones de *Hooke*

representadas por él mismo en un dibujo, y también según observaciones de *Auzout* representadas por *Petit* en otro dibujo, era una quinta parte de la distancia entre la estrella A y la primera de Aries, es decir, 15′ 57″. Y la distancia del cometa hasta la línea que unía la estrella A con la primera de Aries era una cuarta parte de esa quinta parte, es decir, 4′. Por tanto, el cometa se hallaba en ♈ 28.º 29′ 46″, con latitud norte de 8.º 12′ 36″.

En marzo 1d. 7h. 0′, de *Londres*, esto es, marzo Id. 8h. 16′ de *Dantzig*, el cometa fue observado cerca de la segunda estrella de Aries, a una distancia entre ellos que era a la distancia entre la primera y la segunda de Aries, esto es, a 1.º 33′ como 4 a 45 según *Hooke*, y como 2 a 23 según *Gottignies*. Por lo que la distancia del cometa a la segunda de Aries era de 8′ 16″ según *Hooke*, o de 8′ 5″ según *Gottignies*, o la media de 8′ 10″. Pero el cometa, según *Gottignies*, ya había rebasado la segunda de Aries en un espacio de casi la cuarta o quinta parte del recorrido de un día, es decir, 1′ 35″ aproximadamente (con lo que concuerda bien *Auzout*) o algo menos según *Hooke*, quizá 1′. Por lo cual si a la longitud de la primera de Aries se añade 1′ y a su latitud 8′ 10″, se tendrá la longitud del cometa en ♈ 29.º 18′ y la latitud en 8.º 36′ 26″ norte.

En marzo 7d. 7h. 30′ de *París* (es decir, marzo, 7d. 8h. 37′ de *Dantzig*) según las observaciones de *Auzout* la distancia del cometa de la segunda de Aries era igual a la distancia de la segunda de Aries a la estrella A, esto es, 52′ 29″. Y la diferencia de longitudes del cometa y de la segunda de Aries era de 45′ ó 46′ o quizá la media de 45′ 30″. Por lo cual el cometa se hallaba en ♉ 0.º 2′ 48″. Del dibujo de las observaciones de *Azout*, construido por *Petit*, *Hewelcke* dedujo la latitud del cometa en 8.º 54′. Pero el dibujante curvó irregularmente el camino del cometa hacia el final de su movimiento, y en el dibujo de las observaciones de *Azout* construido por *Hewelcke*, éste corrigió la curvatura irregular, y así hizo que la latitud del cometa fuera de 8.º 55′ 30″. Y corrigiendo la irregularidad un poco más, la latitud vendría a: resultar de 8.º 56′ o quizá 8.º 57.

Este cometa fue visto también el 9 de marzo, y entonces debió situarse en ♉ 0.º 18′, con latitud boreal de 9.º 3½ aproximadamente.

Este cometa fue visible durante tres meses y recorrió casi seis signos, y en un día se desplazó casi veinte grados. Su curso se desvió del Círculo Máximo, curvándose notablemente hacia el norte; y su movimiento al final, de retrógrado pasó a directo. Y no obstante un curso tan insólito, la teoría de principio a fin no vino a concordar menos con las observaciones de lo que suelen concordar las teorías de los planetas con las observaciones, como verá quien examine la tabla. Sin embargo hay que restar casi dos minutos primeros, cuando el cometa fue más veloz; cosa que se puede hacer restando doce minutos segundos del ángulo entre el nodo ascendente y el perihelio, o estableciendo dicho ángulo en 49.º 27′ 18″. La paralaje anual de uno y otro cometa (éste y el anteriormente mencionado) fue grande, y por ello se demuestra el movimiento anual de la Tierra en el Orbe Magno.

Se confirma también la teoría por el movimiento del cometa que apareció el año

1683. Este fue retrógrado, en una órbita cuyo plano casi formaba ángulo recto con el plano de la eclíptica. Su nodo ascendente (por cálculos de *Halley*) estaba en ♍ 23.º 23′; la inclinación de la órbita hacia la eclíptica era de 83.º 11′; el perihelio en ♊ 25.º 29′ 30″; la distancia del perihelio al Sol de 56 020, para un radio del Orbe Magno de 100 000, y el momento del perihelio el día 2 de julio a las 3h. 50′. Los lugares del cometa en esta órbita calculados por *Halley* y comparados con los observados por *Flamsteed*, se muestran en la tabla siguiente.

1683 Tiempo ecuatorial		Lugar del Sol	Longitud calculada del cometa	Latitud norte calculada	Longitud observada del cometa	Latitud norte observada	Diferencia longitud	Diferencia latitud
d h m		° ′ ″	° ′ ″	° ′ ″	° ′ ″	° ′ ″	′ ″	′ ″
Jul.	13,12,55	♌ 1, 2,30	♋ 13, 5,42	29,28,13	♋ 13, 6,42	29,28,20	+1, 0	+0, 7
	15,11,15	2,53,12	11,37,48	29,34, 0	11,39,43	29,34,50	+1,55	+0,50
	17,10,20	4,45,45	10, 7, 6	29,33,30	10, 8,40	29,34, 0	+1,34	+0,30
	23,13,40	10,38,21	5,10,27	28,51,42	5,11,30	28,50,28	+1, 3	−1,14
	25,14, 5	12,35,28	3,27,53	24,24,47	3,27, 0	28,23,40	−0,53	−1, 7
	31, 9,42	18, 9,22	♊ 27,55, 3	26,22,52	♊ 27,54,24	26,22,25	−0,39	−0,27
	31,14,55	18,21,53	27,41, 7	26,16,57	27,41, 8	26,14,50	+0, 1	−2, 7
Ag.	2,14,56	20,17,16	25,29,32	25,16,19	25,28,46	25,17,28	−0,46	+1, 9
	4,10,49	22, 2,50	23,18,20	24,10,49	23,16,55	24,12,19	−1,25	+1,30
	6,10, 9	23,56,45	20,42,23	22,47, 5	20,40,32	22,49,05	−1,51	+2, 0
	9,10,26	26,50,52	16, 7,57	20, 6,36	16, 5,55	20, 6,10	+2, 2	−0,27
	15,14, 1	♎ 2,47,13	3,30,48	11,37,33	3,26,18	11,32, 1	−4,30	−5,32
	16,15,10	3,48, 2	0,43, 7	9,34,16	0,41,55	9,34,13	−1,12	−0, 3
	18,15,44	5,45,33	♉ 24,52,53	5,11,15	♉ 24,49, 5	5, 9,11	−3,48	−2, 4
				Sur		Sur		
	22,14,44	9,35,49	11, 7,14	5,16,58	11, 7,12	5,16,58	−0, 2	−0, 3
	23,15,52	10,36,48	7, 2,18	8,17,09	7, 1,17	8,16,41	−1, 1	−0,28
	26,16, 2	13,31,10	♈ 24,45,31	16,38, 0	♈ 24,44, 0	16,38,20	−1,31	+0,20

También se confirma la teoría por el movimiento del cometa retrógrado que apareció el año 1682. Su nodo ascendente (calculado por *Halley*) se hallaba en ♉ 21.º 16′ 30″. La inclinación de la órbita hacia el plano de la eclíptica era de 17.º 56′ 0″. El perihelio en ♒ 2.º 52′ 50″. La distancia del perihelio al Sol de 58 328 para un radio del Orbe Magno de 100 000. Y el tiempo igualado del perihelio el 4 de septiembre a las 7h. 39′. Pero los lugares calculados, a partir de las observaciones de *Flamsteed*, comparados con los calculados mediante la teoría se muestran en la tabla siguiente. (*Vid. página* 776).

Y también se confirma la teoría por el movimiento retrógrado del cometa que apareció el año 1723. Su nodo ascendente (calculado por *Bradley*, profesor *Saviliano* de astronomía en *Oxford*) estaba en ♈ 14.º 16′. La inclinación de la órbita hacia el plano de la eclíptica de 49.º 59′. El perihelio en ♉ 12.º 15′ 20″. La distancia del perihelio al Sol de 998 651, para un radio del Orbe Magno de 1 000 000, y el tiempo igualado del perihelio el 16 de septiembre a las 16h. 10′. Y los lugares del cometa en esta órbita calculados por *Bradley*, comparados con los observados por él mismo, por su tío *Pound* y por *Halley* se muestran en la tabla siguiente. (*Vid. página* 777).

Con estos ejemplos queda harto claro que los movimientos de los cometas se ponen de manifiesto mediante la teoría recién expuesta con no menor exactitud que la conseguida por las teorías de los movimientos de los planetas. Y por lo tanto, las órbitas de los cometas pueden ser calculadas mediante esta teoría, igual que conocer, por fin, el tiempo periódico de un cometa que gira en una órbita cualquiera, y se conocerán entonces los lados transversales de las órbitas elípticas al igual que las alturas de los afelios.

El cometa retrógrado que apareció el año 1607 describió una órbita cuyo nodo ascendente (calculado por *Halley*) estaba en ♉ 20.º 21′; la inclinación del plano de la órbita hacia el plano de la eclíptica era de 17.º 2′; el perihelio en ♒ 2.º 16′; y la distancia del perihelio al Sol era de 58 680 para un radio del Orbe Magno de 100 000. Y el cometa estaba en el perihelio el 16 de octubre a las 3h. 50′. Concuerda muy aproximadamente esta órbita con la órbita del cometa que apareció en 1682. Si estos dos cometas fueran uno y el mismo, este cometa giraría en un período de 75 años, y el eje mayor de dicha órbita será al eje mayor del Orbe Magno como la raíz cúbica de 75 x 75 a 1, o como 1778 a 100 aproximadamente. Y la distancia del afelio de este cometa al Sol será a la distancia media de la Tierra al Sol como 35 a 1 aproximadamente. Conociendo estas cosas no será muy difícil determinar la órbita elíptica de este cometa. Y esto será así si resulta que después de un espacio de 75 años el cometa regresara de nuevo por la misma órbita. Los demás cometas parece que giran en períodos más largos y ascienden más arriba.

1682 Tiempo ap.	Lugar del Sol	Longitud calculada del cometa	Latitud norte calculada	Longitud observada del cometa	Latitud norte observada	Diferencia longitud	Diferencia latitud
d h m	° ′ ″	° ′ ″	° ′ ″	° ′ ″	° ′ ″	′ ″	′ ″
Ag. 19,16,38	♍ 7, 0, 7	♌ 18,14,28	25,50, 7	♌ 18,14,40	25,49,55	−0,12	+0,12
20,15,38	7,55,52	24,46,23	26,14,42	24,46,22	26,12,52	+0, 1	+1,50
21, 8,21	8,36,14	29,37,15	26,20, 3	29,38, 2	26,17,37	−0,47	+2,26
22, 8, 8	9,33,55	♍ 6,29,53	26, 8,42	♍ 6,30, 3	26, 7,12	−0,10	+1,30
29, 8,20	16,22,40	♎ 12,37,54	18,37,47	♎ 12,37,49	18,34, 5	+0, 5	+3,42
30, 7,45	17,19,41	15,36, 1	17,26,43	15,35,18	17,27,17	+0,43	−0,34
Sept. 1, 7,33	19,16, 9	20,30,53	15,13, 0	20,27, 4	15, 9,49	+3,49	+3,11
4, 7,22	22,11,28	25,42, 0	12,23,48	25,40,58	12,22, 0	+1, 2	+1,48
5, 7,32	23,10,29	27, 0,46	11,33,08	26,59,24	11,33,51	+1,22	−0,43
8, 7,16	26, 5,58	29,58,44	9,26,46	29,58,45	9,26,43	−0, 1	+0, 3
9, 7,26	27, 5, 9	♏ 0,44,10	8,49,10	♏ 0,44, 4	8,48,25	+0, 6	+0,45

1723 Tiempo ecuatorial	Longitud observada del cometa	Latitud norte observada	Longitud calculada del cometa	Latitud norte calculada	Diferencia longitud	Diferencia latitud
d h m	° ′ ″	° ′ ″	° ′ ″	° ′ ″	″	″
Oct. 9,8, 5	♒ 7,22,15	5, 2, 0	♒ 7,21,26	5, 2,47	+49	−47
10,6,21	6,41,12	7,44,13	6,41,42	7,43,18	−50	+55
12,7,22	5,39,58	11,55, 0	5,40,19	11,54,55	−21	+05
14,8,57	4,59,49	14,43,50	5, 0,37	14,44, 1	−48	−11
15,6,35	4,47,41	15,40,51	4,47,45	15,40,55	− 4	− 4
21,6,22	4, 2,32	19,41,49	4, 2,21	19,42, 3	+11	−14
22,6,24	3,59, 2	20, 8,12	3,59,10	20, 8,17	− 8	− 5
25,8, 2	3,55,29	20,55,18	3,55,11	20,55, 9	+18	+ 9
29,8,56	3,56,17	22,20,27	3,56,42	22,20,10	−25	+17
30,6,20	3,58, 9	22,32,28	3,58,17	22,32,12	− 8	+16
Nov. 5,5,53	4,16,30	23,38,33	4,16,23	23,38, 7	+ 7	+26
8,7, 6	4,29,36	24, 4,30	4,29,54	24, 4,40	−18	−10
14,6,20	5, 2,16	24,48,46	5, 2,51	24,48,16	−35	+30
20,7,45	5,42,20	25,24,45	5,43,13	25,25,17	−53	−32
Dic. 7,6,45	8, 4,13	26,54,18	8, 3,55	26,53,42	+18	+36

Por lo demás, los cometas, debido a su gran número, y a la gran distancia de los afelios hasta el Sol así como a su gran permanencia en los afelios, deben sufrir perturbaciones no pequeñas entre ellos por causa de las gravedades y aumentar unas veces o disminuir otras, tanto sus excentricidades como sus tiempos periódicos. Por consiguiente, no hay que esperar que el mismo cometa regrese por la misma órbita y en los mismos tiempos periódicos de modo exacto. Basta con que no sobrevengan mutaciones mayores que las que puedan deberse a las causas susodichas.

Y con esto se da cuenta de por qué los cometas no estén comprendidos en el zodíaco como los planetas, sino que salgan de él y con movimientos variados se dirijan hacia todas las regiones del cielo. Y ello con objeto de que en sus afelios, cuando se mueven muy lentamente, disten entre ellos lo más posible y se atraigan entre ellos cuanto menos. Por esta causa, los cometas que más descienden y por lo mismo se mueven en los afelios más lentamente, deberán ascender más.

El cometa que apareció el año 1680 distaba en su perihelio del Sol menos de la sexta parte del diámetro del Sol; y debido a la gran velocidad en esas proximidades y a alguna densidad de la atmósfera solar, debió sufrir alguna resistencia y retardarse un tanto y acercarse más al Sol: y en cada una de las revoluciones acercándose un poco, al fin acabará cayendo en el cuerpo solar. Pero en el afelio, cuando se mueve muy lentamente, puede a veces ser retardado por la atracción de otros cometas y después caer hacia el Sol. También de este modo las estrellas fijas, que poco a poco se extinguen en luz y vapores pueden realimentarse con cometas que caen en ellas, y repuestas con ese nuevo alimento ser tenidas por estrellas nuevas. De esta clase son las estrellas fijas que aparecen de repente y, al principio, brillan enormemente y después poco a poco se desvanecen. Tal fue la estrella de la silla de Casiopea que no vio *Cornelio Gemma* el 8 de noviembre de 1572 observando esa parte del cielo en una noche diáfana; pero a la noche siguiente (9 de noviembre) la vio más brillante que todas las fijas y apenas de menos luz que Venus. *Tycho Brahe* la vio el día 11 del mismo mes cuando más brillaba; y a partir de ese momento decreciendo poco a poco la vio desaparecer en el espacio de 16 meses. En el mes de noviembre, recién aparecida, igualaba en brillo a Venus. En diciembre, un tanto disminuida, parecía igualar a Júpiter. El año 1573, en enero era menor que Júpiter y mayor que Sirio, a la cual ya resultaba igual en febrero y primeros de marzo. En los meses de abril y mayo parecía igual a las estrellas de segunda magnitud, en junio, julio y agosto a las estrellas de tercera magnitud, en septiembre, octubre y noviembre a las de cuarta, en diciembre y en enero de 1574 a las de quinta, en febrero a las de sexta, y en marzo desapareció de vista. Al principio su color era claro, blanquecino y brillante y después amarillo, y el año 1573 en marzo rojiza como la de Marte o Aldebarán; en mayo adquirió una palidez blanquecina, como la que vemos en Saturno, color que conservó hasta el fin, pero haciéndose cada vez más oscura. Así fue también la estrella del pie derecho de Serpentario, observada por los discípulos de *Kepler* el año 1604 el día 30 de septiembre, calendario antiguo, cuando comenzó a aparecer, y

observaron que superaba a Júpiter con su luz, siendo así que la noche precedente no se veía. Desde ese momento empezó a decrecer y en el espacio de 15 ó 16 meses desapareció de la vista. Por una estrella nueva semejante y extremadamente brillante se dice que se vio impulsado *Hiparco* para observar las fijas y confeccionar un catálogo. Pero las fijas que aparecen y desaparecen alternativamente y que crecen poco a poco y apenas superan nunca a las de tercera magnitud, parecen ser de otra clase, y al girar muestran alternativamente sus partes oscuras y sus partes luminosas. Pero los vapores que surgen del Sol, de las estrellas fijas y de las colas de los cometas, pueden caer por su gravedad sobre las atmósferas de los planetas y condensarse allí y convertirse en agua y espíritus húmedos y después por calentamiento lento pasar poco a poco a sales, sulfuros, tinturas, limo, barro, arcilla, arena, piedras, corales y otras sustancias terrestres.

ESCOLIO GENERAL[20]

La hipótesis de los vórtices se ve acosada por muchas dificultades. Para que cada planeta, con un radio trazado hasta el Sol, describa áreas proporcionales a los tiempos, los tiempos periódicos de las partes del vórtice deberían estar en razón cuadrada de las distancias al Sol. Para que los tiempos periódicos de los planetas estén en razón de la potencia $^{2}/_{3}$ de las distancias al Sol, los tiempos periódicos de las partes del vórtice deberían estar en razón de la potencia $^{2}/_{3}$ de las distancias. Para que los vórtices menores en giro en torno a Saturno, Júpiter y otros planetas se conserven y floten tranquilamente en el vórtice del Sol, los tiempos periódicos de las partes del vórtice solar deben ser iguales. Las revoluciones del Sol y de los planetas en torno a sus ejes, que deberían concordar con los movimientos de los vórtices, discrepan de todas estas proporciones. Los movimientos de los cometas son sumamente regulares y observan las mismas leyes que los movimientos de los planetas, y no pueden explicarse por los vórtices. Los cometas se desplazan con movimientos muy excéntricos hacia todas las partes de los cielos, cosa que no podría ocurrir si no se suprimen los vórtices.

Los proyectiles, en nuestra atmósfera, sufren únicamente la resistencia del aire. Suprimiendo el aire, como ocurre en el vacío de *Boyle*, cesa la resistencia, de suerte que una leve pluma y el denso oro caen en este vacío con la misma velocidad. E igual es la cosa en los espacios celestes que están más allá de la atmósfera terrestre. Todos los cuerpos en estos espacios deben moverse con entera libertad; y por lo mismo los planetas y los cometas deberán girar perpetuamente según las leyes expuestas más arriba en órbitas de especie y posición dadas. Ciertamente perseverarán en sus órbitas por las leyes de la gravedad, pero de ningún modo pudieron por estas leyes adquirir inicialmente la situación regular de las órbitas.

Los seis planetas principales giran en torno al Sol en círculos concéntricos al Sol, con la misma dirección de movimiento y aproximadamente en el mismo plano. Diez lunas giran en torno a la Tierra, Júpiter y Saturno en círculos concéntricos, con la misma dirección de movimiento, en los planos de las órbitas de los planetas muy aproximadamente. Y todos estos movimientos regulares no tienen un origen debido a causas mecánicas; toda vez que los cometas circulan en órbitas muy excéntricas libremente y en todas direcciones del firmamento. Con este tipo de movimiento los cometas pasan rápida y fácilmente por las órbitas de los planetas, y en sus afelios,

cuando sus movimientos son más lentos y se detienen por más tiempo, distan entre ellos inmensamente, para que sea mínima la atracción mutua. Tan elegante combinación de Sol, planetas y cometas sólo pudo tener origen en la inteligencia y poder de un ente inteligente y poderoso. Y si las estrellas fijas fueren centros de sistemas semejantes, todos ellos construidos con un esquema similar, estarán sometidos al dominio de *Uno*: sobre todo si la luz de las fijas es de la misma naturaleza que la luz del Sol y todos los sistemas emiten luz hacia todos mutuamente. Y para que los sistemas de las fijas no caigan por la gravedad uno sobre otro, él los habría colocado a inmensas distancias uno de otro.

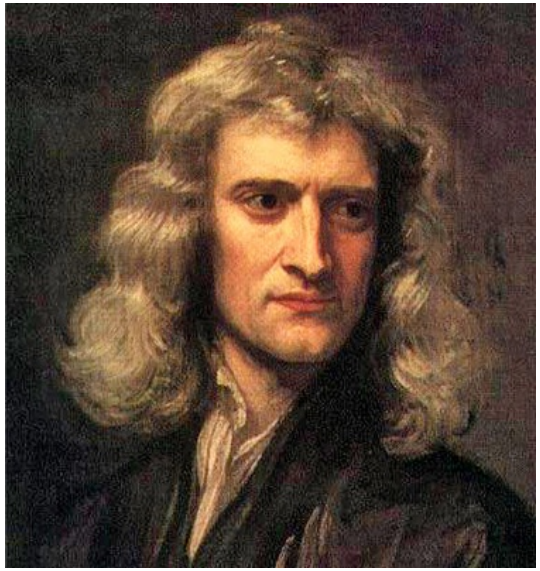
El lo rige todo, no como alma del mundo, sino como dueño de todos. Y por su dominio, suele ser llamado señor dios «*παντσκράτωρ*[a]». Pues dios es una palabra relativa y está en relación con los siervos: y deidad es la dominación de dios, no sobre su propio cuerpo, como creen aquellos para quienes dios es el alma del mundo, sino sobre los siervos. Dios sumo es un ente eterno, infinito, absolutamente perfecto: pero un ente cualquiera perfecto sin dominio no es dios señor. Pues decimos, dios mío, dios vuestro, dios de *Israel*, dios de dioses, y señor de señores; pero no decimos eterno mío, eterno vuestro, eterno de *Israel*, eterno de dioses; no decimos infinito mío, o perfecto mío. Estas denominaciones no tienen relación con los siervos. La voz dios[b] significa con frecuencia dueño: pero todo dueño no es dios. La dominación de un ente espiritual constituye un dios, la verdadera al verdadero, la suma al sumo, la ficticia al ficticio. Y de la verdadera dominación se sigue que un dios verdadero es vivo, inteligente y poderoso; de las demás perfecciones que es sumo o sumamente perfecto. Es eterno e infinito, omnipotente y omnisciente, es decir, dura desde la eternidad hasta la eternidad y está presente desde el principio hasta el infinito: lo rige todo; lo conoce todo, lo que sucede y lo que puede suceder. No es la eternidad y la infinitud, sino eterno e infinito; no es la duración y el espacio, sino que dura y está presente. Dura siempre y está presente en todo lugar, y existiendo siempre y en todo lugar, constituye a la duración y al espacio. Puesto que cada partícula de espacio existe *siempre*, y cada momento indivisible de duración está en *algún lugar*, ciertamente el constructor y señor de todas las cosas no será *nunca, ningún lugar*. Toda alma siente en distintos tiempos y en diversos órganos de los sentidos y de los movimientos y es la misma persona indivisible. Las partes se dan sucesivamente en la duración, coexistentes en el espacio, pero ni unas ni otras en la persona humana o en su principio pensante; y mucho menos en la sustancia pensante de dios. Todo hombre, en tanto que cosa sentiente, es uno y el mismo hombre durante su vida en todos y cada uno de los órganos de sus sentidos. Dios es uno y el mismo dios siempre y en todo lugar. Es omnipotente no sólo *virtualmente* sino *sustancialmente*: pues lo virtual no puede subsistir sin la sustancia. En él[c] se hallan contenidas y se mueven todas las cosas, pero sin mutua interferencia. Dios nada sufre por el movimiento de los cuerpos: éstos no experimentan resistencia alguna por la omnipresencia de dios. Está reconocido que un dios sumo existe necesariamente: y con la misma necesidad existe

siempre y en *todo lugar*. De donde también es todo él semejante a sí mismo, todo ojo, todo oído, todo cerebro, todo brazo, todo fuerza de sentir, de entender, de actuar, pero en modo alguno a la manera humana, o a la manera corporal, sino de una manera totalmente desconocida para nosotros. Como el ciego no tiene idea de los colores, de igual modo nosotros no tenemos idea de los modos con los que dios sapientísimo siente y entiende todas las cosas. Absolutamente desprovisto de todo cuerpo y figura corporal, no puede por ello ser visto ni oído, ni tocado, ni debe ser venerado bajo forma de cosa corpórea alguna. Tenemos ideas de sus atributos, pero que sea la sustancia de alguna cosa lo ignoramos por completo. Solamente vemos las formas y colores de los cuerpos, sólo oímos los sonidos, sólo tocamos las superficies externas, olemos los meros olores y gustamos los sabores: pero las sustancias interiores no las conocemos con ningún sentido, ni con ninguna acción refleja; mucho menos tenemos una idea de la sustancia de dios. A éste le conocemos tan sólo por sus propiedades y atributos y por las sapientísimas y óptimas estructuras y causas finales de las cosas y le admiramos por las perfecciones, pero le veneramos y damos culto por el dominio. Pero le damos culto como siervos, y un dios sin dominio, providencia y causas finales no es nada más que hado y naturaleza. De la ciega necesidad metafísica, que es también la misma siempre y en todo lugar, no surge ninguna variación de las cosas. Toda la variedad de cosas, establecidas según los lugares y los tiempos, solamente pudo originarse de las ideas y voluntad de un ente necesariamente existente. Sé dice alegóricamente que dios ve, oye, habla, ríe, ama, tiene odio, desea, da, recibe, se alegra, está airado, lucha, fabrica, funda, construye. Pues todo discurso sobre dios se produce mediante alguna semejanza a partir de las cosas humanas, ciertamente no perfecta, pero algo semejante. Esto respecto a dios de quien, efectivamente, corresponde hablar en filosofía natural a partir de los fenómenos.

Hasta aquí he expuesto los fenómenos de los cielos y de nuestro mar por la fuerza de la gravedad, pero todavía no he asignado causa a la gravedad. Efectivamente esta fuerza surge de alguna causa que penetra hasta los centros del Sol y de los planetas sin disminución de la fuerza; y la cual actúa, no según la cantidad de las *superficies* de las partículas hacia las cuales actúa (como suelen hacer las causas mecánicas) sino según la cantidad de materia *sólida*; y cuya acción se extiende por todas partes hasta distancias inmensas, decreciendo siempre como el cuadrado de las distancias. La gravedad hacia el Sol se compone de las gravedades hacia cada una de las partículas del Sol, y separándose del Sol decrece exactamente en razón del cuadrado de las distancias hasta más allá de la órbita de Saturno, como se evidencia por el reposo de los afelios de los planetas, y hasta los últimos afelios de los cometas, si semejantes afelios están en reposo. Pero no he podido todavía deducir a partir de los fenómenos la razón de estas propiedades de la gravedad y yo no imagino hipótesis. Pues, lo que no se deduce de los fenómenos, ha de ser llamado *Hipotesis*; y las hipótesis, bien metafísicas, bien físicas, o de cualidades ocultas, o mecánicas, no tienen lugar dentro de la *Filosofía experimental*. En esta filosofía las proposiciones se deducen de los

fenómenos, y se convierten en generales por inducción. Así, la impenetrabilidad, la movilidad, el ímpetu de los cuerpos y las leyes de los movimientos y de la gravedad, llegaron a ser esclarecidas. Y bastante es que la gravedad exista de hecho y actúe según las leyes expuestas por nosotros y sea suficiente para todos los movimientos de los cuerpos celestes y de nuestro mar.

Bien podríamos ahora añadir algo de cierto espíritu sutilísimo que atraviesa todos los cuerpos gruesos y permanece latente en ellos; por cuya fuerza y acciones las partículas de los cuerpos se atraen entre ellas a las mínimas distancias y una vez que están contiguas permanecen unidas; y los cuerpos eléctricos actúan a distancias mayores, tanto repeliendo como atrayendo a los corpúsculos vecinos; y la luz se emite, se refleja, se refracta e inflexiona y calienta a los cuerpos; y toda sensación es excitada, y los miembros de los animales se mueven a voluntad, a saber mediante las vibraciones de ese espíritu propagadas por los filamentos sólidos de los nervios desde los órganos externos de los sentidos hasta el cerebro y desde el cerebro hacia los músculos. Pero esto no puede exponerse en pocas palabras; y tampoco está disponible un número suficiente de experimentos mediante los cuales deben determinarse y mostrarse exactamente las leyes de las acciones de este espíritu.



Sir ISAAC NEWTON (Woolsthorpe, Lincolnshire; 25 de diciembre de 1642 (4 de enero de 1643) -Kensington, Londres; 20 de marzo (31 de marzo de 1727). Fue un físico, filósofo, teólogo, inventor, alquimista y matemático inglés. Es autor de los *Philosophiæ naturalis principia mathematica*, más conocidos como los Principia, donde describe la ley de la gravitación universal y estableció las bases de la mecánica clásica mediante las leyes que llevan su nombre. Entre sus otros descubrimientos científicos destacan los trabajos sobre la naturaleza de la luz y la óptica (que se presentan principalmente en su obra *Opticks*) y el desarrollo del cálculo matemático.

Newton comparte con Gottfried Leibniz el crédito por el desarrollo del cálculo integral y diferencial, que utilizó para formular sus leyes de la física. También contribuyó en otras áreas de la matemática, desarrollando el teorema del binomio y las fórmulas de Newton-Cotes.

Entre sus hallazgos científicos se encuentran el descubrimiento de que el espectro de color que se observa cuando la luz blanca pasa por un prisma es inherente a esa luz, en lugar de provenir del prisma (como había sido postulado por Roger Bacon en el siglo XIII); su argumentación sobre la posibilidad de que la luz estuviera compuesta por partículas; su desarrollo de una ley de convección térmica, que describe la tasa de enfriamiento de los objetos expuestos al aire; sus estudios sobre la velocidad del sonido en el aire; y su propuesta de una teoría sobre el origen de las estrellas. Fue también un pionero de la mecánica de fluidos, estableciendo una ley sobre la viscosidad.

Newton fue el primero en demostrar que las leyes naturales que gobiernan el movimiento en la Tierra y las que gobiernan el movimiento de los cuerpos celestes son las mismas. Es, a menudo, calificado como el científico más grande de todos los tiempos, y su obra como la culminación de la revolución científica. El matemático y físico matemático Joseph Louis Lagrange (1736-1813), dijo que «Newton fue el más grande genio que ha existido y también el más afortunado dado que sólo se puede encontrar una vez un sistema que rija el mundo».

Notas de la introducción

[1] Véanse las *Biografías*, reseñadas en la Bibliografía, pero nos hemos ceñido a las más destacadas. Datos menores, aunque significativos habría que tomarlos de las biografías de algunos de sus contemporáneos, en particular de las de Hooke, Whiston, Flamsteed. <<

[2] Al menos 16 retratos y 6 medallas en bronce se hicieron de Newton en vida del mismo. Empezó posando para Kneller en 1689 y lo hizo de nuevo en 1702, 1720 y 1723. Para Gandy en 1706, para Jervas en 1703. Para Richter en 1741, para Murray en 1718, para Thornhill en 1710 (bis), para Vanderbank, en 1725 (bis) y 1726, además de dos anónimos y un boceto de Stukeley de 1719, para Seeman en 1724. Aunque no son retratos geniales como el retratado, permiten entre todos entrever cierta altivez y fría lejanía en Newton. Pero no hay que olvidar que se encargaban para algún fin. Y esto era lo primero que debía intentar el retrato. <<

[3] Los apuntes de Stukeley, William, se han publicado con el título *Memoirs of sir Isaac Newton's Life*, editados por A. Hastings White, en Londres 1936. Las notas de Conduitt se hallan en los manuscritos *Keynes*, en la biblioteca del King College de Cambridge y no han sido publicados más que fragmentariamente en citas diversas. De esa procedencia son la mayor parte de anécdotas sobre su vida en Londres y también las noticias que recibió Fontenelle para su *Elogio*. <<

[4] Aparece en el manuscrito *Keynes*, caligrafía de Conduitt. Ms. 130-132, págs. 17-18. <<

[5] Gracias a un cuaderno de notas de Newton de esta época que se conserva en la Pierpont Morgan Library de Nueva York, sabemos que uno de los libros que centraron su interés en esta época y que inspiraron su genio mecánico como constructor de artefactos fue *The Mysteries of Nature and Art*, de John Bate. Vde. E. N. da C. Andrade, «Newton's Early Notebook», en *Nature*, 135 (1935), pág. 360. <<

[6] Cfr. Stukeley, o. c., pág. 43 y Conduitt, *Keynes Ms*. 130-132, pág. 24. El reloj de agua y otro de sol junto con una reproducción a escala del molino, llamaron la atención de todos. <<

[7] *Ibíd*. Su afición por la pintura y el dibujo le valió luego en el diseño de los diagramas de sus experimentos. <<

[8] Entre los textos que manejó Newton hubo uno de orientación aristotélica cuyo autor era un clásico en el siglo XVII inglés. Johannes Margirus. Su *Physiologiae peripateticae* no satisfacía a Newton y menos sus datos. Entre líneas escribe: «Galileo dice que el diámetro aparente de las estrellas de primera magnitud es de 5″…etc. y M. Azout estima el diámetro de…, etc». Cuando Newton conoce estos otros valores y formas de hacer astronomía abandona a Margirus y con él la cosmología aristotélica. <<

[9] Recientemente publicado por J. E. McGuire y Martin Tamy; *Certain Philosophical Questions, Newton's Trinity Notebook*, Cambridge U. Press. 1983. Con un amplio comentario inicial de 325 páginas y el texto con la transcripción al inglés actual. <<

[10] Los tópicos que Newton incluye en el texto citado en la nota anterior son: 1) la materia prima, 2) los átomos, 3) el vacío y los átomos, 4) de la cantidad, 5) de la conjunción de cuerpos, 6) del lugar, 7) del tiempo y la eternidad, 8) del movimiento, 9) de la materia celeste y de los orbes, 10) del sol, las estrellas, los planetas y cometas, 11) de la rareza, la densidad, la rarefacción y condensación, 12) transparencia y opacidad, 13) de la fluidez, estabilidad y sequedad, de la blandura, la dureza, la flexibilidad y tractilidad, 14) del calor y el frío, 15) de la gravedad y levedad, 16) del calor y frío (bis), 17) del movimiento violento, 18) de la reflexión, ondulación y refracción, 19) del fuego, 20) del aire, 21) del aire y la sal, 22) de la tierra y de la atracción magnética, 23) movimiento perpetuo, 24) vegetales, 25) filtración y electricidad, 26) de la luz, 27) de la sensación, 28) de la especie visible, 29) de la visión, 30) de los colores, 31) de los sonidos y, tras algunos sobre cuestiones de imaginación y fantasía, vuelve a los minerales, a los meteoros, al flujo y reflujo del mar, al movimiento violento, de nuevo, a los cometas (tablas de Snell) y siguen sus observaciones del cometa de 1664 con dibujos, como antes con el movimiento, con un largo estudio de la luz y del movimiento que alcanza a sus estudios de 1665-1666. En este conjunto de textos se halla una buena pista de lo que estamos diciendo respecto a los programas de investigación de Newton. <<

[11] Sin duda, el mejor estudio de los desarrollos matemáticos de Newton se puede hallar en la introducción y comentarios de D. T. Whiteside, contenidos en su monumental edición de *The Mathematical Papers of Isaac Newton*, 8 vol. Cambridge U. Press, 1967-1982, pero para una breve y, a la vez, precisa exposición de este aspecto de los hallazgos y progresos de Newton, será bueno remitir al lector al Cap. IV («Resolving Problems by motion») de la Biografía de R. S. Westfall, *Never at Rest*, Cambridge U. Press, 1980, rep. 81, 82, 83 y 84. <<

[12] Confr. Westfall, o. c., págs. 135-137. <<

[13] Véase el manuscrito 3958.5 de los complementarios en la Biblioteca de la U. de Cambridge, y ya publicado en *The Background to Newton's Principia*, Oxford, 1965, págs. 195-197. Esta relación del inverso del cuadrado, que dista mucho de la formulación final, le dio pie para sostener frente a Hooke la independencia y prioridad de su teoría de la gravitación y de la ley del inverso. <<

[14] Uno de los datos más relevantes es la colaboración entre maestro y discípulo en la publicación de las *Lectiones XVIII, Cantabrigiae in scholis publicis habitae; in quibus phaenomenom opticorum genuinae radones investigantur*…, etc. de Barrow, en las cuales la mano de Newton introdujo algunas brillantes demostraciones, etc. Pero no sólo fue en este caso cuando Newton colaboró con, o a instancias de, Barrow en trabajos editoriales; también lo hizo en matemáticas, geografía, etc. <<

[15] Este aspecto del perfil newtoniano se va haciendo más explícito, a la vez que se va matizando más, en la etapa de las revisiones y reediciones. El contraste con Descartes y Leibniz no debe hacer olvidar a los otros interlocutores, sobre todo a los compatriotas como Locke, Boyle, More, etc. Las cartas a Bentley son de especial interés. <<

[16] Para el lector de lengua castellana que desee una ampliación a estas consideraciones relacionadas con la *Optica*, es posible acudir a la edición que de la misma hizo Carlos Solís en Madrid, Alfaguara, 1977, cuya introducción y notas ofrecen muchas pistas sobre este tema, que en Newton es una especie de cruce de hilos entre sus diferentes investigaciones. <<

[17] En realidad, salvo algunas insinuaciones en la *Optica* y unas cartas a Boyle y algunas notas más como el «De natura acidorum» (cfr. I. B. Cohen, editor. *Papers and Letters on Natural Philosophy*, Harvard U. Press, 1977, págs. 259-268), casi nada se halla publicado de los cuantiosos manuscritos alquímicos de Newton. Pero de que no tengan interés científico no se debe inferir que no tengan interés filosófico. Para un acercamiento a este aspecto puede verse, además del clásico libro de B. J. T. Dobbs: *The Foundations of Newton's Alchemy: The Hunting ofthe Greene Lyon*, Cambridge, 1975, un artículo de Karin Figala en *History of Science*, 15 (1977), 102-137 «Newton as Alchemist» y otro de R. y M. Boas Hall en *Archives internationales d'histoire des Sciences*, num. 11, 1958: 113-152 «Newton's Chemical Experiments». <<

[18] Buen ejemplo es el *Tractatus de methodis serierum et fluxionum* (véase el vol. III de los *Mathe. Papers*, págs. 32-328) que, reescrito y ampliado (sin título) en esta época, constituye un amplio y sistemático tratado de gran importancia para la sistematización posterior del cálculo. Pero no lo publicó tampoco. <<

[19] Se pueden ver estas cartas en la *Correspondence*, vol. I, págs. 412 y 416 respectivamente. La de Hooke es cortés y hasta humilde, la de Newton correcta pero firme. Llama a Descartes y a Hooke «gigantes» sobre cuyas espaldas él, Newton, se ha subido, pero, por ello, ha visto más lejos. <<

[20] Se publicó en Londres en 1665 y tuvo gran eco en el mundo científico de su tiempo. Es, quizá, el mejor exponente del talante científico de Hooke: su viva imaginación hace gala de ingenio e intuición anticipando muchas cosas que luego tendrán una confirmación más adecuada. Así, define el calor como «la propiedad de un cuerpo que es producida por el movimiento y agitación de sus partes», etc. Por otra parte su capacidad de observador y experimentador queda bien probada en esta obra. Por lo demás, puede verse el tono de la reivindicación de Hooke en I. B. Cohen, *Papers and Letters*, págs. 198-199. <<

[21] Pueden verse estas cartas en el vol. II de la *Correspondence*, págs. 297 y siguientes. Esta primera está fechada en «Gresham College, 24 de noviembre de 1979». <<

[22] Cfr. *Correspondence*, pág. 300. <<

[23] Se publicó en las *Phil. Trans.*, 1 (1668), págs. 684 y siguientes, y junto con un estudio del problema y de su estado de desarrollo histórico en ese momento puede verse en A. Koyré: *Chute des corps et mouvement de la Terre*, Vrin, París, 1973. <<

[24] Nos referimos a *Varia Opera Mathematica*, Tolosa, 1679. Pero los *Cogitata* y con ellos parte del problema ya eran públicos desde 1644. <<

[25] Vid. *Cor respóndeme*, vol. II, pág. 304. Véase también un ensayo muy bien documentado de Koyré en *Études Newtoniennes*, Gallimard, París, 1968, págs. 269-313, con el título «Une lettre inédite de Robert Hooke á Isaac Newton». <<

[26] Cfr. Westfall, *Never at Rest*, págs. 386-387 y notas 142 y 145. El argumento de la impotencia de Hooke para el tratamiento matemático de su idea ya lo esgrimió Newton, como Westfall y Cohen, por ejemplo, pero no acaba de convencer por completo. Einstein tampoco era un matemático brillante y, sin embargo, no se regatea su genio. <<

[27] En *Corresp*. II, págs. 430 y siguientes. <<

[28] Recientemente se ha creído haber identificado el manuscrito en el que Newton plasmó el descubrimiento del teorema de la elipse. Véase Herivel, *The Background to Newtorts Principia*, págs. 247-253. En 1961, Herivel publicó un artículo en *Archives internationales d'histoire des Sciences*, 14, págs. 23-33, en el que pretendía mostrar que el manuscrito en cuestión era el originario descubrimiento de Newton: «Newtonian Studies III. The Originàis of the Two Propositions Discovered by Newton in December 1679?». En *Archives* 16 (1963), págs. 23-28, R. y M. Hall se declaran contrarios a tal identificación. Westfall en *ibíd*. 22 (1969), págs. 51-60, parece dar su acuerdo a Herivel. Whitheside, en *Id*. 194 (1970), 2-3 y en *Math. Papers* VI, 14 y 553-554 se opone a los argumentos de Westfall. La copia no estaba entre los papeles de Newton solamente, sino que se hallaba entre los de Locke, copiada por su amanuense en 1689-1690.

En la correspondencia con Flamsteed de 1685-1686 (vide *Corresp*. II, 338 y 346). Newton se fiaba de las observaciones de Flamsteed pero no tanto de sus interpretaciones. Aunque la idea del retorno desde el Sol de los cometas no era aparentemente compatible con las velocidades y direcciones observadas, aún era menos compatible con las ideas de Newton la explicación de Flamsteed de ese cambio de dirección tan repentino y violento. Newton necesitó tiempo y observar por sí mismo el paso de los cometas de 1680 y 1682 para aceptar el retorno y para tratar de extender hasta él la teoría de la gravitación. <<

[29] De esta lectura Newton no dijo ni palabra hasta mucho después, pero en su *Waste Book* (Mss. adic. 4004, f. 103r y 103v) aparece el resumen y análisis de este opúsculo, precisamente hacia 1681-1682. Como suele acontecer en los escritos de Hooke, en este tampoco hay soluciones satisfactorias, ni siquiera para él mismo, a los problemas que se plantea, pero se hace tantas y tales preguntas que provoca a sus lectores. Llega a decir que su trayectoria no es fácil de determinar pero no más de lo que suele costar determinar la de los planetas y que su gravitación es parecida. <<

[30] Mss. adicionales, 4007, f. 107rv. Biblioteca de la Universidad de Cambridge. <<

[31] Carta de Halley a Newton de 29 de junio 1686; *Corresp*. II, 441-442. <<

[32] Aunque la correspondencia de Halley es con Hooke y la S. R., ella da pie a que se conozca y critique la obra de Cassini y su teoría de grandes órbitas cometarias. Halley está entre ambos y Newton lee y anota el libro de Cassini como se desprende de J. R. Harrison, *The Library of Isaac Newton*, Cambridge, 1978, pág. 116. La estima de Newton por las observaciones de Cassini, aunque menor por su idea de la Tierra, de su forma y figura geométrica, es evidente en los *Principia*. <<

[33] I. B. Cohen en su magnífica *Introduction to Newton's Principia*, Cambridge U. Press, 1971, pág. 50, se plantea la cuestión de cuál pudo ser la pregunta de Halley. El análisis que seguimos es similar. El manuscrito (f. 101) a que hacemos referencia puede verse en el Apéndice I de la o. c., págs. 293 y 294. En el mismo (folio 106) Newton hace la enumeración que sigue inmediatamente en el texto, pág. 295. o. c. <<

[34] l. c. en nota anterior. <<

[35] Ibíd. passim. <<

[36] El *Horologium Oscilatorium sive de Motu Pendulorum ad Horologia Aptato Demostrationes Geometricae* pudo estar presente en Newton, además de con su contenido, también con su forma. La parte primera es una descripción del reloj de péndulo, pero desde la segunda se configura con tres Hypótesis y continua con XXVI Propositiones y un Lema; y por si fuera poco la parte tercera se abre también con IV Definitiones y XI Propositiones y todavía la parte IV añade XIII Definitiones y dos Hypótesis antes de sus XXV Propositiones. En la quinta parte muy breve, describe otra forma de construir un reloj de péndulo y añade los 13 *Teoremata de vi centrífugo ex motu circulori*. <<

[37] Herivel incluye algunos ensayos y fragmentos desde esta lecha y pueden se en su *The Background*… sobre todo a partir de la pág. 147 y también puede se un resumen de este inicial planteamiento en Westfall, Never at Rest, 148-150. <<

[38] Pertenece al memorándum Conduitt-De Moivre y se halla transcrito y publicado en extracto por Cohen en su *Introduction.*… Apéndice I, págs. 297-298. <<

[39] Hablamos del «pequeño tratado» de la S. R. por dos razones: primera porque Newton no le dio título y segunda porque más tarde, en abril de 1686 y marzo-abril de 1687 haría llegar también el manuscrito M del «gran tratado», los *Principia*, esta vez con título. Respecto al «pequeño» cabe decir que fue hallado por Stephen Peter Rigaud, quien lo publicó con el título «Isaaci Newtoni Propositiones de Motu» como Apéndice I de su *Historical Essay on the First Publication of Sir Isaac Newton's Principia*, Oxford, 1838. Después ha sido traducido al inglés y publicado por Herivel, R. Ball, R. y M. Hall, etc. <<

[40] Rigaud, o. c., pág. 7 del Apéndice. <<

[41] I. B. Cohen, *The Newtonian Revolution*, Cambridge, 1980, pág. 242. Naturalmente Cohen no afirma que Newton fuera consciente en ese momento de dicha equivalencia; para serlo fue preciso un largo camino. <<

[42] Rigaud, o. c., pág. 12. <<

[43] Cfr. Henry Pemberton, *A View of Sir Isaac Newton's Philosophy*, Londres, 1728, Prólogo. Pero es preciso añadir inmediatamente que la «leyenda» newtoniana, más allá de las notas recogidas por De Moivre-Conduitt-Stukeley y sin olvidar las sutiles pinceladas del propio Newton, sobre todo con motivo de sus polémicas de prioridad habidas con Hooke y Leibniz, ha gozado del favor británico. Basta recordar que el famoso manzano de la leyenda aún produce pequeñas astillas para turistas y devotos. Cohen en su *Introduction*…, Apéndices I y II, ofrece algunas notas autobiográficas del proceso de escritura de los *Principia*, y los recuerdos de Humphrey Newton sobre *sir* Isaac. <<

[44] Se conserva un texto manuscrito con correcciones de otra mano distinta de la de Newton y, por supuesto, distinta de la de Humphrey, que apunta a Halley como consultor, aunque no haya para ello otra base que la calidad de las correcciones propuestas y el parecido de los rasgos de la escritura. Confer. Cohen, *Introduction*…, Apéndice VII. <<

[45] La clasificación, descripción e interdependencia de todos estos borradores, ensayos o pre*Principia*, se halla suficientemente expuesta en Cohen, *Introduction.…*, hasta el punto de que el lector puede ver el mapa temporal y ordinal de los desarrollos, equivalencias y ampliaciones que Newton va haciendo desde que inicia su trabajo hasta que acaba la redacción, junto con piezas no incluidas y revisiones o titubeos en la redacción. Lo mismo para las dos reediciones de 1713 y 1726. <<

[46] Hacia 1676 Mercator publica sus *Institutionum astronomicarum libri duo*; en la pág. 286 y siguientes ofrece un resumen de una carta (hoy perdida) de Newton en la que se presenta una explicación de los movimientos libratorios de la Luna. Bien puede suponerse que una explicación más completa se halla en la Proposición XVII del Libro III de los *Principia*. <<

[47] Rigaud, o. c., Apéndice I. Cohen, siguiendo en esto a Rigaud y a los demás editores del tratado lo denomina «de Motu», pero al ser varios los tratados que se conservan con este título, preferimos referirnos a él con una perífrasis como «pequeño tratado» de la S. R. De las otras copias da cuenta Cohen en *Introduction*…, pág. 56. <<

[48] Cohen, por razones de comodidad y brevedad, propone en su *Introduction*…, un cierto código literal con subíndices para representar las ediciones, los manuscritos, etc. l. c., pág. XXVII. <<

[49] La base textual para analizar esta presunta reconversión de Newton al clasicismo geométrico puede encontrarse en el vol. IV de los *Math. Papers*, aunque el editor sostiene que tal reconversión es sólo superficial y declarativa, si atiende al contenido de sus trabajos, particularmente los ligados a los «loci». Una parte de estos estudios reaparecerán en los *Principia*, sección V, Lemas XVII-XXL Proposición XXII. <<

[50] Sería largo y complejo justificar esta sospecha. Nos lleva a ella una analogía con lo que ocurre en matemáticas: en el ensayo «Veterum loca sólida restituta» (*Math. Papers*, IV, pág. 277) muestra tal rechazo de las vías analíticas y tanto aprecio por los griegos y sus métodos de proporciones simples que induce a considerar hasta qué punto el método de las «ratios» y «proportios» no se deberá a una actitud semejante en la física. <<

[51] Cfr. *Marginalia in Newtoni Principia Mathematica*, editadas por E. A. Fellman con traducción al francés de J. F. Courtine. Vrin, París. 1973. Y en relación con este punto, véase el comentario a *M* 48 en pág. 76 (versión alemana) y pág. 98 (versión francesa). La nota marginal 48 corresponde a la Proposición X del Libro I. <<

[52] Para una exposición ordenada de estas elaboraciones teóricas puede verse René Dugas, *A History of Machanics*, Neuchatel, 1955. La tercera parte, pág. 231 y siguientes, se halla consagrada a los desarrollos de la Mecánica Clásica y en ella se comprenden los Bernoulli, Euler, D'Alembert, etc. <<

[53] Sus protestas en este sentido son muchas y algunas hallará el lector en los *Principia* (la más explícita en el Escolio General), pero es oportuno recordar aquí la firmeza de su negativa a suscribir esa opinión en carta a Bentley de 1693 donde dice: «… a veces habla usted de la gravedad como esencial e inherente a la materia. Ruego que tal idea no se me atribuya, pues la causa de la gravedad es cosa que no pretendo conocer y además necesitaría mucho más tiempo para considerarla». Cfr. *Corresp*. III, 234. <<

[54] Cohen en su *Introduction…* se extiende en el análisis de la generación cronológica de cada una de las Proposiciones de los *Principia* a partir de este pequeño tratado y de las sucesivas extensiones que se conservan depositadas, por lo general, en la Biblioteca de la Universidad de Cambridge como parte del archivo de Newton. Con gran paciencia Cohen trata de conectar todos esos manuscritos entre sí, aunque a veces le faltan piezas al *puzzle*. Vide o. c. págs. 68-129. <<

[55] Se trata del mismo fragmento citado en la nota 33 (mss. adic. 3968, f. 101) en cuya parte final hace una crítica de los métodos analíticos de los modernos y una valoración positiva de los métodos sintéticos de los antiguos mediante los cuales «construían sus descubrimientos para el público». <<

[56] De especial importancia son los corolarios de las Proposiciones IV y XVI. En su conjunto estas tres primeras secciones responden a los orígenes del problema y del tratado. Por tanto, respecto al núcleo problemático inicial, se puede decir que son el origen programático, y no sólo sistemático, de los *Principia*. <<

[57] Toda esta sección V responde a los estudios newtonianos sobre los «loci» (vide nota 49) y se desencadenaron como respuesta a algunas publicaciones recién aparecidas en París, entre ellas, Fermat (vide nota 24), los *Nouveaux èléments des sections coniques*, de la Hire, así como la recuperación de los Libros VII y VIII de Pappo de Alejandría gracias a la traducción latina (incompleta) de Commandino desde 1588, cuya sutileza llamaba la atención de los matemáticos. Una parte de la *Gèométrie* de Descartes se ocupa del llamado «problema de Pappo», Wallis en 1688 encuentra un fragmento del Libro II en Oxford y lo traduce y publica, y Newton en esta sección V, Lema XIX «construye» una nueva solución y aprovecha la ocasión (Corolario 2) para aclarar que este problema «de las cuatro líneas… se muestra no como un cálculo, sino como una composición geométrica, como la que buscaban los antiguos». <<

[58] El citado Lema y la Proposición siguiente se enfrentan con el llamado «Problema de Kepler» que ha resistido al análisis directo de los matemáticos y sólo mediante aproximaciones (como Newton hizo) ha podido hallar solución. Hasta el punto de que existen tablas de aproximaciones para primeras, segundas, terceras derivadas, etc. El valor más importante del Lema radica en su enunciado negativo (y consiguiente demostración) como muestra de la moderna concepción del análisis que poseía Newton. <<

[59] Newton pensó en publicar su tratado «de Quadratura Curvarum» junto con los *Principia*. Quizá esto explique que empiece a dar por supuesto algo que, al no estar publicado ni ser conocido más que por unos pocos (Craig, Gregory Collins, Halley) iniciados, exigía un acto de fe por parte de sus lectores, esto en 1687; otro era el caso en 1713 con la segunda edición. <<

[60] La correspondencia con Halley durante junio-julio 1686 (ya citada) trata de los detalles de la publicación de los *Principia*, cuyo Libro I ya está desde abril en poder de Halley, y también trata de las reclamaciones de Hooke sobre su paternidad de la ley de la gravitación. Newton en esos momentos se halla recomponiendo y acabando el Libro II y empezando a reescribir el Libro III. <<

[61] La redondez de la Tierra ofrecía varios modelos posibles (dado que sólo se había recorrido en dirección Este-Oeste) y ya Colón había sugerido el modelo de una «pera», pero a la sazón se creía en el modelo «husiforme» de Cassini. En algún momento Voltaire comentó que Maupertuis había aplastado a la vez a la Tierra y a Cassini, pero cuando se vio enredado en una controversia con Maupertuis le aderezó los ingeniosos versos:

> «Vous avez confirmé dans les lieux pleins d'ennui
> Ce que Newton connut sans sortir de chez lui».

(Discours sur la Modération) aludiendo a su viaje a Laponia de 1736-1737 junto con Clairaut y otros. <<

[62] Véase el estudio de Koyré en *Études newtoniennes*, Gallimard, París, 1968, «Une lettre inédite de Robert Hooke á Isaac Newton», págs. 269-313. <<

[63] Conf. Cohen, *Introduction.…*, Apéndice II, del memorándum de Stukeley recogido de Humphrey Newton. <<

[64] Aunque son mejor conocidos estos años de triunfo social, en los que también hizo filosofía para las princesas, como él mismo dijera de Leibniz, no podemos alargarnos en ellos y remitimos al lector a las biografías que enumeramos en otra parte. <<

[65] Su correspondencia, ya publicada por Edleston en 1850, puede verse también en *Corresp.*…. vol. V *passim*. Item Apéndice V de la edición Koyré-Cohen de los *Principia*, vol. II, págs. 817-826, en extracto. <<

[66] Cfr. W. W. R. Ball, *An Essay on Newton's Principia*, Londres, 1893, pág. 74. <<

[67] *Sir* David Brewster en *Memoirs of the Life, Writings, and Discoveries of Sir Isaac Newton*, Edimburgo, 1855, vol. II, págs. 549 y siguientes ofrece una lista de los cambios y alteraciones más notables de la tercera edición respecto a la segunda, debidos a Pemberton sobre todo. Y pese a su gran número, ninguna es relevante. Vide Cohen-Koyré, vol. II de la edición de los *Principia*, Apéndice VI, págs. 827-847. <<

[68] Entrando en la Abadía, en el trascoro, a la izquierda se alza un grandioso monumento a su memoria erigido por sus sobrinos en 1731 con un epitafio resumen de sus contribuciones científicas:

Texto latino	Texto castellano
Hic situs est	Aquí yace.
Isaacus Newton, Eques Auratus,	Isaac Newton, Caballero,
Qui animi vi prope divina,	que, con fuerza de espíritu casi divina,
Planetarum Motus, Figuras,	los movimientos de los planetas, las figuras,
Cometarum semitas, Oceanique Aestus,	las sendas de los cometas, las mareas del océano
Sua Mathesi facem preferente,	con sus matemáticas como antorcha
Primus demonstravit.	fue el primero en demostrar.
Radiorum Lucís dissimilitudines,	Las diferencias de los rayos de luz,
Colorumque inde nascentium proprietates,	y las propiedades de los colores de ellos nacientes
Quas nemo antea vel suspicatus erat, pervestigavit.	que antes nadie ni hubiese sospechado, investigó con rigor.
Naturae, Antiquitatis, S. Scripturae,	De la naturaleza, de la antigüedad, de la S. Escritura
Sedulus, sagax, fidus Interpres,	asiduo, sutil y fiel intérprete,
Dei Opt. Max. Majestatem philosophia asseruit,	afirmó con la filosofía la majestad de Dios Opt. Max.
Evangelii simplicitatem moribus expressit.	expresó con sus costumbres la simplicidad del Evangelio.
Sibi gratulentur Mortales, tale tantumque extitisse	Congratúlense los mortales de que existiese tal y tan grande
HUMANI GENERIS DECUS.	ORNAMENTO DEL GÉNERO HUMANO.
Natus XXV. Decemb. MDCXLII. Obiit XX. Mar. MDCCXXVII.	Nació el XXV diciembre MDCXLII. Murió XX marzo MDCCXXVII.

<<

Notas del libro primero

[1] Este poema de Halley, en el que tratamos de retener algo de su ritmo y sintaxis latinos, es la traducción del texto de la tercera edición. Para la segunda Richard Bentley propuso algunas modificaciones estilísticas (principalmente) que no afectan al sentido en nada importante. Más significativo puede ser el hecho de que Halley se viera impulsado a este canto, y también inspirado en algunas expresiones, por la Egloga que precedía al *Horologium Oscillatorium* de Huygens, en la cual las alabanzas mitológicamente enlazadas también se cierran con un «Nec Phaebum ac Pimplae fas est contemnere Divas»; similar al exámetro de Halley «Nec fas est propius mortali attingere divos». <<

[2] Puso fecha en la segunda edición. Y se puede pensar que ésta no fue la fecha en que redactó este prefacio. Pero en 1713 ya sabía que las fechas tenían su importancia. <<

[3] Este argumento, casi literal, le fue dado a Cotes por Newton ante una consulta de aquél relativa a la expresión «cum attractio omnis mutua sit» del Corolario I, Proposición V, Libro III. Cfr. Edleston, J. en *Corresponde*…, pág. 154. <<

[4] Este párrafo y los siguientes están dirigidos contra Leibniz y las críticas procedentes de su entorno (J. Bernoulli, Huygens y Wolf cuyo órgano eran las *Acta eruditorum* principalmente) y aparece aquí en uno de sus episodios. Pero la guerra ya había empezado mucho antes. En 1703 George Cheyne, un discípulo de Gregory, había publicado un ensayo con el título *On the inverse method of fluxions* en el que se insinúa que todo el cálculo era una creación de Newton. La fase siguiente —1705— es una recensión anónima (de Leibniz) en las *Acta* dedicada a la primera edición de la *Optica* y de los tratados que la acompañaban (*Ennumeario Curvarum y De Cuadratura*) en la que Leibniz aclara que el método de fluxiones es una aplicación en términos de «fluxiones y fuentes» de su método diferencial. Pero la guerra estalla en 1710 con la aparición en las *Phil. Trans*. (número correspondiente a 1708) de un artículo de John Keill «On the Laws of Centripetal Forces» en el que acusa a Leibniz de plagiario. Por lo demás existe otro contencioso, metafísico esta vez, sobre la naturaleza de la fuerza, del espacio y tiempo, de Dios, etc. del que Leibniz en la Polémica con Clarke deja constancia, de los puntos principales. A este respecto véase A. R. Hall, *Philosophers at War. The quarrel between Newton and Leibniz*. Cambridge U. Press., 1980. <<

[5] Ya hemos mencionado la dificultad, y posible circularidad, que la crítica ha visto en esta Definición. Parte del problema puede residir en el concepto de partícula elemental que manejaba Newton. A ello puede atribuirse el hecho de que su definición de densidad no tenga un carácter cuantitativo sino sólo proporcional. En el Corolario 4 de la Proposición VI del Libro III da una definición de *densidad*. Pero como las partículas contenidas en un mismo volumen, aunque tengan la misma densidad, no tienen la misma forma, no se puede asignar el mismo número de ellas al mismo volumen. Respecto a la naturaleza de las partículas véase la Quest. 31 de la *Optica*. <<

[6] Esta es una definición que no fue siempre bien entendida. En el sentido en que Newton la propone equivale al concepto moderno de «momento» o lo que se llamó «fuerza muerta» ($F = mv$) en contraposición a la «fuerza viva» ($\frac{1}{2}mv^2$). <<

[7] En su copia de la segunda edición añadió Newton: «No entiendo por fuerza de inercia la de Kepler por la cual los cuerpos tienden al reposo, sino la fuerza de mantenerse en el mismo estado de reposo o de movimiento». Pero no lo incluyó en el texto de la tercera edición. <<

[8] Las discusiones generadas por este Escolio se iniciaron con Berkeley (principalmente en el *Treatise on the Principies of Human Knowledye*, 1710, *De Motu*, 1720) continuaron en la Polémica con Clarke desde el lado leibniziano y, con el impulso de las geometrías no euclídeas, llegó a su mayor dimensión con Mach y la teoría de la Relatividad. Puede verse un amplio debate en A. [*Grünbaum*], *Philosophical Problems of Space and Time*, Reidel, Dordrecht/Boston, 2.ª, 1973. <<

[9] Salvo correcciones de estilo insignificantes, estas Leyes no ofrecieron problemas a Newton en las revisiones a que sometió su texto inicial. La I.ª (de «inercia») no ofreció problemas debido a su eficacia práctica en un sistema inercial como el de nuestra galaxia. Pero ¿se puede hablar de algún cuerpo en reposo absoluto? La II.ª depende de la definición de «cantidad de movimiento» (*m. v.*). El cambio de cantidad es proporcional a la fuerza motriz impresa y la medida de ésta se obtiene por la cantidad de cambio del valor de *m. v*. Pero se ha demostrado experimentalmente que la masa de un electrón aumenta cuando la velocidad del mismo se va acercando a la de la luz (Kaufmann, 1901). Pero esta relatividad de la masa no es relevante a velocidades alejadas de la de la luz. Por ello, la teoría de la relatividad no es de aplicación necesaria en la física ordinaria de masas y velocidades ordinarias. En cambio evidencia el alcance limitado de la Ley de Newton. Hacia 1692 Newton anotó en una de sus copias una aclaración añadida al texto: Así, si la gravedad de un cuerpo que cae fuese uniforme, ésta, actuando igualmente en cada partícula igual de tiempo, imprimirá fuerzas iguales sobre dicho cuerpo, y generará velocidades iguales; y por tanto la fuerza total impresa sobre el cuerpo que cae y la velocidad total generada siempre serán como el tiempo total de caída. Y así en los demás. (Tachada esta frase final). <<

[10] Entre los ensayos correctores de 1690 y siguientes, ya mencionados, destaca uno relativo a este primer Corolario de las Leyes. El tema era de fondo: una contradicción aparente entre la idea de fuerza como sucesión de impulsos discretos o como acción continua. Este problema se hace igualmente patente en sus revisiones de la Ley II (vide, Cohen, *Introduct.*…, pág. 163). Al final los cambios fueron insignificantes: introdujo las frases «impresa en el punto A»… «con movimiento uniforme»… «Seguirá, pues, un movimiento rectilíneo de A a D por la Ley I». <<

[11] Este Corolario incorpora en síntesis los estudios que en 1665-1666 inició Newton sobre la teoría cartesiana del choque y reflexión de cuerpos perfectamente elásticos. Cfr. R. S. Westfall, *Force in Newton's Physics*, Londres, 1971, principalmente Cap. 2 a 5. El testimonio se halla en el «Waste Book» con el título «On Reflections» y ha sido publicado por Herivel en *The Background of Newton's «Principia»*, pág. 132-182. <<

[12] Desde «Para un cuerpo que cae…» hasta «… AB cuadrado» aparece sólo en esta tercera edición que ofrecemos. <<

[13] Se refiere a *Traité* de la percusión *ou choc des corps* que apareció en París en 1673 y cuya tercera edición de 1684 incluye los experimentos completos. <<

[14] Este párrafo aparece en la tercera edición. Su tema fue expuesto a Cotes como se dijo en nota 3, supra y también la figura, obviamente. <<

[15] Una de las críticas primeras (mejor intencionada que fundada) que merecieron los *Principia* y a la cual Newton respondió con todo respeto, provenía de un antiguo y ya retirado profesor de matemáticas de nombre Gilbert Clerk (vde. *Correspondencia* 2, 152-154 y 490) obedecía a que el uso de esta terminología ofrecía dudas. Newton la utiliza con los sentidos siguientes:

Duplicata ratio = razón cuadrada = $(a/b)^2$.

Subduplicata ratio = raíz cuadrada de la razón = $\sqrt{a/b}$

Triplicata ratio = razón cúbica = $(a/b)^3$.

Subtriplicata ratio = raíz cúbica de la razón = $\sqrt[3]{a/b}$

Sesquiplicata ratio = potencia $\frac{3}{2}$ de la razón = $(a/b)^{\frac{3}{2}} = \sqrt[2]{(a/b)^3}$

Subsesquiplicata ratio = potencia $\frac{2}{3}$ de la razón = $(a/b)^{\frac{2}{3}} = \sqrt[3]{(a/b)^2}$

Sesquialtera ratio = razón sesquiáltera = $3a/2a = \frac{3}{2}a$

Proportio = razón (a veces) = nuestra razón a/b.

Con esta aclaración por delante nos permitiremos usar indistintamente la expresión arcaica y la moderna. <<

[16] Esta sección II, junto con la siguiente, merecieron gran atención por parte de Newton en sus ensayos revisionistas. Llegó a redactar unos enunciados nuevos para las Proposiciones I y II y proyectó Corolarios adicionales y sustituyó otros, además de reordenar las Proposiciones IV y V. Al final se limitó a 1) trasladar los Corolarios 1 y 2 de la Proposición I al lugar actual de la Proposición II con algunas variaciones sin importancia. 2) Añadir 6 nuevos Corolarios a la Proposición I de los cuales el 2 y el 4 le permiten simplificar la prueba de la Proposición IV. 3) Simplificar y reescribir la prueba de la Proposición IV. 4) En los Corolarios de esta Proposición IV aumenta su número de 7 a 9; el 7 de la nueva serie es debido a Fado de Duillier probablemente. 5) La anterior prueba de la Proposición VI se convierte en Corolario 1 de dicha Proposición, y el anterior Corolario de la misma será el 5. 6) El resultado es una nueva Proposición VI que pasa de un Corolario a 5. 7) Pruebas alternativas para las Proposiciones VII, IX y X con 3 nuevos Corolarios para la Proposición VI y una notable expansión del Escolio final de la sección II. Pero los ensayos fueron mucho más amplios y evidencian un interés de Newton por reescribir estas secciones quizá en términos de fuerzas y no de áreas, y a la vez presentar su línea argumental con mayor elegancia y brevedad. Vde. Cohen, *Introduc*., págs. 167-171. <<

[17] Pasar de un polígono de infinitos lados a una curva continua es pasar de una serie a su límite y esto matemáticamente no ofrecía ya dificultad alguna, ni siquiera era preciso acudir al Lema III. Pero esto comportaba, de paso, asimilar a la fuerza continua la fuerza discreta compuesta de impulsos sucesivos. Quizá aquí estuviera la razón de los repetidos ensayos revisionistas de esta sección. <<

[18] Los 7 Corolarios incluidos aquí en la primera edición decían como sigue:

COROLARIO 1. De aquí que las fuerzas centrípetas son como los cuadrados de las velocidades divididos por los radios de los círculos.

COROLARIO 2. E inversamente, las fuerzas entre sí son como los cuadrados de los tiempos periódicos divididos por los radios. Esto es (hablando como los geómetras), estas fuerzas están en razón compuesta directamente de la razón cuadrada de las velocidades y de la razón simple de los radios inversamente y también inversamente de la razón cuadrada de los tiempos periódicos.

COROLARIO 3. De donde, si los tiempos periódicos son iguales, tanto las fuerzas centrípetas como las velocidades serán como los radios, y viceversa.

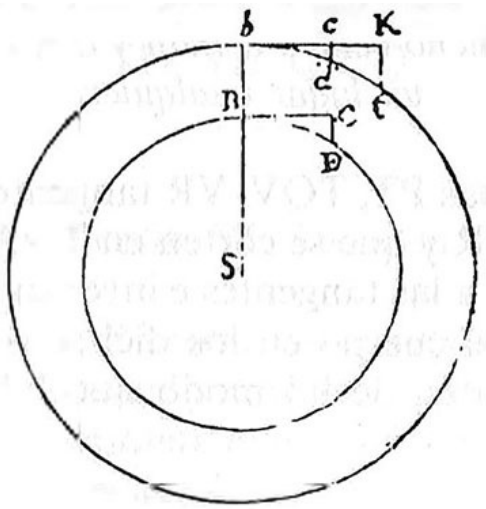
COROLARIO 4. Si los cuadrados de los tiempos periódicos son como los radios, las fuerzas centrípetas son iguales, y las velocidades están en razón de la potencia 1 de los radios, y viceversa.

COROLARIO 5. Si los cuadrados de los tiempos periódicos son como los cuadrados de los radios, las fuerzas centrípetas son inversamente como los radios y las velocidades iguales, y viceversa.

COROLARIO 6. Si los cuadrados de los tiempos periódicos son como los cubos de los radios, las fuerzas centrípetas son inversamente como los cuadrados de los radios, mientras que las velocidades son como la potencia ½ de los radios; y viceversa.

COROLARIO 7. Todas las cosas sobre tiempos, velocidades y fuerzas con los que los cuerpos describen partes semejantes de cualesquiera figuras semejantes con centros semejantemente situados, se siguen de la demostración de las precedentes aplicada a estos casos.

Por otra parte hay que recordar que esta Proposición IV tenía otra demostración basada en dos círculos concéntricos, demostración que daba pie a la anterior serie de Corolarios. Una vez construida la figura, la demostración se cifra en que la figura *tkb* es semejante a la figura DCB y por el Lema V, la línea CD será a la línea *kt* como el arco BD al arco *bt*: o sea, por el Lema XI, la línea naciente *tk* a la línea naciente *de* como bt^2 a bd^2 y, por tanto, la línea DC a la línea *dc* como BD·*bt* a bd^2, o lo que es igual, como $\frac{BD \cdot bt}{Sb}$ a $\frac{bd^2}{Sb}$ y por ende $\left(\text{al ser iguales las razones } \frac{bt}{Sb} \text{ y } \frac{BD}{SB}\right)$ como $\frac{BD^2}{SB}$ a $\frac{bd^2}{Sb}$. Q. E. D.



Como consecuencia de la supresión de la anterior prueba hubo de alterar también un párrafo del Escolio siguiente para ponerlo en relación con el nuevo sistema de Corolarios. Al principio del Escolio hay un párrafo añadido por Newton con motivo de la correspondencia con Halley en 1686. En él muestra cómo puede hallar la ley (al menos para el movimiento circular) incluso antes de la publicación del *Horologium Oscillatorium* de Huygens. <<

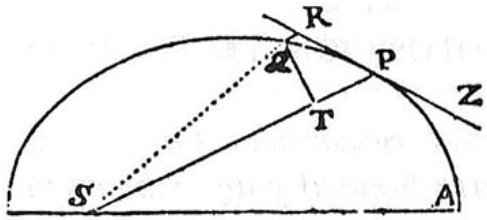
[19] Los cambios introducidos en esta Proposición V son (una vez abandonada la idea de renumerarla como Proposición IV, Problema I): Cambiar la redacción del segundo párrafo que decía «Puesto que el cuerpo en P y Q con radios trazados al centro describe áreas proporcionales a los tiempos, y dichas áreas descritas a la vez sean como las velocidades en P y Q multiplicadas, respectivamente, por las perpendiculares trazadas desde el centro hasta las tangentes PT y QT, serán dichas perpendiculares como las velocidades inversamente; y por tanto…». <<

[20] En la Proposición VI, como ya hemos adelantado, los cambios son más amplios:

1) El antiguo enunciado de la Proposición pasa ahora a Corolario 1.

2) La figura era más simple.

3) La prueba de la anterior Proposición (ahora Corolario) era: Pues en la figura infinitamente pequeña QRPT la línea naciente QR, en un tiempo dado, es como la fuerza centrípeta (por la Ley II) y con fuerza dada como el cuadrado del tiempo (por el Lema X) y, por tanto, no dado ninguno, como la fuerza centrípeta y el cuadrado del tiempo conjuntamente y, por tanto, la fuerza centrípeta directamente como la línea QR y como el cuadrado del tiempo inversamente. Pero el tiempo es como el área SPQ o su doble SP·QT, esto es como SP y QT conjuntamente, y por tanto la fuerza centrípeta es como QR directamente y $SP^2 \cdot QT^2$ inversamente, esto es, como $\frac{SP^2 \cdot QT^2}{QR}$ inversamente. Q. E. D.



4) El único Corolario pasa con mínimos cambios a ser el 5. Además de estos cambios Newton pensó otros más radicales y en papeles sueltos intercalados en sus copias de la primera edición llegó a redactar una Proposición nueva. <<

[21] De nuevo, aparecen las correcciones. Newton proyectó rehacer esta Proposición como Teorema, con el siguiente enunciado: «Si ordenadas proporcionales de dos órbitas caen sobre abscisas proporcionales y los centros de fuerzas se sitúan semejantemente en las abscisas, los cuerpos describirán las partes correspondientes de órbitas en tiempos proporcionales, y las fuerzas centrípetas serán como las alturas de los cuerpos directamente y los cuadrados de los tiempos inversamente». De hecho dejó el Problema tal cual y se limitó a cambiar la figura y dos párrafos de la solución que la hacen más directa. Pero ahora añade todo lo que sigue a «*Lo mismo de otro modo*». Varias son las interpretaciones posibles para estas indecisiones; desde intenciones de elegancia matemática hasta titubeos en la presentación bajo forma geométrica o bajo forma dinámica, e incluso la percepción de que había un impasse entre planteamientos dinámicos y cinemáticos en toda esta (y otras) sección(es). <<

[22] Aunque casi todos los párrafos que introducen pruebas alternativas («Lo mismo de otro modo») son más elegantes, éste llama la atención por su brevedad y estilo directo. Por otra parte este problema (la pregunta de Halley o la respuesta a la pregunta) fue quizá la causa de su retraso en la respuesta escrita prometida. En todo caso esta Proposición debe entenderse como formando parte de un todo junto con la XVI; y quizá toda esta sección sea un modelo de análisis de un solo problema desplegado en sus diferentes casos y aspectos. Tampoco hay que olvidar que Newton elaboró estas pruebas alternativas en la década de los noventa y primeros años de 1700. Su lectura de la primera redacción le fue llevando a formulaciones más precisas y más concisas. <<

[23] Se cierra esta sección y, con ella, los intentos de Newton de mejorar la estructura inicial de los *Principia*. En la primera edición no hay Escolio al final de la Proposición; Newton además de añadirlo pensó, y redactó con ese fin, añadir un Corolario 5 que decía: «Puesto que en la solución de este Problema se enumeran todos los casos en los que un cuerpo puede partir de un lugar dado, con una velocidad dada y según una recta dada, y en todos ellos resulta una sección cónica que tiene su foco en el centro de fuerzas (también se sigue del primer Corolario de la Proposición XIII —entre paréntesis y tachado—), las fuerzas centrípetas del cuerpo que gira en la sección hallada (por las Proposiciones XI, XII y XIII) están en razón inversa del cuadrado de la distancia al centro: los cuerpos con fuerzas centrípetas que están en razón inversa del cuadrado de la distancia al centro se moverán siempre en secciones cónicas». Era redundante y por ello y porque el Escolio siguiente resume mejor el problema decidió tal vez suprimirlo. Por otra parte, la conclusión necesitaría precisar para qué velocidades iniciales es válida la generalización. <<

[24] En nota marginal añade a la última frase del Escolio: «y esta cantidad $\frac{CG^3}{RP^2}$ cuando el centro R de fuerzas se halla hacia la parte convexa de la órbita resulta negativa y expresa una fuerza centrífuga. Con este solo teorema se evidencia la fuerza con la cual un cuerpo en una sección cónica cualquiera puede moverse en torno a cualquier centro de fuerzas». Otro dato interesante es seguir los sucesivos desarrollos de la figura que aparece en la Proposición VI, sobre la cual, con algunas variantes, se va desarrollando la construcción argumental de estas dos secciones. Dicha figura admite como caso particular el del Corolario 4 de la Proposición XVI, caso que por incluir los mismos valores para «radio» y para «distancia media» desde el foco pudo constituir el tropiezo con que topó Newton cuando «rehizo» los cálculos después de la visita de Halley. El hecho de que las velocidades, en el Corolario 4, Proposición IV y Corolario 4, Proposición XVI sean inversamente como las raíces cuadradas de distancias iguales (solo iguales en esa situación) pudo dar lugar al tropiezo. (Aunque Newton hable de un diagrama trazado con poco cuidado). Con la nueva redacción de los Corolarios de la Proposición IV se deja más claro el proceso generalizador por el cual la Proposición se va extendiendo en su alcance desde el círculo y su radio hasta las distancias medias de un punto de una órbita al foco, y como hace al fin explícito en el Corolario 8 de esa Proposición IV. Por ello en el Escolio final de esta sección III parece dispuesto a generalizar el problema y la solución como lo hace. <<

[25] Se refiere Newton a una obra perdida de Apolonio de Perga cuyo título *επφαί* (Tangencias) nos fue trasmitido por Pappo junto con algunas noticias sobre su contenido entre el cual tuvo importancia este tema de los «tres elementos» (líneas o puntos): elegir cualquier combinación de tres de ellos y trazar un círculo tangente a los tres. Vieta en 1595 al final de su «Responsum» a Adriano Romano y refiriéndose a «las Tangentes» propone el problema de trazar un círculo tangente a otros tres dados. Ni A. Romano ni Regiomontano creían que pudiera resolverse con regla y compás. En 1600 Vieta da una solución euclidiana satisfactoria. Posiblemente Newton conocía la edición de Schooten de (Leiden) 1646. <<

[26] Este Lema ha sido objeto de estudios minuciosos por parte de estudiosos del desarrollo de la Geometría. Sin duda, este es el primer paso en el estudio de las conversiones geométricas, basado en una idea newtoniana relacionada con su visión de las ecuaciones como expresiones de la naturaleza de las figuras. Estas son equivalentes si las ecuaciones que las definen también lo son y viceversa. Como aplicación inmediata Newton propone la conversión de figuras por traslado de los parámetros que definen una figura respecto a alguno de sus elementos tomado como eje de conversión hasta otro lugar en el cual dicho eje, aunque cambie las direcciones de su posición, permite reproducir los valores de las ecuaciones de la primera figura y dar lugar a otra equivalente. Aunque en las curvas de orden superior es muy difícil imaginar la evolución de cada punto para llegar a construir con procedimientos geométricos la figura transformada, en este Lema, genuinamente newtoniano, se abre la puerta para ulteriores desarrollos con procedimientos analíticos. <<

[27] Transcribimos en notación actual la anterior demostración: por las cónicas,

$$\frac{hc^2}{(ad\cdot hb)} = \frac{ic^2}{id^2} = \frac{ke^2}{kd^2} = \frac{el^2}{(al\cdot lb)}$$

y por tanto también están en dicha razón subduplicada

$$\frac{hc}{\sqrt{(ah\cdot hb)}} = \frac{ic}{id} = \frac{ke}{kd} = \frac{el}{\sqrt{(al\cdot lb)}} =$$

$$\frac{hc + ic + ke + el}{\sqrt{(ah\cdot hb)} + id + kd + \sqrt{al\cdot lb}} = \frac{ih + kl}{\sqrt{(ah\cdot hb)} + ik + \sqrt{(al\cdot lb)}}\text{»}$$

Por otra parte conviene recordar que la expresión «lado cuadrado de un rectángulo» significa raíz cuadrada del área del mismo; si un rectángulo AB x BC = M, entonces «lado cuadrado» de M = √M = L, y L^2 = M. <<

[28] Transcrito sería así:

$$\frac{EC}{CA} = \frac{CA}{CL} \text{ y por tanto } \frac{EC}{CA} = \frac{EC - CA}{CA - CL} = \frac{EA}{AL}$$

y también

$$\frac{EA}{EA + AL} = \frac{EC}{EC + CA} = \frac{EA}{EL} = \frac{EC}{EB} \ldots <<$$

[29] Este Lema también ha sido objeto de discusión entre analistas. F. Cajori en el Apéndice de su edición, pág. 647, menciona a H. Brougham y E. J. Routh, *Analytical View of Sir Isaac Newton's Principia*, Londres, 1855, págs. 72-74, y aduce el ejemplo o contraejemplo con el cual éstos apoyan su desacuerdo con el Lema. El punto en cuestión se centra en la frase «… la ecuación por la que se halla la intersección de dos líneas contiene todas las intersecciones entre ellas mediante otras tantas raíces y alcanza, por tanto, tantas dimensiones como intersecciones haya». Pero la cuestión planteada por Newton recae sobre la inexistencia de una ecuación canónica que resuelva de modo general la cuadratura de cualquier sección oval valiéndose de coeficientes racionales. En cambio, el contraejemplo mencionado por Cajori permite afirmar que la conclusión expresada por Newton, «no existe figura oval cuya área cortada por rectas cualesquiera pueda ser determinada de modo general por medio de ecuación alguna de este tipo» está formulada imprecisamente. Quizá la aserción hubiera sido menos falsable si se hubiese escrito: «no existe ecuación general alguna de este tipo que pueda permitir resolver la cuadratura de cualquier figura oval cortada…», para lo cual no bastaría con resolver un caso singular dado.

Por lo demás, Leibniz y Huygens también juzgaron la conclusión y el enunciado arriesgados. Leibniz en sus «Marginaba» escribió «Error» al final del enunciado mientras Huygens le hacía ver que Newton no había definido con precisión lo que entendía por «figura oval» y Leibniz creía poder ofrecer un contraejemplo que cita en su Marginaba: vide Marginaba, págs. 101-104; y Gerhardt, Mathematische Schriften II/84.

Por otra parte Newton omitió en la primera edición la palabra «cuadrado» en la frase «cuadrado de la recta dentro del ovalo», del primer párrafo, y sólo lo cambió tras ser advertido por Fatio y Gregory, aunque Leibniz y Huygens ya se habían percatado. <<

[30] Newton reescribió este Escolio para la segunda edición, quizá porque la primitiva redacción era poco directa e implicaba unas aproximaciones demasiado prolijas. Se basaba en ella en las relaciones respecto al tiempo que existen, al pasar por los ábsides de una elipse, entre las áreas barridas por un vector —el radio de un círculo y el de la elipse— las velocidades y los sectores nacientes o evanescentes en ambas. Pero el método, que se aproximaba bastante en los ábsides, alcanzaba un error de «casi la cincuentava parte del área total de la elipse en las cuadraturas, o algo más»; lo que exigía una ulterior aproximación. En la nueva redacción la aproximación de ángulos mediante unas relaciones trigonométricas resulta más directa y exacta. <<

[31] Este método «bien conocido» era explicitado en la primera edición: «el método bien conocido del Dtr. Seth Ward obispo de Salisbury de mi mayor veneración…». <<

[32] Newton en las copias de la primera edición añadió aquí algunos párrafos que luego no trasladó a las ediciones siguientes. Primero pensó un Corolario 4 que decía: «Concedidas las cuadraturas de las curvas, se pueden hallar las líneas curvas según las cuales los cuerpos cayendo o bien se acercan continuamente al horizonte, o bien observan otras leyes asignadas de movimiento».

Y quizá quiso añadir un Escolio en el cual perfilaba más el problema: Escolio: Hemos mostrado en la Proposición IX del Libro I que un cuerpo con una fuerza centrípeta inversamente proporcional al cubo de la distancia, se puede mover en una espiral que corta todos los radios en un ángulo dado. Pero el cuerpo, según mis diferentes velocidades, se moverá en curvas distintas. Si la velocidad fuese aquélla con la que el cuerpo pudiera girar a una distancia en torno al centro de fuerzas, se moverá en dicha espiral. Pero si la velocidad es menor, el cuerpo se moverá en una espiral (VPQ) que con infinitas espirales descenderá al centro pero no ascenderá al infinito. Si la velocidad es mayor, el cuerpo se moverá en una curva (VPQ) que tiene una asíntota, y según su brazo hiperbólico se va al infinito… <<

[33] En el Ejemplo II de esta Proposición, ofrece Newton una primera noticia de su método de series convergentes. Aquí deja de lado el método «sintético» y acude al análisis, aunque sea con carácter auxiliar y transitorio. Debe tenerse presente la figura principal de la Proposición anterior, que no incluíamos en esta Proposición porque no aparece reproducida tampoco en el original latino. <<

[34] El párrafo que sigue es añadido en la segunda edición sustituyendo a otro más largo y con razonamiento menos preciso. Incluía también algún problema de términos; decía «ratio dimidiata» en lugar de «subduplicata». <<

[35] La introducción de esta sección XI es notable por varias razones. En primer término, por el reconocimiento que implica de la inadecuación de la hipótesis matemática de «un centro inmóvil» cuando se proyecta sobre puntos físicos. La interacción entre cuerpos exigida por la Ley tercera hace que el modelo matemático, relativamente lineal y sencillo, se presente necesitado de una complejidad creciente si es que quiere seguir dando cuenta de la naturaleza de las cosas.

Otro problema que no se escapa a la consideración de Newton es el de que los centros de atracción se hallan a efectos cinemáticos en los centros comunes de gravedad (centros de giro) lo cual exige nuevamente una mayor complejidad.

Finalmente, Newton se hace cargo de la ambivalencia en que ha de moverse al hablar de atracciones al referirse a las fuerzas centrípetas. ¿Son acciones continuas o impulsos? Newton —vide Proposición LVII y siguientes— siempre habla de atracciones, cuerpos atrayentes, etc.; pero por otra parte —Proposición LXVI y siguientes— también aparece este problema, junto con el de la pluralidad de cuerpos, en la medida en que el tratamiento de las atracciones es un tratamiento de acción continua global, sin poder pasar a un tratamiento por impulsos, entre otras razones porque la Tercera Ley implicada en el juego exige simultaneidad entre acciones y reacciones. Sin embargo, los impulsos parecen exigir un tiempo y un mecanismo de transmisión, y ello, si fueran los impulsos inversamente proporcionales a los cuadrados de las distancias, haría que unas veces llegaran más frecuentemente —cuando los cuerpos se acercan— y otras más espaciadamente —cuando se alejan—, y lo que es aún más duro de admitir: suponen un intermediario físico. Newton afirma que «ahora nos movemos en matemáticas» y no entra en disputas físicas. <<

[36] No sólo se plantea ahora el problema de los dos —o tres— cuerpos de la Proposición LXVI y siguientes sino que se adelanta una condición que aparece tomo Hipótesis I en el Libro III. Newton no conoce movimientos de desplazamiento del Sol, aunque sí conoce movimientos oscilatorios del mismo. «El movimiento absoluto de los cuerpos en el espacio inmóvil» debe, pues, entenderse ion la perspectiva histórica pertinente, independientemente de la cuestión filosófica n ontológica sobre la naturaleza del espacio. No obstante, en esta Proposición, en imito que condición matemática no ofrece problemas filosóficos. Más bien hace posible abordar el problema mediante la restricción que ella introduce al eliminar, aunque sea por hipótesis de trabajo —si bien aquí sea algo más que eso— otros movimientos de fondo. <<

Notas del libro segundo

[37] Esta es la segunda vez que Newton —la primera en la Proposición XLV anterior— aplica su método de fluxiones. <<

[1] Newton añade en la segunda edición los Corolarios 1 y 2, con lo cual los cinco primitivos llevan los números 3 a 7. <<

[2] Newton hacia 1712, en un prefacio que pensó anteponer y del cual se conservan diferentes borradores, o estadios del borrador —vide *Math. Papers*, VIH, 625 y siguientes— dice que utilizó en los *Principia* el método de las cuadraturas muy ampliamente, «eodemque plurium usus sum in componendo libro precedente de *Philosophiae Naturalis Principiis Mathematicis…*»; pero además de los dos casos del Libro I ya mencionados, sólo es posible citar unos pocos más en los cuales debe tratar velocidades instantáneas (Proposiciones II, III, V-IX y XI-XIV del Libro II), en las cuales se ve obligado a tratar V*x* y V*y* (*dx/dt*) y (*dy/dt*). En todos los demás casos, pese a las diferentes afirmaciones, a lo largo de muchos textos aderezados contra Leibniz en su disputa, no queda rastro alguno de que hubiese utilizado «su método» para obtener ni sus resultados ni las demostraciones de los mismos. Antes bien, D. T. Whiteside en *Math. Papers*, 8, pág. 68, afirma que «sobre la base de un examen del texto de los *Principia* y de los manuscritos auxiliares de los mismos se apoya nuestra conclusión de que por el contrario la mayor parte de sus Proposiciones fueron derivadas esencialmente en la misma forma en que fueron impresas».

Pero este Lema II siempre fue aducido por Newton como argumento central de su prioridad y de su utilización del método infinitesimal tanto en los *Principia* como antes. En él presenta el principio fundamental del cálculo. Pero hay en torno un problema histórico que puede ser de alguna utilidad recordar aquí. Se trata de un Escolio que apareció en la primera edición al final del Lema. Volvió a aparecer en la segunda edición, aunque Newton redactó algunos arreglos y versiones alternativas, que luego descartó, y que finalmente desapareció en la última edición. Ya había muerto Leibniz en noviembre de 1716 y ya habían circulado más que de sobra las razones de uno y de otro. Conf. A. R. Hall *Philosophers at War*, Cambridge U. Press, 1980. El Escolio decía: «En cartas que hace diez años mediaron entre el sabio geómetra G. G. Leibniz y yo mismo, como manifestare que yo era poseedor de un método para determinar máximos y mínimos, para tratar tangentes y para hacer cosas semejantes, método que servía con términos irracionales, tanto como con términos racionales, y mediante trasposiciones literales que implicaban la siguiente Proposición (“dada una ecuación cualquiera que contenga cantidades fluentes, hallar las fluxiones, y viceversa”) la ocultare; me contestó el muy preclaro varón que también él había llegado a un método semejante, y comunicó un método apenas diferente al mío más que en las formas verbales y de notación. El fundamento de ambos métodos se halla en este Lema». Para el episodio completo, véase Hall, o. c. y *Math. Papers*, VIII. <<

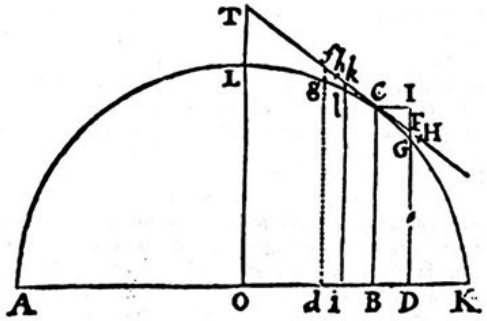
[3] Aparecían en la 1.ª edición (E_1) dos Corolarios más: «Corolario 4: pero la partícula de tiempo, en que la partícula mínima NKLO de espacio es descrita es como el rectángulo KN·PQ, pues como el espacio NKLO es como la velocidad multiplicada por la partícula de tiempo; la partícula de tiempo será como dicho espacio dividido por la velocidad, es decir, como el rectángulo mínimamente pequeño KN·KL dividido por AP. Antes era KL como AP·PQ. Luego la partícula de tiempo es como KN·PQ o lo que es lo mismo, como PQ/CK. Q. E. D.

COROLARIO 5. Por el mismo argumento la partícula de tiempo en la que la partícula *nklo* de espacio es descrita en el ascenso es como PQ/CK». La figura era notablemente distinta y equivalía a la superposición de la de la Proposición VIII con la de la IX. <<

[4] En la primera edición Newton cometió un error que no detectó ni corrigió en la revisión para la segunda edición. Tampoco Cotes en su propio repaso. Cuando ya estaba impreso el cuadernillo correspondiente a esta Proposición (confer. *Correspondencia*, V, 951, pág. 347). Nicolás Bernuilli llegó de improviso a Londres y a primeros de octubre de 1712 hizo saber a Newton que su tío Bernuilli había encontrado un error en esta Proposición. Newton puso manos a la obra, paró la impresión que Cotes estaba realizando y reconstruyó el razonamiento. Decía la versión inicial: «Sea AK el plano perpendicular al plano de la figura; ACK la línea curva; C el cuerpo que se mueve en ella; y FCf la recta tangente a la misma en C. Imagínese que el cuerpo C ora progresa de A a K por la línea ACK, ora regresa por la misma línea; y en el progreso es impedido por el medio, al igual que en el regreso es impelido por él, de tal modo que en los mismos lugares siempre sea igual la velocidad del cuerpo en la ida o en la vuelta. Pero en tiempos iguales describa el cuerpo al ir un arco mínimo CG y el regresar el arco C*q*; y sean CH, *Ch* longitudes rectas iguales, que describirían los cuerpos partiendo de C en dichos tiempos y sin las acciones del medio y de la gravedad: y desde los puntos C, G, *g*, trácense sobre el plano horizontal las perpendiculares CB, GD, *gd* de las cuales GD, *gd* que toquen a la tangente en F y f. Por la resistencia del medio ocurre que el cuerpo que progresa en lugar de la longitud CH describe solamente la longitud CF; y por la fuerza de la gravedad el cuerpo es trasladado desde F a G: y por tanto, la pequeña línea HF debida a la fuerza de la resistencia y la pequeña línea FG debida a la fuerza de la gravedad se generan a la vez. Por lo tanto (por el Lema X del Libro I), la pequeña línea FG es como la fuerza de la gravedad y el cuadrado del tiempo conjuntamente, y (por estar dada la gravedad) como el cuadrado del tiempo; y la pequeña línea HF como la resistencia y la pequeña línea FG. Y de aquí que la resistencia sea como HF directamente y como FG inversamente o como HF/FG. Esto es así en segmentos nacientes. Pues en segmentos de magnitud finita estas razones no son exactas.

Y por un argumento semejante fg es como el cuadrado del tiempo y por ser iguales los tiempos, igual a FG; y el impulso con que es urgido el cuerpo que regresa es como hf/fg. Pero el impulso del cuerpo que regresa y del cuerpo que progresa, al comienzo mismo del movimiento, son iguales y, por tanto, también son iguales los proporcionales a los mismos hf/fg y HF/FG; y por la igualdad de fg y FG. también son iguales hf y HF. y también CF, CH (o *c*H) y Cf están en progresión aritmética, y de aquí que HF es la semidiferencia de las propias cf y CF; y la resistencia que antes fue como HF/FG es como (Cf - CF)/FG. Pero la resistencia es como la densidad del medio y el cuadrado de la velocidad. Y la velocidad es como la longitud recorrida CF directamente y el tiempo $\sqrt{\text{FG}}$ inversamente, esto es como CF/$\sqrt{\text{FG}}$, y por tanto el

cuadrado de la velocidad como CF^2/FG. Por lo cual la resistencia, y su proporcional $(Cf - CF)/FG$ es como la densidad del medio y CF^2/FG conjuntamente; y de aquí que la densidad del medio es como $(Cf - CF)/FG$ directamente y CF^2/FG inversamente, esto es, como $(Cf - CF)/CF^2$. Q. E. I.



Corolario 1. De aquí se sigue que si en Cf se toma Ck igual a CF y sobre el plano horizontal AK se deja caer la perpendicular *ki* secante de la curva ACK en *b*; la densidad del medio será como $(FG - kl)/CF \times (FG + kl)$. Pues fC será a *k*C como $\sqrt{fg}$ o *de* $\sqrt{FG}$ a $\sqrt{kl}$. Y por partes fk a kC, esto es, $Cf - CF$ a CF como $\sqrt{FG} + \sqrt{kl}$ a $\sqrt{kl}$, esto es (si ambos términos se multiplican por la razón primera naciente $\sqrt{FG} + \sqrt{kl}$) como $FG - kl$ a $kl + \sqrt{FG} \times kl$ o también $FG + kl$. Pues la razón primera de las nacientes $kl + \sqrt{FG} \times kl$ y $FG + kl$ es de igualdad. Escríbase, por tanto $(FG - kl)/(FG + kl)$ en lugar de $Cf - CF$; y la densidad del medio que fue como $(Cf - CF)/CF^2$ resultará como $(FG - kl)/CF \times FG + kl$.

Corolario 2. De donde, puesto que 2HF y $Cf - CF$ son iguales, y FG y *kl* (por la razón de igualdad) compongan 2FG; será 2HG a CF como $FG - kl$ a 2FG; y por ello HF a FG, esto es, la resistencia a la gravedad, como el rectángulo $CF \times FG - kl$ a $4FG^2$».

Corolario 3. Es el actual Corolario 2.

Hemos introducido subrayados para significar la línea de argumentación errónea. 1) «…la pequeña línea FG debida a la fuerza de la gravedad…» FG no sólo es generada por la gravedad, sino también por la componente (aquí negativa) de la resistencia al movimiento según la linea CFH. Por ser negativa tiende hacia abajo coincidiendo con la gravedad y es segunda derivada de un tercer término en θ^3: 2) «Como la fuerza de la gravedad y el cuadrado del tiempo conjuntamente». Aquí se olvida de que FG es además función de la componente resistencial del paso anterior, y por ello 3) la resistencia resultaría como HF/2FG. Y así en 4) no son iguales fg y FG; la diferencia entre ambos (el tercer término de la serie que no es despreciable) es $\frac{1}{3}(gP/v)\theta^3$. Esta diferencia desconocida u olvidada por Newton le lleva a considerarlos iguales en 5) donde iguala las raíces $\sqrt{fg}$ y $\sqrt{FG}$. De no ocurrir esto y haber seguido operando con $\sqrt{fg}$ habría llegado a un resultado exacto en este Corolario 1. Bernuilli y Leibniz tomaron pie en este error para mantener que Newton no conocía en 1687 la técnica diferencial puesto que había tropezado en el desarrollo de la serie al no llegar a una derivada segunda. Whiteside en *Mat. Papers*, 8, 312-419 hace una exposición completa del contencioso, a la vez que muestra hasta dónde llegaba la competencia analítica de Newton. Por otra parte, en el Ejemplo 1 suprimió un pequeño párrafo en el que introducía los valores del problema por las variables de la serie porque en ellos aparecía un error a consecuencia de lo anterior, pero el método es el mismo. <<

[5] En las dos ediciones anteriores faltaban las figuras que ahora aparecen en esta Proposición, pero aparecían como figuras de los Casos 1, 2 y 3 de la Proposición XIII. En la tercera edición se trasladan todas a la Proposición XIV y se introducen las de la Proposición XIII tomando sólo parte de las antiguas, como observará el lector. <<

[6] Escolio: se introduce en la tercera edición. <<

[7] En la primera edición el enunciado era: «Las partículas que huyen unas de otras con fuerzas que son inversamente proporcionales a las distancias de sus centros componen un fluido elástico cuya densidad es proporcional a la compresión. Y viceversa, si la densidad de un fluido compuesto de partículas que huyen unas de otras es como la compresión, las fuerzas centrífugas de las partículas son inversamente proporcionales a las distancias de los centros».

En el Escolio suprimido en la segunda edición, en el antepenúltimo párrafo, un fragmento decía «De modo que si cada partícula con su fuerza, que sea inversamente como la distancia al centro de cada lugar, repele a todas las otras hasta el infinito; las fuerzas con las cuales el fluido puede ser comprimido y condensado igualmente en vasos semejantes, serán como los cuadrados de los diámetros de los vasos; y por ello la fuerza con la cual el fluido se comprime en el mismo vaso, será inversamente como el “latus cubicum” a la potencia 5 de la densidad». <<

[8] El Escolio general con que se cierra esta Sección VI se hallaba en la primera edición al final de la Sección VII. En su correspondencia con Cotes (Corresp. V-784, pág. 42) se limita a decir que «deseo que se imprima al final de la sexta sección inmediatamente después de la Proposición XXXI» y aunque hay algunas correcciones mínimas no hay ningún cambio sustancial en el aparato experimental incluido en el Escolio. En cambio suprimió un párrafo que seguía inmediatamente a la 4.ª tabla que decía:

«La resistencia aquí nunca aumenta más que la razón del cuadrado de la velocidad. Y es verosímil que eso ocurriera en un péndulo mayor, siempre que el área aumente en razón del péndulo. Pues la resistencia tanto en el aire como en el agua, si la velocidad aumenta por pasos hasta el infinito, deberá aumentar al fin en una razón mayor que el cuadrado, toda vez que en los experimentos aquí descritos la resistencia es menor que la razón demostrada para los cuerpos muy veloces en las Proposiciones XXXVI y XXXVIII de este libro. Pues los cuerpos muy veloces dejan a su espalda un espacio vacío y, por tanto, la resistencia que sufren en su parte delantera, en nada es disminuida por la presión del medio en sus partes posteriores».

Hay algunos otros párrafos suprimidos que amplían la idea del anterior o en los que explica cómo hacerla más precisa mediante péndulos cónicos, cuyas resistencias están más acordes con la idea principal. Y finalmente (pág. 524) suprime dos pequeños párrafos. El primero (a continuación de «fluido sutil») decía: «pero opino que la causa ha de ser muy otra. Pues los tiempos de oscilación de la caja llena son menores que los tiempos de oscilación de la caja vacía y, por tanto, la resistencia de la caja llena en la superficie externa es mayor, según la velocidad y la longitud del espacio descrito oscilando, que la de la caja vacía. Al ser esto así, la resistencia de las cajas en sus partes internas o es nula o claramente insensible».

El segundo párrafo (después del último punto y aparte) decía: «Con el mismo método con el cual hemos hallado la resistencia de los cuerpos esféricos en el agua y en el mercurio puede hallarse la resistencia de cuerpos con otras figuras; y así las distintas figuras de naves construidas en modelos pequeños pueden compararse entre sí, para probar con pocos gastos cuales sean las naves más aptas para navegar». <<

[9] La revisión de esta sección VII tuvo su origen en las objeciones de Fatio (vide Proposición XXXVI, nota 12) y se remonta a los primeros años de la década anterior. En carta de Cotes a Newton (Corresp. V, 805, pág. 65) aquél acusa el recibo de la revisión de toda esta parte. También hay que destacar la discusión de Cotes sobre esta nueva revisión —en la carta mencionada y en las siguientes— porque aclaran notablemente las nuevas consideraciones de otros experimentos que llevan a Newton a la formulación actual (Torricelli, Halley, etc.).

La reforma consiste en:

a) Las Proposiciones XXXII y XXXIII son prácticamente las mismas.

b) En la Proposición XXXIII se suprimen los Corolarios 6, 7 y 8. El Corolario 9 pasa a ser, por tanto, el 6.º

c) Se suprime la Proposición XXXIV y la primitiva XXXV se renumera con el núm. XXXIV.

d) El resto de la sección, Proposición XXXV a XL y Escolio, se redactan de nuevo.

No incluimos aquí la traducción de este largo texto suprimido. Puede verse en Ap. I del tomo II de la edición de Koyré-Cohen. Cabe añadir que esta redacción junto con los experimentos correspondientes llevó a Newton largo tiempo por cuanto que Cotes rehizo los experimentos de Newton y halló objeciones que honestamente propuso. La respuesta de Newton (24 de marzo de 1711) incluye un largo ensayo relativo a estas Proposiciones XXXVI y siguientes (Corresp. V, 826, pág. 103, etc.). <<

[10] Corolarios: los corolarios suprimidos eran:

Corolario 6. «Mas como las partículas de los fluidos, debido a las fuerzas por las que mutuamente se repelen, no pueden moverse sin que a la vez agiten a otras partículas en torno y, por tanto, se muevan más difícilmente entre ellas que si careciesen de estas fuerzas; y cuanto mayores sean sus fuerzas centrífugas, tanto más difícilmente se muevan entre ellas: parece evidente que un proyectil en semejante fluido se moverá más difícilmente cuanto más intensas sean dichas fuerzas; y por tanto si la velocidad de un cuerpo muy veloz de los Corolarios anteriores se disminuyese, puesto que la resistencia disminuiría como el cuadrado de la velocidad, siempre que las fuerzas de las partículas disminuyesen en la misma razón cuadrada; y en cambio no se disminuirían las fuerzas, está claro que la resistencia disminuiría en menor razón que el cuadrado de la velocidad.

Corolario 7. Además, puesto que las fuerzas centrífugas llevan a aumentar la resistencia en la medida en que las partículas propagan su movimiento por el fluido mediante dichas fuerzas a mayor distancia mutua; y puesto que dicha distancia está en menor proporción respecto a cuerpos mayores; es evidente que el aumento de resistencia debido a dichas fuerzas para cuerpos mayores sea de menor importancia; y por tanto cuanto mayores sean los cuerpos tanto más exactamente la resistencia de los cuerpos en desaceleración decrece en razón cuadrada a la velocidad.

Corolario 8. De donde también la susodicha razón cuadrada se cumplirá más exactamente en los fluidos que, a igual densidad y fuerza elástica, consten de partículas más pequeñas. Pues si se disminuyen aquellos cuerpos mayores y las partículas del fluido, manteniendo sus densidades y su fuerza elástica, disminuyen en dicha razón, se mantendrá la misma razón de resistencia que antes; como se colige fácilmente de lo anterior». <<

[11] Era la XXXV de la primera edición. La primitiva XXXIV decía:

«PROPOSICIÓN XXXIV. TEOREMA XXVII.

Lo demostrado en las dos Proposiciones anteriores no se cumple cuando las partículas de los sistemas se suceden continuamente unas a otras, siempre que dichas partículas sean de gran lubricidad.

Imagínese que las partículas mutuamente se repelen con cualesquiera fuerzas y dichas fuerzas aumenten hacia la superficie de las partículas, hasta el infinito, mientras al alejarse de ellas disminuyen muy rápidamente y desaparecen de pronto. Imagínese también que los sistemas se comprimen de modo que sus partes se toquen salvo en la medida en que dichas fuerzas impiden el contacto. Pero sean los espacios por los que se difunden las fuerzas de las partículas sumamente angostos, de modo que las partículas se acerque mutuamente lo más posible. Se deslizarán con la misma facilidad que si fuesen sumamente lubrificadas, y si se apoyan entre ellas rebotaran unas de otras gracias a las fuerzas susodichas como si fuesen elásticas. En consecuencia, los movimientos serán los mismos en ambos casos, salvo en la medida en que la pequeña distancia de las partículas no contiguas entre ellas genere una diferencia, la cual puede disminuirse hasta el infinito disminuyendo los intersticios entre partículas. Ahora bien, lo demostrado en las dos Proposiciones precedentes se cumple en partículas no contiguas entre sí, y esto aunque los intervalos entre partículas, disminuyendo los espacios por los que se difunden las fuerza, disminuyen infinitamente. Y por tanto, lo mismo se cumple en partículas que se tocan, exceptuando las diferencias que al final resulten menores que unas diferencias dadas. Por tanto, digo que se cumplen exactamente. Si lo niegas asigna una diferencia para un caso cualquiera. Pero ya se ha probado que la diferencia es menor que cualquiera dada. Luego la diferencia se asigna erróneamente y, por tanto, no existe. Q. E. D.

Corolario 1. Por tanto, si todas las partes de dos sistemas reposan entre ellas excepto las que son mayores que las otras y son entre sí correspondientes y situadas semejantemente entre las demás. Estas, lanzadas de cualquier modo según líneas semejantemente dispuestas suscitarán en los sistemas movimientos semejantes y en tiempos proporcionales completarán espacios semejantes proporcionales a sus diámetros; y serán resistidas en razón compuesta del cuadrado de las velocidades y el cuadrado de los diámetros y en razón de la densidad de los sistemas.

Corolario 2. De donde, si dichos sistemas son dos fluidos semejantes y sus dos partes mayores son cuerpos proyectados dentro de ellos: pero las partículas de los fluidos son de gran lubricidad y proporcionales a los cuerpos en cuanto a magnitud y densidad; los cuerpos describirán en tiempos proporcionales espacios semejantes y

proporcionales a sus diámetros, y serán resistidos en la razón expuesta en el Corolario anterior.

Corolario 3. Por tanto, en dicho fluido un proyectil dado en magnitud es resistido en razón del cuadrado de la velocidad.

Corolario 4. Pero si las partículas del fluido no son de la misma lubricidad o se agitan mutuamente con cualesquiera fuerzas que disminuyen la libertad de los movimientos; los proyectiles más lentos superarán la resistencia más difícilmente y, por tanto, serán resistidos en razón mayor que la del cuadrado de la velocidad». <<

[12] Esta Proposición —correspondiente a la XXXVII de la primera edición— primero fue objeto de una profunda corrección y después de una total reconstrucción. El motivo parece haber sido la crítica que de la primitiva hizo Fatio a Newton. Fatio escribió al margen en su copia de la primera edición: «No he podido librar a Newton de los errores contenidos en esta Proposición más que mediante el experimento; a saber, construyendo un vaso destinado a este propósito». Conf. Cohen: *Introd*., pág. 182. <<

[13] Cotes discute y analiza la nueva redacción de Newton con gran penetración. El lector interesado en seguir de cerca la evolución de todo este tema puede ver *Corresp*. V, 829 y siguientes, págs. 107 y siguientes, y también *Math. Papers*, VI, págs. 456 y siguientes. <<

[14] Escolio.

De nuevo aparecen correcciones en la base experimental, quizá debido a la premura con que inicialmente se hicieron, pero que a la vez son signo de la «debilidad» del aparato teórico con que Newton abordó toda la mecánica de fluidos. De hecho ésta no logró entrar en el buen camino, pese a las contribuciones de Daniel Bernuilli, Clairaut y D'Alambert, hasta que Euler —«Mechanica sive motus scientia analityce expósita», en 1736, y las tres memorias de 1753-1755. «Principes Generaux de l'etat d'equilibre des fluides», «Principes Generaux du moviment des fluides» y «Continuation des reserches sur la theorie du moviment des fluides»— estableció las bases definitivas de la hidrodinámica.

En este Escolio Newton había incluido unos párrafos que luego suprimió en la segunda edición que seguían al primer párrafo que concluye con «979 pies». Añadió: Escribe Mersenne en su *Balística*, Proposición XXXV, que él, hechos los experimentos oportunos, había hallado que el sonido en cinco segundos recorre 1150 toesas francesas (esto es, 6900 pies franceses). De donde toda vez que el pie francés es al inglés como 1068 a 1000, el sonido en un tiempo de un segundo debería recorrer 1474 pies ingleses. También escribe Mersenne que Robervall, geómetra insigne, había observado en el asedio de *Theodon* que el sonido de los cañones se había oído 13 ó 14 segundos después de ver el fuego, cuando se hallaba apenas a media legua de los cañones. La legua francesa tiene 2500 toesas y por tanto, en el tiempo de 13 o 14 segundos según la observación de Robervall, recorrió 7500 pies parisinos, y en un segundo 560 pies parisinos y casi 600 ingleses. Estas observaciones difieres mucho entre sí y nuestro cálculo se sitúa en lugar intermedio. En el pórtico de nuestro colegio de 208 pies de largo, el sonido producido en cualquiera de los extremos con cuatro rechazos produce cuatro ecos. Pero hechos los experimentos hallé que en cada recurso del sonido un péndulo de seis o siete pulgadas de largo oscilaba al primer choque con la ida y al segundo con la vuelta. No podía medir con exactitud la longitud del péndulo pero con la longitud de cuatro pulgadas las oscilaciones eran demasiado rápidas y con nueve pulgadas parecían demasiado lentas. De donde, el sonido yendo y viniendo completó 416 pies en menos tiempo que la oscilación del péndulo de nueve pulgadas y mayor que uno de cuatro pulgadas; esto es, en un tiempo menor que 28¾‴ y mayor que 19⅙‴ , y por tanto, en el tiempo de un segundo completa en pies ingleses más de 866 y menos que 1272, y por lo tanto es más veloz que lo observado por Robervall y menos que lo observado por Mersenne. Pero además determiné después con observaciones más exactas que la longitud del péndulo debe ser mayor que cinco y media pulgadas y menor que ocho pulgadas; y por tanto que el sonido en el tiempo de un segundo recorre más de 920 pies ingleses y

menos que 1085. Por tanto, los movimientos de los sonidos, según el cálculo geométrico expuesto más arriba, situado entre dichos límites, concuerda con los fenómenos hasta donde fue posible hasta ahora intentarlo. Por tanto, puesto que este movimiento depende de la densidad del aire, se sigue que los sonidos consisten en la agitación de todo el aire y no del éter o de otro aire más sutil.

Algunos experimentos con el sonido propagado en vasos vacíos de aire parecen estar en contra, pero los vasos difícilmente pueden vaciarse de todo el aire; y cuando se vacían bastante los sonidos suelen disminuir bastante; por ejemplo, si de todo el aire permanece en el vaso sólo la centésima parte, el sonido deberá ser cien veces más tenue y, por tanto, no debe oírse menos que si alguien escuchando el mismo sonido emitido en el aire libre, se alejase después a una distancia diez veces mayor del objeto sonoro. Por tanto, habrá que comparar dos cuerpos igualmente sonoros y cuyas distancias del oyente estén en razón de la mitad de las densidades del aire; y si el sonido del primer cuerpo no supera al del segundo, desaparecerá la objeción.

Conocida la velocidad de los sonidos, se conocen también los intervalos de las pulsaciones. Escribe Mersenne (Libro I de los «de Harmonía». Proposición IV) que él (hechos ciertos experimentos que allí describe) encontró que una cuerda tensa en un segundo vibra 104 veces cuando está al unísono con una flauta de órgano de cuatro pies abierta o de dos pies cerrada a la que los organistas llaman C fa ut. Por tanto, son 104 pulsaciones en el espacio de 968 pies los que el sonido describe en el tiempo de un segundo y, por tanto, una pulsación ocupa un espacio de 9¼ pies aproximadamente; es decir, casi el doble de la flauta. De donde es verosímil que las anchuras de las pulsaciones en los sonidos de todas las flautas abiertas sean iguales al doble de las longitudes de las flautas. <<

[15] Añadía en el manuscrito original un Corolario 12 que, sin duda, canceló por redundante. «Corolario 12. Y por un argumento semejante es evidente que muchos orbes sólidos y concéntricos de los cuales los exteriores arrastren a los interiores no pueden permanecer en movimiento cualquiera sin aceleración y retardación, salvo que todos a la vez giren a modo de un solo cuerpo sólido sin traslación mutua o que el exterior y el interior, merced a alguna fuerza externa impuesta continuamente, sean obligados perseverar en sus movimientos; en cuyo caso los orbes intermedios girarán en movimientos intermedios, salvo que con fuerzas impuestas también externas les impida moverse o se les haga moverse de otro modo. Suprímanse semejantes fuerzas externas, y los orbes, frenados por el rozamiento mutuo, poco a poco reposarán entre ellos». <<

[1] Newton cambió notablemente este Epígrafe: se esquematizan los cambios en el cuadro adjunto.

1.ª edición	1.ª edición	1.ª edición
Hipótesis 1	Regla I	Regla I
Hipótesis 2	Regla II	Regla II
Hipótesis 3	Suprimida	Suprimida
—	Regla III	Regla III
—	—	Regla IV
Hipótesis 4	Hipótesis I	Hipótesis I
Hipótesis 5	Fenómeno I	Fenómeno I
—	Fenómeno II	Fenómeno II
Hipótesis 6	Fenómeno III	Fenómeno III
Hipótesis 7	Fenómeno IV	Fenómeno IV
Hipótesis 8	Fenómeno V	Fenómeno V
Hipótesis 9	Fenómeno VI	Fenómeno VI

Las Hipótesis 1 y 2 apenas difieren de las Reglas I y II. Pero la Hipótesis 3. única suprimida por completo, como se ve por la tabla anterior, decía: «Todo cuerpo puede transformarse en un cuerpo de otro género cualquiera, y adoptar sucesivamente todos los grados intermedios de cualidades».

Respecto a las diferencias —de índole epistemológica sobre todo— que introduce en las Reglas I y II respecto a las primitivas Hipótesis cabe señalar: En la Regla I introduce las frases: «Ya dicen los filósofos: la naturaleza nada hace en vano, y vano sería hacer mediante mucho lo que se puede hacer mediante poco…». En la Regla II en el enunciado mismo hace aparecer un matiz nuevo: la Hipótesis decía: «Y por tanto las causas son las mismas para los efectos naturales del mismo género». Y respecto a la Hipótesis 4 cambia de sitio y como Hipótesis 1 la hallará el lector a continuación de la Proposición X.

Hasta qué punto esta Regla III es un «manifiesto» filosófico por una parte y un «caveat» metodológico por otro, es un problema discutido. No obstante, y poniéndolo en relación con la introducción de Cotes, puede verse en ella una respuesta a las críticas «continentales». Cuando afirma que «éste es el fundamento de toda la filosofía» cabe entender que se refiere a la filosofía natural de su tratado. Pero si se extendiese a toda sana filosofía frente a la que «fantasea temerariamente sueños», ha de verse en esta Regla una toma de posición, por parte de Newton que es más

explícita aquí que lo fuera en 1685-1687. Y, con todo, el lector sigue preguntándose si Newton ha hecho «principios matemáticos» a partir de la física o ha hecho física a partir de «principios matemáticos», en cuyo caso, parece que algo no cuadra con el discurso de esta Regla III. Sobre todo porque los hechos que «constan» experimentalmente y que recoge en el último párrafo, no parece que consten tan experimentalmente como se exige en la primera parte de la Regla. Y ello además, porque «gravitar» no es un observable, como Newton sabía muy bien. La Regla IV ofrece aún mayores motivos para la reflexión. Lo dejamos para el lector. <<

[2] Las tablas de observaciones varían bastante de la primera a las otras ediciones. Lo más notable, sin embargo, es la eliminación de las observaciones de Flamsteed y de toda referencia al mismo que era, en la primera edición, muy importante. Además, los resultados de Flamsteed eran los más próximos a los que Newton asume al final del texto del fenómeno. <<

[3] Este cálculo parece que es el que Newton realizó hacia 1666. El Escolio se añade en la tercera edición. Esta Proposición, de responder a esta conjetura, habría sido la primera de todas las que Newton se autoformuló, aunque no fuera satisfactoria por los «malos» datos manejados. Esto tampoco es muy seguro. Newton conocía valores para el semidiámetro terrestre y para la distancia de la Luna muy cercanos a las que aquí maneja. Snell ya en 1617 había fijado en 66j millas inglesas el valor del grado de latitud. Varen en la *Geografía* que editó Newton da unos datos erróneos que Newton corrige de acuerdo con los valores de Snell. <<

[4] Convendría advertir al lector que Newton no sólo introduce aquí la gravedad —previa la licencia que se autoconcede en el Escolio de la Proposición anterior— sino que también, aunque enunciado condicionalmente, introduce el vacío entre partículas. En la segunda edición añadió la última frase del Corolario 3 y todo el Corolario 4, en el que afirma que «se da de: hecho el vacío». Pero el lector puede aceptar el condicional o no. Si no lo acepta se verá en graves problemas —Corolario 3— y si lo acepta habrá aceptado también, junto con el vacío, unos corpúsculos, no sólo igualmente duros sino también infinitamente duros, por cuanto que su densidad, al carecer de poros no puede aumentarse y tampoco disminuirse en cada uno de ellos. <<

[5] Principio de la Gravitación Universal, formulado como gravitación de todos con todos, partes con todos y partes con partes. De la Proposición anterior y de ésta cabría inferir que, puesto que se postula una unidad mínima de materia, el corpúsculo, también correlativamente habría de aparecer la unidad de acción gravitatoria correspondiente al corpúsculo. Y aunque no fuese empíricamente determinable (tampoco lo era el corpúsculo ni su masa en el momento) Newton no se atrevió a postular esa unidad de acción gravitatoria. Hacerlo hubiera significado aceptar que la gravedad es una suma de impulsos, con los problemas anexos. <<

[6] Tenía 5 Corolarios que fueron objeto de revisiones y sucesivos cálculos. De los Corolarios 1 y 2 hizo el Corolario 1 y de los otros, siempre con correcciones, hizo los Corolarios 2, 3 y 4 de la tercera edición. La mayor parte de los cambios se refieren a las distintas magnitudes aparentes de los diámetros, las elongaciones, etc., debidas a Cassini, Flamsteed, Riccioli, Vandelin, etc., y los cálculos que con esas bases le permitían atribuir pesos, densidades, etc. a los diferentes cuerpos celestes. Pero el 5.º Corolario, transformado en el 4.º actual, decía: «Pero las densidades de los planetas están entre sí casi en razón compuesta de la razón de las distancias al Sol y de la razón de las raíces cuadradas (ratio dimidiata, correg. a subduplicata) de los diámetros aparentes vistos desde el Sol. A saber, las densidades de Saturno, Júpiter, Tierra y Luna de 60, 76, 387 y 700, son casi como los inversos de las distancias 1/9538, 1/5201, 1/1000 y 1/1000 multiplicado por las raíces de los diámetros aparentes 18″, 39½″, 40″ y 11″. Ya dijimos en el Corolario 2 que la gravedad en la superficie de los planetas está muy aproximadamente en razón de la raíz cuadrada de las distancias aparentes vistas desde el Sol; y en el Lema IV que las densidades son como dichas gravedades divididas por los diámetros verdaderos: y por tanto, las densidades son casi como las raíces de los diámetros aparentes divididas por los diámetros aparentes, esto es, inversamente como las distancias de los planetas al Sol multiplicadas por las raíces de los diámetros aparentes. Por tanto Dios situó a los planetas a diferentes distancias del Sol para que cada uno, según el grado de densidad, disfrutase de un grado mayor o menor de calor solar». Newton eliminó la referencia directa a Dios, aunque no la indirecta: sigue diciendo que: «…había que situar a distintas distancias del Sol a los planetas para que cada uno…». <<

[7] Introducido Junto con la última frase del Corolario 2, en la segunda edición. <<

[8] Respecto a las medidas que aparecen en esta Proposición, y a sus autores, hay cambios abundantes pero mínimamente relevantes para el argumento. Noorwood, Picard, Cassini son medidores del arco de meridiano. Sus valores comparados con los teóricos, y con los pendulares, permiten a Newton establecer con suficiente garantía sus conclusiones respecto a la relación entre ejes terrestres.

Por otra parte es éste uno de los test a que sería sometida la teoría de Newton y originó las expediciones medidoras del «grado» terrestre en el siglo XVIII, Clairant, Maupertuis y nuestros Jorge Juan y Ulloa. <<

[9] Aunque son numerosas las correcciones introducidas por Newton tanto para la segunda edición como para la tercera, no merecen mayor atención aquí por cuanto sería descender a detalles mínimos.

No obstante, a veces, introduce o suprime alguna frase que vale la pena tener presente. A continuación del párrafo que concluye con «hacia los polos», decía en E2:

«Y que la desigualdad de los diámetros de la Tierra se puede inferir más fácil y más ciertamente mediante experimentos pendulares o incluso mediante los eclipses de la Luna, que mediante arcos geográficamente medidos en el meridiano.

Esto es así bajo la hipótesis de que la Tierra consta de materia uniforme. Pues si la materia fuese hacia el centro algo más densa que en la superficie, las diferencias de los péndulos y de los grados de meridiano serán algo mayores que en la tabla anterior, debido a que si la materia sobrante hacia el centro con la cual la densidad es allí mayor se resta y se considera por sí misma, la gravedad hacia la Tierra restante uniformemente densa, será inversamente como la distancia del peso al centro; pero hacia la materia redundante será inversamente como el cuadrado de la distancia a dicha materia, muy aproximadamente. Luego la gravedad en el Ecuador hacia dicha materia redundante es menor que en lo computado anteriormente; y por tanto, la Tierra allí, por la disminución de la gravedad sube un poco más alta, y el exceso de las longitudes de los péndulos y de los grados hacia los polos serán algo mayores de lo que se definió anteriormente».

Estas y otras consideraciones fueron suprimidas en la tercera edición.

Las demás modificaciones están relacionadas con las medidas y sus diferentes formas y exactitudes, pero muestran la amplitud de su información y su interés por los datos observacionales. <<

[10] De nuevo el problema de los tres cuerpos. La teoría de la Luna es un ejemplo. Pero Newton disponía de otros, Sol, Júpiter, Saturno; Júpiter y sus satélites, etc. Intentos posteriores tampoco han alcanzado una solución totalmente satisfactoria, pese a los nombres que han dado amplia bibliografía al problema: D'Alambert, Clairaut, Euler, Lagrange, Laplace, etc. <<

[11] Este Escolio y la «Carta» de Machín fueron incluidos en la tercera edición, aunque nada añaden que fuese nuevo a esas alturas de los cálculos y de las observaciones disponibles. Pero sus disputas con Flamsteed y otros clientelismos pudieran dar alguna razón de esta inclusión. <<

[12] En la primera edición este Escolio era más breve y se detenía en mostrar la proximidad o lejanía entre valores calculados y observados. Pero ambas adolecían de inexactitudes, y por ello este nuevo intento, en el que Newton pone de manifiesto las posibilidades que encierra la teoría de la gravitación. La frase introductoria no deja de ser admirable: «que los movimientos lunares puedan calcularse mediante la teoría de la gravedad a partir de sus causas», es decir, de la naturaleza de la gravedad que es una causa de movimiento y, por tanto, más que un fenómeno cuyos valores matemáticos se hallan recogidos en la teoría. En realidad todo este Escolio, pese a las perífrasis y cautelas verbales, destila el convencimiento de que la gravedad es una fuerza y causa de movimiento y no una mera condición matemática de la cinemática de los cuerpos celestes. En suma, que existe, actúa como causa y se mide por sus efectos. Pero no se sabe cuál es su naturaleza. <<

[13] En toda esta teoría de la Luna de nuevo aparecen ajustes en los cálculos y en las observaciones que van cambiando poco a poco hasta la redacción que aparece como texto en la tercera edición. Newton supo siempre que el cálculo de los movimientos e interacciones entre Luna, Tierra y Sol no era preciso. En definitiva, se trataba del problema de los tres cuerpos en una u otra forma. Por eso nunca dejó de mano este tema y, lo que el lector puede hallar a pie de página en la edición de Koyré-Cohen, son los restos y deshechos de ese trabajo constante con este problema. Quizá creyó que unas buenas observaciones podrían resolver el problema (Flamsteed fue un poco víctima de esa necesidad newtoniana), pero era más bien la premisa de su método matemático la que hubiera de haber revisado aquí. Aunque tampoco esto habría bastado. <<

[14] En esta edición el Lema I de la primera se subdividió en dos (Lemas I y II). El Lema II pasó entonces a Lema III, mientras el Lema III se convierte en Hipótesis II. <<

[15] Hay otro largo arreglo en esta Proposición que responde a los mismos motivos. Newton cambió su organización argumental de acuerdo con los nuevos datos mejorados. <<

[16] Se añade el Corolario 4 que no aparecía en la primera edición. <<

[17] Tanto el Corolario como el Escolio no aparecían en la primera edición. <<

[18] Antes de pasar a la segunda tabla en la primera edición añadía: «En estas observaciones Flamsteed puso tanto cuidado que después de observar dos veces la distancia del cometa a una estrella fija, y luego otra vez por duplicado la distancia a otra estrella fija, volvía a la estrella primera y observaba de nuevo la distancia hasta ella del cometa, y esto también dos veces; y después a partir del incremento o decremento proporcional al tiempo de su distancia, infería la distancia media en el tiempo intermedio, cuando observaba la distancia a la otra estrella. Flamsteed primero participó a los amigos los lugares del cometa calculados rápidamente a partir de estas observaciones, y después reexaminados de nuevo con cálculos más cuidados los corrigió. Nosotros hemos propuesto aquí los lugares corregidos».

Hay desde luego algunas variantes en las columnas 2.ª, 4.ª y 5.ª tanto para la segunda como para la tercera edición. Y también hay algunas mínimas variantes para la tabla segunda, fruto, tanto unas como otras de correcciones introducidas en los cálculos primitivos.

Pero es digno de notarse que en medio de todas estas correcciones hay una cuestión histórica de cierto interés para los historiadores. Flamsteed y Newton habían intercambiado cartas sobre los cometas y sus trayectorias desde antes de 1648-1685 y en esa correspondencia fue Flamsteed quien adelantó la idea de que se trataba de un solo cometa que se veía antes de y después de pasar por el Sol. Ahora Newton aprovecha la ocasión para apartar a Flamsteed del centro de la historia y, por ello, diluye cuanto puede o suprime las referencias al mismo. Así, por ejemplo, suprimió un párrafo que seguía a la exposición de sus propias observaciones que decía: «Además, como el ilustre Flamsteed en cierta ocasión me arguyó por carta, que el cometa aparecido en el mes de noviembre era el mismo cometa de los meses siguientes y trazase una trayectoria no muy diferente de esta órbita parabólica, me pareció oportuno calcular, los lugares del cometa en el mes de noviembre en esta órbita y compararlos con las observaciones. Las observaciones son como sigue».

Por lo demás recordemos que la unidad del cometa de 1680-1681 le fue sugerida por Flamsteed vía Crompton. De ello hay buena constancia en el «Waste Book». <<

[19] Tras este último ejercicio comprobatorio y las conclusiones a que llega, bueno será que el lector conozca el final que coronaba el texto antes del Escolio general y que decía (añadido en la segunda edición):

«Pues al decrecer el cuerpo del Sol, los movimientos medios de los planetas en torno al Sol poco a poco se retardan, y al crecer la Tierra el movimiento medio de la Luna poco a poco aumentará. Y comparadas, en efecto, las observaciones babilónicas de los eclipses con las de Al-Batani y con las actuales Halley concluyó, el primero que yo sepa, que comparado el movimiento medio de la Luna con el movimiento diurno de la Tierra, aquél se aceleraba un poco.» Todas las máquinas acaban desajustándose o buscando equilibrios dinámicos. La filosofía mecánica encontró aquí otro de sus temas clásicos. <<

[20] Existen algunas variantes para determinados párrafos que muestran la preocupación de Newton por expresar con propiedad, a la vez que prevenir críticas, su pensamiento acerca de la naturaleza y propiedades de ese Uno, Dios, dominador, etc. y de sus relaciones con el mundo concebido ahora como sistema o sistema de sistemas.

Antes de «Es eterno e infinito…» decía en uno de sus borradores: «Esta era la significación de las palabras *θεον* y Dei entre todos los griegos y latinos antiguos [y nosotros, cambiadas las significaciones de las voces, incorrectamente (corrompidamente) hablamos en sus lenguas (tachado por Newton)]. Pertenece, pues, el dominio al significado de la voz Dios. Pero el Dios Summo, del que aquí efectivamente se trata es Eterno e Infinito…».

Otro párrafo en lugar de «Todo hombre…» decía: «Uno y el mismo soy yo durante la vida en todos los órganos de los sentidos. Y el mismo es Dios siempre y en todas partes…».

Pero el resultado de esas dudas iniciales fue el texto principal que permanece casi intacto en la tercera edición, sólo con retoques de matiz.

En una de sus copias de la segunda edición añade a la última palabra del Escolio («espíritu») los calificativos de «eléctrico y elástico» que no incluyó en el texto de la tercera edición, aunque algunas traducciones dependientes de la de Motte los han incorporado al texto principal. Finalmente, en el penúltimo párrafo de este Escolio general aparece la célebre frase «et hipotheses non fingo» sobre la cual se han dado muchas interpretaciones. En el texto no ofrece, creemos, dificultad:

a) Habla en un presente la primera persona que tiene valor enfático: yo no hago ficciones, yo no me dedico a contar fábulas, etc. lo que lleva a pensar que otros sí lo hacen o puede que sí lo hagan.

b) Las «hipotheses» son cuentos, narraciones, discursos, propuestas, teorías, o lo que se quiera de carácter imaginario. Recordamos al lector que en otro pasaje (Regla 3 de este Libro III) dice: «Certe, contra experimentorum tenorem *somnia* temere *contingenda* non sunt», donde a las «hipotheses» se las caracteriza como «sueños» y la fabulación o fabricación de sueños-hipótesis lo denomina con la misma raíz verbal: «con-fingere». Por consiguiente, el contexto histórico de las «hypotheses» en uso o en desuso es el que deberá aclarar el sentido completo de tan rotunda afirmación. Conf. también Koyre: *Etudes Newtoniennes*, págs. 60 y siguientes passim.

Las notas siguientes: a, b, c, son marginales de Newton y pertenecen al texto. <<

[a] Esto es, Emperador universal. <<

[b] Nuestro *Pocock* deriva la voz «dei» de la voz árabe «du» (y en caso oblicuo «di») que significa señor. Y con este sentido los príncipes son llamados «dii», en el *Salmo* LXXXII, ver 6, y en *Juan*, X. 45. Y *Moisés* es llamado «dios» de su hermano *Aarón*, y «dios» del rey *Faraón* (Exod. IV, 16 y VII. 1). Y con el mismo sentido los gentiles llamaban «dii» a las almas de los príncipes muertos, pero falsamente por la falta de dominio. <<

[c] Así pensaban los antiguos, como *Pitágoras*, en *Cicer*, «de Natura deorum», lib. I. *Tales*, Anaxágoras, *Virgilio*, Georg., lib. IV, ver 220 y «Eneida», lib. VI, ver 721. *Filón*, al principio del libro I de las Alegorías, *Arato*, al comienzo de su «Fenómenos». Así también escritores sagrados como *Pablo*, en «Hechos», XVII, 27, 28, *Juan*, en Evang., XIV. 2. *Moisés*, en Deut., IV. 39, y en X. 14; *David*, «Salm». CXXXIX. 7, 8, 9. *Salomón*, en «I Reyes», VIII. 27; *Job*, XXII. 12, 13, 14; Jeremías, XXIII. 23, 24. Y los idólatras imaginaban que el Sol, la Luna y los astros, las almas de los hombres y otras partes del mundo eran parte del sumo dios y por lo tanto habían de ser veneradas, pero es falso. <<

El camino a la realidad

Una guía completa de las leyes del universo

ROGER PENROSE

Traducción de
Javier García Sanz

DEBATE

Título original: *The Road to Reality*
Publicado originariamente por Jonathan Cape, Londres, 2004

Primera edición: octubre de 2006

© 2004, Roger Penrose
© 2006, de la presente edición en castellano para todo el mundo:
Random House Mondadori, S. A.
Travessera de Gràcia, 47-49. 08021 Barcelona
© 2006, Javier García Sanz, por la traducción

Quedan prohibidos, dentro de los límites establecidos en la ley y bajo los apercibimientos legalmente previstos, la reproducción total o parcial de esta obra por cualquier medio o procedimiento, ya sea electrónico o mecánico, el tratamiento informático, el alquiler o cualquier otra forma de cesión de la obra sin la autorización previa y por escrito de los titulares del *copyright*.

Printed in Spain – Impreso en España

ISBN-13: 978-84-8306-6681-2
ISBN-10: 84-8306-681-5
Depósito legal: B. 35.801-2006

Compuesto en Fotocomposición 2000, S. A.
Impreso y encuadernado en Liberdúplex S. L. U.
Sant Llorenç d'Hortons (Barcelona)

C 846812

Dedico este libro a la memoria de
DENNIS SCIAMA,
que me mostró la emoción de la física

Índice

PREFACIO 21
AGRADECIMIENTOS 31
NOTACIÓN 35
PRÓLOGO 39

1. *Las raíces de la ciencia* 47
   1.1. La búsqueda de las fuerzas que configuran el mundo . . 47
   1.2. La verdad matemática 50
   1.3. ¿Es «real» el mundo matemático de Platón? 53
   1.4. Tres mundos y tres profundos misterios 61
   1.5. Lo bueno, lo verdadero y lo bello 66
2. *Un teorema antiguo y una pregunta moderna* 71
   2.1. El teorema de Pitágoras 71
   2.2. Los postulados de Euclides 75
   2.3. La demostración del teorema de Pitágoras por áreas semejantes 78
   2.4. Geometría hiperbólica: imagen conforme 81
   2.5. Otras representaciones de la geometría hiperbólica . . . 86
   2.6. Aspectos históricos de la geometría hiperbólica 92
   2.7. ¿Relación con el espacio físico? 97
3. *Tipos de números en el mundo físico* 105
   3.1. ¿Una catástrofe pitagórica? 105
   3.2. El sistema de los números reales 109
   3.3. Los números reales en el mundo físico 116
   3.4. ¿Necesitan los números naturales al mundo físico? . . . 120
   3.5. Números discretos en el mundo físico 123

4. *Los mágicos números complejos* 131
4.1. El mágico número «i» 131
4.2. Resolviendo ecuaciones con números complejos 135
4.3. Convergencia de series de potencias 138
4.4. El plano complejo de Caspar Wessel 143
4.5. Cómo se construye el conjunto de Mandelbrot 147
5. *Geometría de logaritmos, potencias y raíces* 151
5.1. La geometría del álgebra compleja 151
5.2. La idea del logaritmo complejo 156
5.3. Multivaluación, logaritmos naturales 159
5.4. Potencias complejas 164
5.5. Algunas relaciones con la física de partículas moderna . 168
6. *Cálculo infinitesimal con números reales* 173
6.1. ¿Qué hace respetable a una función? 173
6.2. Pendientes de funciones 176
6.3. Derivadas de orden superior; funciones C^∞-suaves . . . 178
6.4. ¿La noción «euleriana» de función? 184
6.5. Las reglas de diferenciación 186
6.6. Integración 189
7. *Cálculo infinitesimal con números complejos* 197
7.1. Suavidad compleja; funciones holomorfas 197
7.2. Integración de contorno 199
7.3. Series de potencias a partir de la suavidad compleja . . 203
7.4. Prolongación analítica 206
8. *Superficies de Riemann y aplicaciones complejas* 213
8.1. La idea de una superficie de Riemann 213
8.2. Aplicaciones conformes 218
8.3. La esfera de Riemann 222
8.4. El género de una superficie de Riemann compacta . . . 226
8.5. El teorema de la aplicación de Riemann 230
9. *Descomposición de Fourier e hiperfunciones* 235
9.1. Series de Fourier 235
9.2. Funciones sobre un círculo 240
9.3. Separación de frecuencias sobre la esfera de Riemann . 245
9.4. La transformada de Fourier 249
9.5. Separación de frecuencias a partir de la transformada de Fourier 251

9.6. ¿Qué tipo de función es apropiada? 254
9.7. Hiperfunciones 258
10. *Superficies* 269
10.1. Dimensiones complejas y dimensiones reales 269
10.2. Suavidad, derivadas parciales 271
10.3. Campos vectoriales y 1-formas 277
10.4. Componentes, productos escalares 284
10.5. Las ecuaciones de Cauchy-Riemann 286
11. *Números hipercomplejos* 293
11.1. El álgebra de los cuaterniones 293
11.2. ¿Hay un papel físico para los cuaterniones? 296
11.3. Geometría de cuaterniones 299
11.4. ¿Cómo componer rotaciones? 304
11.5. Álgebras de Clifford 306
11.6. Álgebras de Grassmann 310
12. *Variedades de* n *dimensiones* 317
12.1. ¿Por qué estudiar variedades de dimensiones más altas? 317
12.2. Variedades y cartas de coordenadas 322
12.3. Escalares, vectores y covectores 324
12.4. Productos de Grassmann 329
12.5. Integrales de formas 332
12.6. Derivada exterior 335
12.7. El elemento de volumen; convenio de suma 341
12.8. Tensores; notación de índices abstractos y notación diagramática 344
12.9. Variedades complejas 348
13. *Grupos de simetría* 355
13.1. Grupos de transformaciones 355
13.2. Subgrupos y grupos simples 359
13.3. Transformaciones lineales y matrices 364
13.4. Determinantes y trazas 372
13.5. Autovalores y autovectores 374
13.6. Teoría de la representación y álgebras de Lie 378
13.7. Espacios de representación tensoriales; reducibilidad 383
13.8. Grupos ortogonales 389

13.9. Grupos unitarios 396
13.10. Grupos simplécticos 402
14. *Cálculo infinitesimal en variedades* 411
14.1. ¿Diferenciación en una variedad? 411
14.2. Transporte paralelo 414
14.3. Derivada covariante 418
14.4. Curvatura y torsión 422
14.5. Geodésicas, paralelogramos y curvatura 425
14.6. Derivada de Lie 432
14.7. Lo que una métrica puede hacer por usted 441
14.8. Variedades simplécticas 446
15. *Fibrados y conexiones gauge* 451
15.1. Algunas motivaciones físicas para los fibrados 451
15.2. La idea matemática de un fibrado 454
15.3. Secciones transversales de fibrados 459
15.4. El fibrado de Clifford 462
15.5. Fibrados vectoriales complejos, fibrados (co)tangentes 467
15.6. Espacios proyectivos 471
15.7. No trivialidad en una conexión fibrada 476
15.8. Curvatura fibrada 481
16. *La escalera del infinito* 491
16.1. Campos finitos 491
16.2. ¿Una geometría finita o una geometría infinita para la física? 494
16.3. Diferentes tamaños de infinito 500
16.4. El corte diagonal de Cantor 504
16.5. Enigmas en los fundamentos de las matemáticas . . 509
16.6. Las máquinas de Turing y el teorema de Gödel . . . 513
16.7. Tamaños de infinitos en física 518
17. *Espaciotiempo* 525
17.1. El espaciotiempo de la física aristotélica 525
17.2. El espaciotiempo para la relatividad galileana 528
17.3. La dinámica newtoniana en términos del espaciotiempo 531
17.4 El principio de equivalencia 535
17.5. El «espaciotiempo newtoniano» de Cartan 539

17.6. La velocidad finita y fija de la luz 546
17.7. Conos de luz 548
17.8. El abandono del tiempo absoluto 552
17.9. El espaciotiempo de la relatividad general de Einstein 557
18. *Geometría minkowskiana* 563
18.1. Los 4-espacios euclídeo y minkowskiano 563
18.2. Los grupos de simetría del espacio de Minkowski . 567
18.3. Ortogonalidad lorentziana; la «paradoja del reloj» . 570
18.4. Geometría hiperbólica en el espacio de Minkowski 576
18.5. La esfera celeste como una esfera de Riemann . . . 582
18.6. Energía y momento (angular) newtonianos 587
18.7. Energía y momento (angular) relativistas 590
19. *Los campos clásicos de Maxwell y Einstein* 599
19.1. Evolución fuera de la dinámica newtoniana 599
19.2. La teoría electromagnética de Maxwell 602
19.3. Leyes de conservación y de flujo en la teoría de Maxwell 607
19.4. El campo de Maxwell como curvatura gauge 610
19.5. El tensor energía-momento 617
19.6. La ecuación de campo de Einstein 622
19.7. Cuestiones adicionales: la constante cosmológica, el tensor de Weyl 627
19.8. La energía del campo gravitatorio 629
20. *Lagrangianos y hamiltonianos* 639
20.1. El mágico formalismo lagrangiano 639
20.2. La más simétrica imagen hamiltoniana 644
20.3. Pequeñas oscilaciones 648
20.4. La dinámica hamiltoniana como geometría simpléctica 654
20.5. Tratamiento lagrangiano de los campos 658
20.6. Cómo impulsan los lagrangianos la teoría moderna 661
21. *La partícula cuántica* 667
21.1. Variables no conmutativas 667
21.2. Hamiltonianos cuánticos 671
21.3. La ecuación de Schrödinger 674
21.4. La base experimental de la teoría cuántica 676

21.5. Comprendiendo la dualidad onda-partícula 682
21.6. ¿Qué es la «realidad» cuántica? 685
21.7. La naturaleza «holística» de una función de onda .. 691
21.8. Los misteriosos «saltos cuánticos» 696
21.9. Distribución de probabilidad en una función de onda 698
21.10. Estados de posición 701
21.11. Descripción en el espacio de momentos 703
22. *Álgebra, geometría y espín cuánticos* 711
22.1. Los procedimientos cuánticos **U** y **R** 711
22.2. La linealidad de **U** y sus problemas para **R** 715
22.3. Estructura unitaria, espacio de Hilbert, notación de Dirac 718
22.4. Evolución unitaria: Schrödinger y Heisenberg ... 722
22.5. «Observables» cuánticos 726
22.6. Medidas sí/NO, proyectores 730
22.7. Medidas nulas, helicidad 733
22.8. Espín y espinores 739
22.9. La esfera de Riemann de los sistemas de dos estados 745
22.10. Espín más alto: la imagen de Majorana 752
22.11. Armónicos esféricos 755
22.12. Momento angular cuántico relativista 761
22.13. El objeto cuántico aislado general 765
23. *El entrelazado mundo cuántico* 777
23.1. Mecánica cuántica de sistemas de muchas partículas 777
23.2. La enormidad del espacio de estados de muchas partículas 780
23.3. Entrelazamiento cuántico; desigualdades de Bell .. 783
23.4. Experimentos EPR tipo Bohm 787
23.5. El ejemplo EPR de Hardy: casi libre de probabilidad 792
23.6. Dos misterios del entrelazamiento cuántico 795
23.7. Bosones y fermiones 798
23.8. Los estados cuánticos de los bosones y los fermiones 801
23.9. Teleportación cuántica 804
23.10. Cuanlazamiento 810
24. *El electrón y las antipartículas de Dirac* 819
24.1. Tensión entre la teoría cuántica y la relatividad ... 819

24.2. ¿Por qué las antipartículas implican campos cuánticos? 821
24.3. Positividad de la energía en mecánica cuántica ... 823
24.4. Dificultades con la fórmula de la energía relativista 826
24.5. La no invariancia de $\partial/\partial t$ 828
24.6. La raíz cuadrada de Clifford-Dirac de un operador de ondas 831
24.7. La ecuación de Dirac 833
24.8. La ruta de Dirac al positrón 836
25. *El modelo estándar de la física de partículas* 843
25.1. Los orígenes de la moderna física de partículas ... 843
25.2. La imagen zig-zag del electrón 845
25.3. Interacciones electrodébiles, asimetría de reflexión 850
25.4. Conjugación de carga, paridad e inversión temporal 857
25.5. El grupo de simetría electrodébil 860
25.6. Partículas fuertemente interactuantes 866
25.7. «Quarks coloreados» 870
25.8. ¿Más allá del modelo estándar? 874
26. *Teoría cuántica de campos* 881
26.1. El estatus fundamental de la QFT en la teoría moderna 881
26.2. Operadores de creación y aniquilación 883
26.3. Álgebras de dimensión infinita 887
26.4. Antipartículas en QFT 890
26.5. Vacíos alternativos 892
26.6. Interacciones: lagrangianos e integrales de camino . 894
26.7. Integrales de camino divergentes: la respuesta de Feynman 900
26.8. Construyendo diagramas de Feynman; la matriz S .. 903
26.9. Renormalización 907
26.10. Diagramas de Feynman a partir de lagrangianos .. 913
26.11. Los diagramas de Feynman y la elección del vacío . 915
27. *El big bang y su legado termodinámico* 923
27.1. Simetría temporal en la evolución dinámica 923
27.2. Ingredientes submicroscópicos 925
27.3. Entropía 928
27.4. El carácter robusto del concepto de entropía 931

27.5. Derivación de la segunda ley... ¿o no? 936
27.6. ¿Es el universo en su conjunto un «sistema aislado»? 940
27.7. El papel del big bang 944
27.8. Agujeros negros 951
27.9. Horizontes de sucesos y singularidades espaciotemporales 957
27.10. Entropía de agujero negro 960
27.11. Cosmología 963
27.12. Diagramas conformes 971
27.13. Nuestro extraordinariamente especial big bang ... 975
28. *Teorías especulativas del universo primitivo* 987
28.1. Ruptura espontánea de simetría en el universo primitivo 987
28.2. Defectos topológicos cósmicos 992
28.3. Problemas para la ruptura de simetría en el universo primitivo 997
28.4. Cosmología inflacionaria 1002
28.5. ¿Son válidas las motivaciones para la inflación? ... 1010
28.6. El principio antrópico 1016
28.7. La naturaleza especial del big bang: ¿una clave antrópica? 1022
28.8. La hipótesis de curvatura de Weyl 1026
28.9. La propuesta de «ausencia de frontera» de Hartle-Hawking 1030
28.10. Parámetros cosmológicos: ¿estatus observacional? . 1034
29. *La paradoja de la medida* 1049
29.1. Las ontologías convencionales de la mecánica cuántica 1049
29.2. Ontologías no convencionales para la mecánica cuántica 1054
29.3. La matriz densidad 1061
29.4. Matrices densidad para espín 1/2: la esfera de Bloch 1064
29.5. La matriz densidad en situaciones EPR 1069
29.6. Filosofía FAPP de la decoherencia por el entorno . 1075
29.7. El gato de Schrödinger con la ontología «de Copenhague» 1078
29.8. ¿Pueden las ontologías (b) y (c) resolver el «gato»? . 1081

29.9. ¿Qué ontologías no convencionales pueden ayudar? 1085
30. *El papel de la gravedad en la reducción del estado cuántico* 1093
30.1. ¿Va a quedarse aquí la teoría cuántica actual? 1093
30.2. Claves de una asimetría temporal cosmológica ... 1095
30.3. Asimetría temporal en la reducción del estado cuántico 1098
30.4. Temperatura del agujero negro de Hawking 1103
30.5. Temperatura del agujero negro a partir de periodicidad compleja 1108
30.6. Vectores de Killing, flujo de energía... ¡y viaje en el tiempo! 1115
30.7. Flujo de energía saliente de órbitas de energía negativa 1119
30.8. Explosiones de Hawking 1122
30.9. Una perspectiva más radical 1127
30.10. El bulto de Schrödinger 1132
30.11. Conflicto fundamental con los principios de Einstein 1136
30.12. ¿Estados de Schrödinger-Newton preferidos? 1141
30.13. La propuesta FELIX y otras relacionadas 1144
30.14. Origen de las fluctuaciones en el universo primitivo 1151
31. *Supersimetría, supradimensionalidad y cuerdas* 1163
31.1. Parámetros inexplicados 1163
31.2. Supersimetría 1168
31.3. El álgebra y la geometría de la supersimetría 1173
31.4. Espaciotiempo de dimensiones más altas 1177
31.5. La teoría de cuerdas hadrónica original 1182
31.6. Hacia una teoría de cuerdas del universo 1187
31.7. Motivación de cuerdas para dimensiones espaciotemporales extra 1191
31.8. ¿La teoría de cuerdas como gravedad cuántica? ... 1193
31.9. Dinámica de cuerdas 1197
31.10. ¿Por qué no vemos las dimensiones espaciales extra? 1201
31.11. ¿Deberíamos aceptar el argumento de la estabilidad cuántica? 1207
31.12. Inestabilidad clásica de las dimensiones extra 1211

31.13. ¿Es finita la QFT de cuerdas? 1215
31.14. Los mágicos espacios de Calabi-Yau; la teoría M . . 1218
31.15. Cuerdas y entropía de agujero negro 1225
31.16. El «principio holográfico» 1231
31.17. La perspectiva de la D-brana 1234
31.18. ¿El estatus físico de la teoría de cuerdas? 1238
32. *El sendero más estrecho de Einstein; variables de lazo* 1251
32.1. Gravedad cuántica canónica 1251
32.2. El ingrediente quiral de las variables de Ashtekar . . 1253
32.3. La forma de las variables de Ashtekar 1257
32.4. Variables de lazo 1260
32.5. Las matemáticas de nudos y enlaces 1264
32.6. Redes de espín 1267
32.7. ¿El estatus de la gravedad cuántica de lazo? 1275
33. *Perspectivas más radicales: la teoría de twistores* 1283
33.1. Teorías donde la geometría tiene elementos discretos 1283
33.2. Los twistores como rayos de luz 1288
33.3. El grupo conforme; el espacio de Minkowski compactificado 1296
33.4. Los twistores como espinores de dimensión superior 1301
33.5. Geometría twistorial básica y coordenadas 1304
33.6. Geometría de twistores como partículas sin masa con espín 1309
33.7. Teoría cuántica twistorial 1314
33.8. Descripción twistorial de campos sin masa 1318
33.9. Cohomología de haces twistorial 1321
33.10. Los twistores y la separación en frecuencia positiva/negativa 1327
33.11. El gravitón no lineal 1330
33.12. Twistores y relatividad general 1336
33.13. Hacia una teoría twistorial de la física de partículas 1339
33.14. ¿El futuro de la teoría de twistores? 1340
34. *¿Dónde está el camino a la realidad?* 1351
34.1. Las grandes teorías de la física del siglo XX... ¿y más allá? 1351
34.2. Física fundamental matemáticamente dirigida 1356

34.3. El papel de las modas en la teoría física 1361
34.4. ¿Puede refutarse experimentalmente una teoría errónea? 1365
34.5. ¿Dónde podemos esperar nuestra próxima revolución en física? 1371
34.6. ¿Qué es la realidad? 1375
34.7. Los papeles de la mentalidad en la teoría física 1378
34.8. Nuestro largo camino matemático a la realidad . . . 1383
34.9. Belleza y milagros 1389
34.10. Preguntas profundas respondidas, preguntas más profundas planteadas 1395

EPÍLOGO 1403
BIBLIOGRAFÍA 1405
ÍNDICE ALFABÉTICO 1457

Prefacio

Este libro tiene como objetivo transmitir al lector una idea de lo que es ciertamente uno de los viajes de descubrimiento más importantes y apasionantes en los que se ha embarcado la humanidad. Se trata de la búsqueda de los principios subyacentes que rigen el comportamiento de nuestro universo. Es un viaje que ha durado más de dos mil quinientos años, de modo que no debería sorprendernos que al final se hayan hecho progresos sustanciales. Pero este viaje se ha mostrado muy difícil, y en la mayoría de los casos, el conocimiento real ha llegado lentamente. Esta dificultad intrínseca nos ha llevado en muchas direcciones falsas, y de ello deberíamos aprender a ser cautos. Pero el siglo XX nos ha revelado nuevas y extraordinarias ideas, algunas tan impresionantes que muchos científicos actuales han expresado la opinión de que podríamos estar cerca de una comprensión básica de *todos* los principios subyacentes en la física. En mis descripciones de las teorías fundamentales vigentes cuando escribo esto, recién concluido el siglo XX, trataré de adoptar un punto de vista más modesto. Quizá no todas mis opiniones sean bien recibidas por los «optimistas», pero espero cambios futuros de dirección aún mayores que los que se han dado en el último siglo.

El lector encontrará que en este libro no he rehuido presentar fórmulas matemáticas, pese a las reiteradas advertencias acerca de la drástica reducción de lectores que esto implicaría. He pensado seriamente sobre esto y he llegado a la conclusión de que lo que tengo que decir no puede transmitirse razonablemente sin cierta cantidad de notación matemática y la exploración de genuinos conceptos matemáticos. El conocimiento que tenemos de los principios que realmente subyacen

en el comportamiento de nuestro mundo físico depende, de hecho, de una apreciación de sus matemáticas. Quizá para algunas personas esto pueda ser un motivo de desesperación, pues se habrán formado la idea de que no tienen ninguna capacidad para las matemáticas, por muy elementales que sean. Sin duda se preguntarán cómo van a poder comprender la investigación que se está haciendo en la misma frontera de la teoría física si ni siquiera dominan la forma de tratar las *fracciones*. Veo la dificultad, por supuesto.

Pese a todo, soy optimista en cuestiones de transmisión del conocimiento. Tal vez sea un optimista incurable. Me pregunto si esos lectores potenciales que no pueden manipular fracciones —o que dicen que no pueden manipular fracciones— no se están engañando a sí mismos, al menos un poco, y una buena proporción de ellos tienen realmente una capacidad en esta dirección de la que no son conscientes. Sin duda hay algunos que, cuando se enfrentan a una línea de símbolos matemáticos, independientemente de la sencillez con que estén presentados, solo pueden ver el rostro severo de un padre o un profesor que trataba de inculcarles a la fuerza una aparente competencia sin contenido y similar a la de un papagayo —una obligación, y solo una obligación— sin que pudiera traslucir ningún indicio de la magia o belleza del tema. Tal vez para algunos sea demasiado tarde; pero, como digo, soy un optimista, y creo que todavía quedan muchos, incluso entre aquellos que nunca pudieron dominar la manipulación de fracciones, que tienen la capacidad de vislumbrar algo de un mundo maravilloso que pienso que debe ser, en un grado significativo, genuinamente accesible para ellos.

Una de las mejores amigas de juventud de mi madre era una de esas personas que no podían comprender las fracciones. Ella misma me lo contó en cierta ocasión, una vez que se había retirado de una carrera exitosa como bailarina. Yo aún era joven, y todavía no me había lanzado plenamente a mi actividad como matemático, pero se me reconocía como alguien que disfrutaba trabajando en este tema. «Es todo eso de la simplificación —me decía—, nunca le cogí el tranquillo a la simplificación.» Era una mujer elegante y muy inteligente, y en mi opinión no hay ninguna duda de que las cualidades mentales que se necesitan para comprender la sofisticada coreografía que es fundamen-

tal para el ballet no son en modo alguno inferiores a las que deben ejercitarse con un problema matemático. Así, sobrestimando ampliamente mi capacidad expositiva, intenté, como otros lo habían hecho antes, explicarle la simplicidad y la naturaleza lógica del procedimiento de «simplificación».

Creo que mis esfuerzos fueron tan infructuosos como lo habían sido los de los otros que lo habían intentado. (Dicho sea de paso, su padre había sido un eminente geólogo y miembro de la Royal Society, de modo que ella debía de haber tenido una formación adecuada para la comprensión de cuestiones científicas. Quizá aquí podría haber intervenido un factor del tipo «rostro severo». No lo sé.) Pero reflexionando sobre ello, me pregunto ahora si ella, y muchas personas como ella, no tenían un complejo inhibidor más racional, uno que yo ni siquiera había advertido con toda mi verborrea matemática. Hay, en realidad, una cuestión profunda con la que uno tropieza una y otra vez en matemáticas y en física matemática, y que encuentra por primera vez en la noción aparentemente inocente de cancelación de un factor común en el numerador y el denominador de una fracción numérica ordinaria.

Aquellos para los que la acción de simplificar se ha convertido en algo automático, debido a la familiaridad repetida con tales operaciones, pueden ser insensibles a una dificultad que en realidad acecha tras este procedimiento aparentemente sencillo. Quizá muchos de los que encuentran misteriosa la simplificación están viendo cierto punto con más profundidad que aquellos de nosotros que pasamos de largo de forma displicente, pareciendo ignorarlo. ¿De qué punto se trata? Concierne a la forma misma en que los matemáticos pueden ofrecer una existencia a sus entidades matemáticas y a cómo pueden relacionarse tales entidades con la realidad física.

Recuerdo que cuando estaba en la escuela, más o menos a los once años, me quedé muy sorprendido cuando el maestro preguntó a la clase qué es realmente una fracción (tal como 3/8, pongamos por caso). Hubo varias sugerencias, como dividir un pastel en porciones y cosas así, pero fueron rechazadas por el profesor sobre la base (válida) de que estas se referían simplemente a situaciones físicas imprecisas a las que tenía que *aplicarse* la noción matemática precisa de una fracción; pero no

nos decían cuál *es* dicha noción matemática precisa. Siguieron otras sugerencias, tales como 3/8 es «algo con un 3 arriba y un 8 abajo y con una línea horizontal en medio», ¡y me quedé sorprendido al descubrir que el profesor parecía tomar estas sugerencias en serio! No recuerdo muy bien cómo se resolvió finalmente la cuestión, pero con la intuición adquirida de muchas experiencias posteriores como estudiante de licenciatura en matemáticas, podría conjeturar que mi maestro de escuela estaba haciendo un valiente intento por decirnos la definición de una fracción en términos de la ubicua noción matemática de *clase de equivalencia*.

¿Cuál es esta noción? ¿Cómo puede aplicarse en el caso de una fracción, y decirnos qué es realmente una fracción? Empecemos con el «algo con un 3 arriba y un 8 abajo» de mi compañero de clase. Básicamente, esto nos sugiere que una fracción está especificada por un par ordenado de números enteros, en este caso los números 3 y 8. Pero es evidente que no podemos considerar que la fracción *es* dicho par ordenado porque, por ejemplo, la fracción 6/16 es el mismo número que la fracción 3/8, mientras que el par (6,16) no es ciertamente el mismo que el par (3, 8). Esta es precisamente una cuestión de simplificación; en efecto, podemos escribir 6/16 como $\frac{3 \times 2}{8 \times 2}$ y luego cancelar el 2 de arriba con el 2 de abajo, para obtener 3/8. ¿Por qué se nos permite hacer esto y con ello «igualar», en cierto sentido, el par (6, 16) con el par (3, 8)? La respuesta matemática —que muy bien puede sonar como un «escaqueo»— tiene la regla de simplificación incorporada en la definición de una fracción: se considera que un par de números enteros $(a \times n, b \times n)$ representa la misma fracción que el par (a, b) siempre que n sea un número entero distinto de cero (y donde tampoco admitimos que b sea cero).

Pero esto tampoco nos dice qué es una fracción; simplemente nos dice algo sobre la forma en que representamos las fracciones. ¿Qué *es*, entonces, una fracción? Según la noción de «clase de equivalencia» de los matemáticos, la fracción 3/8, por ejemplo, es simplemente la colección infinita de todos los pares

$$(3, 8), (-3, -8), (6, 16), (-6, -16), (9, 24), (-9, -24), (12, 32), \ldots,$$

donde cada par puede obtenerse de cada uno de los otros pares de la lista por aplicación repetida de la regla de simplificación anterior.* También necesitamos definiciones que nos digan cómo sumar, restar y multiplicar estas colecciones infinitas de pares de números enteros, donde son válidas las reglas normales del álgebra, y cómo identificar los propios números enteros como tipos particulares de fracción.

Esta definición cubre todo lo que necesitamos matemáticamente de las fracciones (tal como que 1/2 es un número que sumado a sí mismo da el número 1, etc.), y la operación de simplificación está, como hemos visto, incorporada en la definición. Pero todo parece muy formal y podemos preguntarnos si realmente recoge la noción intuitiva de lo que es una fracción. Aunque este ubicuo procedimiento de las clases de equivalencia, de la que el ejemplo anterior es un caso particular, es muy potente como pura herramienta matemática para establecer la consistencia y la existencia matemática, puede proporcionarnos entidades con una apariencia demasiado pesada. ¡Difícilmente nos transmite la noción intuitiva de lo que es 3/8, por ejemplo! No es extraño que la amiga de mi madre estuviera confundida.

En mis descripciones de nociones matemáticas trataré de evitar, hasta donde pueda, el tipo de pedantería matemática que nos lleva a definir una fracción en términos de una «clase infinita de pares», incluso si, por supuesto, tiene su valor en rigor y precisión matemáticos. En mis descripciones me interesaré más en transmitir la idea —y la belleza y la magia— inherente a muchas nociones matemáticas importantes. La idea de una fracción tal como 3/8 consiste simplemente en que es cierto tipo de entidad que tiene la propiedad de que cuando se suma a sí misma 8 veces da 3. La magia está en que la idea de una fracción funciona pese al hecho de que en el mundo físico no experimentamos directamente cosas que estén cuantificadas exactamente por fracciones: las porciones de pastel solo conducen a aproximaciones. (Esto es muy diferente del caso de los números naturales, tales como 1, 2, 3, que cuantifican de forma precisa muchas entidades de nuestra experiencia

* Esto se denomina una «clase de equivalencia» porque realmente es una clase de entidades (que, en este caso concreto, son pares de números enteros), cada uno de cuyos miembros se considera equivalente, en un sentido específico, a cada uno de los demás miembros.

directa.) Una forma de ver que las fracciones tienen un sentido consistente es utilizar la «definición» en términos de colecciones infinitas de pares de enteros, como se ha indicado antes. Pero esto no significa que 3/8 sea realmente una colección semejante. Es mejor pensar en 3/8 como una entidad con un tipo de existencia (platónica) propia, y que la colección infinita de pares es simplemente una forma de llegar a entender la consistencia de este tipo de entidad. A medida que nos familiarizamos, empezamos a creer que podemos captar con facilidad una noción tal como 3/8 como algo que tiene su propio tipo de existencia, y la idea de una «colección infinita de pares» es simplemente un artificio pedante, un artificio que enseguida se retira de nuestra imaginación una vez que lo hemos captado. Buena parte de las matemáticas es así.

Para los matemáticos (al menos para la mayoría de ellos, por lo que puedo entender), las matemáticas no son solo una actividad cultural que hemos creado nosotros mismos, sino que tienen vida propia, y buena parte de ellas está en sorprendente armonía con el universo físico. No podemos tener una comprensión profunda de las leyes que rigen el mundo físico sin entrar en el mundo de las matemáticas. En particular, la noción anterior de una clase de equivalencia es relevante no solo para una gran cantidad de matemáticas importantes (aunque confusas), sino también para una gran cantidad de física importante (y confusa), tal como la teoría de la relatividad general de Einstein y los principios de las «teorías gauge» que describen las fuerzas de la naturaleza según la moderna física de partículas. En la física moderna, uno no puede evitar el enfrentarse a las sutilezas de muchas matemáticas sofisticadas. Por esta razón, he dedicado los dieciséis primeros capítulos de esta exposición a la descripción de ideas matemáticas.

¿Qué consejo puedo dar al lector para hacer frente a esto? Este libro tiene cuatro niveles diferentes de lectura. Quizá sea usted un lector, en un extremo de la escala, que sencillamente se da la vuelta cuando se le presenta una fórmula matemática (y algunos de estos lectores pueden muy bien tener dificultades en entender las fracciones). Si es así, creo que todavía puede usted aprender mucho de este libro saltándose simplemente todas las fórmulas y leyendo solo las palabras. Supongo que esto será muy parecido a la forma en que yo solía hojear las revistas de ajedrez que estaban desperdigadas por mi casa cuando era

niño. El ajedrez constituía buena parte de las vidas de mis padres y hermanos, pero yo me interesé muy poco por él, excepto en que disfrutaba leyendo las hazañas de aquellos excepcionales, y con frecuencia extraños, personajes que se dedicaban a ese juego. De esa lectura, me hacía una idea de la brillantez de las jugadas que solían hacer, incluso si no las comprendía ni hacía ningún intento por seguir las diferentes posiciones a través de las notaciones. Pese a todo, descubrí que era una actividad agradable e ilustrativa que podía mantener mi atención. Del mismo modo, espero que las exposiciones matemáticas que ofrezco aquí puedan transmitir algo de su interés incluso a algunos lectores poco interesados en las matemáticas que, por valentía o curiosidad, deciden acompañarme en mi viaje de exploración de las ideas matemáticas y físicas que parecen subyacer en el universo físico. No tenga miedo en saltarse ecuaciones (yo mismo lo hago con frecuencia) y, si lo desea, saltarse capítulos enteros o partes de capítulos cuando empiecen a hacerse demasiado rimbombantes. Hay mucha variedad en las dificultades y tecnicismos del material, y quizá haya algo más de su gusto en otro lugar. Puede decidir simplemente introducirse y mirar. Mi esperanza es que las extensas referencias cruzadas puedan ilustrar bien nociones poco familiares, de modo que debería ser posible localizar los conceptos y la notación necesarios regresando a secciones anteriores no leídas en busca de aclaración.

En un segundo nivel, quizá sea usted un lector que está dispuesto a echar una ojeada a las fórmulas matemáticas, cuando quiera que se le presenten, pero no siente la inclinación (ni quizá tiene tiempo) de verificar por usted mismo las afirmaciones que hago. Las confirmaciones de muchas de estas afirmaciones constituyen las soluciones a los ejercicios que he desperdigado por las partes matemáticas de este libro. He señalado tres niveles de dificultad mediante los iconos

quad muy sencillo

quad necesita un poco de reflexión

quad no debe abordarse con ligereza

Es perfectamente razonable fiarse de estas, si lo desea, y no hay pérdida de continuidad si decide adoptar esta postura.

Si, por el contrario, usted es un lector que quiere ejercitarse con estas diversas (e importantes) nociones matemáticas, pero para quien las ideas que describo no son familiares, espero que el trabajo con estos ejercicios le ofrecerá una ayuda importante para reforzar tales habilidades. Con las matemáticas, siempre sucede que una pequeña experiencia directa de reflexión propia sobre las cosas puede proporcionar una comprensión mucho más profunda que la mera lectura sobre ellas. (Si necesita las soluciones, consulte la página web www.roadsolutions.ox.ac.uk.)

Por último, quizá usted sea ya un experto, en cuyo caso no debería tener dificultades con las matemáticas (la mayoría de las cuales le serán muy familiares) y tal vez no quiera perder el tiempo con los ejercicios. Pero aún puede descubrir que hay algo que extraer de mis propios enfoques sobre diversos temas, que es probable que sean algo diferentes (a veces muy diferentes) de los habituales. Quizá usted tenga cierta curiosidad por mis opiniones concernientes a varias teorías modernas (por ejemplo, supersimetría, cosmología inflacionaria, la naturaleza del big bang, agujeros negros, teoría de cuerdas o teoría M, variables de lazo en gravedad cuántica, teoría de twistores, incluso los propios fundamentos de la teoría cuántica). Sin duda encontrará mucho en lo que discrepar conmigo en bastantes de estos temas. Pero la controversia es una parte importante del desarrollo de la ciencia, y por eso no tengo reparos en presentar opiniones que puedan considerarse parcialmente reñidas con algunas de las actividades de la corriente principal de la física teórica moderna.

Puede decirse que este libro trata realmente de la relación entre las matemáticas y la física, y de cómo el diálogo entre ambas influye poderosamente en los impulsos que subyacen en nuestra búsqueda de una mejor teoría del universo. Un ingrediente esencial de estos impulsos en muchos desarrollos modernos procede de juicios sobre la belleza, profundidad y sofisticación matemáticas. Es evidente que tales influencias matemáticas pueden ser de importancia vital, como sucede con algunos de los éxitos más impresionantes de la física del siglo XX: la ecuación de Dirac para el electrón, el armazón general de la mecánica cuántica y la relatividad general de Einstein. Pero en todos estos casos,

las consideraciones físicas —en última instancia las observacionales— han proporcionado el criterio primordial de aceptación. En muchas de las ideas modernas para avanzar de manera fundamental en nuestra comprensión de las leyes del universo, no se dispone de criterios físicos adecuados —i.e., datos experimentales, o siquiera la posibilidad de investigación experimental—. Por ello, podemos cuestionarnos si los desiderata matemáticos accesibles son suficientes para permitirnos estimar las probabilidades de éxito de estas ideas. El tema es delicado, y trataré de plantear cuestiones que, en mi opinión, no han sido suficientemente discutidas en otros lugares.

Aunque en algunos pasajes presentaré opiniones que pueden considerarse controvertidas, me he preocupado por aclarar al lector cuándo me estoy tomando realmente tales libertades. En consecuencia, este libro puede utilizarse como una guía genuina para las ideas (y las preguntas) fundamentales de la física moderna. Puede utilizarse en las aulas como una honesta introducción a la física moderna, tal como se entiende ahora, cuando nos movemos en los primeros años del tercer milenio.

Agradecimientos

Al tratarse de un libro de esta magnitud, cuya redacción me ha llevado aproximadamente ocho años, es inevitable que haya muchas personas a quienes debo dar las gracias. Resulta casi igual de inevitable que, entre ellas, haya algunas cuya valiosa contribución no quede reconocida debido a mi desorganización y despiste congénitos. Permítanme, en primer lugar, expresar mi especial agradecimiento, y también mis disculpas, a esas personas que me han dado su generosa ayuda pero cuyos nombres no me vienen a la memoria. Pasando a las diversas informaciones y ayudas concretas que puedo recordar con más claridad, doy las gracias a Michael Atiyah, John Baez, Michael Berry, Dorje Brody, Robert Bryant, Hong-Mo Chan, Joy Christian, Andrew Duggins, Maciej Dunajski, Freeman Dyson, Arthur Ekert, David Fowler, Margaret Gleason, Jeremy Gray, Stuart Hameroff, Keith Hannabuss, Lucien Hardy, Jim Hartle, Tom Hawkins, Nigel Hitchin, Andrew Hodges, Dipankar Home, Jim Howie, Chris Isham, Ted Jacobson, Bernard Kay, William Marshall, Lionel Mason, Charles Misner, Tristan Needham, Stelios Negrepontis, Sarah Jones Nelson, Ezra (Ted) Newman, Charles Oakley, Daniel Oi, Robert Osserman, Don Page, Oliver Penrose, Alan Rendall, Wolfgang Rindler, Engelbert Schücking, Bernard Schutz, Joseph Silk, Christoph Simon, George Sparling, John Stachel, Henry Stapp, Richard Thomas, Gerard 't Hooft, Paul Tod, James Vickers, Robert Wald, Rainer Weiss, Ronny Wells, Gerald Westheimer, John Wheeler, Nick Woodhouse y Anton Zeilinger. Agradecimientos particulares merecen Lee Smolin, Kelly Stelle y Lane Hughston por su ayuda en numerosos y variados puntos. Estoy en deuda especial con Florence Tsou (Sheung Tsun), por su inmensa ayuda en temas de física de partículas; con Fay Dowker,

por su ayuda y evaluación sobre varios temas, muy especialmente la presentación de ciertas cuestiones mecanocuánticas; con Subir Sarkar, por la valiosa información concerniente a datos cosmológicos y su interpretación; con Vahe Gurzadyan, por lo mismo y por alguna información anticipada de sus hallazgos cosmológicos concernientes a la geometría global del universo, y particularmente con Abhay Ashtekar, por su completa información sobre la teoría de variables de lazo y también sobre varios puntos detallados concernientes a la teoría de cuerdas.

Agradezco a la National Science Foundation su apoyo mediante las becas PHY 93-96246 y 00-90091, y a la Leverhulme Foundation por una beca de dos años Leverhulme Emeritus Fellowship, durante 2000-2002. Los nombramientos a tiempo parcial en el Gresham College, Londres (1998-2001), y el Center for Gravitational Physics and Geometry en la Universidad del Estado de Pennsylvania, Estados Unidos, me han sido muy valiosos para escribir este libro, como lo ha sido la ayuda administrativa (muy especialmente, la de Ruth Preston) y de espacio en el Mathematical Institute de la Universidad de Oxford. También ha sido incalculable el especial apoyo por la parte editorial, compaginando duras restricciones de horario con un autor de hábitos de trabajo erráticos. La primera ayuda editorial de Eddie Mizzi fue vital para iniciar el proceso de convertir mis caóticos escritos en un libro real, y Richard Lawrence, con su experiencia y su paciente y sensible persistencia, ha sido un factor crucial para llevar a término este proyecto. Teniendo que trabajar con esta complicada reescritura, John Holmes ha hecho un trabajo excelente al proporcionar un índice. Asimismo, estoy especialmente agradecido a William Shaw, por acudir en nuestra ayuda en el tramo final para generar excelentes gráficas por ordenador (Figs. 1.2 y 2.19, y por la implementación de la transformación incluida en las Figs. 2.16 y 2.19), utilizadas aquí para el conjunto de Mandelbrot y el plano hiperbólico. En cuanto a Jacob Foster, todas las gracias que pueda expresarle por su trabajo hercúleo en la búsqueda y obtención de referencias y por revisar el manuscrito completo en un tiempo extraordinariamente breve y rellenar innumerables lagunas, no podrán hacer justicia de ninguna manera a la magnitud de su ayuda. Su impronta personal en un enorme número de notas finales les da una calidad especial. Por supuesto, ninguna de las personas a las que doy las gracias son culpables de los errores y omisiones que pueda haber, cuya única responsabilidad recae sobre mí.

Debo expresar una gratitud especial a la M.C. Escher Company, en Holanda, por el permiso para reproducir obras de Escher en las Figs. 2.11, 2.12, 2.16 y 2.22, y en particular por permitir las modificaciones de la Fig. 2.11 que se han utilizado en las Figs. 2.12 y 2.16, la última de las cuales es una transformación matemática explícita. Todas las obras de Escher utilizadas en este libro tienen *copyright* (2004) The M.C. Escher Company. Gracias también al Institute of Theoretical Physics, la Universidad de Heidelberg y a Charles H. Lineweaver, por su permiso para reproducir las gráficas respectivas en las Figs. 27.19 y 28.19.

Por último, mi gratitud ilimitada se dirige a mi querida esposa Vanessa, no solo por suministrarme gráficas de ordenador al instante (Figs. 4.1, 4.2, 5.7, 6.2-6.8, 8.15, 9.1, 9.2, 9.8, 9.12, 21.3b, 21.10, 27.5, 27.14, 27.15, y los poliedros en la Fig. 1.1), sino por su continuado amor y cariño, y su profunda comprensión y sensibilidad, pese a los en apariencia interminables años en que su marido solo estaba presente mentalmente en parte. Y también Max, que durante toda su vida solo ha tenido la oportunidad de conocerme en ese estado distraído, merece mi más calurosa gratitud, no solo por retrasar la redacción de este libro (que hizo posible alargarse e incluir dos elementos de información importantes que, de otra forma, no habría tenido), sino por la continua alegría y el optimismo que transmite y que me ha ayudado a mantenerme en buena forma. Después de todo, es mediante la renovación de la vida, tal como él mismo representa, como vendrán las nuevas fuentes de ideas e intuiciones necesarias para un genuino progreso futuro a la búsqueda de esas leyes más profundas que realmente gobiernan el universo en que vivimos.

Notación

(No lo lea antes de que se haya familiarizado con los conceptos, ¡aunque quizá encuentre los tipos de letra confusos!)

He tratado de ser razonablemente coherente en el uso de tipos de letra especiales a lo largo de todo el libro. Pero puesto que no todo esto es estándar, puede resultar útil para el lector que haga explícitos los usos que he adoptado.

Las letras itálicas (griegas o latinas), tales como en w^2, p^n, $\log z$, $\cos\theta$, $e^{i\theta}$, o e^x se utilizan de la manera convencional para variables matemáticas que son cantidades numéricas o escalares; pero las constantes numéricas establecidas, tales como e, i, o π o las funciones establecidas tales como sen, cos o log se denotan mediante redondas. Sin embargo, las constantes físicas estándar tales como c, G, h, $\hbar$, g o k son itálicas.

Una cantidad vectorial o tensorial, cuando se considera en su totalidad (abstracta), se denota por una letra itálica negrilla, tal como $\boldsymbol{R}$ para el tensor de curvatura de Riemann, mientras que su conjunto de componentes podría escribirse con itálicas (tanto para el núcleo como para sus índices), como R_{abcd}. De acuerdo con la notación de índices abstractos, introducida aquí en §12.8, la cantidad R_{abcd} puede representar alternativamente el tensor entero $\boldsymbol{R}$, si esta interpretación es apropiada, y esto debería hacerse claro en el texto. Las transformaciones lineales abstractas son tipos de tensores, y letras itálicas negrillas tales como $\boldsymbol{T}$ se utilizan también para tales entidades. También se utiliza aquí, cuando es apropiado, la forma de índices abstractos $T^a{}_b$ para una transformación lineal abstracta, donde el escalonamiento de índices hace clara la conexión precisa con el ordenamiento de la multiplicación de matrices. Así, la expresión de índices abstractos $S^a{}_b T^b{}_c$ representa el producto $\boldsymbol{ST}$ de transformaciones lineales. Como sucede con los tensores generales, los símbolos $S^a{}_b$ y $T^b{}_c$ pueden representar alternativamente (según el contexto o la especificación explícita en el texto) los correspondientes conjuntos de componentes —que son *matrices*— para los que también pueden utilizarse letras redondas negrillas **S** y **T**. En ese caso, **ST** denota el correspondiente producto matricial. Esta interpretación «ambivalente» de símbolos tales como R_{abcd} o $S^a{}_b$ (ya representen el conjunto de componentes o el propio tensor abstracto) no debería causar confusión, pues las relaciones algebraicas (o diferenciales) a las que están sujetos estos símbolos son idénticas para ambas interpretaciones. A veces se utiliza también aquí una tercera notación para dichas cantidades —la *notación diagramática*— que se describe en las Figs. 12.17, 12.18, 14,6, 14.7, 14.21, 19.1 y otros lugares en el libro.

Hay lugares en este libro donde necesito distinguir las entidades espaciotemporales 4-dimensionales de la teoría de la relatividad de las correspondientes entidades espaciales ordinarias puramente 3-dimensionales. Así, aunque podría utilizarse una notación itálica negrilla, como antes, tal como $\boldsymbol{p}$ o $\boldsymbol{x}$, para el 4-momento o la 4-posición, respectivamente, las correspondientes entidades puramente espaciales 3-dimensionales se denotarán por las correspondientes letras negrillas **p** o **x**. Por analogía con la notación **T** para una matriz, en oposición a $\boldsymbol{T}$ para una transformación lineal abstracta, se considera que las cantidades **p** y **x** «representan» las tres componentes espaciales, en cada caso, mientras que $\boldsymbol{p}$ y $\boldsymbol{x}$ podrían verse como una interpretación más abstracta libre de componentes (aunque yo no seré especialmente estricto en esto). La «longitud» euclídea de una cantidad 3-vectorial $\mathbf{a} = (a_1, a_2, a_3)$ puede escribirse a, donde $a^2 = a_1^2 + a_2^2 + a_3^2$, y el producto escalar de **a** por $\mathbf{b} = (b_1, b_2, b_3)$, se escribe $\mathbf{a} \bullet \mathbf{b} = a_1 b_1 + a_2 b_2 + a_3 b_3$. Esta notación «punto» para productos escalares se aplica también, en el contexto n-dimensional general, para el producto escalar (o interno) $\boldsymbol{\alpha} \bullet \boldsymbol{\xi}$ de un covector abstracto $\boldsymbol{\alpha}$ por un vector $\boldsymbol{\xi}$.

Sin embargo, surge una complicación notacional con la mecánica cuántica, puesto que las magnitudes físicas, en esta disciplina, suelen representarse como operadores lineales. No adopto el procedimiento totalmente estándar en este contexto de colocar «sombreros» (circunflejos) sobre las letras que representan las versiones como operadores cuánti-

cos de las familiares magnitudes clásicas, pues creo que esto conduce a una innecesaria acumulación de signos. (En su lugar, tenderé a adoptar un punto de vista filosófico según el cual las magnitudes clásicas y cuánticas son realmente las «mismas» —y por ello es justo utilizar los mismos símbolos para ambas—, excepto que en el caso clásico uno está justificado al ignorar cantidades del orden de $\hbar$, de modo que las propiedades de conmutación clásica $ab = ba$ pueden ser válidas, mientras que en mecánica cuántica ab podría diferir de ba en algo del orden de $\hbar$.) Por consistencia con lo anterior, parecería que tales operadores lineales tienen que ser denotados por letras itálicas negrillas (como $\boldsymbol{T}$), pero eso anularía la filosofía y las distinciones invocadas en el párrafo anterior. Por consiguiente, con respecto a magnitudes específicas, tal como el momento $\mathbf{p}$ o $\boldsymbol{p}$, o la posición $\mathbf{x}$ o $\boldsymbol{x}$, tenderé a utilizar la misma notación que en el caso clásico, en la misma línea con lo que se ha dicho antes en este párrafo. Pero para operadores cuánticos menos específicos se utilizarán letras itálicas negrillas tales como $\boldsymbol{Q}$.

Las letras $\mathbb{N}$, $\mathbb{Z}$, $\mathbb{R}$, $\mathbb{C}$ y $\mathbb{F}_q$, respectivamente, para el sistema de los números naturales (i.e., enteros no negativos), los enteros, los números reales, los números complejos y el campo finito con q elementos (siendo q una potencia de un número primo, véase §16.1), son ahora estándares en matemáticas, como lo son los correspondientes $\mathbb{N}^n$, $\mathbb{Z}^n$, $\mathbb{R}^n$, $\mathbb{C}^n$ y $\mathbb{F}_q^n$, para los sistemas de n-tuplas ordenadas de tales números. Estas son entidades matemáticas canónicas de uso estándar. En este libro, esta notación se extiende a otras estructuras matemáticas estándar tales como un 3-espacio euclídeo $\mathbb{E}^3$ o, con más generalidad, un n-espacio euclídeo $\mathbb{E}^n$. En este libro aparece con frecuencia el espaciotiempo de Minkowski plano 4-dimensional estándar, que es en sí mismo un tipo de espacio «pseudo»-euclídeo, de modo que utilizo la letra $\mathbb{M}$ para este espacio (con $\mathbb{M}^n$ para denotar la versión n-dimensional, un espaciotiempo «lorentziano» con 1 tiempo y $(n-1)$ dimensiones espaciales). A veces utilizo $\mathbb{C}$ como un adjetivo, para denotar «complexificado», de modo que podríamos considerar el 4-espacio euclídeo complejo, por ejemplo, denotado por $\mathbb{CE}^n$. La letra $\mathbb{P}$ puede utilizarse también como un adjetivo, para denotar «proyectivo» (véase §15.6), o como un sustantivo, cuando $\mathbb{P}^n$ denota el n-espacio proyectivo (utilizo $\mathbb{RP}^n$ o $\mathbb{CP}^n$ si hay que dejar claro que estamos interesados en un n-espacio proyectivo real o complejo, respectivamente). En teoría de twistores (capítulo 33), existe el 4-espacio complejo $\mathbb{T}$, que está relacionado con $\mathbb{M}$ (o su complexificación $\mathbb{CM}$) de una manera canónica, y existe también la versión proyectiva $\mathbb{PT}$. En esta teoría existe también un espacio $\mathbb{N}$ de twistores nulos (el doble papel que tiene esta letra no causa aquí conflicto), y su versión proyectiva $\mathbb{PN}$.

El papel adjetivo de la letra $\mathbb{C}$ no debería confundirse con el del tipo C, que aquí representa «conjugado complejo de» (como se utiliza en §§13.1,2). Esto es básicamente similar a otro uso de C en física de partículas, a saber la *conjugación de carga*, que es la operación que intercambia cada partícula con su antipartícula (véanse los capítulos 24 y 25). Esta operación se considera habitualmente junto con otras dos operaciones básicas en física de partículas, a saber, P para *paridad*, que se refiere a la operación de reflexión en un espejo, y T, que se refiere a la *inversión temporal*. Las letras sans serif que son negrillas sirven aquí para un objetivo diferente, el de etiquetar *espacios vectoriales*, siendo las letras $\mathbf{V}$, $\mathbf{W}$ y $\mathbf{H}$ las utilizadas con más frecuencia con este objetivo. El uso de $\mathbf{H}$ es específico para los espacios de Hilbert de la mecánica cuántica, y $\mathbf{H}^n$ representaría un espacio de Hilbert de n dimensiones complejas. Los espacios vectoriales son, en un sentido claro, planos. Los espacios que son (o podrían ser) *curvos* se denotan por letras script, tales como $\mathcal{M}$, $\mathcal{S}$ o $\mathcal{T}$, donde hay un uso especial para el tipo concreto $\mathscr{I}$ que denota el *infinito nulo*. Además, sigo el convenio común que utiliza letras script para lagrangianos ($\mathcal{L}$) y hamiltonianos ($\mathcal{H}$), en vista de su estatus muy especial en la teoría física.

Prólogo

Am-tep, un artista de habilidades consumadas, era el maestro artesano del rey. Esa noche estaba durmiendo en el sofá de su taller, cansado después de una tarde de trabajo generosamente productivo, pero no podía conciliar el sueño, quizá por una tensión intangible que se respiraba en el aire. En realidad, ni siquiera estaba seguro de estar dormido cuando sucedió. De repente, había amanecido, aunque él tenía la sensación de que todavía debía ser de noche.

Se levantó con brusquedad. Algo muy extraño estaba sucediendo. La luz del alba no podía venir del norte; y, sin embargo, una luz roja resplandecía alarmantemente en su amplia ventana que daba al norte sobre el mar. Se acercó a la ventana y miró al exterior con estupor e incredulidad. ¡Nunca antes había salido el Sol por el norte! Aturdido como estaba, necesitó algún tiempo para darse cuenta de que eso no podía ser el Sol. Era un rayo de luz distante, de un intenso rojo vivo, proyectado verticalmente desde las aguas hacia el cielo.

Mientras permanecía allí, sobre el haz apareció una nube más oscura que daba al conjunto la apariencia de una sombrilla gigantesca y lejana, de un brillo maligno, con un bastón llameante y lleno de humo. La capucha de la sombrilla empezó a ensancharse y oscurecerse: parecía un demonio del mundo subterráneo. La noche se había aclarado, pero ahora las estrellas desaparecían una tras otra engullidas por el avance de esa monstruosa criatura del infierno.

Aunque su reacción natural debería haber sido de terror, él permaneció inmóvil, paralizado durante varios minutos ante la perfecta simetría y la impresionante belleza de la escena. Pero entonces la terrible nube empezó a desviarse ligeramente hacia el este, empujada por el

viento. Quizá eso le alivió algo y el hechizo se rompió momentáneamente. Pero de inmediato volvió a sentir aprensión cuando le pareció experimentar una extraña alteración en el suelo, acompañada de un ruido sordo, inquietante, de una naturaleza completamente desconocida para él. Empezó a preguntarse qué podía haber provocado tanta furia. Nunca antes había sido testigo de la ira divina de una magnitud semejante.

Su primera reacción fue culparse a sí mismo por el diseño de la copa sacrificial que acababa de terminar: eso le había tenido preocupado. ¿Quizá su representación del dios-toro no había sido suficientemente terrorífica? ¿Se había ofendido el dios? Pero pronto comprendió lo absurdo de su idea. La furia de la que había sido testigo no podía ser el resultado de una acción tan trivial como la suya, y seguramente no iba dirigida contra él en particular. Pero sabía que habría problemas en el Gran Palacio. El rey sacerdote se apresuraría a tratar de apaciguar a ese dios demonio. Habría sacrificios. Las tradicionales ofrendas de frutos o incluso animales no serían suficientes para aplacar una ira de esa magnitud. Los sacrificios serían humanos.

De repente, y para su sorpresa, se vio lanzado hacia el fondo de la habitación por un golpe de aire seguido de un viento violento. El ruido era tan atronador que por un momento le dejó sordo. Muchos de sus vasos de arcilla bellamente ornamentados fueron barridos de las estanterías y se hicieron añicos contra la pared trasera. Tendido en el suelo en el apartado rincón al que había sido lanzado por el golpe de aire, empezó a recobrar el sentido y vio que la habitación estaba en completo desorden. Quedó horrorizado al ver que una de sus grandes urnas favoritas estaba destrozada y ya no existían los preciosos diseños que tan primorosamente había trabajado.

Am-tep se levantó tambaleándose y al cabo de un rato se acercó de nuevo a la ventana, esta vez con gran agitación, para contemplar de nuevo aquella lejana y terrible escena en medio del mar. Entonces creyó ver una perturbación que se dirigía hacia él, iluminada por ese horno lejano. Parecía una enorme depresión que se movía con rapidez hacia la costa, seguida por un muro de agua que semejaba un acantilado. De nuevo quedó paralizado, observando cómo la ola que se aproximaba alcanzaba proporciones gigantescas. Finalmente, la perturbación lle-

gó a la costa y la parte del mar que había inmediatamente ante él se vació, dejando numerosos barcos encallados en la playa recién formada. Luego la ola acantilado entró en la zona vaciada y la golpeó con terrible violencia. Todos los barcos quedaron hechos pedazos, y muchas casas próximas fueron destruidas en un instante. Aunque el agua alcanzó una gran altura en el espacio que había delante de él, su propia casa se salvó de la destrucción gracias a que estaba situada en un terreno elevado y a gran distancia del mar.

El Gran Palacio también se salvó. Pero Am-tep temía que lo peor estaba por llegar, y tenía razón —aunque no podía imaginar hasta qué punto—. Sabía, no obstante, que ahora no bastaría con ningún sacrificio humano ordinario de un esclavo. Sería necesario algo más para aplacar la ira tempestuosa de ese dios terriblemente enfurecido. Pensó en sus hijos e hijas, y en su nieto recién nacido. Ni siquiera ellos estarían a salvo.

Am-tep estaba en lo cierto al temer nuevos sacrificios humanos. Pronto fueron apresados una joven y un joven de buena cuna, y llevados a un templo cercano, a gran altura en la falda de una montaña. Se estaba procediendo al correspondiente ritual cuando sobrevino otra catástrofe. El suelo tembló con una violencia devastadora, y el techo del templo se vino abajo, matando al instante a todos los sacerdotes y a sus presuntas víctimas sacrificiales. Allí, atrapados en mitad del ritual, ¡yacerían enterrados durante tres mil quinientos años!

La devastación fue espantosa, pero no absoluta. Muchas de las islas donde vivían Am-tep y su pueblo sobrevivieron al terrible terremoto, aunque el Gran Palacio quedó destruido casi por completo. Se reconstruyeron muchas cosas en el curso de los años. Incluso el palacio, construido sobre las ruinas del antiguo, iba a recuperar mucho de su esplendor original. Pese a todo, Am-tep se había jurado abandonar la isla. Su mundo había cambiado irremisiblemente.

En el mundo que él conoció se habían dado mil años de paz, prosperidad y cultura, durante los cuales había reinado la diosa tierra. Había florecido un arte maravilloso. Existía un gran comercio con las islas vecinas. El magnífico Gran Palacio era un enorme y lujoso laberinto, prácticamente una ciudad en sí mismo, adornado con soberbios frescos de animales y plantas. Había agua corriente, un excelente sistema de al-

cantarillado y cisternas. La guerra era casi desconocida y las defensas innecesarias. Pero ahora Am-tep tenía la sensación de que la diosa tierra había sido derrocada por un ser con valores completamente diferentes.

Pasaron algunos años antes de que Am-tep, acompañado de su familia superviviente, abandonase definitivamente la isla en un barco reconstruido por su hijo más joven, que era un hábil carpintero y marino. El nieto de Am-tep había crecido y se había convertido en un muchacho despierto, interesado en todo lo que le rodeaba. El viaje duró varios días, pero el tiempo era sumamente apacible. Una noche clara, Am-tep estaba explicando a su nieto las figuras que formaban las estrellas cuando le asaltó una extraña idea: *Las figuras que formaban las estrellas no habían sufrido la más mínima alteración con respecto a las que eran antes de la catástrofe de la emergencia del terrible demonio.*

Am-tep conocía muy bien esas figuras, pues tenía la visión profunda de un artista. Si sus vasijas habían quedado destrozadas y su gran urna se había hecho añicos, pensaba él, ¿no deberían aquellas minúsculas candelas en el cielo haber sido apartadas, aunque fuera ligeramente, de sus posiciones por la violencia de aquella noche? La Luna también había mantenido su cara, igual que antes, y su ruta a través del cielo lleno de estrellas no había cambiado un ápice, hasta donde Am-tep podía afirmar. Durante muchas lunas posteriores a la catástrofe, los cielos habían parecido en efecto diferentes. Hubo oscuridad y nubes extrañas, y la Luna y el Sol habían mostrado a veces colores inusuales. Pero ahora que eso había pasado, sus movimientos parecían ser exactamente los mismos que habían sido antes. Y, de igual forma, las minúsculas estrellas no se habían movido en absoluto.

Si los cielos, que tienen una estatura mucho mayor que la de ese terrible demonio, habían mostrado tan poco interés por la catástrofe, pensó Am-tep, ¿por qué las fuerzas que controlaban al propio demonio habían de mostrar interés por lo que estaba haciendo el pequeño pueblo de la isla, con sus ridículos rituales y sacrificios humanos? Se sintió avergonzado por las absurdas ideas que había tenido entonces, cuando pensó que el demonio podría estar interesado en las sencillas figuras de sus vasijas.

Pero Am-tep seguía preocupado por la pregunta: ¿por qué? ¿Qué profundas fuerzas controlan el comportamiento del mundo, y por qué

a veces estallan de formas violentas y aparentemente incomprensibles? Compartía estas preguntas con su nieto, pero no había respuestas.

Pasó un siglo, y luego un milenio, y aún no había respuestas.

Amphos el artesano había vivido toda su vida en el mismo pueblo que su padre, y que el padre de su padre antes de este, y el padre del padre de su padre aun antes de eso. Se ganaba la vida haciendo brazaletes de oro bellamente decorados, pendientes, copas ceremoniales y otros finos productos fruto de sus habilidades artísticas. Ese trabajo había sido la ocupación de la familia durante cuarenta generaciones: una línea ininterrumpida desde que Am-tep se hubiera establecido allí mil cien años antes.

Pero no eran solo las habilidades artísticas las que se habían transmitido de una generación a otra. Las preguntas de Am-tep preocupaban a Amphos como habían preocupado al propio Am-tep en una época anterior. La gran historia de la catástrofe que destruyó a una antigua y pacífica civilización se había transmitido de padres a hijos. La percepción que tuvo Am-tep de la catástrofe también había sobrevivido con sus descendientes. Amphos, asimismo, comprendía que los cielos tenían una magnitud y estatura tan enormes que estarían completamente desinteresados por aquel terrible suceso. En cualquier caso, el suceso había tenido un efecto catastrófico sobre la pequeña población con sus ciudades y sus sacrificios humanos y sus insignificantes rituales religiosos. Por comparación, el propio suceso debía haber sido el resultado de fuerzas enormes completamente indiferentes a tales acciones triviales de los seres humanos. Pero la naturaleza de dichas fuerzas era tan desconocida en la época de Amphos como lo era para Am-tep.

Amphos había estudiado la estructura de las plantas, los insectos y otros pequeños animales, así como de las rocas cristalinas. Su habilidad para la observación también le había sido útil para sus dibujos decorativos. Se interesó por la agricultura y quedó fascinado por el crecimiento del trigo y otras plantas a partir del grano. Pero nada de esto le decía «¿por qué?», y se sentía insatisfecho. Creía que había una razón subyacente en las pautas de la naturaleza, pero no estaba preparado para descubrir dichas razones.

Una noche clara, Amphos levantó la vista al cielo y, a partir de las pautas de las estrellas, trató de construir las figuras de aquellos héroes y heroínas que formaban las constelaciones en el cielo. Para su humilde ojo de artista, los parecidos de aquellas formas eran muy pobres. Él mismo podría haber dispuesto las estrellas de forma mucho más convincente. ¿Por qué los dioses no han dispuesto las estrellas de una forma más adecuada?, se preguntaba. Tal como estaban, las disposiciones se parecían más a granos diseminados, sembrados al azar por un granjero, que a un diseño deliberado de un dios. Entonces le asaltó una extraña idea: *No busques razones en las pautas concretas de las estrellas, o en otras disposiciones desordenadas de objetos; busca en su lugar un orden universal más profundo en el comportamiento de los objetos.*

Amphos razonaba que, después de todo, no encontramos orden en las figuras que forman las semillas dispersas cuando caen al suelo, sino en la forma milagrosa en que cada una de ellas puede desarrollarse hasta formar una planta viva, con una soberbia estructura, y cada una de ellas similar en los detalles a las demás. Nosotros no trataríamos de buscar significado en las disposiciones de las semillas dispersas en el suelo; pese a todo, debe de haber un significado en el misterio oculto de las fuerzas internas que controlan el crecimiento de cada semilla individual, de tal modo que cada una sigue esencialmente el mismo curso maravilloso. En realidad, las leyes de la naturaleza deben de tener una soberbia precisión para que esto sea posible.

Amphos se convenció de que sin precisión en las leyes subyacentes no podría haber orden en el mundo, mientras que se percibe mucho orden en el comportamiento de las cosas. Más aún, debe haber precisión en nuestros modos de pensar acerca de estas cuestiones si no queremos extraviarnos sin remedio.

Sucedió que Amphos tuvo noticias de un sabio que vivía en otro lugar de la tierra, y cuyas creencias parecían estar en armonía con las suyas. Según este sabio, uno no podía basarse en las enseñanzas y tradiciones del pasado. Para estar seguro de las propias creencias, era necesario llegar a conclusiones precisas mediante el uso de una razón indiscutible. Esta precisión tenía que ser de naturaleza matemática, dependiente en definitiva de la noción de *número* y su aplicación a las formas geométricas. En consecuencia, debían ser número y geome-

tría, y no mito y superstición, los que gobernaran el comportamiento del mundo.

Igual que había hecho Am-tep once siglos antes, Amphos se hizo a la mar. Encontró su camino a la ciudad de Crotona, donde el sabio y su fraternidad de 571 hombres sabios y 28 mujeres sabias estudiaban en busca de la verdad. Al cabo de un tiempo, Amphos fue aceptado en la fraternidad. El nombre del sabio era Pitágoras.

1

Las raíces de la ciencia

1.1. La búsqueda de las fuerzas que configuran el mundo

¿Qué leyes rigen nuestro universo? ¿Cómo las conoceremos? ¿Cómo puede servirnos este conocimiento para comprender el mundo y con ello orientar sus acciones en nuestro provecho?

Desde los albores de la humanidad, los hombres se han sentido profundamente intrigados por preguntas como estas. Al principio trataron de dar sentido a las fuerzas que controlan el mundo aferrándose al tipo de conocimiento que les era accesible a partir de sus propias vidas. Imaginaban que cualquier cosa o quienquiera que fuera lo que controlaba su entorno lo haría de la misma forma en que ellos se esforzaban para controlar las cosas: originalmente habían creído que su destino estaba bajo la influencia de seres que actuaban de acuerdo con sus propios y variados impulsos humanos. Tales fuerzas impulsoras podían ser el orgullo, el amor, la ambición, la rabia, el miedo, la venganza, la pasión, el castigo, la lealtad o el arte. Por consiguiente, el curso de los fenómenos naturales —como el Sol, la lluvia, las tormentas, el hambre, la enfermedad o la pestilencia— se entendía como el capricho de dioses o diosas motivados por tales impulsos humanos. Y lo único que se podía hacer para influir en estos acontecimientos era apaciguar a las figuras divinas.

Pero poco a poco se empezó a establecer la fiabilidad de otro tipo de pautas. La precisión del movimiento del Sol en el cielo y su evidente relación con la alternancia del día y la noche ofrecía el ejemplo más obvio; pero también la posición del Sol respecto a las estrellas del orbe celeste aparecía estrechamente asociada al cambio y a la implacable re-

gularidad de las estaciones, y a la clara influencia en el clima que la acompañaba y, en consecuencia, en la vegetación y el comportamiento animal. También el movimiento de la Luna parecía firmemente regulado, y sus fases determinadas por su relación geométrica con el Sol. Se advirtió que en aquellos lugares de la Tierra en los que los océanos abiertos se encuentran con la tierra, las mareas tenían una regularidad rígidamente gobernada por la posición (y la fase) de la Luna. Por último, incluso los mucho más complicados movimientos aparentes de los planetas empezaron a ceder sus secretos, revelando una regularidad y una inmensa precisión subyacente. Si los cielos estaban realmente controlados por los caprichos de los dioses, entonces estos mismos dioses parecían estar bajo el hechizo de leyes matemáticas exactas.

Del mismo modo, las leyes que controlaban algunos fenómenos terrestres —tales como los cambios diarios y anuales de temperatura, el flujo y reflujo de los océanos, y el crecimiento de las plantas— que, al menos en ese aspecto, se veían influidos por los cielos, compartían esa misma regularidad matemática que parecía guiar a los dioses. Pero este tipo de relación entre el comportamiento de los cuerpos celestes y los terrestres iba a ser a veces exagerado o mal entendido, e iba a cobrar una importancia desmesurada, que llevaría a las connotaciones ocultas y místicas de la astrología. Pasaron muchos siglos antes de que el rigor del conocimiento científico hiciera posible desenredar las verdaderas influencias de los cielos de las puramente hipotéticas y místicas. Pese a todo, desde los tiempos más remotos había estado claro que aquellas influencias existían realmente y que, en consecuencia, las leyes matemáticas de los cielos debían tener relevancia también aquí en la Tierra.

De forma en apariencia independiente se percibieron otras regularidades en el comportamiento de los objetos terrestres. Una de ellas era la tendencia de todas las cosas en una vecindad a moverse en la misma dirección hacia abajo, bajo la influencia de lo que ahora llamamos *gravedad*. Se observó que a veces la materia se transformaba de una forma en otra, tal como ocurría en la fusión del hielo o la disolución de la sal, aunque la cantidad total de materia nunca parecía cambiar, lo que refleja la ley que ahora conocemos como *conservación de la masa*. Además, se advirtió que hay muchos cuerpos materiales con la importante propiedad de que conservan su forma, de donde surgió el concepto de

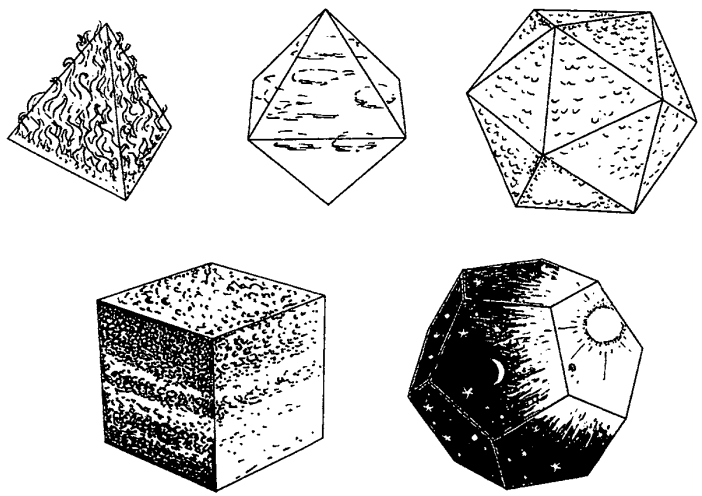


Fig. 1.1. Una asociación fantástica, hecha por los antiguos griegos, entre los cinco sólidos platónicos y los cuatro «elementos» (fuego, aire, agua y tierra), junto con el firmamento celeste representado por el dodecaedro.

movimiento espacial rígido; y se hizo posible comprender las relaciones espaciales en términos de una geometría precisa y bien definida: la geometría tridimensional que ahora denominamos *euclídea*. Más aún, la noción de «línea recta» en esta geometría resultó ser la misma que la que proporcionaban los rayos luminosos (o las líneas visuales). Sin duda, había una extraordinaria precisión y belleza en estas ideas, que despertaban una gran fascinación en los antiguos, igual que la despiertan hoy en nosotros.

Sin embargo, y en relación con nuestras vidas cotidianas, las implicaciones de esta precisión matemática para las acciones del mundo parecían con frecuencia poco excitantes y limitadas, pese al hecho de que las propias matemáticas parecían representar una verdad profunda. En consecuencia, en tiempos antiguos muchas personas iban a permitir que su imaginación se dejara llevar por su fascinación por el tema y les condujese mucho más allá de lo que era adecuado. En astrología, por ejemplo, las figuras geométricas también solían generar connotaciones místicas y ocultas, como era el caso de las supuestas potencias mágicas

de pentagramas y heptagramas. Y había una supuesta asociación completamente hipotética entre los sólidos platónicos y los estados elementales de la materia (véase la Fig. 1.1). Tardarían muchos siglos en llegar los conocimientos más profundos que tenemos en la actualidad, concernientes a las relaciones reales entre la masa, la gravedad, la geometría, el movimiento planetario y el comportamiento de la luz.

1.2. La verdad matemática

Los primeros pasos hacia una comprensión de las influencias reales que controlan la naturaleza requerían desenredar lo verdadero de lo puramente hipotético. Pero antes de que estuvieran en situación de hacer esto de forma fiable para su conocimiento de la naturaleza, los antiguos necesitaban algo más. Lo primero que tenían que hacer era descubrir la forma de desenredar lo verdadero de lo hipotético en *matemáticas*. Se necesitaba un procedimiento para decir si se puede confiar o no en la verdad de una afirmación matemática dada. Hasta que no quedara establecida de forma razonable esta cuestión preliminar, habría pocas esperanzas de abordar con seriedad aquellos problemas más difíciles concernientes a las fuerzas que controlan el comportamiento del mundo y cualesquiera que pudieran ser sus relaciones con la verdad matemática. Esta comprensión de que la clave para entender la naturaleza reside en unas matemáticas incuestionables fue quizá el primer avance trascendental en la ciencia.

Aunque ya desde los tiempos antiguos de Egipto y Babilonia se habían supuesto todo tipo de verdades matemáticas, solo cuando los filósofos griegos Tales (*c.* 625-547 a.C.) y Pitágoras[1]* de Samos (*c.* 572-497 a.C.) empezaron a introducir la idea de *demostración matemática* se colocó la primera piedra fundacional firme del conocimiento matemático —y, por consiguiente, de la propia ciencia—. Quizá fuera Tales el primero en introducir esta idea de demostración, pero parece que fueron los pitagóricos quienes hicieron por primera vez un uso importante de la misma para establecer cosas que, de otro modo, no eran obvias. Parece que Pitágoras también tuvo una fuerte intuición de la importancia del *número*, y de los conceptos aritméticos, en el gobierno de las acciones del mundo físico. Se dice que un factor importante en esta comprensión fue el darse cuenta de que las armonías más bellas producidas por liras o flautas correspondían a las razones más simples entre las longitudes de las cuerdas vibrantes o los tubos. También se dice que él introdujo la «escala pitagórica», cuyas razones numéricas sabemos ahora que son las frecuencias que determinan los intervalos principales en los que se basa esencialmente la música occidental.[2] El famoso *teorema de Pitágoras*, que afirma que el cuadrado de la hipotenusa de un triángulo rectángulo es igual a la suma de los cuadrados de los otros dos lados, mostró, quizá más que cualquier otra cosa, que existe una relación precisa entre la aritmética de los números y la geometría del espacio físico (véase el capítulo 2).

Pitágoras tuvo un número considerable de seguidores —los *pitagóricos*— establecidos en la ciudad de Crotona, en lo que hoy es el sur de Italia, pero su influencia en el mundo exterior se vio dificultada por el hecho de que todos los miembros de la fraternidad pitagórica hacían un juramento de secreto. Por ello, casi todas sus conclusiones detalladas se han perdido. De todas formas, algunas de estas conclusiones se filtraron, con consecuencias desafortunadas para los «topos», que, al menos en una ocasión, ¡sufrieron el castigo de morir ahogados!

A la larga, la influencia de los pitagóricos sobre el progreso del pensamiento humano ha sido enorme. Por primera vez, con demostración matemática, era posible hacer afirmaciones significativas de un carácter incuestionable, de modo que seguirían siendo tan verdaderas hoy como en la época en que se hicieron, con independencia de cuánto haya progresado nuestro conocimiento del mundo desde entonces. Empezaba a revelarse la naturaleza verdaderamente intemporal de las matemáticas.

Pero ¿qué es una demostración matemática? En matemáticas, una demostración es un argumento impecable, basado solo en los métodos del razonamiento puramente lógico, que permite inferir la validez de una afirmación matemática dada a partir de la validez preestablecida de otras afirmaciones matemáticas, o de ciertas afirmaciones concretas

* Las notas, indicadas en el texto mediante superíndices, se recogen al final de cada capítulo.

primitivas —los *axiomas*— cuya validez se considera evidente. Una vez que tal afirmación matemática ha quedado establecida de esta forma, se conoce como un *teorema*.

Muchos de los teoremas que interesaban a los pitagóricos eran de naturaleza geométrica; otros eran solo afirmaciones sobre números. Aquellos que concernían puramente a los números tienen hoy una validez inequívoca, igual que la tenían en los tiempos de Pitágoras. ¿Qué ocurre con los teoremas *geométricos* que los pitagóricos habían obtenido utilizando sus procedimientos de demostración matemática? También estos tienen hoy una clara validez, pero ahora surge una cuestión que complica las cosas. Se trata de una cuestión cuya naturaleza es más obvia para nosotros desde nuestro punto de vista moderno que lo era en el tiempo de Pitágoras. Los antiguos solo conocían un tipo de geometría, a saber, la que ahora llamamos *geometría euclídea*, pero ahora conocemos otros muchos tipos. Así pues, al considerar los teoremas geométricos de la época griega antigua es importante especificar que la noción de geometría a la que nos referimos es en realidad la geometría de Euclides. (Seré más explícito sobre estas cuestiones en §2.4, donde se dará un ejemplo importante de geometría no euclídea.)

La geometría euclídea es una estructura matemática específica, con sus propios axiomas específicos (incluidas algunas afirmaciones menos seguras conocidas como postulados), que proporciona una excelente aproximación a un aspecto concreto del mundo físico. Este era el aspecto de realidad, muy familiar para los antiguos griegos, que remitía a las leyes que gobiernan la geometría de los objetos rígidos y sus relaciones con otros objetos rígidos cuando se movían en el espacio tridimensional. Algunas de estas propiedades eran tan familiares y autoconsistentes que tendían a ser consideradas como verdades matemáticas «evidentes» y se tomaban como axiomas (o postulados). Como veremos en los capítulos 17-19 y en §§27.8,11, la relatividad general de Einstein —e incluso el espaciotiempo minkowskiano de la relatividad especial— proporciona geometrías para el universo físico que son diferentes de, e incluso más precisas que, la geometría de Euclides, pese al hecho de que la geometría euclídea de los antiguos era ya extraordinariamente precisa. Así pues, a la hora de considerar afirmaciones geométricas debemos tener cuidado si confiamos en los «axiomas» como si fueran, en cualquier sentido, realmente *verdaderos*.

Pero ¿qué significa «verdadero» en este contexto? La dificultad fue apreciada por el gran filósofo griego Platón, que vivió en Atenas desde *c.* 429 hasta 347 a.C., aproximadamente un siglo y medio después de Pitágoras. Platón dejó claro que las proposiciones matemáticas —las cosas que podían considerarse como incuestionablemente verdaderas— no se refieren a objetos físicos reales (como los cuadrados, triángulos, círculos, esferas y cubos aproximados que podrían construirse con marcas en la arena, o con piedra o madera), sino a ciertas entidades idealizadas. Él imaginaba que esas entidades ideales habitaban en un mundo diferente, distinto del mundo físico. Hoy día podríamos llamar a este mundo el *mundo platónico de las formas matemáticas*. Las estructuras físicas, tales como los cuadrados, los círculos o los triángulos recortados en papiro, o marcados en una superficie plana, o quizá los cubos, los tetraedros o las esferas esculpidas en mármol, podrían ajustarse estrechamente a estos ideales, pero solo de forma aproximada. Los cuadrados, cubos, círculos, esferas, triángulos, etc., *matemáticos* reales no serían parte del mundo físico, sino que serían habitantes del mundo platónico de las formas matemáticas idealizadas.

1.3. ¿Es «real» el mundo matemático de Platón?

Esta era una idea extraordinaria para su época, y ha resultado ser una idea muy fecunda. Pero ¿existe realmente el mundo matemático platónico, en cualquier sentido significativo? Muchas personas, incluidos los filósofos, podrían considerar que un «mundo» semejante es una completa ficción, un mero producto de nuestra imaginación desbordante. Pese a todo, el punto de vista platónico es inmensamente valioso. Nos dice que debemos ser cuidadosos en distinguir las entidades matemáticas precisas de las aproximaciones que vemos a nuestro alrededor en el mundo de los objetos físicos. Más aún, nos proporciona el esquema con el que ha procedido la ciencia desde entonces. Los científicos propondrán modelos del mundo —o, mejor, de ciertos aspectos del mundo— y estos modelos pueden ser puestos a prueba frente a observa-

ciones previas y frente a los resultados de experimentos cuidadosamente diseñados. Los modelos se juzgan apropiados si sobreviven a este examen riguroso y si, además, son estructuras con consistencia interna. Para nuestra discusión actual, el punto importante en estos modelos es que son básicamente modelos *matemáticos* puramente abstractos. En particular, la cuestión misma de la consistencia interna de un modelo científico requiere que el modelo esté especificado de forma precisa. La precisión requerida exige que el modelo sea matemático, pues de lo contrario no se puede estar seguro de que estas preguntas tengan respuestas bien definidas.

Si hay que atribuir algún tipo de «existencia» al propio modelo, entonces dicha existencia está localizada dentro del mundo platónico de las formas matemáticas. Por supuesto, se podría adoptar un punto de vista opuesto: que el modelo va a tener existencia solo dentro de nuestras diversas *mentes*, antes que aceptar que el mundo de Platón sea en algún sentido absoluto y «real». Pese a todo, se gana algo importante al considerar que las estructuras matemáticas poseen una realidad por sí mismas. En efecto, nuestras mentes individuales son notoriamente imprecisas, poco fiables e inconsistentes en sus juicios. La precisión, fiabilidad y consistencia que requieren nuestras teorías científicas exige algo más allá de cualquiera de nuestras mentes individuales (poco dignas de confianza). En las matemáticas encontramos una solidez mucho mayor que la que puede localizarse en cualquier mente concreta. ¿No apunta esto a algo exterior a nosotros mismos, con una realidad que está más allá de lo que cada individuo puede alcanzar?

De todas formas, aún se podría adoptar el punto de vista alternativo según el cual el mundo matemático no tiene existencia independiente y consiste meramente en algunas ideas que han sido destiladas de nuestras diversas mentes, que se han mostrado totalmente dignas de confianza y en las que todos coinciden. Pero incluso este punto de vista parece dejarnos muy lejos de lo que se necesita. ¿Queremos decir «en las que todos coinciden», por ejemplo, o «en las que coinciden quienes están en su sano juicio», o «en las que coinciden todos aquellos que tienen un doctorado en matemáticas (poco frecuente en la época de Platón) y que tienen derecho a aventurar una opinión autorizada»? Parece que aquí hay un peligro de circularidad; pues juzgar si alguien

está o no «en su sano juicio» requiere algún patrón externo. Lo mismo sucede con el significado de «autorizada», a menos que se adoptara algún canon de naturaleza acientífica tal como la «opinión de la mayoría» (y debería quedar claro que la opinión de la mayoría, por importante que pueda ser para un gobierno democrático, no debería ser utilizada en modo alguno como criterio de aceptabilidad científica). Las propias matemáticas parecen tener realmente una solidez que va mucho más allá de lo que cualquier matemático individual es capaz de percibir. Aquellos que trabajan en esta disciplina, ya estén implicados activamente en la investigación matemática o bien utilicen resultados que han sido obtenidos por otros, sienten normalmente que son meros exploradores de un mundo que está mucho más allá de ellos mismos, un mundo que posee una objetividad que trasciende la mera opinión, ya sea dicha opinión la suya propia o la propuesta de otros, con independencia de cuán expertos pudieran ser esos otros.

Quizá pueda ayudar el que yo plantee de una forma diferente el caso de la existencia real del mundo platónico. Lo que entiendo por esta «existencia» es tan solo la objetividad de la verdad matemática. La existencia platónica, tal como yo la veo, se refiere a la existencia de un canon externo objetivo que no depende de nuestras opiniones individuales ni de nuestra cultura concreta. Tal «existencia» podría también referirse a objetos distintos de las matemáticas, tales como la moralidad o la estética (cf. §1.5), pero aquí estoy interesado solo en la objetividad matemática, que parece ser una cuestión mucho más clara.

Permítaseme ilustrar este punto considerando un ejemplo famoso de verdad matemática, y relacionarlo con la cuestión de la «objetividad». En 1637, Pierre de Fermat hizo su famosa afirmación conocida hoy día como el «último teorema de Fermat» (que ninguna potencia n-ésima[3] positiva de un número entero puede ser la suma de otras dos potencias n-ésimas positivas si n es un número entero mayor que 2), que él escribió en un margen de su copia de la *Arithmetica*, libro escrito en el siglo III por el matemático griego Diofanto. En este margen, Fermat anotó también: «He encontrado una demostración de esto verdaderamente maravillosa, que no cabe en este estrecho margen». La afirmación matemática de Fermat quedó sin confirmar durante más de trescientos cincuenta años, pese a que aunó los esfuerzos de muchos

matemáticos destacados. Finalmente, Andrew Wiles publicó una demostración en 1995 (que se basaba en el trabajo previo de otros matemáticos), y esta demostración ha sido ahora aceptada como un argumento válido por la comunidad matemática.

Ahora bien, ¿aceptamos el punto de vista de que la afirmación de Fermat fue siempre verdadera, mucho antes de que este la hiciera en realidad, o es su validez una cuestión puramente cultural, dependiente de cuáles pudieran ser los cánones subjetivos de la comunidad de matemáticos humanos? Supongamos que la validez de la afirmación de Fermat es, de hecho, una cuestión subjetiva. Entonces no sería un absurdo que un matemático X hubiera dado con un contraejemplo real y concreto de la afirmación de Fermat, siempre que X lo hubiera hecho antes de 1995.[4] En tal caso, la comunidad matemática tendría que aceptar la corrección del contraejemplo de X. A partir de entonces, cualquier esfuerzo por parte de Wiles de demostrar la afirmación de Fermat tendría que ser infructuoso, por la sencilla razón de que X había obtenido su argumento primero y, en vista de ello, ¡la afirmación de Fermat sería ahora falsa! Más aún, podríamos plantear la pregunta adicional acerca de si, de acuerdo con la corrección del contraejemplo que iba a dar X, el propio Fermat habría estado necesariamente equivocado al creer en la validez de su «demostración verdaderamente maravillosa», en el instante en que escribió su nota en el margen. En el punto de vista subjetivo de la verdad matemática hubiera podido darse el caso de que Fermat tuviera una demostración válida (que habría sido aceptada como tal por sus pares en la época, si él la hubiera revelado), ¡y que fue el secretismo de Fermat el que permitió la posibilidad de que X obtuviese más tarde un contraejemplo! Creo que prácticamente todos los matemáticos, con independencia de las actitudes que profesen hacia el «platonismo», considerarán que tales posibilidades son manifiestamente absurdas.

Por supuesto, aún podría darse el caso de que el argumento de Wiles contenga un error y que la afirmación de Fermat fuera en realidad falsa. O que pudiera haber un error fundamental en el argumento de Wiles, pero que la afirmación de Fermat sea en cualquier caso verdadera. O podría ser que el argumento de Wiles sea correcto en sus líneas esenciales aunque contenga «pasos no rigurosos» que no superarían el

canon de algunas reglas futuras de aceptabilidad matemática. Pero estas cuestiones no abordan el punto que estoy señalando aquí. La cuestión es la objetividad de la propia afirmación de Fermat, y no si la demostración (o la negación) particular de la misma que hiciera alguien podría resultar convincente para la comunidad matemática de cualquier época concreta.

Quizá habría que mencionar que, desde el punto de vista de la lógica matemática, la afirmación de Fermat es en realidad un enunciado matemático de un tipo particularmente simple,[5] cuya objetividad es especialmente evidente. Solo una pequeñísima minoría de matemáticos[6] consideraría que la verdad de tales afirmaciones es de algún modo «subjetiva» —aunque podría haber cierta subjetividad acerca de los tipos de argumentos que se considerarían convincentes—. Sin embargo, hay otros tipos de afirmaciones matemáticas cuya verdad podría considerarse plausiblemente como una «cuestión de opinión». Tal vez la más conocida de dichas afirmaciones sea el *axioma de elección*. No es importante, por el momento, que sepamos qué es el axioma de elección. (Lo describiré en §16.3.) Aquí se cita solo como ejemplo. Probablemente la mayoría de los matemáticos considerarán que el axioma de elección es «obviamente verdadero», mientras que otros pueden considerarlo una afirmación algo cuestionable que incluso podría ser falsa (y yo mismo me inclino, en cierta medida, hacia este segundo punto de vista). Otros aún podrían tomarlo como una afirmación cuya «verdad» es una mera cuestión de opinión o, más bien, como algo que puede tomarse de un modo o de otro, dependiendo de a qué sistemas de axiomas y reglas de inferencia (un «sistema formal»; véase §16.6) decida uno adherirse. Los matemáticos que defienden este último punto de vista (pero aceptan la objetividad de la verdad de enunciados matemáticos particularmente nítidos, como la afirmación de Fermat que he mencionado antes) serían platonistas relativamente débiles. Aquellos que se adhieren a la objetividad con respecto a la verdad del axioma de elección serían platonistas más fuertes.

Volveré al axioma de elección en §16.3, pues tiene cierta relevancia para las matemáticas subyacentes en el comportamiento del mundo físico, pese al hecho de que no se aborda mucho en la teoría física. Por el momento será mejor que no nos preocupemos demasiado por esta

cuestión. Si el axioma de elección puede ser dilucidado en un sentido u otro mediante alguna forma apropiada de razonamiento matemático incuestionable,[7] entonces su verdad es en realidad una cuestión totalmente objetiva, y o bien el axioma pertenece al mundo platónico o bien lo hace su negación, en el sentido en que estoy interpretando este «mundo platónico». Si, por el contrario, el axioma de elección es una simple cuestión de opinión o de decisión arbitraria, entonces el mundo platónico de las formas matemáticas absolutas no contiene axioma de elección ni su negación (aunque podría contener afirmaciones de la forma «tal y cual se sigue del axioma de elección», o «el axioma de elección es un teorema de acuerdo con las reglas de tal y cual sistema matemático»).

Los enunciados matemáticos que pueden pertenecer al mundo de Platón son precisamente aquellos que son objetivamente verdaderos. De hecho, yo consideraría que la objetividad matemática es realmente el objeto del platonismo matemático. Decir que una afirmación matemática tiene una existencia platónica es sencillamente decir que es verdadera en un sentido objetivo. Un comentario similar es aplicable a las *nociones* matemáticas —tales como el concepto del número 7, por ejemplo, o la regla para la multiplicación de números enteros, o la idea de que cierto conjunto contiene infinitos elementos—, todas las cuales tienen una existencia platónica porque son nociones objetivas. En mi opinión, la existencia platónica es simplemente una cuestión de objetividad y, en consecuencia, no debería verse como algo «místico» o «acientífico», pese a que así la consideran algunos.

No obstante, como sucede con el axioma de elección, las preguntas acerca de si debe considerarse o no que cierta propuesta concreta de una entidad matemática tiene una existencia objetiva pueden ser delicadas y a veces muy técnicas. Pese a ello, ciertamente no necesitamos ser matemáticos para apreciar la solidez general de muchos conceptos matemáticos. En la Fig. 1.2 he representado varias porciones pequeñas de esa famosa entidad matemática conocida como el *conjunto de Mandelbrot*. El conjunto tiene una estructura extraordinariamente complicada, pero no se debe a ningún diseño humano. Lo realmente notable es que esta estructura está definida por una regla matemática particularmente simple. Llegaremos a ella explícitamente en §4.5, pues

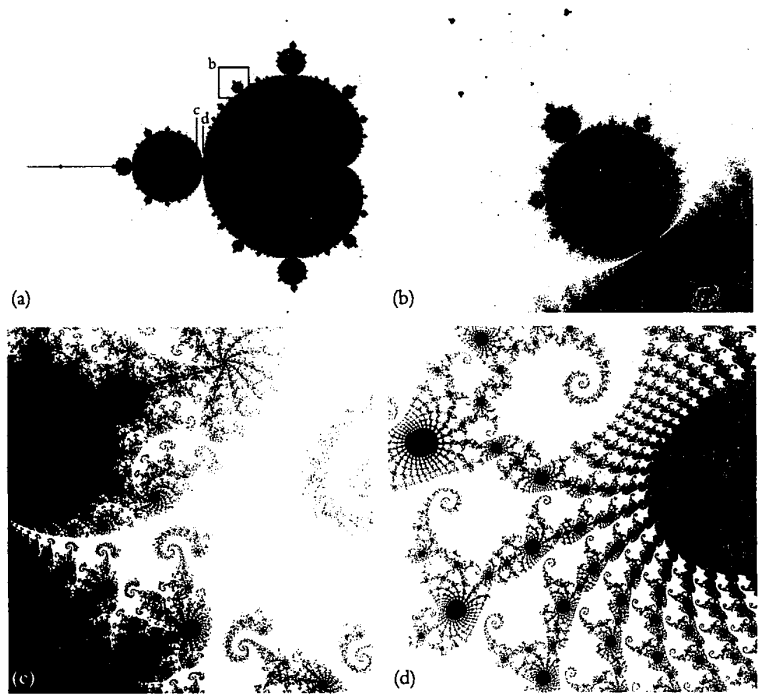


Fig. 1.2. (a) El conjunto de Mandelbrot. (b), (c) y (d) Algunos detalles que ilustran ampliaciones de las regiones correspondientemente marcadas en la Fig. 1.2a, aumentadas por factores lineales respectivos 11,6, 168,9 y 1.042.

nos distraeríamos de nuestros propósitos actuales si tratase ahora de ofrecer esta regla en detalle.

El punto que deseo señalar es que nadie, ni siquiera el propio Mandelbrot cuando vio por primera vez las increíbles complicaciones en los detalles finos del conjunto, tuvo ninguna preconcepción real de la extraordinaria riqueza del conjunto. El conjunto de Mandelbrot no fue invención de ninguna mente humana: sencillamente, está ahí de manera objetiva, en las propias matemáticas. Si tiene significado atribuir una existencia real al conjunto de Mandelbrot, entonces dicha existencia no está dentro de nuestras mentes, pues nadie puede abarcar por completo la inacabable variedad y la ilimitada complejidad del conjunto.

Y su existencia tampoco puede residir dentro de la multitud de representaciones gráficas impresas por un computador que empiezan a captar algo de su increíble sofisticación y detalle, pues, en el mejor de los casos, tales representaciones gráficas recogen tan solo una sombra de una aproximación al propio conjunto. Pese a todo, tiene una solidez que está más allá de cualquier duda, pues la misma estructura se revela —en todos sus detalles perceptibles, con finura cada vez mayor cuanto más de cerca se examina— independientemente del matemático o computador que la examine. Su existencia solo puede estar dentro del mundo platónico de las formas matemáticas.

Soy consciente de que aún habrá muchos lectores que encuentren difícil atribuir cualquier tipo de existencia real a las estructuras matemáticas. Rogaría a tales lectores que amplíen su idea de lo que la palabra «existencia» puede significar para ellos. Las formas matemáticas del mundo de Platón no tienen evidentemente el mismo tipo de existencia que los objetos físicos ordinarios tales como las mesas y las sillas. No tienen localización espacial; no existen en el tiempo. Hay que pensar en las nociones matemáticas objetivas como entidades intemporales, y no debe considerarse que nacieron en el instante en que fueron humanamente percibidas por primera vez. Las espirales concretas del conjunto de Mandelbrot que se muestran en las Figs. 1.2c o 1.2d no alcanzaron su existencia en el instante en que se vieron por primera vez en la pantalla o la impresora de un computador. Ni surgieron cuando la idea general que hay tras el conjunto de Mandelbrot fue propuesta por primera vez por un ser humano —no por Mandelbrot, tal como sucedió, sino por R. Brooks y J. P. Matelski, en 1981, o quizá antes—. Pues ciertamente ni Brooks ni Matelski, ni siquiera al principio el propio Mandelbrot, tenían ninguna concepción real de los diseños detallados y complicados que vemos en las Figs. 1.2c y 1.2d. Dichos diseños ya «existían» desde el principio de los tiempos, en el sentido potencial e intemporal con que necesariamente se iban a revelar en la forma exacta en que hoy los percibimos, con independencia de qué momento o qué lugar eligiera cualquier ser perceptivo para examinarlos.

1.4. Tres mundos y tres profundos misterios

Así pues, la existencia matemática es diferente no solo de la existencia física, sino también de una existencia que es atribuida por nuestras percepciones mentales. Pese a todo, hay una conexión misteriosa y profunda con cada una de esas otras dos formas de existencia: la física y la mental. En la Fig. 1.3 he mostrado de manera esquemática estas tres formas de existencia —la física, la mental y la matemático-platónica— como entidades que pertenecen a tres «mundos» separados, representados esquemáticamente como esferas. También están indicadas las misteriosas conexiones entre los mundos, y al dibujar el diagrama he impuesto al lector algunas de mis creencias, o prejuicios, acerca de tales misterios.

Con respecto al *primero* de esos misterios —que relaciona el mundo matemático-platónico con el mundo físico—, puede advertirse que estoy admitiendo que solo una pequeña parte del conjunto de las matemáticas tiene que tener relevancia para el funcionamiento del mundo físico. Sucede ciertamente que la gran mayoría de las actividades actuales de los matemáticos puros no tienen una conexión obvia con la física, ni con ninguna otra ciencia (cf. §34.9), aunque con frecuencia nos veamos sorprendidos por aplicaciones importantes e inesperadas. Análogamente, en relación con el *segundo* misterio, por el que la men-

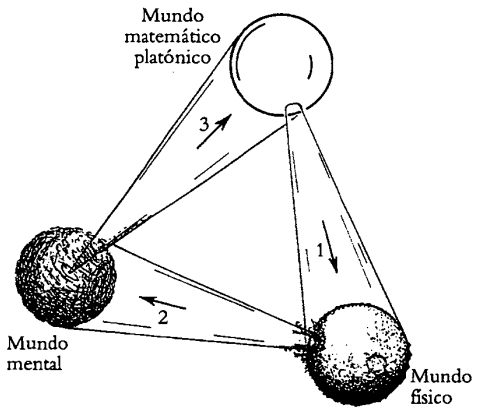


Fig. 1.3. Tres «mundos» —el matemático-platónico, el físico y el mental— y los tres profundos misterios en las conexiones entre ellos.

talidad entra en asociación con ciertas estructuras físicas (más concretamente, los cerebros humanos vivos, sanos y despiertos), no estoy insistiendo en que la mayoría de las estructuras físicas tengan que inducir mentalidad. Aunque el cerebro de un gato puede evocar realmente cualidades mentales, no estoy exigiendo lo mismo de una piedra. Por último, respecto al *tercer* misterio, ¡considero evidente que solo una pequeña fracción de nuestra actividad mental tiene que estar interesada en la verdad matemática absoluta! (Es más probable que estemos interesados en las múltiples irritaciones, placeres, preocupaciones, emociones y sensaciones por el estilo que llenan nuestras vidas cotidianas.) Estos tres hechos están representados en el pequeño tamaño de la base de la conexión de cada mundo con el siguiente, tomando los mundos del diagrama en el sentido de las agujas del reloj. Sin embargo, es en el hecho de englobar cada mundo entero dentro del ámbito de su conexión con el mundo que le precede donde estoy mostrando mis propios prejuicios.

Así pues, según la Fig. 1.3, todo el mundo físico se representa gobernado de acuerdo con leyes matemáticas. En capítulos posteriores veremos que hay una evidencia muy fuerte (aunque incompleta) que apoya esta opinión. Desde este punto de vista, todo lo que hay en el universo físico está realmente gobernado en todos sus detalles por principios matemáticos, quizá por ecuaciones, tales como las que trataremos en los capítulos que siguen, o quizá por algunas nociones matemáticas futuras fundamentalmente diferentes de aquellas que hoy etiquetamos con el término «ecuaciones». Si esto es así, entonces incluso nuestras propias acciones físicas estarían enteramente sujetas a semejante control matemático último, donde «control» podría admitir todavía cierto comportamiento aleatorio gobernado por principios probabilistas estrictos.

Muchas personas se sienten incómodas con este tipo de ideas, y debo confesar que a mí también me producen cierta desazón. De todas formas, mis prejuicios personales están realmente a favor de un punto de vista de este carácter general, puesto que es difícil ver cómo podría trazarse una línea que separe las acciones físicas bajo control matemático de aquellas que pudieran estar más allá de él. A mi modo de ver, la desazón que muchos lectores puedan compartir conmigo acerca de

esta cuestión surge en parte de una noción muy limitada de lo que pudiera entrañar el «control matemático». Parte del objetivo de este libro es señalar, y revelar al lector, algo de la extraordinaria riqueza, poder y belleza que pueden brotar una vez que se ha dado con las nociones matemáticas correctas.

Ya en el conjunto de Mandelbrot, tal como se ilustra en la Fig. 1.2, podemos empezar a vislumbrar el alcance y la belleza inherentes en tales objetos. Pero incluso estas estructuras habitan en un rincón muy limitado del conjunto de las matemáticas, donde el comportamiento está gobernado por un control computacional estricto. Más allá de este rincón, hay una increíble riqueza potencial. ¿Cómo me siento realmente al considerar la posibilidad de que todas mis acciones, y las de mis amigos, estén gobernadas, en última instancia, por principios matemáticos de este tipo? Puedo aceptarlo. De hecho, preferiría que estas acciones estuviesen controladas por algo que residiera en algún aspecto semejante del fabuloso mundo matemático de Platón a que estuvieran sujetas al tipo de motivos primarios simples, tales como la búsqueda del placer, la codicia personal o la violencia agresiva, que muchos argumentarán que son las consecuencias de una posición estrictamente científica.

Pese a todo, imagino que muchos lectores seguirán teniendo dificultades para aceptar que tales acciones en el universo puedan estar enteramente sujetas a leyes matemáticas. Análogamente, muchos podrán poner objeciones a otros dos de mis prejuicios que están implícitos en la Fig. 1.3. Podrían pensar, por ejemplo, que estoy adoptando una actitud científica demasiado fría al dibujar mi diagrama de una forma que implica que toda mentalidad tiene sus raíces en la fisicidad. Esto es en realidad un prejuicio, pues aunque es cierto que no tenemos evidencia científica razonable de la existencia de «mentes» que no tengan una base física, no podemos estar completamente seguros de ello. Más aún, muchas personas con convicciones religiosas defenderán con vehemencia la posibilidad de mentes independientes de lo físico, y podrían apelar a lo que ellos consideran evidencia poderosa de un tipo diferente de la que se revela por la ciencia ordinaria.

Otro de mis prejuicios se refleja en el hecho de que en la Fig. 1.3 he representado todo el mundo platónico dentro del ámbito de la

mentalidad. Con esto pretendo indicar que, al menos en principio, no hay verdades matemáticas que estén más allá del alcance de la razón. Por supuesto, hay enunciados matemáticos (incluso simples sumas aritméticas) que son tan enormemente complicados que nadie podría tener la fortaleza mental para llevar a cabo el razonamiento necesario. Sin embargo, tales objetos estarían *potencialmente* dentro del alcance de la mentalidad (humana), y serían compatibles con el significado de la Fig. 1.3, tal y como he pretendido representar. En cualquier caso, uno debe considerar que podría haber otros enunciados matemáticos que están incluso fuera del alcance potencial de la razón, y estos violarían la pretensión que hay tras la Fig. 1.3. (Esta cuestión será considerada más extensamente en §16.6, donde se examinará su relación con el famoso teorema de la incompletitud de Gödel.)[8]

En la Fig. 1.4, y como concesión a aquellos que no comparten todos mis prejuicios personales sobre estas cuestiones, he vuelto a dibujar las conexiones entre los tres mundos para admitir las tres posibles violaciones de mis prejuicios. En consecuencia, ahora se tiene en cuenta la posibilidad de acción física más allá del alcance del control matemático. El diagrama admite también la creencia de que pudiera haber mentalidad que no estuviera enraizada en estructuras físicas. Finalmente, permite la existencia de enunciados matemáticos verdaderos cuya verdad es en principio inaccesible mediante la razón y la intuición.

Esta imagen ampliada presenta otros misterios potenciales que van incluso más allá de aquellos que he admitido en mi imagen favorita del mundo, como se representa en la Fig. 1.3. En mi opinión, el punto de vista científico más firmemente organizado de la Fig. 1.3 tiene suficientes misterios. Estos misterios no desaparecen al pasar al esquema más relajado de la Fig. 1.4, pues sigue siendo un profundo enigma por qué tendrían que aplicarse las leyes matemáticas al mundo físico con tan extraordinaria precisión. (Vislumbraremos algo de la extraordinaria exactitud de las teorías físicas básicas en §19.8, §26.7 y §27.13.) Además, no es solo la precisión, sino también la sofisticación sutil y la belleza matemática de estas acertadas teorías lo que es profundamente misterioso. Hay asimismo un profundo e indudable misterio en cómo puede llegar a suceder que la materia física adecuadamente organizada —y aquí me refiero en concreto a cerebros humanos (o animales)

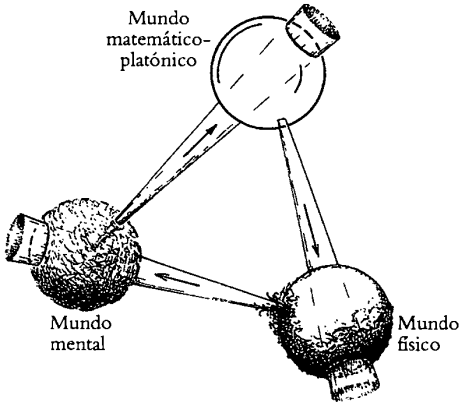


Fig. 1.4. Un nuevo dibujo de la Fig. 1.3 en el que se admiten violaciones de tres de los prejuicios del autor.

vivos— pueda evocar de algún modo la cualidad mental del conocimiento consciente. Por último, hay también un misterio en cómo percibimos la verdad matemática. No se trata solamente de que nuestros cerebros estén programados para «calcular» de manera fiable. Hay algo mucho más profundo que eso en las intuiciones que incluso los más humildes de entre nosotros tenemos cuando apreciamos, por ejemplo, los significados reales de los términos «cero», «uno», «dos», «tres», «cuatro», etc.[9]

Algunas de las cuestiones que surgen en conexión con este tercer misterio serán objeto de nuestro interés en el capítulo siguiente (y más explícitamente en §§16.5,6) en relación con la noción de *demostración matemática*. Pero el impulso principal de este libro tiene que ver con el primero de estos misterios: la notable relación entre las matemáticas y el comportamiento real del mundo físico. No se puede alcanzar una apreciación adecuada del extraordinario poder de la ciencia moderna sin al menos cierta familiaridad con estas ideas matemáticas. Sin duda, muchos lectores pueden asustarse ante la perspectiva de tener que entender semejantes matemáticas para llegar a esta apreciación. Pese a todo, soy optimista, y creo que quizá se darán cuenta de que estas cosas no son tan terribles como ellos temen. Más aún, espero poder persuadir a muchos lectores de que, pese a lo que hayan podido creer previamente, ¡las matemáticas pueden ser divertidas!

Aquí no me interesaré especialmente por el segundo de los miste-

rios mostrados en las Figs. 1.3 y 1.4, a saber, la cuestión de cómo la mentalidad —más en concreto, el conocimiento consciente— puede darse en asociación con estructuras físicas apropiadas (aunque tocaré esta profunda cuestión en §34.7). Tendremos ocupación más que suficiente en la exploración del universo físico y sus leyes matemáticas asociadas. Además, las cuestiones concernientes a la mentalidad son profundamente controvertidas, y si nos concentráramos en ellas nos distraería del objetivo de este libro. Sin embargo, quizá no esté de más hacer algún comentario. En mi opinión, se trata de que hay pocas posibilidades de que podamos tener una profunda comprensión de la naturaleza de la mente sin que antes aprendamos mucho más sobre las bases mismas de la realidad física. Como quedará claro en las discusiones que presentaré en capítulos posteriores, creo que se requieren revoluciones importantes en nuestra comprensión física. Hasta que no se hayan producido tales revoluciones, es muy optimista esperar que puedan hacerse demasiados progresos reales en la comprensión de la naturaleza real de los procesos mentales.[10]

1.5. Lo bueno, lo verdadero y lo bello

En relación con esto, hay otra serie de cuestiones planteadas por las Figs. 1.3 y 1.4. He tomado la noción de Platón de un «mundo de formas ideales» solo en el sentido limitado de formas matemáticas. Las matemáticas se interesan crucialmente en el ideal concreto de *verdad*. El propio Platón habría insistido en que hay otros dos ideales fundamentales y absolutos, a saber, los de lo *bello* y lo *bueno*. No me niego ni mucho menos a admitir la existencia de tales ideales y a permitir que se amplíe el mundo platónico para contener absolutos de esta naturaleza.

De hecho, más adelante encontraremos algunas notables interrelaciones entre verdad y belleza que iluminan y confunden a la vez las cuestiones del descubrimiento y la aceptación de las teorías físicas (véase §§34.2, 3,9 en particular; véase también la Fig. 34.1). Más aún, aparte del indudable (aunque a menudo ambiguo) papel de la belleza en las matemáticas subyacentes en las acciones del mundo físico, los criterios

estéticos son fundamentales para el desarrollo de ideas matemáticas por sí mismas, al aportar tanto el impulso hacia el descubrimiento como una poderosa guía a la verdad. Incluso conjeturaría que un elemento importante en la convicción común que tienen los matemáticos en que un mundo platónico externo tiene una existencia independiente de nosotros mismos procede de la extraordinaria e inesperada belleza oculta que tan a menudo revelan las ideas mismas.

De relevancia menos obvia aquí —pero de evidente importancia en un contexto más amplio— es la cuestión de un ideal absoluto de moralidad: ¿qué es bueno y qué es malo, y cómo perciben nuestras mentes dichos valores? La moralidad tiene una profunda conexión con el mundo mental, puesto que está íntimamente relacionada con los valores asignados por seres conscientes y, lo que es más importante, con la presencia misma de la propia consciencia. Es difícil ver qué podría significar la moralidad en ausencia de seres conscientes. A medida que progresan la ciencia y la tecnología, se hace cada vez más relevante una comprensión de las circunstancias físicas bajo las que se manifiesta la mentalidad. Creo que, en la cultura tecnológica de hoy día, es más importante que nunca que las cuestiones científicas no se separen de sus implicaciones morales. Pero estas cuestiones nos alejarían demasiado del alcance inmediato de este libro. Necesitamos abordar la cuestión de separar lo verdadero de lo falso antes de que podamos intentar de forma adecuada una aplicación de tal comprensión a separar el bien del mal.

Por último, hay otro misterio concerniente a la Fig. 1.3 que he dejado para el final. He dibujado deliberadamente la figura para ilustrar una paradoja. ¿Cómo es posible que, de acuerdo con mis prejuicios, cada mundo parezca englobar al siguiente en su totalidad? No creo que esto sea una razón para abandonar mis prejuicios, sino meramente para demostrar la presencia de un misterio aún más profundo que trasciende a aquellos que he señalado antes. Quizá haya un sentido en el que los tres mundos no sean en absoluto independientes, sino que meramente reflejen, individualmente, aspectos de una verdad más profunda sobre el mundo como un todo de la que tenemos muy poca idea en el momento presente. Tenemos un largo camino que recorrer antes de que tales cuestiones puedan ser iluminadas adecuadamente.

Me he permitido alejarme demasiado de las cuestiones centrales que nos interesarán en este libro. El objetivo principal de esta sección ha sido el de acentuar la importancia capital que tienen las matemáticas en la ciencia, tanto antigua como moderna. Echemos ahora una ojeada al mundo de Platón, al menos a una parte relativamente pequeña pero importante de dicho mundo, de especial relevancia para la naturaleza de la realidad física.

Notas

Sección 1.2

1.1. Por desgracia, no se conoce casi nada fiable sobre Pitágoras, su vida, sus seguidores o su trabajo, aparte de su existencia misma y el reconocimiento por parte de Pitágoras del papel de las razones simples en la armonía musical. Véase Burkert (1972). Pese a todo, muchas cosas de gran importancia se atribuyen habitualmente a los pitagóricos. En consecuencia, utilizaré el término «pitagórico» solo como una etiqueta, sin ninguna pretensión de exactitud histórica.

1.2. Esta es la «escala diatónica» pura en la que las frecuencias (inversamente proporcionales a las longitudes de los elementos vibrantes) están en las razones 24 : 27 : 30 : 32 : 36 : 40 : 45 : 48, que presentan muchos casos de razones simples, que subyacen a las armonías que resultan agradables al oído. Las «teclas blancas» de un piano moderno están afinadas (siguiendo un compromiso entre la pureza pitagórica de la armonía y la facilidad de los cambios de clave) como aproximaciones a estas razones pitagóricas, según la escala *uniformemente temperada*, con frecuencias relativas $1 : \alpha^2 : \alpha^4 : \alpha^5 : \alpha^7 : \alpha^9 : \alpha^{11} : \alpha^{12}$, donde $\alpha = \sqrt[12]{2} = 1{,}05946\ldots$ (Nota: α^5 significa la quinta potencia de α, i. e. $\alpha \times \alpha \times \alpha \times \alpha \times \alpha$. La cantidad $\sqrt[12]{2}$ es la raíz duodécima de 2, que es el número cuya duodécima potencia es 2, i. e. $2^{1/12}$, de modo que $\alpha^{12} = 2$. Véanse la nota 1.3 y §5.2.)

Sección 1.3

1.3. Recuérdese de la nota 1.2 que la potencia n-ésima de un número es dicho número multiplicado por sí mismo n veces. Así, la tercera potencia de 5 es 125, y se escribe $5^3 = 125$; la cuarta potencia de 3 es 81, escrito $3^4 = 81$; etc.

1.4. De hecho, mientras Wiles estaba tratando de corregir una «laguna» en su demostración del último teorema de Fermat que se había hecho evidente tras su presentación inicial en Cambridge en junio de 1993, se extendió por la comunidad matemática el rumor de que el matemático Noam Elkies había encontrado un contraejemplo de la afirmación de Fermat. Previamente, en 1988, Elkies había hallado un contraejemplo de la conjetura de Euler —que no hay soluciones enteras de la ecuación $x^4 + y^4 + z^4 = w^4$—, demostrando con ello que era falsa. No era inverosímil, por consiguiente, que él hubiera demostrado que la afirmación también fuera falsa. Sin embargo, el correo electrónico que inició el rumor tenía fecha de 1 de abril y se descubrió que era una broma de Henri Darmon; véase Singh (1997), p. 293.*

1.5. Técnicamente es una Π_1-sentencia; véase §16.6.

1.6. Me doy cuenta de que, en cierto sentido, estoy cayendo en mi propia trampa al hacer una afirmación semejante. No se trata en realidad de si los matemáticos que adoptan un punto de vista tan extremadamente subjetivo constituyen una pequeñísima minoría o no (y la verdad es que no he realizado una encuesta fiable entre los matemáticos sobre este punto); de lo que se trata es de si una posición tan extrema debe tomarse realmente en serio. Dejo esto a juicio del lector.

1.7. Quizá algunos lectores conozcan los resultados de Gödel y Cohen, según los cuales el axioma de elección es independiente de los axiomas más básicos de la teoría de conjuntos estándar (el sistema axiomático de Zermelo-Frankel). Debería quedar claro que el argumento de Gödel-Cohen no establece por sí mismo que el axioma de elección nunca será dilucidado en un sentido o en otro. Este punto es resaltado, por ejemplo, en la sección final del libro de Paul Cohen (Cohen, 1966, cap. 14, §13), salvo que Cohen está más interesado de forma explícita en la *hipótesis del continuo* que en el axioma de elección; véase §16.5.

Sección 1.4

1.8. Quizá se da aquí la ironía de que un antiplatónico hecho y derecho, que crea que las matemáticas están «todas en la mente», también debe creer —así parece— que no hay enunciados matemáticos verdaderos que estén en principio más allá de la razón. Por ejemplo, si el último teorema de Fermat hubiera sido inaccesible (en principio) a la razón, entonces

* El día 1 de abril es en el mundo anglosajón el equivalente al Día de los Inocentes. *(N. del T.)*

esta visión antiplatónica no admitiría la validez de su verdad ni de su falsedad, ya que tal validez solamente viene del acto mental de percibir una demostración o una refutación.

1.9. Véase, por ejemplo, Penrose (1997b).

1.10. Mis propias ideas sobre el tipo de cambio que será necesario en nuestra visión del mundo física para que pueda acomodarse la mentalidad consciente se exponen en Penrose (1989, 1994, 1997a, 1997b).

2

Un teorema antiguo y una pregunta moderna

2.1. El teorema de Pitágoras

Consideremos la cuestión de la geometría. ¿Cuáles son, realmente, los diferentes «tipos de geometría» a los que he aludido en el capítulo anterior? Para abordar esta cuestión, volveremos a nuestro encuentro con Pitágoras y consideraremos el famoso teorema que lleva su nombre:[1] para cualquier triángulo rectángulo, el cuadrado de la longitud de la hipotenusa (el lado opuesto al ángulo recto) es igual a la suma de los cuadrados de las longitudes de los otros dos lados (Fig. 2.1). ¿Qué razones tenemos para creer que esta afirmación es cierta? ¿Cómo podemos «demostrar» realmente el teorema de Pitágoras? Se conocen muchos argumentos. Quiero considerar dos de ellos, elegidos por su especial transparencia, que ponen el acento en puntos diferentes.

En cuanto al primero, consideremos la estructura que se ilustra en la Fig. 2.2. Está compuesta enteramente por cuadrados de dos tamaños diferentes. Puede considerarse «obvio» que esta estructura puede prolongarse indefinidamente y que el plano entero queda así recubierto de forma regular y repetitiva, sin huecos ni solapamientos, por cuadrados

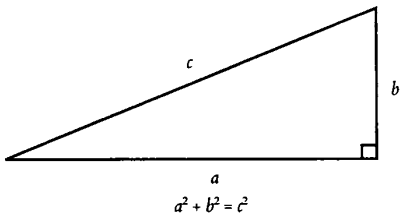


Fig. 2.1. El teorema de Pitágoras: para cualquier triángulo rectángulo, el cuadrado de la hipotenusa c es la suma de los cuadrados de los otros dos lados a y b.

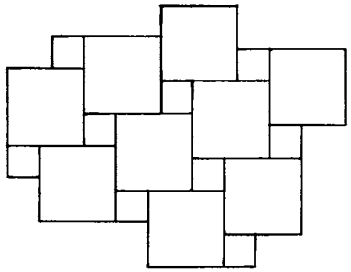


Fig. 2.2. Una teselación del plano por cuadrados de dos tamaños diferentes.

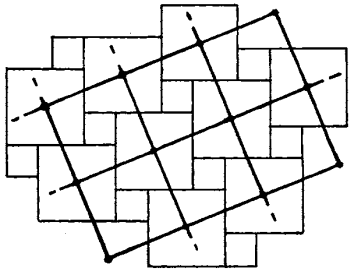


Fig. 2.3. Los centros de los (por ejemplo) cuadrados mayores forman los vértices de un retículo de cuadrados aún mayores, inclinados en un determinado ángulo.

de estos dos tamaños. La naturaleza repetitiva de esta estructura se hace manifiesta por el hecho de que si marcamos los centros de los cuadrados más grandes, dichos centros constituyen los vértices de otro sistema de cuadrados, de un tamaño algo mayor que cualquiera de los otros, pero inclinados en un cierto ángulo respecto a los originales (Fig. 2.3) y que por sí solos recubrirán el plano entero. Cada uno de estos cuadrados inclinados tiene exactamente las mismas marcas, de modo que las líneas interiores de estos cuadrados encajan para formar la estructura de dos cuadrados original. Lo mismo se aplicaría si, en lugar de tomar los centros de los cuadrados más grandes de entre los dos tipos de cuadrados de la estructura original, elegimos cualquier otro punto, junto con su conjunto de puntos correspondientes a lo largo de toda la estructura. La nueva estructura de cuadrados inclinados es exactamente la misma que antes pero desplazada sin rotación —i.e., mediante un movimiento que se conoce como una *traslación*—. Por simplicidad, podemos escoger ahora nuestro punto de partida en una de las esquinas de la estructura original (véase la Fig. 2.4).

Debería quedar claro que el área del cuadrado inclinado debe ser igual a la suma de las áreas de los dos triángulos más pequeños: de hecho, para cualquier punto de partida de los cuadrados inclinados, las piezas en que el cuadrado mayor queda subdividido por las líneas interiores pueden ser desplazadas sin rotación hasta que encajen para formar los dos cuadrados más pequeños (por ejemplo, la Fig. 2.5). Además, resulta evidente de la Fig. 2.4 que la longitud del lado del cuadrado inclinado grande es la hipotenusa de un triángulo rectángulo cuyos otros dos lados tienen longitudes iguales a los lados de los dos cuadrados más pequeños. Hemos establecido así el teorema de Pitágoras: el cuadrado de la hipotenusa es igual a la suma de los cuadrados de los otros dos lados.

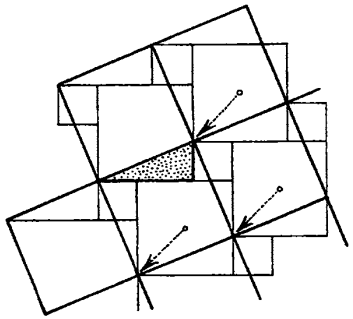


Fig. 2.4. El retículo de cuadrados inclinados puede desplazarse por una traslación, de modo que los vértices del retículo inclinado están sobre los vértices del retículo de dos cuadrados original, lo que muestra que el lado de un cuadrado inclinado es la hipotenusa de un triángulo rectángulo (sombreado) cuyos otros dos lados son los de los dos cuadrados originales.

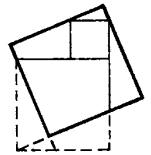


Fig. 2.5. Para cualquier punto de partida particular en el cuadrado inclinado, como el que se muestra, el cuadrado inclinado queda dividido en piezas que encajan para formar los dos cuadrados más pequeños.

El argumento anterior proporciona los elementos esenciales de una demostración sencilla de este teorema y, además, nos ofrece una «razón» para creer que el teorema tiene que ser verdadero, lo que quizá no fuera tan obvio en el caso de un argumento más formal construido mediante una sucesión de pasos lógicos sin un motivo claro. Sin embargo, habría que señalar que en este argumento han intervenido varias hipótesis implícitas. Una de ellas, y no la menos importante, es la hipótesis de que la estructura en apariencia obvia de cuadrados repetidos que se muestra en la Fig. 2.2 o incluso en la Fig. 2.6 es geométricamente posible —o incluso, y más críticamente, ¡que un *cuadrado* es algo geométricamente posible!—. Después de todo, ¿qué entendemos por

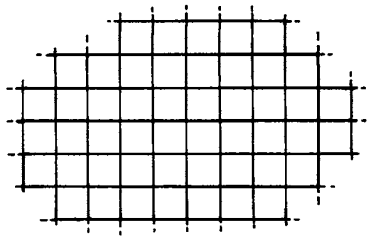


Fig. 2.6. El retículo familiar de cuadrados iguales. ¿Cómo sabemos que existe?

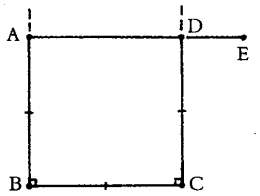


Fig. 2.7. Tratemos de construir un cuadrado. Tomemos ABC y BCD como ángulos rectos, con AB = BC = CD. ¿Se sigue de esto que DA es también igual a estas longitudes y que DAB y CDA son también ángulos rectos?

un «cuadrado»? Normalmente pensamos en un cuadrado como una figura plana cuyos lados son todos iguales y cuyos ángulos son todos rectos. ¿Qué es un ángulo recto? Bien, podemos imaginar dos líneas rectas que se cortan en un punto formando cuatro ángulos que son todos iguales. Cada uno de estos ángulos iguales es entonces un ángulo recto.

Tratemos ahora de construir un cuadrado. Tomemos tres segmentos de recta iguales AB, BC y CD, donde ABC y BCD son ángulos rectos, y D y A están en el mismo lado de la línea BC, como en la Fig. 2.7. Surge la pregunta: ¿tiene AD la misma longitud que los otros tres segmentos? Más aún: ¿son los ángulos DAB y CDA también ángulos rectos? Estos ángulos deberían ser iguales entre sí por la simetría izquierda-derecha de la figura, pero ¿son realmente ángulos rectos? El hecho de que lo sean solo parece obvio a causa de nuestra familiaridad con los cuadrados, o quizá porque podemos recordar de nuestros días escolares alguna proposición de Euclides que puede ser utilizada para deducir que los lados BA y CD tendrían que ser «paralelos» entre sí, y alguna proposición acerca de que cualquier recta «transversal» a un par de paralelas tiene que formar ángulos correspondientes iguales allí donde corta a las dos paralelas. De esto se sigue que el ángulo DAB tendría que ser igual al ángulo complementario de ADC (i.e., al ángulo EDC, en la Fig. 2.7, siendo ADE una recta) además de ser, como se ha señalado antes, igual al ángulo ADC. Un ángulo (ADC) solo puede ser

igual a su complementario (EDC) si es un ángulo recto. Debemos demostrar también que el lado AD tiene la misma longitud que BC, pero ahora esto se sigue también, por ejemplo, de las propiedades de las rectas transversales a las paralelas BA y CD. Por lo tanto, es cierto que a partir de un argumento euclídeo de este tipo podemos demostrar que realmente sí existen cuadrados hechos de ángulos rectos. Pero aquí se esconde una cuestión profunda.

2.2. Los postulados de Euclides

Al construir su noción de geometría, Euclides puso mucho cuidado en ver de qué hipótesis dependía su demostración.[2] En particular, tuvo cuidado en hacer una distinción entre ciertas afirmaciones llamadas *axiomas* —que se tomaban como verdades autoevidentes, y que básicamente eran definiciones de lo que él entendía por puntos, líneas, etc.— y los cinco *postulados*, que eran hipótesis cuya validez parecía menos segura, pero que parecían ser ciertas para la geometría de nuestro mundo. La última de estas hipótesis, conocida como el quinto postulado de Euclides, se consideraba menos obvia que las otras, y durante varios siglos se tuvo la sensación de que debería ser posible encontrar una forma de demostrarla a partir de los otros postulados más evidentes. El quinto postulado de Euclides se conoce habitualmente como *el postulado de las paralelas*, y yo seguiré esta práctica aquí.

Antes de discutir el postulado de las paralelas, vale la pena señalar la naturaleza de los otros cuatro postulados de Euclides. Los postulados conciernen a la geometría del plano (euclídeo), aunque Euclides también consideró el espacio tridimensional en sus obras posteriores. Los elementos básicos de su geometría plana son puntos, líneas rectas y círculos. Aquí consideraré que una «línea recta» (o simplemente una «recta») se extiende indefinida en ambas direcciones; si no es así, me referiré a un «segmento de recta». El *primer* postulado de Euclides afirma, en efecto, que existe un (único) segmento de recta que conecta dos puntos cualesquiera. Su *segundo* postulado afirma la prolongabilidad ilimitada de cualquier segmento de recta. Su *tercer* postulado afirma la existencia de un círculo con un centro cualquiera y con cualquier va-

lor para su radio. Finalmente, su *cuarto* postulado afirma la igualdad de todos los ángulos rectos.[3]

Desde una perspectiva moderna, algunos de estos postulados parecen un poco extraños, en particular el cuarto, pero debemos tener en cuenta el origen de las ideas que subyacen en la geometría de Euclides. Él estaba interesado básicamente en el movimiento de cuerpos rígidos idealizados y la noción de *congruencia* que se manifestaba cuando uno de tales cuerpos rígidos idealizados se movía hasta coincidir con otro. La igualdad entre un ángulo recto en un cuerpo y un ángulo en otro cuerpo tenía que ver con la posibilidad de mover un cuerpo de modo que las líneas que formaban su ángulo recto coincidieran con las líneas que formaban el ángulo recto en el otro. De hecho, el cuarto postulado está afirmando la isotropía y la homogeneidad del espacio, de modo que una figura en un lugar podría tener la «misma» (i.e., congruente) forma geométrica que una figura en otro lugar. Los postulados segundo y tercero expresan la idea de que el espacio es indefinidamente prolongable y sin «huecos» en su interior, mientras que el primero expresa la naturaleza básica de un segmento de línea recta. Aunque Euclides consideraba la geometría de un modo bastante diferente de como la consideramos hoy, sus cuatro primeros postulados recogen básicamente nuestra noción actual de un espacio métrico (bidimensional) con completa homogeneidad e isotropía, e infinito en extensión. De hecho, semejante imagen parece estar en estrecho acuerdo con la naturaleza espacial a muy gran escala del universo real, de acuerdo con la cosmología moderna, como veremos en §27.11 y §28.10.

¿Cuál es, entonces, la naturaleza del quinto postulado de Euclides, el postulado de las paralelas? Tal como Euclides formulaba esencialmente dicho postulado, este afirma que si dos segmentos de recta a y b en un plano cortan a otra línea recta c (de modo que c es lo que se denomina *transversal* a a y b) de tal forma que la suma de los ángulos interiores en el mismo lado de c es menor que dos ángulos rectos, entonces a y b, cuando se prolongan a suficiente distancia en ese lado de c, se cortarán en alguna parte (véase la Fig. 2.8a). Una forma equivalente de este postulado (a veces conocida como *axioma de Playfair*) afirma que, para cualquier línea recta y para cualquier punto que no esté en dicha línea, existe una única línea recta que pasa por dicho punto y

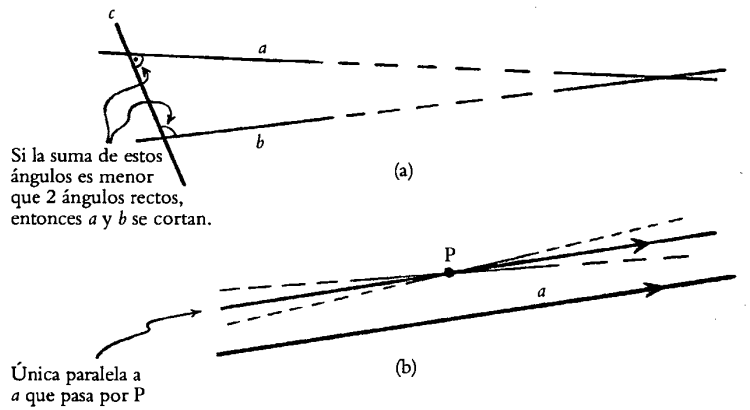


Fig. 2.8. (a) Postulado de las paralelas de Euclides. Las líneas a y b son trasversales a una tercera línea c, tal que los ángulos interiores donde a y b cortan a c suman menos que dos ángulos rectos. Entonces a y b (suponiendo que se prolongan lo suficiente) se cortarán en última instancia. (b) Axioma (equivalente) de Playfair: si a es una línea en un plano y P un punto del plano que no está en a, entonces hay solo una línea paralela a a que pasa por P en el plano.

es paralela a la primera recta (véase la Fig. 2.8b). Aquí, rectas «paralelas» serían dos líneas rectas coplanares que no se cortan (y recordemos que *mis* «rectas» son entidades completamente prolongadas, y no los «segmentos de línea recta» de Euclides.[2.1]

Una vez que tenemos el postulado de las paralelas, podemos proceder a establecer la propiedad necesaria para la existencia de un cuadrado. Si una recta transversal a un par de líneas rectas corta a estas de modo que los ángulos internos a un lado de la transversal suman dos ángulos rectos, entonces se puede demostrar que las líneas que forman el par son realmente paralelas. Además, se sigue inmediatamente que cualquier otra transversal al par tiene exactamente la misma propiedad angular. Esto es lo que necesitábamos para el argumento dado antes para la construcción de nuestro cuadrado. Vemos que es precisamente el postulado de las paralelas el que debemos utilizar para demostrar que

[2.1] Demuestre que si es válida la forma de Euclides del postulado de las paralelas, entonces debe seguirse la conclusión de Playfair de la unicidad de las paralelas.

nuestra construcción da realmente un cuadrado, con todos sus ángulos rectos y todos sus lados iguales. Sin el postulado de las paralelas no podemos establecer que existen realmente los cuadrados (en el sentido habitual de tener todos sus ángulos rectos).

Puede parecer que es una mera cuestión de pedantería matemática el preocuparse por cuáles son exactamente las hipótesis necesarias para proporcionar una «prueba rigurosa» de la existencia de un objeto tan obvio como un cuadrado. ¿Por qué deberíamos interesarnos en cuestiones tan pedantes, cuando un «cuadrado» es simplemente esa figura familiar que todos conocemos? Bien, pronto veremos que Euclides dio realmente pruebas de una extraordinaria perspicacia al preocuparse por tales cuestiones. La pedantería de Euclides está relacionada con una cuestión profunda que tiene mucho que decir sobre la geometría real del universo, y en más de una forma. En particular, no es en absoluto obvio que existan «cuadrados» físicos a una escala cosmológica en el universo real. Esta es una cuestión que se tiene que resolver mediante la observación, y por el momento la evidencia parece contradictoria (véanse §2.7 y §28.10).

2.3. La demostración del teorema de Pitágoras por áreas semejantes

En la próxima sección volveré a la trascendencia matemática que tiene el *no* dar por supuesto el postulado de las paralelas. Las cuestiones físicas relevantes serán examinadas en §18.4, §27.11, §28.10 y §34.4. Pero antes de discutir tales cuestiones, será instructivo dirigir nuestra atención a la otra demostración del teorema de Pitágoras que ya había prometido con anterioridad.

Una de las formas más sencillas de ver que la afirmación de Pitágoras es cierta en la geometría euclídea es considerar la configuración consistente en el triángulo rectángulo dado subdividido en dos triángulos más pequeños por una recta perpendicular a la hipotenusa trazada desde el ángulo recto (Fig. 2.9). Ahora tenemos tres triángulos: el original y los dos en que este ha sido subdividido. Evidentemente, el área del triángulo original es la suma de las áreas de los dos más pequeños.

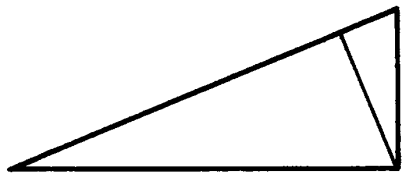


Fig. 2.9. Demostración del teorema de Pitágoras utilizando triángulos semejantes. Tómese un triángulo rectángulo y trácese una perpendicular desde su ángulo recto a su hipotenusa. Los dos triángulos en el que ahora está dividido el triángulo original tienen áreas cuya suma es la del triángulo original. Los tres triángulos son semejantes, de modo que sus áreas son proporcionales a los cuadrados de sus hipotenusas respectivas. De ello se sigue el teorema de Pitágoras.

Ahora es sencillo ver que estos tres triángulos son *semejantes* entre sí. Esto significa que todos ellos tienen la misma *forma* (aunque diferentes tamaños), i.e., se obtienen unos a partir de otros mediante una dilatación o contracción uniforme, junto con un movimiento rígido. Esto se sigue porque cada uno de los tres triángulos posee exactamente los mismos ángulos, en cierto orden. Cada uno de los dos triángulos más pequeños tiene un ángulo en común con el más grande, y uno de los ángulos de cada triángulo es un ángulo recto. El tercer ángulo también debe coincidir porque la suma de los ángulos de cualquier triángulo es siempre la misma. Ahora bien, una propiedad general de las figuras planas semejantes es que sus áreas son proporcionales a los cuadrados de sus correspondientes dimensiones lineales. Para cada triángulo, podemos considerar que esta dimensión lineal es el lado más largo, i.e., su hipotenusa. Notemos que la hipotenusa de cada uno de los triángulos más pequeños coincide con uno de los lados (no hipotenusa) del triángulo original. Así pues, se sigue al mismo tiempo (del hecho de que el área del triángulo original es la suma de las áreas de los otros dos) que el cuadrado de la hipotenusa del triángulo original es realmente la suma de los cuadrados de los otros dos lados: *¡el teorema de Pitágoras!*

Una vez más, tendremos que examinar algunas hipótesis concretas que intervienen en este argumento. Un ingrediente importante del argumento es el hecho de que la suma de los ángulos de un triángulo

tiene siempre el mismo valor. (Este valor de la suma es, por supuesto, 180°, pero Euclides hubiera dicho «dos ángulos rectos». La descripción matemática «natural» y más moderna consiste en decir que los ángulos de un rectángulo, en la geometría de Euclides, suman π. Esto supone utilizar radianes para la medida absoluta de un ángulo, donde el símbolo para el grado «°» cuenta como $\pi/180$, de modo que podemos escribir $180° = \pi$.) La demostración habitual se representa en la Fig. 2.10. Prolonguemos CA hasta E y tracemos una recta AD que pasa por A y es paralela a CB. Entonces (como se sigue del postulado de las paralelas), los ángulos EAD y ACB son iguales, y también DAB y CBA son iguales. Puesto que los ángulos EAD, DAB y BAC suman π (o 180°, o dos ángulos rectos), así deben hacerlo también los tres ángulos ACB, CBA y BAC del triángulo, tal como queríamos demostrar. Pero nótese que aquí se ha utilizado el postulado de las paralelas.

Esta demostración del teorema de Pitágoras también hace uso del hecho de que las áreas de figuras semejantes son proporcionales a los cuadrados de cualquier medida lineal de sus tamaños. (Aquí escogemos la hipotenusa de cada triángulo como representación de su medida lineal.) Este hecho no solo depende de la existencia misma de figuras semejantes de diferentes tamaños —que en el caso de los triángulos de la Fig. 2.9 se han establecido utilizando el postulado de las paralelas—, sino también de algunas cuestiones más sofisticadas que se relacionan con el modo en que definimos realmente el «área» para formas no rectangulares. Estas cuestiones generales se abordan en términos de procedimientos de paso al límite y no quiero entrar por el momento en esta clase de discusión. Nos llevará a cuestiones más profundas relacionadas con el tipo de números que se utilizan en geometría. Volveremos a la cuestión en §§3.1-3.

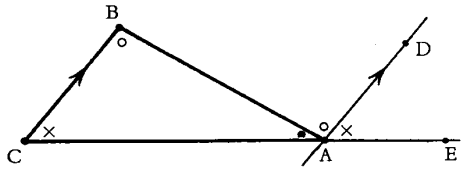


Fig. 2.10. Demostración de que la suma de los ángulos de un triángulo ABC es π (= = 180° = dos ángulos rectos). Prolónguese CA hasta E; dibújese AD paralela a CB. Se sigue del postulado de las paralelas que los ángulos EAD y ACB son iguales y que los ángulos DAB y CBA son iguales. Puesto que los ángulos EAD, DAB y BAC suman π, también lo hacen los ángulos ACB, CBA y BAC.

Un mensaje importante de la discusión de las secciones precedentes es que el teorema de Pitágoras parece depender del postulado de las paralelas. ¿Es así realmente? Supongamos que el postulado de las paralelas fuera falso. ¿Significa eso que el propio teorema de Pitágoras podría ser falso? ¿Tiene sentido semejante posibilidad? Tratemos de abordar la cuestión de lo que sucedería si se admite que el postulado de las paralelas sea considerado falso. Parecerá que estemos entrando en un mundo de fantasía, en el que la geometría que aprendimos en la escuela se pone patas arriba. Así es, pero también encontraremos que aquí hay un objetivo más profundo.

2.4. Geometría hiperbólica: imagen conforme

Echemos una ojeada a la imagen de la Fig. 2.11. Es una reproducción de uno de los grabados en madera de M. C. Escher, llamado *Límite circular I*. En realidad, nos proporciona una representación muy aproximada de un tipo de geometría —llamada geometría *hiperbólica* (o a veces *lobachevskiana*)— en la que el postulado de las paralelas es falso, el teorema de Pitágoras deja de ser válido y los ángulos de un triángulo no suman π. Además, para una figura de un tamaño dado no existe, en general, una figura semejante de tamaño mayor.

En la Fig. 2.11, Escher ha utilizado una representación concreta de la geometría hiperbólica en la que el «universo» entero del plano hiperbólico está «comprimido» en el interior de un círculo en un plano euclídeo ordinario. La circunferencia que limita al círculo representa el «infinito» para este universo hiperbólico. Podemos ver que, en la imagen de Escher, los peces parecen apretarse mucho a medida que se acercan a dicha frontera. Pero debemos considerar esto como una ilusión. Imagínese que usted es uno de los peces. Entonces, ya esté situado próximo al borde de la imagen de Escher o próximo a su centro, el

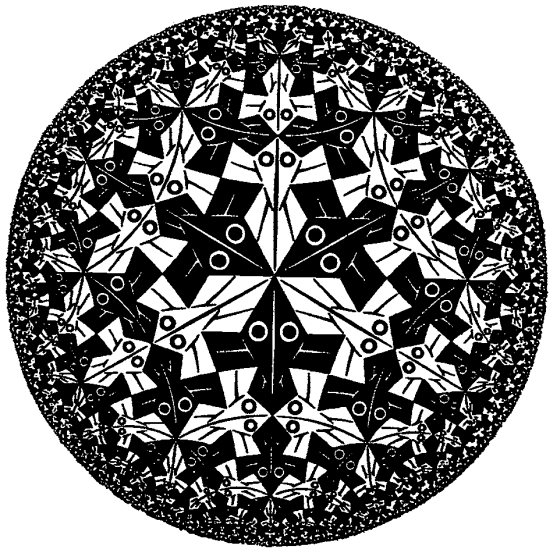


Fig. 2.11. *Límite circular I*, grabado en madera de M.C. Escher, que ilustra la representación conforme del plano hiperbólico.

universo (hiperbólico) entero tendrá la misma apariencia para usted. La noción de «distancia» en esta geometría no coincide con la del plano euclídeo en cuyos términos ha sido representada. Cuando miramos la imagen de Escher desde nuestra perspectiva euclídea, los peces próximos a la frontera parecen hacérsenos minúsculos. Pero desde su propia perspectiva «hiperbólica», los peces blancos o negros piensan que tienen exactamente la misma forma y tamaño que los que están próximos al centro. Más aún, aunque desde nuestra perspectiva euclídea exterior ellos parecen acercarse cada vez más a la propia frontera, desde su propia perspectiva hiperbólica dicha frontera siempre queda infinitamente lejos. Ni el círculo frontera ni nada del espacio «euclídeo» exterior tiene existencia para ellos. Su universo entero consiste en lo que para nosotros parece estar estrictamente dentro del círculo.

En términos más matemáticos, ¿cómo está construida esta imagen de la geometría hiperbólica? Consideremos un círculo cualquiera en el plano euclídeo. El conjunto de puntos interiores a este círculo va a re-

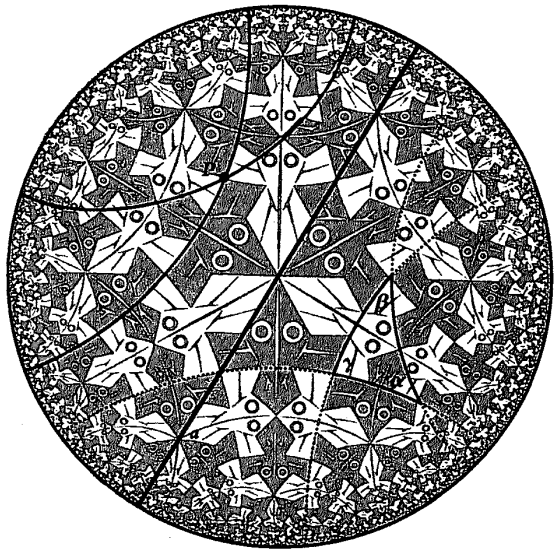


Fig. 2.12. La misma imagen de Escher que en la Fig. 2.11, pero con líneas rectas hiperbólicas (círculos euclídeos o líneas que cortan ortogonalmente al círculo frontera) y un triángulo hiperbólico. Los ángulos hiperbólicos coinciden con los euclídeos. Evidentemente se viola el postulado de las paralelas y los ángulos de un triángulo suman menos que π.

presentar el conjunto de puntos en el plano hiperbólico entero. Las líneas rectas se representan, de acuerdo con la geometría hiperbólica, como círculos euclídeos que cortan *ortogonalmente* —lo que significa a ángulos rectos— al círculo frontera. Ahora resulta que la noción hiperbólica de *ángulo* entre dos curvas cualesquiera, en su punto de intersección, es exactamente la misma que la medida euclídea del ángulo entre las dos curvas en el punto de intersección. Una representación de esta naturaleza se denomina representación *conforme*. Por esta razón, a la representación concreta de la geometría hiperbólica que utilizó Escher se le llama a veces el modelo *conforme* del plano hiperbólico. (También suele llamársele *disco de Poincaré*. La dudosa justificación histórica de esta terminología será examinada en §2.6.)

Estamos ahora en situación de ver si los ángulos de un triángulo, en

la geometría hiperbólica, suman o no π. Una rápida ojeada a la Fig. 2.12 nos lleva a creer que no lo hacen y que suman algo menos. De hecho, la suma de los ángulos de un triángulo en la geometría hiperbólica es siempre menor que π. Podríamos considerar esto como una característica desagradable de la geometría hiperbólica, ya que parece que no tenemos una respuesta «clara» para la suma de los ángulos de un triángulo. Sin embargo, cuando sumamos los ángulos de un triángulo hiperbólico sucede algo extraordinario y particularmente elegante: el déficit es siempre proporcional al área del triángulo. De forma más explícita, si los tres ángulos del triángulo son α, β y γ, entonces tenemos la fórmula (descubierta por Johann Heinrich Lambert, 1728-1777)

$$\pi - (\alpha + \beta + \gamma) = C\Delta,$$

donde Δ es el área del triángulo y C es una constante. Esta constante depende de las «unidades» que se escojan para medir longitudes y áreas. Siempre podemos fijar la escala de modo que $C = 1$. Es un hecho realmente notable que el área de un triángulo pueda expresarse de una forma tan simple en la geometría hiperbólica. En la geometría euclídea no hay modo de expresar el área de un triángulo simplemente en función de sus ángulos, y la expresión para el área de un triángulo en función de las longitudes de sus lados es bastante más complicada.

De hecho, aún no he acabado mi descripción de la geometría hiperbólica en términos de esta representación conforme, puesto que todavía no he descrito cómo va a definirse la *distancia* hiperbólica entre dos puntos (y convendría saber qué es «distancia» antes de que podamos hablar realmente de áreas). Permítaseme dar una expresión para la distancia hiperbólica entre dos puntos A y B interiores al círculo. Esta expresión es

$$\log \frac{\mathrm{QA} \cdot \mathrm{PB}}{\mathrm{QB} \cdot \mathrm{PA}},$$

donde P y Q son los puntos donde el círculo euclídeo (i.e., la línea recta hiperbólica) que pasa por A y B y es ortogonal al círculo frontera *corta* a dicho círculo frontera, y donde «QA», etc., se refieren a distancias euclídeas (véase la Fig. 2.13). Si se quiere incluir la «C» de la fórmula del área de Lambert (con $C \neq 1$), solo hay que multiplicar la expresión

anterior para la distancia por $C^{-1/2}$ (el recíproco de la raíz cuadrada de C).[4][2.2] Por razones que espero que se aclaren más adelante, me referiré a la cantidad $C^{-1/2}$ como el *pseudorradio* de la geometría.

Si expresiones matemáticas como la fórmula «log» anterior le parecen disuasorias, por favor, no se preocupe. Solo la doy para aquellos a quienes les gusta ver las cosas explícitamente. En cualquier caso, no voy a explicar por qué funciona la expresión (por ejemplo, por qué la distancia hiperbólica más corta entre dos puntos, así definida, se mide realmente a lo largo de una línea recta hiperbólica, o por qué las distancias a lo largo de una línea recta hiperbólica «se suman» de la forma adecuada).[2.3] También pido disculpas por el «log» (logaritmo), pero así son las cosas. De hecho, este es un *logaritmo natural* («log en base e») y tendré mucho que decir sobre ello en §§5.2,3. Veremos que los logaritmos son entidades realmente muy bellas y misteriosas, además de ser importantes en muchos contextos diferentes.

La geometría hiperbólica, con esta definición de distancia, resulta tener todas las propiedades de la geometría euclídea salvo aquellas que necesitan el postulado de las paralelas. Podemos construir triángulos, y otras figuras planas, de diferentes formas y tamaños, y podemos despla-

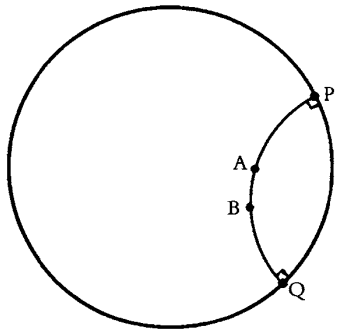


Fig. 2.13. En la representación conforme, la distancia hiperbólica entre A y B es log {QA.PB/QB.PA}, donde QA, etc., son distancias euclídeas, y P y Q son los puntos donde el círculo euclídeo que pasa por A y B y es ortogonal al círculo frontera (línea hiperbólica) corta a este círculo.

[2.2] ¿Puede ver una razón sencilla para ello?

[2.3] Trate de demostrar que, según esta fórmula, si A, B y C son tres puntos sucesivos en una línea recta hiperbólica, entonces las distancias hiperbólicas «AB», etc., satisfacen «AB» + «BC» = «AC». Puede suponer la propiedad general de los logaritmos, $\log(ab) = \log a + \log b$, como se describe en §§5.2,3.

zarlos «rígidamente» (manteniendo invariables sus formas y tamaños hiperbólicos) con la misma libertad con que podemos hacerlo en la geometría euclídea, de modo que, igual que en la geometría euclídea, surge una noción natural de cuándo son «congruentes» dos figuras (donde «congruente» significa «que pueden desplazarse rígidamente hasta que llegan a coincidir»). De hecho, todos los peces blancos del grabado de Escher son mutuamente congruentes, según esta geometría hiperbólica, y también lo son todos los peces negros.

2.5. Otras representaciones de la geometría hiperbólica

Por supuesto, no todos los peces blancos parecen de la misma forma y tamaño, pero ello se debe a que los estamos viendo desde una perspectiva euclídea en lugar de una hiperbólica. La imagen de Escher hace uso simplemente de una *representación* euclídea concreta de la geometría hiperbólica. La propia geometría hiperbólica es algo más abstracto que no depende de ninguna representación euclídea en particular. Sin embargo, tales representaciones son muy útiles para nosotros, pues proporcionan un modo de visualizar la geometría hiperbólica refiriéndola a algo que es más familiar y en apariencia más «concreto» para nosotros, a saber, la geometría euclídea. Además, tales representaciones dejan claro que la geometría hiperbólica es una estructura consistente, y que, en consecuencia, el postulado de las paralelas no puede demostrarse a partir de las otras leyes de la geometría euclídea.

Hay, de hecho, otras representaciones de la geometría hiperbólica en términos de geometría euclídea que son diferentes de la representación conforme que utilizó Escher. Una de ellas es la que se conoce como el modelo *proyectivo*. Aquí, el plano hiperbólico entero es de nuevo representado como el interior de un círculo en un plano euclídeo, pero las líneas rectas hiperbólicas están ahora representadas como líneas *rectas* euclídeas (y no como arcos de círculo). Hay, sin embargo, un precio a pagar por esta aparente simplificación, puesto que los ángulos hiperbólicos no son ahora iguales a los ángulos euclídeos, y muchos considerarían que este es un precio demasiado alto. Para aquellos lectores que estén interesados, la distancia hiperbólica entre dos puntos A y B en esta representación está dada por la fórmula (véase la Fig. 2.14).

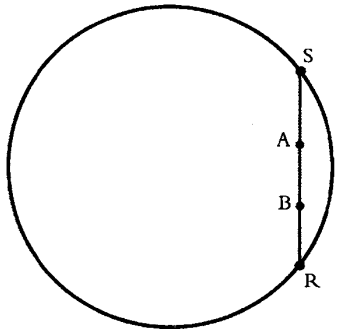


Fig. 2.14. En la representación proyectiva, la fórmula para la distancia hiperbólica es ahora 1/2 log {RA.SB/RB.SA}, donde R y S son las intersecciones de la línea recta euclídea (i.e., hiperbólica) AB con el círculo frontera.

$$\frac{1}{2}\log\frac{RA\cdot SB}{RB\cdot SA}$$

(tomando $C = 1$, que es casi igual que la expresión que teníamos antes para la representación conforme), donde R y S son las intersecciones de la línea recta prolongada AB con el círculo frontera. Esta representación de la geometría hiperbólica puede obtenerse a partir de la conforme por medio de una dilatación radial a partir del centro en una cantidad dada por

$$\frac{2R^2}{R^2 + r_c^2},$$

donde R es el radio del círculo frontera y r_c es la distancia euclídea de un punto al centro (euclídeo) del círculo frontera en la representación

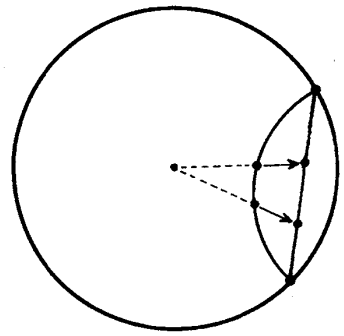


Fig. 2.15. Para pasar de la representación conforme a la proyectiva, se expande a partir del centro en un factor $2R^2/(R^2 + r_c^2)$, donde R es el radio del círculo frontera y r_c es la distancia euclídea al punto en la representación conforme.

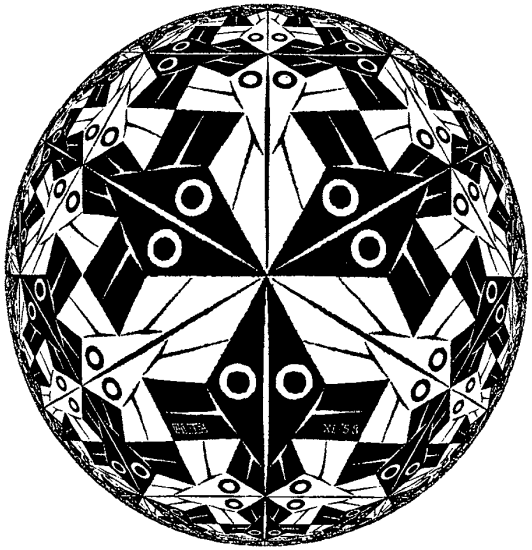


Fig. 2.16. La imagen de Escher de la Fig. 2.11 transformada desde la representación conforme a la proyectiva.

conforme (véase la Fig. 2.15).[2.4] En la Fig. 2.16 la imagen de Escher de la Fig. 2.11 ha sido transformada del modelo conforme al modelo proyectivo utilizando esta fórmula. (Pese a la pérdida de detalle, el arte preciso de Escher sigue siendo evidente.) Aunque resulta menos atractiva de esta manera, ¡presenta un nuevo punto de vista!

Existe una forma más directamente geométrica de relacionar las imágenes conforme y proyectiva, a través de otra representación aún más ingeniosa de la misma geometría. Las tres representaciones se deben al ingenioso geómetra italiano Eugenio Beltrami (1835-1900). Consideremos una esfera S, cuyo ecuador coincide con el círculo frontera de la representación proyectiva de la geometría hiperbólica que se ha dado antes. Ahora vamos a encontrar una representación de la geometría hiperbólica en el *hemisferio norte* S^+ de S, que llamaré representación *hemisférica*. Véase la Fig. 2.17. Para pasar de la representación proyectiva en el plano (considerado horizontal) a la nueva representación en la esfera, simplemente proyectamos verticalmente hacia arriba (Fig. 2.17a). Las líneas rectas en el plano, que representan líneas rectas hiperbólicas, se representan en S^+ por semicírculos que cortan ortogonalmente al ecuador. Ahora, para ir de la representación en S^+ a la representación conforme en el plano proyectamos desde el *polo sur* (Fig. 2.17b). Esta es la que se denomina *proyección estereográfica*, y más ade-

[2.4] Demuéstrelo. (*Sugerencia*: Si lo desea, puede utilizar la geometría de Beltrami, tal como se ilustra en la Fig. 2.17.)

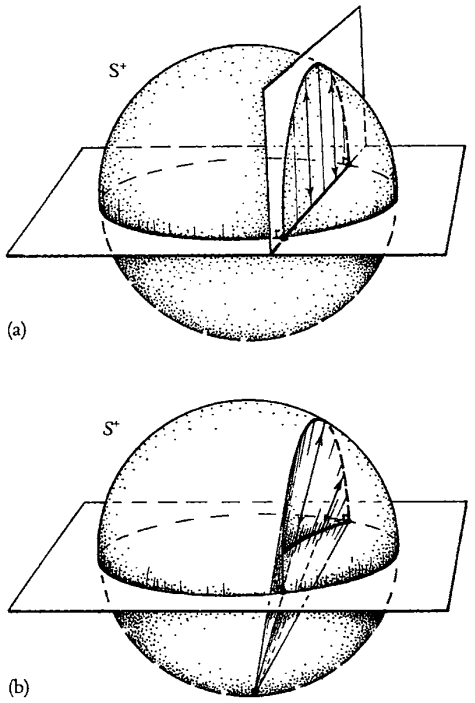


Fig. 2.17. Geometría de Beltrami, que relaciona tres de sus representaciones de la geometría hiperbólica. (a) La representación hemisférica (conforme en el *hemisferio norte* S^+) se proyecta verticalmente en la representación proyectiva sobre el disco ecuatorial. (b) La representación hemisférica proyecta estereográficamente desde el *polo sur* a la representación conforme sobre el disco ecuatorial.

lante desempeñará un papel importante en este libro (cf. §8.3, §18.4, §22.9 y §33.6). Dos propiedades importantes de la representación estereográfica a la que llegaremos en §8.3 son que es *conforme*, de modo que conserva los ángulos, y que hace corresponder círculos en la esfera a círculos (o, excepcionalmente, a líneas rectas) en el plano.[2.5],[2.6]

La existencia de varios modelos diferentes de geometría hiperbólica, expresados en términos de espacio euclídeo, sirve para acentuar el hecho de que en realidad estos son meramente «modelos euclídeos» de geometría hiperbólica y no debemos considerar que nos estén diciendo qué *es* realmente la geometría hiperbólica. La geometría hiperbólica tiene su propia «existencia platónica», igual que la tiene la geometría euclídea (véanse §1.3 y el prefacio). Ninguno de estos modelos debe tomarse como la «representación» correcta de la geometría hiperbólica, en detrimento de los otros. Las representaciones de la misma que hemos estado considerando son muy valiosas como ayudas para nuestra comprensión, pero solo porque el marco euclídeo es aquel al que estamos más acostumbrados. Para una criatura sintiente que haya crecido con una experiencia directa de la geometría hiperbólica (antes que de la euclídea), un modelo de geometría euclídea en términos hiperbólicos parecería la vía más natural. En §18.4 encontraremos aún otro modelo de geometría hiperbólica, esta vez en términos de la geometría minkowskiana de la relatividad especial.

Para terminar esta sección, volvamos a la cuestión de la existencia de cuadrados en la geometría hiperbólica. Aunque en la geometría hiperbólica no existen cuadrados cuyos ángulos sean ángulos rectos, sí existen «cuadrados» de un tipo más general cuyos ángulos son menores que los ángulos rectos. La forma más fácil de construir un cuadrado de este tipo es trazar dos líneas rectas que se cortan a ángulos rectos en un punto O. Nuestro «cuadrado» es ahora el cuadrilátero cuyos cuatro

[2.5] Suponiendo estas dos propiedades enunciadas de la proyección estereográfica, y siendo la representación conforme de la geometría hiperbólica la que se ha establecido en §2.4, demuestre que la representación hemisférica de Beltrami es conforme, con «líneas rectas» hiperbólicas como semicírculos verticales.

[2.6] ¿Puede ver cómo es posible demostrar estas dos propiedades? (*Sugerencia*: Demuestre, en el caso de círculos, que el cono de proyección es intersectado por dos planos de inclinación exactamente opuesta.)

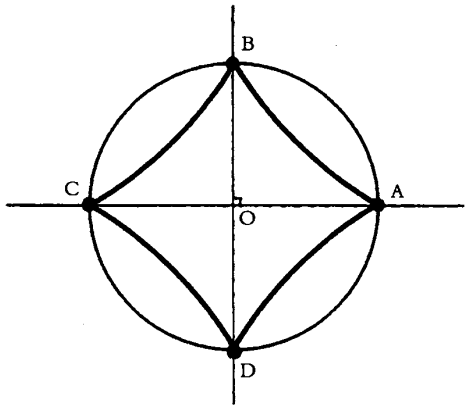


Fig. 2.18. Un «cuadrado» hiperbólico es un cuadrilátero hiperbólico, cuyos vértices son las intersecciones A, B, C, D (tomadas cíclicamente) de dos rectas hiperbólicas perpendiculares que pasan por un punto O con un círculo centrado en O. Debido a la simetría, los cuatro lados de ABCD, así como los cuatro ángulos son iguales. Estos ángulos, no son rectos, pero pueden ser iguales a cualquier ángulo positivo dado menor que $\frac{1}{2}\pi$.

vértices son las intersecciones A, B, C, D (tomadas cíclicamente) de estas dos líneas con un círculo con centro O. Véase la Fig. 2.18. Debido a la simetría de la figura, los cuatro lados del cuadrilátero resultante ABCD son iguales y los cuatro ángulos también deben ser iguales. Pero ¿son rectos estos ángulos? No en la geometría hiperbólica. De hecho, pueden ser cualquier ángulo (positivo) que queramos que sea menor que un ángulo recto, pero no igual a un ángulo recto. Cuanto más grande es el cuadrado (hiperbólico), es decir, cuanto mayor es el círculo en la construcción anterior, menores serán sus ángulos. En la Fig. 2.19a he representado un retículo de cuadrados hiperbólicos, utilizando el modelo conforme, donde hay cinco cuadrados en cada vértice (en lugar de los cuatro euclídeos), de modo que el ángulo es $\frac{2}{5}\pi$, o 72°. En la Fig. 2.19b he dibujado el mismo retículo utilizando el modelo proyectivo. Se verá que este no permite las modificaciones que serían necesarias para el retículo de dos cuadrados de la Fig. 2.2.[2.7]

[2.7] Vea si puede hacer algo similar, pero con pentágonos y cuadrados regulares hiperbólicos.

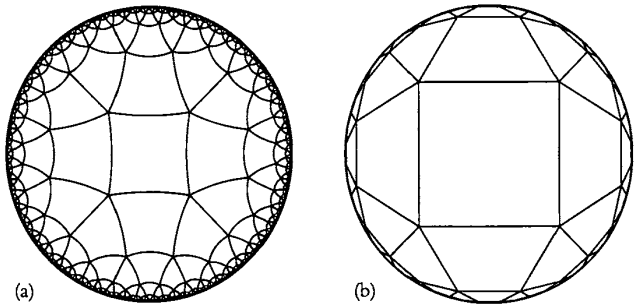


Fig. 2.19. Un retículo de cuadrados, en el espacio hiperbólico, en el que cinco cuadrados se encuentran en cada vértice, de modo que los ángulos del cuadrado son $\frac{2\pi}{5}$, o 72°. (a) Representación conforme. (b) Representación proyectiva.

2.6. Aspectos históricos de la geometría hiperbólica

Aquí es oportuno hacer algunos comentarios históricos concernientes al descubrimiento de la geometría hiperbólica. Durante los siglos siguientes a la publicación de los elementos de Euclides, aproximadamente en el año 300 a.C., varios matemáticos intentaron demostrar el quinto postulado a partir de los otros axiomas y postulados. Estos esfuerzos alcanzaron su culminación con el heroico trabajo del jesuita Girolamo Saccheri, en 1773. Podría parecer que el propio Saccheri debió de pensar que la obra de su vida era, en definitiva, un fracaso, pues se reducía a un intento insatisfactorio de *demostrar* el postulado de las paralelas mostrando que la hipótesis de que la suma de los ángulos de todo triángulo es menor que dos ángulos rectos lleva a una contradicción. Incapaz de hacer esto de forma lógica tras tremendos esfuerzos, concluyó, más bien débilmente:

> La hipótesis del ángulo agudo es absolutamente falsa, porque repugna a la naturaleza de la línea recta.[5]

La hipótesis del «ángulo agudo» afirma que las líneas *a* y *b* de la Fig. 2.8 a veces no se cortan. Es realmente viable, ¡y, de hecho, da la geometría hiperbólica!

¿Cómo puede ser que Saccheri descubriera efectivamente algo que él estaba tratando de demostrar que era imposible? La propuesta de Saccheri para demostrar el quinto postulado de Euclides consistía en formular la hipótesis de que el quinto postulado era falso y obtener entonces una contradicción a partir de dicha hipótesis. De este modo, él proponía hacer uso de uno de los principios más tradicionales y fructíferos que han sido propuestos en matemáticas —muy posiblemente introducido por primera vez por los pitagóricos— llamado *demostración por contradicción* (o *reductio ad absurdum*, para darle su nombre latino). Según este procedimiento, para probar que una afirmación es cierta se formula primero la hipótesis de que la afirmación en cuestión es *falsa*, y luego se argumenta que de ello se sigue una contradicción. Si se llega a encontrar tal contradicción, se deduce que la afirmación debe ser, después de todo, verdadera.[6] La demostración por contradicción proporciona un método muy potente de razonamiento en matemáticas, hoy aplicado con frecuencia. Aquí es apropiada una cita del distinguido matemático G. H. Hardy:

> La *reductio ad absurdum*, que tanto amaba Euclides, es una de las armas matemáticas más valiosas. Es un gambito mucho más fino que cualquier gambito de ajedrez: un jugador de ajedrez puede ofrecer el sacrificio de un peón o incluso una pieza, pero un matemático ofrece *el juego*.[7]

Veremos otros usos de este importante principio más adelante (véanse §3.1 y §§16.4,6).

Sin embargo, Saccheri fracasó en su intento de encontrar una contradicción. Por consiguiente, no pudo obtener una demostración del quinto postulado. Pero al esforzarse en ello, descubrió algo mucho más grande: una nueva geometría, diferente de la de Euclides —la geometría discutida en §§2.4,5 que ahora llamamos *geometría hiperbólica*—. A partir de la hipótesis de que el quinto postulado de Euclides era falso, obtuvo, en lugar de una contradicción real, un montón de teoremas de apariencia extraña y apenas creíble, pero muy interesantes. No obstante, por extraños que parecieran tales resultados, ninguno de ellos era realmente una contradicción. Como sabemos ahora, no había ninguna

posibilidad de que Saccheri encontrara de esta manera una contradicción genuina, por la sencilla razón de que la geometría hiperbólica *existe* realmente, en el sentido matemático de que existe una estructura semejante consistente. En la terminología de §1.3 la geometría hiperbólica habita en el mundo platónico de las formas matemáticas. (La cuestión de la realidad física de la geometría hiperbólica se tocará en §2.7 y §28.10.)

Poco tiempo después de Saccheri, el muy perspicaz matemático Johann Heinrich Lambert (1728-1777) obtuvo también numerosos y fascinantes resultados geométricos a partir de la hipótesis de que el quinto postulado de Euclides es falso, incluyendo el bello resultado mencionado en §2.4 que da el área de un triángulo hiperbólico en función de la suma de sus ángulos. Parece que Lambert pudo haberse formado la opinión, al menos en alguna etapa de su vida, de que realmente podía obtenerse una geometría consistente a partir de la negación del quinto postulado de Euclides. Al parecer, la razón tentativa de Lambert era que podía contemplar la posibilidad teórica de la geometría en una «esfera de radio imaginario», i.e., en una esfera cuyo «radio al cuadrado» es negativo. La fórmula de Lambert $\pi - (\alpha + \beta + \gamma) = C\Delta$ da el área, Δ, de un triángulo hiperbólico, donde α, β, γ son los ángulos del triángulo y C es una constante (siendo $-C$ lo que ahora llamaríamos la «curvatura gaussiana» del plano hiperbólico). Esta fórmula tiene básicamente la misma apariencia que una previamente conocida debida a Thomas Hariot (1560-1621), $\Delta = R^2(\alpha + \beta + \gamma - \pi)$, para el área Δ de un *triángulo esférico*, dibujado con arcos de círculo máximo[8] en una esfera de radio R (véase la Fig. 2.20).[2.8] Para recuperar la fórmula de Lambert, tenemos que poner

$$C = -\frac{1}{R^2}.$$

Pero, para dar el valor *positivo* de C, como sería necesario en la geometría hiperbólica, necesitamos que el radio de la esfera sea «imaginario» (i.e., que sea la raíz cuadrada de un número negativo). Nótese que el radio R está dado por la cantidad imaginaria $(-C)^{-1/2}$. Esto explica el término «pseudorradio», introducido en §2.4, para la cantidad real $C^{-1/2}$. De hecho, el procedimiento de Lambert está perfectamente justificado desde nuestra perspectiva más moderna (véanse el capítulo 4 y §18.4), y el hecho de haberlo previsto revela una gran intuición por su parte.

[2.8] Trate de demostrar esta fórmula del triángulo esférico, utilizando básicamente solo argumentos de simetría y el hecho de que el área total de la esfera es $4\pi R^2$. *Sugerencia*: Empiece encontrando el área de un segmento de una esfera acotado por dos arcos de círculo máximo que conectan un par de puntos antípodas de la esfera; luego corte y pegue y utilice argumentos de simetría. Tenga en cuenta la Fig. 2.20.

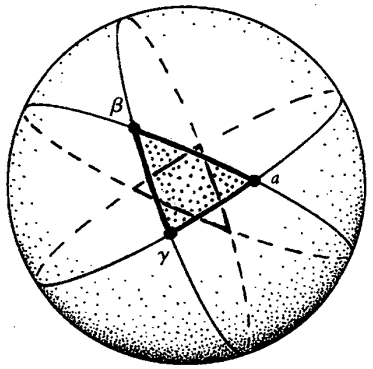


Fig. 2.20. La fórmula de Hariot para el área de un *triángulo esférico*, con ángulos α, β, γ es $\Delta = R^2(\alpha + \beta + \gamma - \pi)$. La fórmula de Lambert, para un triángulo hiperbólico, tiene $C = -1/R^2$.

Sin embargo, el punto de vista convencional (algo injusto, en mi opinión) niega a Lambert el honor de haber construido por primera vez una geometría no euclídea, y considera que (aproximadamente medio siglo más tarde) la primera persona que llegó a una aceptación clara de una geometría completamente consistente, distinta de la de Euclides, en la que el postulado de las paralelas es falso, fue el gran matemático Carl Friedrich Gauss. Al ser un hombre muy cauteloso, y temiendo la controversia que semejante revelación pudiera causar, Gauss no publicó sus hallazgos y se los reservó para sí.[9] Unos treinta años después de que Gauss hubiera empezado a trabajar en ello, la geometría hiperbólica fue redescubierta de forma independiente por otros, entre ellos el húngaro János Bolyai (en 1829) y, muy en especial, el geómetra ruso Nicolái Ivánovich Lobachevski hacia 1826 (de ahí que la geometría hiperbólica sea denominada con frecuencia geometría *lobachevskiana*).

Las realizaciones concretas proyectiva y conforme de la geometría hiperbólica que he descrito antes fueron encontradas por Eugenio Beltrami, y publicadas en 1868, junto con algunas otras elegantes repre-

sentaciones que incluyen la hemisférica mencionada en §2.5. No obstante, la representación conforme se conoce normalmente como el «modelo de Poincaré», porque el redescubrimiento de esta representación que hizo Poincaré en 1882 es mejor conocido que la obra original de Beltrami (básicamente debido al importante uso que hizo Poincaré de este modelo).[10] Análogamente, la representación proyectiva del pobre Beltrami se denomina a veces «representación de Klein». No es infrecuente en matemáticas que el nombre habitualmente asociado a un concepto matemático no sea el de su descubridor original. Al menos, en este caso, Poincaré sí *re*descubrió la representación conforme (como hizo Klein con la proyectiva en 1871). Hay otros ejemplos en matemáticas en los que el (los) matemático(s) cuyo nombre (o nombres) está asociado a un resultado ¡ni siquiera conocía el resultado en cuestión![11]

La representación de la geometría hiperbólica por la que Beltrami es más conocido es otra que él también encontró en 1868. Esta representa la geometría sobre cierta superficie conocida como una *pseudoesfera* (véase la Fig. 2.21). Dicha superficie se obtiene rotando una *tractriz*, una curva investigada por primera vez por Isaac Newton en 1676, alrededor de su «asíntota». La asíntota es una línea recta a la que se aproxima la curva, haciéndose asintóticamente tangente a ella cuando la curva se extiende al infinito. Aquí vamos a imaginar la asíntota dibujada en un plano horizontal de textura rugosa. Imaginemos ahora una varilla ligera, recta y rígida, uno de cuyos extremos, P, lleva unida

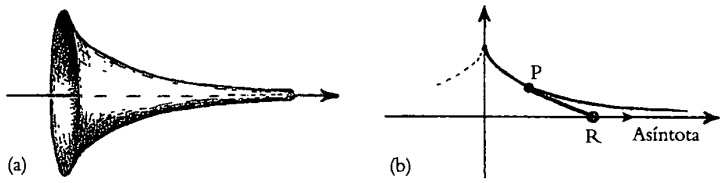


Fig. 2.21. (a) Una *pseudoesfera*. Esta se obtiene rotando una *tractriz* (b) alrededor de su asíntota. Para construir una tractriz, imaginemos que su plano es horizontal, sobre el que se arrastra una varilla ligera, rígida y sin fricción. Un extremo de la varilla es un peso puntual P con fricción, y el otro extremo R se mueve a lo largo de la asíntota (recta).

una masa puntual pesada, y el otro extremo R se mueve a lo largo de la asíntota. El punto P describe entonces una tractriz. Ferdinand Minding descubrió, en 1839, que la pseudoesfera tiene una geometría intrínseca negativa constante, y Beltrami utilizó esto para construir el primer modelo de geometría hiperbólica. Parece que el modelo de la pseudoesfera de Beltrami fue el que convenció a los matemáticos de la consistencia de la geometría hiperbólica plana, puesto que la medida de la distancia hiperbólica coincide con la distancia euclídea a lo largo de la superficie. Sin embargo, es un modelo algo complicado porque representa a la geometría hiperbólica solo localmente, en lugar de presentar toda la geometría de una vez, como hacen los otros modelos de Beltrami.

2.7. ¿Relación con el espacio físico?

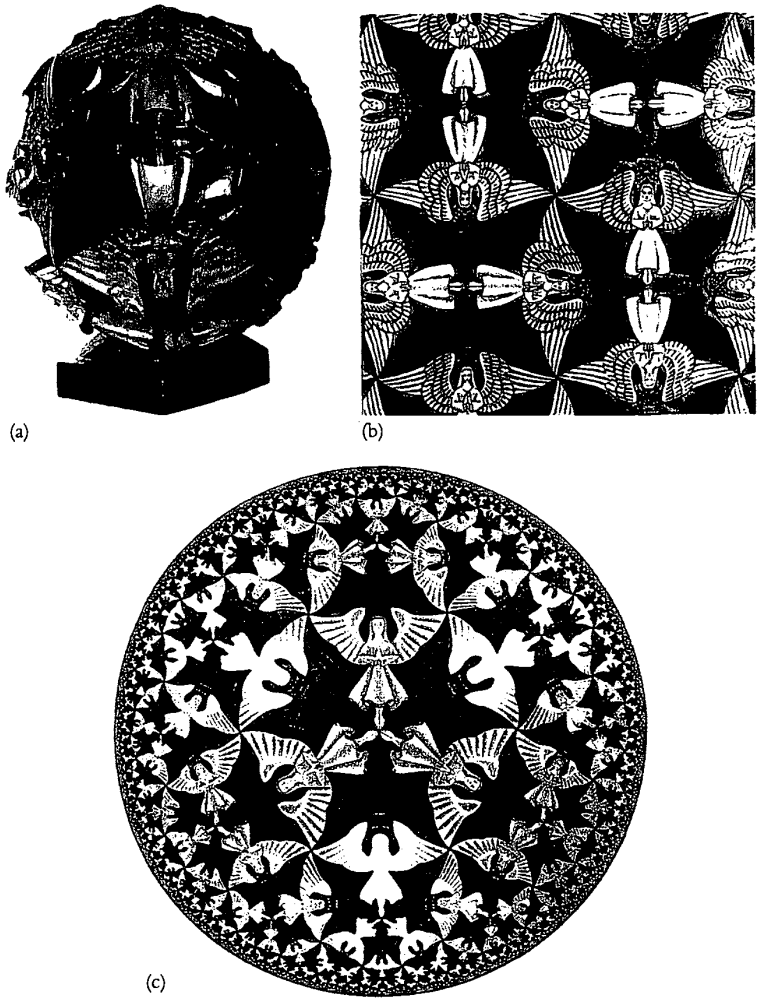
La geometría hiperbólica también funciona perfectamente en dimensiones más altas. Más aún, existen versiones de dimensión superior de los modelos conforme y proyectivo. En el caso de la geometría hiperbólica tridimensional, tenemos una esfera frontera en lugar de un círculo frontera. Toda la geometría hiperbólica tridimensional infinita está representada por el interior de esta esfera euclídea finita. El resto es básicamente igual que lo que teníamos antes. En el modelo conforme, las líneas rectas en esta geometría hiperbólica tridimensional se representan como círculos euclídeos que cortan ortogonalmente a la esfera frontera; los ángulos vienen dados por las medidas euclídeas, y las distancias vienen dadas por la misma fórmula que en el caso bidimensional. En el modelo proyectivo, las líneas rectas hiperbólicas son líneas rectas euclídeas, y las distancias vienen dadas de nuevo por la misma fórmula que en el caso bidimensional.

¿Qué pasa con nuestro universo real a escalas cosmológicas? ¿Esperamos que su geometría espacial sea euclídea, o podría estar en mejor acuerdo con alguna otra geometría, tal como la extraordinaria geometría hiperbólica (aunque en tres dimensiones) que hemos estado examinando en §§2.4-6? Esta es una cuestión realmente importante. Sabemos por la relatividad general de Einstein (a la que llegaremos en

§17.9 y §19.6) que la geometría euclídea es solo una aproximación (extraordinariamente precisa) a la geometría real del espacio físico. Dicha geometría no es ni siquiera exactamente uniforme, al tener pequeños rizos de irregularidad debidos a la presencia de densidad de materia. Pese a todo, y de forma notable, de acuerdo con la mejor evidencia observacional de que hoy disponen los cosmólogos, estos rizos parecen promediarse, en escalas cosmológicas, hasta un grado extraordinariamente preciso (véanse §27.13 y §§28.4-10), y la geometría espacial de nuestro universo real parece concordar extraordinariamente bien con una geometría uniforme (homogénea e isótropa; véase §27.11). Parece que al menos los cuatro primeros postulados de Euclides han superado de forma impresionante la prueba del tiempo.

Aquí es necesario hacer un comentario aclaratorio. Básicamente existen tres tipos de geometría que satisfarían las condiciones de homogeneidad (todos los puntos son iguales) e isotropía (todas las direcciones son iguales), que se conocen como euclídea, hiperbólica y elíptica. La geometría euclídea nos es familiar (y lo ha sido durante unos veintitrés siglos). La geometría hiperbólica ha constituido nuestro interés principal en este capítulo. Pero ¿cuál es la geometría elíptica? Esencialmente, la geometría elíptica plana es la satisfecha por figuras dibujadas en la superficie de una esfera. Apareció en la discusión de la aproximación de Lambert a la geometría hiperbólica, en §2.6. Véanse las Figs. 2.22a,b,c, para la interpretación de Escher de los casos elíptico, euclídeo e hiperbólico, respectivamente, utilizando en los tres casos una teselación similar de ángeles y demonios, la tercera de las cuales ofrece una alternativa interesante a la Fig. 2.11. (Existe también una versión tridimensional de la geometría elíptica, y hay versiones en las que se considera que puntos diametralmente opuestos de la esfera representan el mismo punto. Estas cuestiones se examinarán con algo más de detalle en §27.11.) Sin embargo, podría decirse que el caso elíptico viola los postulados segundo y tercero de Euclides (además del primero). En efecto, se trata de una geometría que es finita en extensión (y en la que más de un segmento de línea une un par de puntos).

¿Cuál es, entonces, el estatus observacional de la geometría espacial a gran escala del universo? Solo se puede decir que todavía no lo sabemos, aunque recientemente se ha dado gran publicidad a afirmaciones



(a) (b) (c)

Fig. 2.22. Los tres tipos básicos de geometría plana uniforme, tal como son ilustrados por Escher utilizando teselaciones de ángeles y demonios. (a) Caso elíptico (curvatura positiva); (b) Caso euclídeo (curvatura cero), y (c) Caso hiperbólico (curvatura negativa), en la representación conforme (*Límite circular IV* de Escher, que debe compararse con la Fig. 2.17).

de que la geometría de Euclides era correcta en todos los niveles, y que su quinto postulado también se cumple, de modo que la geometría espacial promediada es lo que llamamos «euclídea».[12] Por otra parte, existe también evidencia (parte de la cual procede de los mismos experimentos) que parece apuntar firmemente a una geometría global *hiperbólica* para el universo espacial.[13] Además, algunos teóricos han argumentado hace tiempo a favor del caso elíptico, y este no está desde luego descartado por la propia evidencia que se aporta en apoyo del caso euclídeo (véanse las últimas partes de §34.4). Como percibirá el lector, la cuestión está todavía llena de controversia y, como cabría esperar, de discusiones con frecuencia acaloradas. En posteriores capítulos de este libro, trataré de presentar muchas de las ideas que se han propuesto en relación con esto (y no intento ocultar mi propia opinión a favor del caso hiperbólico, aunque tratando de ser tan justo respecto a los otros como sea posible).

Por fortuna para aquellos, como yo mismo, que se sienten atraídos por las bellezas de la geometría hiperbólica, y también por la magnificencia de la física moderna, existe otro papel para esta soberbia geometría que es indiscutiblemente fundamental para nuestra moderna comprensión del universo físico. En efecto, según la moderna teoría de la relatividad, el espacio de *velocidades* es ciertamente una geometría hiperbólica tridimensional (véase §18.4), en lugar de la euclídea que sería válida en la más antigua teoría newtoniana. Esto nos ayuda a entender algunos de los enigmas de la relatividad. Imaginemos, por ejemplo, un proyectil lanzado hacia delante, con velocidad cercana a la de la luz, desde un vehículo que también se mueve hacia delante con una velocidad comparable y pasa frente a un edificio. Pese a todo, con relación a dicho edificio, el proyectil nunca puede superar la velocidad de la luz. Aunque esto parece imposible, veremos en §18.4 que encuentra una explicación directa en términos de geometría hiperbólica. Pero estas materias fascinantes deben esperar hasta capítulos posteriores.

¿Qué pasa con el teorema de Pitágoras, cuyo fallo hemos visto en la geometría hiperbólica? ¿Debemos abandonar el mayor de los regalos concretos que hicieron los pitagóricos a la posteridad? En absoluto, pues la geometría hiperbólica —y, de hecho, todas las geometrías «riemannianas» que generalizan la geometría hiperbólica de una manera

irregularmente curvada (que forma el marco esencial de la teoría de la relatividad general de Einstein; véanse §13.8, §14.7, §18.1 y §19.6)— depende vitalmente de la validez del teorema de Pitágoras en el límite de pequeñas distancias. Además, su enorme influencia impregna otras vastas áreas de las matemáticas y la física (por ejemplo, la estructura métrica «unitaria» de la mecánica cuántica; véase §22.3). A pesar de que este teorema es, en cierto sentido, reemplazado para «grandes» distancias, sigue siendo central para la estructura a pequeña escala de la geometría, encontrando un rango de aplicación que supera muchísimo a aquel para el que fue propuesto originalmente.

Notas

Sección 2.1

2.1. No está muy claro históricamente quién demostró realmente por primera vez lo que ahora conocemos como «teorema de Pitágoras»; véase la nota 1.1. Parece que los antiguos egipcios y babilonios conocían al menos muchos ejemplos de este teorema. El verdadero papel desempeñado por Pitágoras o sus seguidores es básicamente supuesto.

Sección 2.2

2.2. No obstante, incluso con todo este cuidado en la obra de Euclides quedaron varias hipótesis ocultas, que tienen que ver básicamente con lo que ahora llamaríamos cuestiones «topológicas» que habrían parecido «intuitivamente obvias» para Euclides y sus contemporáneos. Estas hipótesis no mencionadas solo fueron advertidas siglos después, en particular por Hilbert a finales del siglo XIX. Las ignoraré en lo que sigue.

2.3. Véase, por ejemplo, Thomas (1939). Compárese también con Schulz (1997), que da una bella exposición axiomática de la geometría espaciotemporal 4-dimensional de Minkowski (§17.8, §18.1).

Sección 2.4

2.4. La notación «exponencial» tal como $C^{-1/2}$ se utiliza con frecuencia en este libro. Como ya se ha dicho en la nota 1.1, a^5 significa $a \times a \times a \times a \times a$; por consiguiente, para un entero positivo n, el producto de a consigo

mismo un total de n veces se escribe a^n. Esta notación se extiende a exponentes negativos, de modo que a^{-1} es el recíproco $1/a$ de a, y a^{-n} es el recíproco $1/a^n$ de a^n, o de forma equivalente $(a^{-1})^n$. De acuerdo con la discusión más general de §5.2, $a^{1/n}$, para un número positivo a, es la «raíz n-ésima de a», que es el número (positivo) que satisface $(a^{1/n})^n = a$ (véase la nota 1.1). Además, $a^{m/n}$ es la potencia m-ésima de $a^{1/n}$.

Sección 2.6

2.5. Saccheri (1733), Prop. XXXIII.

2.6. Existe un punto de vista conocido como *intuicionismo*, mantenido por una minoría (bastante pequeña) de matemáticos, en el que no se acepta el pricipio de «demostración por contradicción». La objeción consiste en que este principio puede ser *no constructivo* en cuanto que a veces lleva a una afirmación de la existencia de cierta entidad matemática, sin que se ofrezca ninguna construcción real de la misma. Esto tiene cierta relevancia para las cuestiones discutidas en §16.6.Véase Heyting (1956).

2.7. Hardy (1940), p. 34.

2.8. Los arcos de círculo máximo son las curvas «más cortas» (geodésicas) sobre la superficie de una esfera; yacen en planos que pasan por el centro de la esfera.

2.9. Es un tema de discusión si Gauss, que estaba profesionalmente interesado en asuntos de geodesia, podría haber tratado de averiguar realmente si hay desviaciones medibles de la geometría euclídea en el espacio físico. Debido a su bien conocida reticencia en cuestiones de geometría no euclídea, es poco probable que lo diera a conocer si en efecto estuviera tratando de hacerlo, especialmente porque (como ahora sabemos) estaba abocado al fracaso, debido a la pequeñez del efecto, según la teoría moderna. Parece que hoy día hay consenso en que él «solo estaba haciendo geodesia», al estar interesado en la curvatura de la Tierra, y no del espacio. Pero encuentro algo difícil creer que él no anduviera también buscando cualquier discrepancia importante con la geometría euclídea; véanse Fauvel y Gray (1987) y Gray (1979).

2.10. La representación denominada «semiplano de Poincaré» (con forma métrica $(dx^2 + dy^2)/y^2$; véase §14.7) se debe también a Beltrami (1868). La curvatura negativa constante de la «métrica de Poincaré» $4(dx^2 + + dy^2)/(1 - x^2 - y^2)^2$ de las Figs. 2.11-13 fue advertida realmente por Riemann.

2.11. Esto parece aplicarse incluso al propio Gauss (que, por otra parte, había

anticipado con mucha frecuencia el trabajo de otros matemáticos). Existe un importante teorema matemático topológico conocido como «teorema de Gauss-Bonnet», que puede demostrarse elegantemente mediante el uso de la denominada «aplicación de Gauss», pero el propio teorema parece deberse en realidad a Blaschke y el elegante método de demostración citado fue encontrado por Olinde Rodrigues. Parece que Gauss y Bonnet no conocieron jamás ni el resultado ni el método de demostración. Existe un teorema de «Gauss-Bonnet» más elemental, correctamente citado en varios textos; véanse Willmore (1959) y Rindler (2001).

Sección 2.7

2.12. La evidencia principal respecto a la estructura global del universo como un todo procede de un análisis detallado de la *radiación cósmica de fondo de microondas* (CMB) que se discutirá en §§27.7,10,11,13, §§28.5,10 y §30.14. Una referencia básica es de Bernardis *et al.* (2000); para datos más recientes y más precisos, véase Netterfield *et al.* (2001) (concernientes a BOOMERanG).Véanse también Hanany *et al.* (2000) (concernientes a MAXIMA), Halverson *et al.* (2001) (concernientes a DASI), y Bennet *et al.* (2003).

2.13. Véanse Gurzadyan y Torres (1997) y Gurzadyan y Kocharyan (1994) para los soportes teóricos, y Gurzadyan y Kocharyan (1992) (para los datos de COBE) y Gurzadyan *et al.* (2002, 2003) (para los datos de BOOMERanG y (2004) para los datos de WMAP) para los correspondientes análisis de los datos CMB reales.

3

Tipos de números en el mundo físico

3.1. ¿Una catástrofe pitagórica?

Volvamos ahora a la cuestión de la demostración por contradicción, el principio que Saccheri trató inútilmente de utilizar en su intento de demostración del quinto postulado de Euclides. Hay muchos casos en las matemáticas clásicas en los que el principio *ha sido* aplicado con éxito. Uno de los más famosos se remonta a los pitagóricos, y zanjó una cuestión matemática en un sentido que les iba a causar grandes problemas. La cuestión era la siguiente: ¿es posible encontrar un número racional (i.e., una fracción) cuyo cuadrado sea exactamente el número 2? Resulta que la respuesta es no, y la afirmación matemática que voy a demostrar dentro de poco es, en efecto, que no existe tal número racional.

¿Por qué estaban los pitagóricos tan molestos por este descubrimiento? Recordemos que una fracción —es decir, un número racional— es algo que puede expresarse como la razón a/b de dos enteros a y b, siendo b distinto de cero. (Véase el prefacio para una discusión de la definición de una fracción.) Los pitagóricos tenían al principio la esperanza de que toda su geometría podría expresarse en términos de longitudes que podrían medirse en términos de números racionales. Los números racionales son cantidades bastante simples, pues son descriptibles y comprensibles en términos finitos simples; pese a todo, pueden utilizarse para especificar distancias tan pequeñas o tan grandes como queramos. Si pudiera hacerse toda la geometría con los racionales, esto haría las cosas relativamente sencillas y fácilmente comprensibles. La noción de un número «irracional», por el contrario, requiere

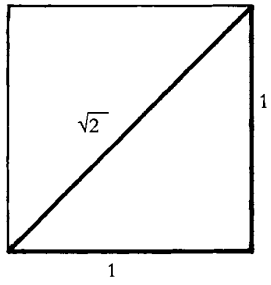


Fig. 3.1. Un cuadrado de lado longtiud unidad tiene diagonal $\sqrt{2}$, por el teorema de Pitágoras.

procesos infinitos, y esto hubiera planteado dificultades considerables a los antiguos (y con toda la razón). ¿Qué dificultad plantea el hecho de que no exista un número racional cuyo cuadrado sea 2? Esta procede del propio teorema de Pitágoras. Si en la geometría euclídea tenemos un cuadrado con lados de longitud unidad, entonces la longitud de su diagonal es un número cuyo cuadrado es $1^2 + 1^2 = 2$ (véase la Fig. 3.1). Sería realmente catastrófico para la geometría que no hubiera ningún número que pudiera describir la longitud de la diagonal de un cuadrado. Al principio, los pitagóricos trataron de arreglárselas con una noción de «número en acto» que pudiera describirse simplemente en términos de razones de números enteros. Veamos por qué esto no puede funcionar.

La cuestión consiste en ver por qué la ecuación

$$\left(\frac{a}{b}\right)^2 = 2$$

no tiene solución para enteros a y b, donde estos enteros se consideran positivos. Utilizaremos las demostración por contradicción para demostrar que no pueden existir tales a y b. Supongamos, por lo tanto, que *sí* existen tales a y b. Multiplicando ambos miembros de la ecuación anterior por b^2, encontramos que se convierte en

$$a^2 = 2b^2$$

y concluimos[1] claramente que $a^2 > b^2 > 0$. Ahora, el segundo miembro, $2b^2$, de la ecuación anterior es par, de donde se sigue que a debe ser par (y no impar, puesto que el cuadrado de cualquier número impar es impar). Así pues, existe un entero positivo c tal que $a = 2c$. Sus-

tituyendo a por $2c$ en la ecuación anterior, y elevando al cuadrado, obtenemos

$$4c^2 = 2b^2,$$

es decir, dividiendo ambos miembros por 2,

$$b^2 = 2c^2,$$

y concluimos que $b^2 > c^2 > 0$. Ahora bien, esta es precisamente la misma ecuación que teníamos antes, excepto que b reemplaza ahora a a y c reemplaza a b. Nótese que los enteros correspondientes son ahora más pequeños que antes. Podemos repetir el argumento una y otra vez, obteniendo una secuencia interminable de ecuaciones

$$a^2 = 2b^2, b^2 = 2c^2, c^2 = 2d^2, d^2 = 2e^2 \ldots,$$

donde

$$a^2 > b^2 > c^2 > d^2 > e^2 > \ldots,$$

siendo todos estos enteros positivos. Pero cualquier secuencia decreciente de enteros positivos debe llegar a un final, lo que contradice el hecho de que esta secuencia es interminable. Esto nos da una contradicción con lo que se había supuesto, a saber, que existe un número racional cuyo cuadrado es 2. De ello se sigue que no existe tal número racional, como queríamos demostrar.[2]

Algunos puntos del argumento anterior merecen un comentario. En primer lugar, de acuerdo con los procedimientos normales de la demostración matemática, en el argumento se ha apelado a ciertas propiedades de los números que se tomaban como «obvias» o como previamente establecidas. Por ejemplo, hemos utilizado el hecho de que el cuadrado de un número impar es siempre impar y, además, que si un entero no es impar, entonces es par. También utilizamos el hecho fundamental de que cada secuencia estrictamente decreciente de números enteros positivos debe llegar a un final.

Una razón por la que puede ser importante identificar las hipótesis precisas que entran en una demostración —incluso si algunas de estas hipótesis pudieran ser cosas perfectamente «obvias»— es que los matemáticos están a menudo interesados en otros tipos de entidades

distintas de aquellas a las que originalmente concernía la demostración. Si estas otras entidades satisfacen las mismas hipótesis, entonces la demostración será extrapolable y se verá que la afirmación que ha sido demostrada tiene una generalidad mayor que la originalmente percibida, puesto que se aplicará también a estas otras entidades. Por el contrario, si alguna de las hipótesis necesarias deja de ser válida para estas entidades alternativas, entonces la afirmación en cuestión puede resultar falsa para dichas entidades. (Por ejemplo, es importante darse cuenta de que en las demostraciones del teorema de Pitágoras dadas en §2.2 se utilizaba el postulado de las paralelas, por lo que el teorema es realmente falso en la geometría hiperbólica.)

En el argumento anterior, las entidades originales son números enteros y estamos interesados en aquellos números —los números racionales— que se construyen como cocientes de enteros. Con tales números se da realmente el caso de que ninguno de ellos tiene 2 como cuadrado. Pero hay otros tipos de números además de los simples enteros y racionales. En realidad, la necesidad para una raíz cuadrada de 2 obligó a los antiguos griegos, muy en contra de su voluntad en aquel tiempo, a salir de los confines de los números enteros y los racionales, los únicos tipos de números que previamente habían estado dispuestos a aceptar. El tipo de número al que se vieron llevados fue el que hoy día llamamos un «número real»: un número que ahora expresamos mediante un desarrollo decimal interminable (aunque una representación semejante no estaba disponible para los antiguos griegos). De hecho, 2 tiene una raíz cuadrada dentro de los números reales, a saber (como ahora la escribiríamos):

$$\sqrt{2} = 1{,}414\ 213\ 562\ 373\ 095\ 048\ 801\ 688\ 72 \ldots$$

En la próxima sección consideraremos con más detalle el estatus *físico* de tales números «reales».

Como curiosidad, podemos preguntar por qué la demostración anterior de la no existencia de una raíz cuadrada de 2 falla para números reales (o para razones de números reales, lo que es equivalente). ¿Qué sucede si reemplazamos «entero» por «número real» en todo el argumento? La diferencia básica es que no es cierto que cualquier sucesión estrictamente decreciente de reales positivos (o incluso de fracciones)

debe llegar a un final, y el argumento se viene abajo en ese punto.[3] (Consideremos, por ejemplo, la secuencia interminable 1, 1/2, 1/4, 1/8, 1/16, 1/32, ..., .) A uno podría preocuparle qué sería un número real «impar» y «par» en este contexto. De hecho, el argumento no encuentra ninguna dificultad en este punto porque *todos* los números reales tendrían que contar como «pares», puesto que para cualquier real a hay siempre un real c tal que $a = 2c$, al ser siempre posible la división por 2 para los reales.

3.2. El sistema de los números reales

Así fue como los griegos se vieron obligados a admitir que los números racionales no son suficientes para desarrollar adecuadamente las ideas de la geometría (de Euclides). Hoy día no nos preocupa excesivamente que cierta cantidad geométrica no pueda ser medida en términos de números racionales simplemente. Esto se debe a que la noción de un «número real» resulta muy familiar para nosotros. Aunque nuestras calculadoras de bolsillo expresan los números mediante un número finito de dígitos, aceptamos que esta es una aproximación a la que nos obliga el hecho de que la calculadora es un objeto finito. Estamos dispuestos a admitir que el número matemático ideal (platónico) podría requerir que el desarrollo decimal se prolongue indefinidamente. Esto se aplica, por supuesto, incluso a la representación decimal de la mayoría de las fracciones, tales como

$$\frac{1}{3} = 0{,}333\ 333\ 333\ \ldots,$$

$$\frac{29}{12} = 2{,}416\ 666\ 666\ \ldots,$$

$$\frac{9}{7} = 1{,}285\ 714\ 285\ 714\ 285\ \ldots,$$

$$\frac{237}{148} = 1{,}601\ 351\ 351\ 35\ \ldots\ .$$

En el caso de una fracción, el desarrollo decimal siempre *acaba siendo periódico*, lo que quiere decir que llega un momento en que la secuencia infinita de dígitos consiste en una secuencia finita que se repite indefinidamente. En los ejemplos anteriores, las secuencias repetidas son 3, 6, 285714 y 135, respectivamente.

Los desarrollos decimales no estaban disponibles para los antiguos griegos, pero ellos tenían sus propios modos de entender los números irracionales. En efecto, lo que adoptaron era un sistema de representación de números en términos de lo que ahora llamamos *fracciones continuas*. No hay necesidad de entrar aquí en detalles, pero es apropiado hacer algunos breves comentarios. Una fracción continua[4] es una expresión finita o infinita $a + (b + (c + (d + \ldots)^{-1})^{-1})^{-1}$, donde $a, b, c, d, \ldots$ son enteros positivos:

$$a + \cfrac{1}{b + \cfrac{1}{c + \cfrac{1}{d + \ldots}}}$$

Cualquier número racional mayor que 1 puede escribirse como una de estas expresiones *terminada* (donde, para evitar ambigüedades, exigimos normalmente que el último entero sea mayor que 1), por ejemplo, $52/9 = 5 + (1 + (3 + (2)^{-1})^{-1})^{-1}$:

$$\frac{52}{9} = 5 + \cfrac{1}{1 + \cfrac{1}{3 + \frac{1}{2}}}$$

y, para representar un racional positivo menor que 1, simplemente admitimos que el primer entero en la expresión sea cero. Para expresar un número real, que no es racional, simplemente[3.1] permitimos que la expresión de la fracción continua siga de manera indefinida, siendo algunos ejemplos[5]

[3.1] Experimente con su calculadora de bolsillo (suponiendo que usted tenga teclas «$\sqrt{}$» y «x^{-1}») para obtener estas expresiones con la precisión disponible. Tome $\pi = 3{,}141\ 592\ 653\ 589\ 793\ldots$ (*Sugerencia*: Apunte la parte entera de cada número, réstela de dicho número y forme luego el recíproco del resto. Repita la operación con el número así obtenido.)

$$\sqrt{2} = 1 + (2 + (2 + (2 + (2 + \ldots)^{-1})^{-1})^{-1})^{-1},$$
$$7 - \sqrt{3} = 5 + (3 + (1 + (2 + (1 + (2 + (1 + (2 + \ldots)^{-1})^{-1})^{-1})^{-1})^{-1})^{-1})^{-1},$$
$$\pi = 3 + (7 + (15 + (1 + (292 + (1 + (1 + (1 + \\ + (2 + \ldots)^{-1})^{-1})^{-1})^{-1})^{-1})^{-1})^{-1})^{-1}.$$

En los dos primeros de estos ejemplos infinitos, las secuencias de números naturales que aparecen —a saber, 1, 2, 2, 2, 2, ... en el primer caso y 5, 3, 1, 2, 1, 2, 1, 2, ... en el segundo— tienen la propiedad de que son finalmente periódicas (repitiéndose indefinidamente el 2 en el primer caso y repitiéndose indefinidamente la secuencia 1, 2 en el segundo).[3.2] Recordemos que, como ya se ha advertido antes, en la notación decimal familiar, son los números *racionales* los que tienen expresiones (finitas o) finalmente periódicas. Por otra parte, podemos considerar que una virtud de la representación de «fracción continua» griega es que los números racionales tienen ahora siempre una descripción finita. Una pregunta natural que se puede plantear en este contexto es: ¿qué números tienen una representación en forma de fracción continua *finalmente periódica*? Un teorema notable, demostrado por primera vez que sepamos por el gran matemático del siglo XVIII Joseph L. Lagrange (cuyas otras ideas más importantes encontraremos más adelante, en particular en el capítulo 20), dice que los números cuyas representaciones en términos de fracciones continuas son finalmente periódicas son los denominados *irracionales cuadráticos*.[6]

¿Qué es un irracional cuadrático y qué importancia tiene para la geometría griega? Es un número que puede escribirse en la forma

$$a + \sqrt{b},$$

donde a y b son fracciones, y donde b no es un cuadrado perfecto. Estos números son importantes en geometría euclídea porque son los números irracionales más inmediatos que se encuentran en las construcciones con regla y compás. (Recordemos el teorema de Pitágoras, que en §3.1 nos ha llevado por primera vez a considerar el problema de $\sqrt{2}$,

[3.2] Suponiendo esta periodicidad final de estas dos expresiones de fracciones continuas, demuestre que los números que representan deben ser las cantidades del miembro izquierdo. (*Sugerencia*: Encuentre una ecuación cuadrática que sea satisfecha por esta cantidad, y remítase a la nota 3.6.)

y otras construcciones sencillas de longitudes euclídeas nos llevaron directamente a otros números de la forma anterior.)

Ejemplos concretos de irracionales cuadráticos son aquellos casos donde $a = 0$ y b es un número natural (no cuadrado) o racional mayor que 1, por ejemplo:

$$\sqrt{2}, \sqrt{3}, \sqrt{5}, \sqrt{6}, \sqrt{7}, \sqrt{8}, \sqrt{10}, \sqrt{11}, \ldots$$

La representación de fracción continua de un número semejante es especialmente sorprendente. La secuencia de números naturales que la define como una fracción continua tiene una curiosa propiedad característica. Empieza con cierto número A, que luego es seguido inmediatamente por una secuencia «palindrómica» (i.e., una que se lee igual hacia atrás), $B, C, D, \ldots, D, C, B$, seguida por $2A$, tras el cual la propia secuencia $B, C, D, \ldots, D, C, B, 2A$ se repite indefinidamente. El número $\sqrt{14}$ es un buen ejemplo, para el que la secuencia es

$$3, 1, 2, 1, 6, 1, 2, 1, 6, 1, 2, 1, 6, 1, 2, 1, 6, \ldots$$

Aquí $A = 3$ y la secuencia palindrómica $B, C, D, \ldots, D, C, B$ es simplemente la secuencia de tres términos 1, 2, 1.

¿Cuánto de esto era conocido por los antiguos griegos? Parece probable que conocían mucho —muy posiblemente *todas* las cosas que he descrito antes (incluido el teorema de Lagrange)—, aunque quizá carecieran de demostraciones rigurosas para todo. Al parecer, gran parte de esto fue establecido por Teeteto, contemporáneo de Platón. Parece incluso que hay alguna evidencia de este conocimiento (incluidas las secuencias palindrómicas repetidas antes mencionadas) manifiesta en la dialéctica de Platón.[7]

Aunque la incorporación de los irracionales cuadráticos nos acerca algo hacia los números adecuados para la geometría euclídea, no hace todo lo que se necesita. En el décimo (y más difícil) libro de Euclides se consideran números como $\sqrt{a + \sqrt{b}}$ (con a y b racionales positivos). Estos *no* son generalmente irracionales cuadráticos, pero se dan, de todas formas, en construcciones de regla y compás. Números suficientes para tales construcciones geométricas serían aquellos que pueden formarse a partir de números naturales por uso repetido de las operaciones de adición, sustracción, multiplicación, división, y toma de la raíz cuadra-

da. Pero operar exclusivamente con tales números se hace muy complicado, y estos números son todavía demasiado limitados para consideraciones de geometría euclídea que van más allá de las construcciones de regla y compás. Es mucho más satisfactorio dar el paso capital —y hasta qué punto es capital se indicará en §§16.3-5— de permitir expresiones de fracciones continuas infinitas que son completamente generales. Esto proporcionó a los griegos una forma de describir números que resulta adecuada para la geometría euclídea.

Estos números son en realidad, en terminología moderna, los denominados «números reales». Aunque se considera que una definición completamente satisfactoria de tales números no fue encontrada hasta el siglo XIX (con el trabajo de Dedekind, Cantor y otros), el gran matemático y astrónomo griego Eudoxo, que había sido uno de los discípulos de Platón, había obtenido las ideas esenciales ya en el siglo IV a.C. Aquí resulta apropiado comentar las ideas de Eudoxo.

En primer lugar, notemos que los números en geometría euclídea pueden expresarse en términos de *razones* de longitudes, más que directamente en términos de longitudes. De esta forma, no se necesitaba ninguna unidad específica de longitud (tal como la «pulgada» o el «dactylos» griego). Además, con razones de longitudes no habría restricciones en cuanto al número de tales razones que podían multiplicarse (obviando la aparente necesidad de «hipervolúmenes» en dimensiones más altas cuando se multiplican más de tres longitudes). El primer paso en la teoría de Eudoxo consistía en proporcionar un criterio acerca de cuándo una razón de longitudes $a : b$ sería mayor que otra razón semejante $c : d$. Este criterio es que existen ciertos enteros positivos M y N tales que la longitud a sumada a sí misma M veces supera a b sumada a sí misma N veces, mientras que también d sumada a sí misma N veces supera a c sumada a sí misma M veces.[3.3] Un criterio correspondiente es válido para expresar la condición de que la razón $a : b$ es *menor* que la razón $c : d$. La condición para la igualdad de ambas razones sería que no se cumpla ninguno de estos criterios. Con esta ingeniosa noción de «igualdad» de tales razones, Eudoxo tenía, en efecto, un concepto abstracto de «número real» en términos de razones de

[3.3] ¿Puede ver por qué funciona esto?

longitudes. Él también proporcionó reglas para la suma y el producto de tales números reales.[3.4]

Sin embargo, había una diferencia básica en el punto de vista entre la noción griega de número real y la moderna, porque los griegos creían que el sistema de números nos estaba básicamente «dado» en términos de la noción de *distancia* en el espacio físico, de modo que el problema consistía en tratar de establecer cómo se comportaban realmente estos números «distancia»; pues el propio «espacio» podría haber parecido un absoluto platónico incluso si los objetos físicos reales existentes en dicho espacio estuvieran inevitablemente muy lejos del ideal platónico.[8] (Sin embargo, en §17.9 y §§19.6,8 veremos que la teoría de la relatividad general de Einstein ha cambiado ahora esta perspectiva del espacio y la materia de un modo fundamental.)

Un objeto físico tal como un cuadrado dibujado en la arena o un cubo esculpido en mármol podría haber sido considerado por los antiguos griegos como una razonable o a veces una excelente aproximación al ideal geométrico platónico. Pese a todo, cualquier objeto semejante proporcionaría en cualquier caso una mera aproximación. Detrás de tales aproximaciones a las formas platónicas —así hubiera parecido— estaría el propio espacio: una entidad de existencia tan abstracta o conceptual que muy bien podría haber sido considerada como una realización directa de una realidad platónica. La medida de distancia en esta geometría ideal sería algo a *determinar*; en consecuencia, sería apropiado tratar de extraer esta noción ideal de número real de una geometría del espacio que se suponía *dada*. De hecho, esto es lo que Eudoxo consiguió hacer.

No obstante, en los siglos XIX y XX había surgido la idea de que la noción matemática de número debería presentarse con independencia de la naturaleza del espacio físico. Puesto que se había demostrado que existían geometrías matemáticamente consistentes diferentes de la de Euclides, resultaba inoportuno insistir en que la noción matemática de «geometría» debería ser extraída necesariamente de la naturaleza supuesta del espacio físico «real». Además, podría ser muy difícil, si no imposible, establecer la naturaleza detallada de esta supuesta «geo-

[3.4] ¿Puede ver cómo formularlas?

metría física platónica» subyacente en términos del comportamiento de objetos físicos imperfectos. Para conocer la naturaleza de los números de acuerdo con los cuales debe definirse la «distancia geométrica», por ejemplo, sería necesario saber qué sucede tanto a distancias infinitamente minúsculas como infinitamente grandes. Incluso hoy, estas cuestiones no tienen una clara respuesta (y las abordaré de nuevo en capítulos posteriores). Así pues, era mucho más oportuno establecer la naturaleza del número de un modo que no remitiera directamente a medidas físicas. En consecuencia, Richard Dedekind y Georg Cantor elaboraron sus ideas de lo que «son» los números reales mediante el uso de nociones que no se refieren directamente a la geometría.

Dedekind define un número real a partir de conjuntos infinitos de números racionales. Lo que se hace, básicamente, es considerar que los números racionales, tanto positivos como negativos (y el cero), están dispuestos en orden de tamaño. Podemos imaginar que este ordenamiento tiene lugar de izquierda a derecha, considerando que los racionales negativos se extienden indefinidamente hacia la izquierda, y los racionales positivos se extienden indefinidamente hacia la derecha, estando 0 en el centro. (Esto es solo para propósitos de visualización; de hecho, el procedimiento de Dedekind es completamente abstracto.) Dedekind imagina un «corte» que divide esta disposición claramente en dos, de modo que aquellos números que están a la izquierda del corte son más pequeños que los que están a la derecha. Cuando el «filo del cuchillo» que hace el corte no «incide» sobre un número racional sino que cae entre ellos, entonces decimos que define un número real *irracional*. Dicho de forma más correcta, esto ocurre cuando los que están a la izquierda no tienen un miembro máximo, y los que están a la derecha no tienen un miembro mínimo. Cuando se añade el sistema de los «irracionales», definidos en términos de tales «cortes», al sistema de los números racionales que ya teníamos, se obtiene la familia completa de los *números reales*.

El procedimiento de Dedekind conduce directamente, por medio de simples definiciones, a las leyes de adición, sustracción, multiplicación y división de números reales. Además, permite ir más allá y definir *límites*, mediante los cuales pueden asignarse significados en nú-

meros reales a cosas tales como la fracción continua infinita que hemos visto antes

$$1 + (2 + (2 + (2 + (2 + \ldots)^{-1})^{-1})^{-1})^{-1}$$

o la suma infinita

$$1 - \frac{1}{3} + \frac{1}{5} - \frac{1}{7} + \frac{1}{9} - \ldots$$

De hecho, la primera nos da el número irracional $\sqrt{2}$ y la segunda, $\frac{1}{4}\pi$. La capacidad de tomar límites es fundamental para muchas nociones matemáticas, y es esto lo que da a los números reales su fuerza especial.[9] (Quizá el lector recuerde que la necesidad de «procedimientos de paso al límite» era un requisito para la definición general de áreas, como se ha indicado en §2.3).

3.3. Los números reales en el mundo físico

Aquí estamos tocando una cuestión profunda. Una fuerza impulsora inicial en el desarrollo de las ideas matemáticas ha sido siempre el intento de encontrar estructuras matemáticas que reflejen de forma precisa el comportamiento del mundo físico. Pero normalmente no es posible examinar el propio mundo físico con un detalle tan preciso que de él puedan extraerse directamente nociones matemáticas de la claridad adecuada. El progreso se produce más bien debido a que las nociones matemáticas tienden a tener un «impulso» propio que parece brotar casi por entero del interior de la propia disciplina. Las ideas matemáticas evolucionan, y varios tipos de problemas parecen surgir de forma natural. Algunos de estos (como sucedió con el problema de encontrar la longitud de la diagonal de un cuadrado) pueden llevar a una ampliación esencial de los propios conceptos matemáticos en cuyos términos se había formulado originalmente el problema. Puede parecer que tales ampliaciones nos vienen obligadas, o también pueden surgir de maneras que parecen ser cuestiones de conveniencia, consistencia o elegancia matemática. En consecuencia, podría parecer que el desarrollo de las matemáticas se aleja de lo que se habían propuesto

conseguir, a saber: reflejar el comportamiento físico. Pero, en muchos casos, este mismo impulso hacia la consistencia y elegancia matemáticas nos lleva a estructuras y conceptos matemáticos que resultan reflejar el mundo físico de una forma mucho más profunda y de mayor alcance que aquellas de las que partimos. Es como si la propia naturaleza se guiara por el mismo tipo de criterios de consistencia y elegancia que guían al pensamiento matemático humano.

Un ejemplo de esto es el propio sistema de los números reales. No tenemos ninguna evidencia directa en la naturaleza de que haya una noción física de «distancia» que se extienda hasta escalas arbitrariamente grandes; menos evidencia hay aún de que semejante noción sea aplicable en el nivel infinitamente minúsculo. En realidad, no hay evidencia de que existan «puntos en el espacio» de acuerdo con una geometría que haga uso precisamente de distancias en números reales. En la época de Euclides había escasa evidencia para apoyar siquiera la pretensión de que tales «distancias» euclídeas se extendían hacia fuera hasta más allá de, digamos, unos 10^{12} metros,[10] o hacia dentro, hasta algo tan pequeño como 10^{-5} metros. Pero, al haber sido impulsadas matemáticamente por la consistencia y elegancia del sistema de los números reales, todas nuestras teorías físicas satisfactorias y de más amplio alcance hasta la fecha han seguido ateniéndose, sin excepción, a esta antigua noción de «número real». Aunque podría parecer que ha habido muy poca justificación para hacer esto a partir de la evidencia de que se disponía en la época de Euclides, nuestra fe en el sistema de los números reales parece haberse visto recompensada. En efecto, nuestras modernas y satisfactorias teorías cosmológicas nos permiten ahora ampliar el rango de nuestras distancias en números reales hasta aproximadamente 10^{26} metros o más, mientras que la precisión de nuestras teorías de la física de partículas extiende este rango hacia dentro, hasta 10^{-17} metros o menos. (La única escala para la que se ha propuesto con seriedad que podría llegar un cambio es de unos dieciocho órdenes de magnitud inferior a esta, a saber 10^{-35} metros, que es la «escala de Planck» de la gravitación cuántica que cobrará una importancia especial en nuestras discusiones posteriores; por ejemplo, §§31.1,6-12,14 y §32.7.) Puede considerarse una notable justificación de nuestro uso de idealizaciones matemáticas el hecho de que el rango de validez del sistema de los números reales se ha ampliado desde un total de aproximadamente 10^{17}, desde lo más pequeño a lo más grande, que parecía adecuado en la época de Euclides, hasta al menos los 10^{43} que nuestras teorías actuales utilizan directamente, lo que supone un extraordinario incremento en un factor de 10^{26}.

Hay mucho más a favor de la validez física del sistema de los números reales. En primer lugar, debemos considerar que también las áreas y los volúmenes son magnitudes para las que convienen medidas muy precisas en números reales. Una medida de volumen es el cubo de una medida de distancia (y un área es el cuadrado de una distancia). En consecuencia, en el caso de los volúmenes podemos considerar que el intervalo relevante es el cubo del anterior. Así, para la época de Euclides esto nos daría un rango de aproximadamente $(10^{17})^3 = 10^{51}$; para las teorías actuales, el intervalo es al menos $(10^{43})^3 = 10^{129}$. Además, existen otras medidas físicas que requieren descripciones en números reales, según las teorías que hoy día resultan satisfactorias. La más digna de mención de estas es el tiempo. Según la teoría de la relatividad, este tiene que añadirse al espacio para que nos proporcione el *espaciotiempo* (que será objeto de nuestro estudio en el capítulo 17). Los volúmenes espaciotemporales son tetradimensionales, y muy bien podría considerarse que el rango temporal (de nuevo de aproximadamente 10^{43} o más en rango total, en nuestras teorías mejor comprobadas) debería ser incorporado también en nuestras consideraciones, dando así un total del orden de 10^{172}. Veremos algunos números reales todavía mucho mayores que este cuando lleguemos a nuestras consideraciones posteriores (véanse §27.13 y §28.7), aunque en algunos casos no esté realmente claro que sea esencial el uso de números reales (antes que, por ejemplo, enteros).

Más importante para la teoría física es el hecho de que, desde Arquímedes hasta Maxwell, Einstein, Schrödinger, Dirac y otros, pasando por Galileo y Newton, un papel crucial del sistema de los números reales ha consistido en proporcionar un marco necesario para la formulación estándar del *cálculo infinitesimal* (véase el capítulo 6). Todas las teorías dinámicas satisfactorias han requerido para su formulación las nociones del cálculo. Ahora bien, el enfoque convencional del cálculo requiere que la naturaleza *infinitesimal* de los reales sea la que es. Es decir, en el extremo inferior de la escala es el rango entero de los núme-

ros reales el que está siendo utilizado. Las ideas del cálculo infinitesimal subyacen en otras nociones físicas, tales como velocidad, momento o energía. En consecuencia, el sistema de los números reales entra en nuestras teorías físicas satisfactorias de una manera fundamental para nuestra descripción de todas aquellas magnitudes. Aquí, como se ha mencionado antes en relación con las áreas, en §2.3 y §3.2, se está invocando el límite infinitesimal de la estructura a pequeña escala del sistema de los números reales.

Pese a todo, podemos seguir preguntándonos si el sistema de los números reales es realmente «correcto» para la descripción de la realidad física en sus niveles más profundos. Cuando empezaron a introducirse las ideas mecanocuánticas a comienzos del siglo XX, existía la sensación de que quizá entonces empezábamos a ser testigos de una naturaleza discreta o granular del mundo físico en sus escalas más pequeñas.[11] Aparentemente, la energía solo podía existir en paquetes discretos —o «cuantos»— y las magnitudes físicas de «acción» y «espín» parecen darse solo en múltiplos discretos de una unidad fundamental (véanse §§20.1,5 para el concepto clásico de *acción* y §26.6 para su contrapartida cuántica; véanse §§22.8-12 para el *espín*). Por ello, varios físicos intentaron construir una imagen alternativa del mundo en la que procesos discretos gobernaban todas las acciones en los niveles más ínfimos.

Sin embargo, y tal como ahora entendemos la mecánica cuántica, esta teoría no nos obliga (ni siquiera nos lleva) a la idea de que hay una naturaleza discreta o granular para el espacio, el tiempo o la energía en sus niveles más ínfimos (véanse los capítulos 21 y 22, en particular la última frase de §22.13). En cualquier caso, nos ha quedado la idea de que quizá haya realmente una discretización tal en la naturaleza, pese al hecho de que la mecánica cuántica, en su formulación estándar, no implica esto ni mucho menos. Por ejemplo, el gran físico cuántico Erwin Schrödinger fue uno de los primeros en sugerir que podría ser necesario un cambio hacia alguna forma de discretización espacial fundamental:[12]

> La idea de un *rango continuo*, tan familiar para los matemáticos de nuestros días, es algo bastante exorbitante, una enorme extrapolación de lo que es accesible para nosotros.

Relacionó dichas propuestas con algunas ideas griegas antiguas relativas a la discretización de la naturaleza. También Einstein sugirió, en sus últimas palabras publicadas, que una teoría («algebraica») basada en la discretización podría ser el camino hacia la física futura:[13]

> Se pueden dar buenas razones por las que la realidad no puede representarse como un campo continuo ... Los fenómenos cuánticos ... deben llevar a un intento de encontrar una teoría puramente algebraica para la descripción de la realidad. Pero nadie sabe cómo obtener la base de una teoría semejante.[14]

Otros[15] también han perseguido ideas de este tipo (véase §33.1). A finales de la década de 1950, yo mismo ensayé algo parecido, llegando a un esquema al que denominé teoría de «redes de espín», en la que la naturaleza discreta del *espín* mecanocuántico se toma como el bloque constituyente fundamental para un enfoque *combinatorio* (i.e., discreto en lugar de basado en números reales) de la física (este esquema se describirá brevemente en §32.6). Aunque mis propias ideas en esta dirección no se desarrollaron hasta convertirse en una teoría global (si bien en cierto sentido se metamorfosearon más tarde en la «teoría de twistores»; véase §33.2), la teoría de redes de espín ha sido ahora importada, por otros, en uno de los programas principales para atacar el problema fundamental de la *gravitación cuántica*.[16] Daré breves descripciones de estas ideas en el capítulo 32. En cualquier caso, tal como hoy se ensaya y pone a prueba la teoría física —y como lo ha sido durante los veinticuatro siglos pasados—, los *números reales* siguen constituyendo un ingrediente esencial en nuestra comprensión del mundo físico.

3.4. ¿Necesitan los números naturales al mundo físico?

En la descripción anterior, en §3.2, de la aproximación de Dedekind al sistema de los números reales, he supuesto que los *números racionales* se daban por «entendidos». De hecho, no es difícil pasar de los enteros a los racionales. Los racionales son simplemente razones de enteros (véase el prefacio). ¿Qué pasa entonces con los enteros propiamente dichos?

¿Están enraizados en ideas físicas? Incluso los enfoques discretos de la física, que se mencionaron en los dos párrafos anteriores, dependen de nuestra noción de *número natural* (i.e., «número para recuento») y su extensión, mediante la inclusión de los números negativos, a los *enteros*. Los griegos no consideraban los números negativos como «números» en acto, así que continuemos nuestras consideraciones preguntando primero por el estatus físico de los propios números naturales.

Los *números naturales* son las cantidades que ahora denotamos por 0, 1, 2, 3, 4, etc., i.e., son los números enteros no negativos. (Hoy se incluye al 0 en esta lista, lo que es apropiado desde el punto de vista matemático, aunque parece que los antiguos griegos no reconocieron el «cero» como un número en acto. Esto tuvo que esperar hasta los matemáticos hindúes de la India, empezando con Brahmagupta en el siglo VII y seguido de Mahavira y Bhaskara en los siglos IX y XII, respectivamente.) El papel de los números naturales es claro e inequívoco. Son los «números para recuento» más elementales, y tienen un papel básico, cualesquiera que puedan ser las leyes de la geometría o de la física. Los números naturales están sujetos a ciertas operaciones familiares, muy en especial las operaciones de *adición* (tal como $37 + 79 = 116$) y *multiplicación* (por ejemplo, $37 \times 79 = 2.923$), que permiten combinar pares de números naturales para producir nuevos números naturales. Estas operaciones son independientes de la geometría del mundo.

No obstante, podemos plantear la cuestión de si los propios números naturales tienen un significado o existencia con independencia de la naturaleza real del mundo físico. Quizá nuestra noción de los números naturales depende de que en nuestro universo haya objetos discretos razonablemente bien definidos y que persisten en el tiempo. Después de todo, los números naturales aparecen inicialmente cuando queremos contar cosas. Pero esto parece depender de que existan realmente «cosas» persistentes y distinguibles en el universo que estén disponibles para ser «contadas». Supongamos, por el contrario, que nuestro universo fuera tal que el número de los objetos tuviese tendencia a variar. ¿Realmente serían los números naturales conceptos «naturales» en un universo semejante? Más aún, quizá el universo contenga solo un número finito de «objetos», en cuyo caso ¡los propios números «naturales» podrían terminar en algún punto! Podemos concebir incluso un universo

que consista solo en una sustancia amorfa e indiferenciada, para la cual la noción misma de cuantificación numérica podría parecer intrínsecamente inadecuada. ¿Tendría la noción de «número natural» la más mínima relevancia para la descripción de universos de este tipo?

Aunque muy bien pudiera darse el caso de que los habitantes de un universo semejante encontraran difícil de captar nuestro concepto matemático presente de «número natural», es difícil imaginar que no siguiera habiendo un papel importante para entidades tan fundamentales. Hay varias formas de introducir los números naturales en las matemáticas puras, y estas no parecen depender en absoluto de la naturaleza real del mundo físico. Básicamente, es la noción de un «conjunto» la que necesita ser invocada, siendo esta una noción abstracta que no parece estar relacionada de ninguna manera esencial con la estructura específica del universo físico. De hecho, existen ciertas sutilezas, concernientes a esta cuestión, y volveré a ello más adelante (en §16.5). Por el momento, será conveniente ignorar tales sutilezas.

Consideremos un modo (anticipado por Cantor y defendido por el destacado matemático John von Neumann) de introducir los números naturales utilizando simplemente la noción abstracta de conjunto. Este procedimiento permite definir lo que se denominan «números ordinales». Al conjunto más simple de todos se le llama «conjunto nulo» o «conjunto vacío», y está caracterizado por el hecho de que ¡no contiene ningún miembro! El conjunto vacío se suele denotar por el símbolo $\varnothing$, y podemos escribir esta definición

$$\varnothing = \{\ \},$$

donde los corchetes delimitan un conjunto, el conjunto específico bajo consideración, que tiene como miembros las cantidades indicadas dentro de los corchetes. En este caso, no hay nada dentro de los corchetes, de modo que el conjunto descrito es realmente el conjunto vacío. Asociemos $\varnothing$ con el número natural 0. Ahora podemos continuar y definir el conjunto cuyo único miembro es $\varnothing$; i.e., el conjunto $\{\varnothing\}$. Es importante darse cuenta de que $\{\varnothing\}$ no es el mismo conjunto que el conjunto vacío $\varnothing$. El conjunto $\{\varnothing\}$ tiene *un* miembro (a saber, $\varnothing$), mientras que el propio $\varnothing$ no tiene ninguno. Asociemos $\{\varnothing\}$ con el número natural 1. A continuación definimos el conjunto cuyos dos miembros

son los dos conjuntos que ya hemos encontrado, a saber ∅ y {∅}, de modo que este nuevo conjunto es {∅, {∅}}, que será asociado con el número natural 2. Luego asociamos con 3 la colección de las tres entidades que hemos encontrado hasta ahora, a saber, el conjunto {∅, {∅}, {∅, {∅}}}, y con 4 el conjunto {∅, {∅}, {∅, {∅}}}, {∅, {∅}, {∅, {∅}}}}, cuyos miembros son de nuevo los conjuntos que hemos encontrado previamente, y así sucesivamente. Quizá no sea así como se consideran normalmente los números naturales, en cuanto a su definición, pero es una de las formas en que los matemáticos pueden llegar al concepto. (Compárese esto con la exposición del prefacio.) Más aún, nos muestra, al menos, que cosas como los números naturales[17] pueden ser extraídas literalmente de la nada, empleando solo la noción abstracta de «conjunto». Obtenemos una secuencia infinita de entidades matemáticas abstractas («platónicas»): conjuntos que contienen, respectivamente, cero, uno, dos, tres, etc., elementos, un conjunto por cada uno de los números naturales, de forma completamente independiente de la naturaleza física real del universo. En la Fig. 1.3 imaginábamos un tipo de «existencia» independiente para las nociones matemáticas platónicas —en este caso, los propios números naturales—, pero esta «existencia» puede ser extraída en apariencia, y ciertamente puede accederse a ella, por el mero ejercicio de nuestra imaginación, sin ninguna referencia a los detalles de la naturaleza del universo físico. Además, la construcción de Dedekind muestra cómo puede llevarse más lejos este tipo de procedimiento «puramente mental», lo que nos permite «construir» el sistema entero de los números reales,[18] sin ninguna referencia tampoco a la naturaleza física real del mundo. Pese a todo, y como se ha señalado antes, los «números reales» parecen tener realmente una relevancia directa para la estructura real del mundo, lo que ilustra la muy misteriosa naturaleza del «primer misterio» representado en la Fig. 1.3.

3.5. Números discretos en el mundo físico

Pero me estoy adelantando un poco a mis propósitos. Podemos recordar que la construcción de Dedekind utilizaba realmente conjuntos de

números *racionales*, y no directamente de números naturales. Como se ha mencionado antes, no es difícil «definir» lo que entendemos por número racional una vez que tenemos la noción de número natural. Pero, como paso intermedio, es oportuno definir la noción de *entero*, que es un número natural o el *negativo* de un número natural (siendo el número cero su propio negativo). En un sentido formal, no hay ninguna dificultad para dar una definición matemática de «negativo»: dicho de forma muy tosca, simplemente añadimos un «signo», escrito como «–», a cada número natural (excepto el 0) y definimos consistentemente todas las reglas aritméticas de adición, sustracción, multiplicación y división (excepto por cero). Sin embargo, esto no aborda la cuestión del «significado físico» de un número negativo. ¿Qué podría significar, por ejemplo, decir que hay menos tres vacas en un campo?

Creo que está claro que, a diferencia de los propios números naturales, no hay contenido físico evidente para la noción de un número negativo de objetos físicos. Ciertamente, los enteros negativos tienen un papel organizador extraordinariamente valioso, como se ve en los balances bancarios y otras transacciones financieras. Pero ¿tienen relevancia directa para el mundo *físico*? Cuando digo «relevancia directa», no me estoy refiriendo a circunstancias donde podría parecer que son los números reales negativos los que constituyen las medidas relevantes, como sucede cuando una distancia medida en una dirección cuenta como positiva mientras que medida en dirección contraria contaría como negativa (o lo mismo con respecto al tiempo, en cuyo caso el tiempo que se extiende hacia el pasado cuenta como negativo). Me estoy refiriendo más bien a números que son cantidades *escalares*, en el sentido de que no hay un aspecto direccional (o temporal) en la magnitud en cuestión. En estas circunstancias, parece ser que es el sistema de los enteros, tanto positivos como negativos, el que tiene relevancia física directa.

Resulta notable que solo en los últimos cien años aproximadamente se ha hecho manifiesto que el sistema de los enteros parece tener realmente una relevancia física directa. El primer ejemplo de una magnitud física que parece estar cuantificada apropiadamente por enteros es la *carga eléctrica*.[19] Hasta donde sabemos (aunque todavía no hay una completa justificación teórica para ello), la carga eléctrica de cual-

quier cuerpo discreto aislado está realmente cuantificada en términos de múltiplos enteros, positivos, negativos o nulos, de un valor concreto, a saber, la carga del protón (o del electrón, que es la negativa de la del protón).[20] Ahora pensamos que los protones son objetos compuestos construidos a partir de entidades más pequeñas conocidas como «quarks» (y otras entidades sin carga denominadas «gluones»). Hay tres quarks en cada protón, que tienen cargas eléctricas con valores respectivos de 2/3, 2/3, −1/3. La suma de estas cargas constituyentes da un valor total de 1 para el protón. Si los quarks son entidades fundamentales, entonces la unidad básica de carga es un tercio de la que parecíamos tener antes. En cualquier caso, sigue siendo cierto que la carga eléctrica se mide en términos de enteros, aunque ahora son múltiplos enteros de un tercio de la carga del protón. (El papel de los quarks y los gluones en la moderna física de partículas se discutirá en §§25.3-7.)

La carga eléctrica es solo un ejemplo de lo que se denomina un *número cuántico aditivo*. Los números cuánticos son cantidades que sirven para caracterizar a las partículas de la naturaleza. Un número cuántico semejante, que consideraré aquí que es un número real de algún tipo, es «aditivo» si para obtener su valor para una entidad compuesta sumamos simplemente los valores individuales de las partículas constituyentes —teniendo en cuenta debidamente, por supuesto, los signos, como en el caso del protón y sus quarks constituyentes que se ha señalado antes—. Es un hecho muy sorprendente, de acuerdo con el estado actual de nuestro conocimiento físico, que todos los números cuánticos aditivos conocidos[21] estén cuantificados realmente en términos del sistema de los enteros, no en términos de números reales en general, ni tampoco de simples números naturales, de modo que realmente se dan los valores negativos.

De hecho, según la física del siglo XX, ahora hay un sentido en el que *tiene* significado referirse a un número negativo de entidades físicas. El gran físico Paul Dirac propuso, en 1929-1931, su teoría de las antipartículas, según la cual (tal como se entendió posteriormente) por cada tipo de partícula existe también su correspondiente *antipartícula* para la que cada número cuántico aditivo tiene exactamente el negativo del valor que tiene para la partícula original (véanse §§24.2,8). Así pues, el sistema de los enteros (incluidos los negativos) parece tener

realmente una clara relevancia para el universo físico, una relevancia física que se ha hecho manifiesta solo en el siglo XX, a pesar de los muchos siglos durante los cuales los enteros han encontrado gran valor en las matemáticas, el comercio y muchas otras actividades humanas.

Llegados a este punto, habría que hacer una matización importante. Aunque es verdad que, en cierto sentido, un antiprotón es un protón negativo, no es realmente «menos un protón». La razón está en que la inversión de signo se refiere solo a los números cuánticos *aditivos*, mientras que la noción de *masa* no es aditiva en la teoría física moderna. Esta cuestión será explicada de manera más detallada en §18.7. «Menos un protón» tendría que ser un antiprotón cuya masa fuera el negativo del valor de la masa de un protón ordinario. Pero la masa de una partícula física real no puede ser negativa. Un antiprotón tiene la misma masa que un protón ordinario, que es una masa positiva. Veremos más adelante que, según las ideas de la teoría cuántica de campos, existen objetos denominados partículas «virtuales» para los que la masa (o, más correctamente, la energía) puede ser negativa. «Menos un protón» sería realmente un antiprotón virtual. Pero una partícula virtual no tiene una existencia independiente como la de una «partícula real».

Ahora nos haremos la correspondiente pregunta acerca de los números racionales. ¿Ha encontrado este sistema de números cualquier relevancia directa para el universo físico? Hasta donde sabemos, no parece que sea así, al menos en lo que respecta a la teoría convencional. Existen algunas curiosidades físicas[22] en las que la familia de los números racionales desempeña su papel, pero sería difícil sostener que esto revela algún papel físico fundamental para los números racionales. Por otra parte, pudiera ser que hubiera un papel especial para los racionales en las probabilidades mecanocuánticas fundamentales (donde una probabilidad racional representa una elección entre alternativas, cada una de las cuales implica solo un número finito de posibilidades). Este tipo de cosas desempeña un papel en la teoría de las redes de espín, como se describirá brevemente en §32.6. De momento, el estatus adecuado de estas ideas no está claro.

Pese a todo, existen otros tipos de números que, según la teoría aceptada, sí parecen desempeñar un papel fundamental en la marcha del universo. Los más importantes y sorprendentes de estos son los *nú-*

meros complejos, en los que se introduce la cantidad aparentemente mística $\sqrt{-1}$, normalmente denotada por «i», y se añade al sistema de los números reales. Encontrados por primera vez en el siglo XVI, pero tratados con desconfianza durante cientos de años, la utilidad matemática de los números complejos impresionó poco a poco a la comunidad matemática en un grado cada vez mayor, hasta que los números complejos se convirtieron en un ingrediente indispensable y casi mágico de nuestro pensamiento matemático. Y ahora encontramos que no solo son fundamentales para las matemáticas: estos extraños números desempeñan también un papel extraordinario y muy básico en el funcionamiento del universo físico en sus escalas más ínfimas. Esto produce asombro, y como ejemplo de la convergencia entre ideas matemáticas y los mecanismos más profundos del universo físico es más sorprendente incluso que el sistema de números reales que hemos estado considerando en esta sección. Vayamos ahora a estos números notables.

Notas

Sección 3.1

3.1. Las notaciones $>$, $<$, $\geq$, $\leq$, frecuentemente utilizadas en este libro, representan, respectivamente, «es mayor que», «es menor que», «es mayor o igual que» y «es menor o igual que».

3.2. Algunos lectores quizá conozcan un argumento aparentemente más corto que empieza exigiendo que a/b esté «en sus menores términos» (i.e., que a y b no tengan factores comunes). Sin embargo, esto supone que esta expresión en términos menores existe siempre, lo que, aunque sea completamente cierto, necesita ser demostrado. Encontrar una expresión en términos menores para una fracción dada A/B (implícita o explícitamente, utilizando el algoritmo de Euclides, pongamos por caso; véanse, por ejemplo, Hardy y Wright, 1945, p. 134; Davenport, 1952, p. 26; Littlewood, 1949, cap. 4, y Penrose, 1989, cap. 2) implica un razonamiento similar al dado en el texto, aunque más complicado.

3.3. Podría objetarse que resulta algo curioso utilizar números reales en la demostración anterior, puesto que los «racionales reales» (i.e., cocientes de reales) serían de nuevo números reales simplemente. Esto no invalida, sin embargo, lo que se acaba de decir. Puede señalarse que, en el argumento original, a y b se tomaban además enteros, y no racionales en sí mismos. En efecto, si a y b fueran meramente racionales, el argumento fallaría en la parte de la «secuencia decreciente», incluso si el propio resultado siguiera siendo cierto.

Sección 3.2

3.4. En una ojeada informal, expresiones como $a + (b + (c + (d + \ldots)^{-1})^{-1})^{-1}$ pueden parecer bastante extrañas. Sin embargo, son muy naturales en el contexto del pensamiento griego antiguo (aunque los griegos no utilizaban esta notación concreta). El *algoritmo de Euclides* se ha mencionado en la nota 3.2 en el contexto de encontrar la forma en términos más sencillos de una fracción. El algoritmo de Euclides (cuando se desenmaraña) lleva precisamente a dicha expresión como fracción continua. Los griegos aplicarían este mismo procedimiento a la razón de dos longitudes geométricas. En el caso más general, el resultado sería una fracción continua *infinita*, del tipo considerado aquí.

3.5. Para más información (con demostraciones) sobre fracciones continuas, véase la elegante exposición dada en el capítulo 4 de Davenport (1952). Puede señalarse que en ciertos aspectos la representación como fracciones continuas de los números reales es más profunda e interesante que la normal en términos de desarrollos decimales, y encuentra aplicaciones en muchas áreas diferentes de las matemáticas modernas, incluida la geometría hiperbólica examinada en §§2.4,5. Por otra parte, las fracciones continuas no son en absoluto apropiadas para (la mayor parte de) los cálculos prácticos, pues la representación decimal convencional es mucho más fácil de utilizar.

3.6. Los irracionales cuadráticos se denominan así porque aparecen en la solución de una ecuación cuadrática general

$$Ax^2 + Bx + C = 0,$$

con A distinto de cero, cuya solución es

$$-\frac{B}{2A} + \sqrt{\left(\frac{B}{2A}\right)^2 - \frac{C}{A}} \text{ y } -\frac{B}{2A} - \sqrt{\left(\frac{B}{2A}\right)^2 - \frac{C}{A}},$$

donde, para permanecer dentro del dominio de los números reales, debemos tener B^2 mayor que $4AC$. Cuando A, B y C son números enteros o racionales, y no hay solución racional para la ecuación, las soluciones son irracionales cuadráticos.

3.7. El profesor Stelios Negrepontis me informa de que esta evidencia se encuentra en el diálogo platónico *El político*, el tercero en la «trilogía» *Teeteto-El sofista-El político*.Véase Negrepontis (2000).
3.8. Para una exposición del pensamiento griego antiguo sobre la naturaleza del espacio, véase Sorabji (1983, 1988).
3.9. Véanse Hardy (1914), Conway (1976) y Burkill (1962).

Sección 3.3

3.10. La notación científica «10^{12}» para «un millón de millones» hace uso de exponentes, como se describe en las notas 1.2 y 2.4. En este libro tenderé a evitar términos verbales tales como «millón», y en especial «billón», dando preferencia a esta notación científica mucho más clara. La palabra «billón» es particularmente confusa, pues en su uso estadounidense —ahora comúnmente adoptado también en el Reino Unido— «billón» se refiere a 10^9, mientras que en el uso más antiguo (y más lógico) en el Reino Unido, de acuerdo con la mayoría de las restantes lenguas europeas, se refiere a 10^{12}. Los exponentes negativos, tales como en 10^{-6} (que se refiere a una «millonésima»), se utilizan también aquí de acuerdo con la notación científica normal.

La distancia 10^{12} metros es aproximadamente 7 veces la separación entre la Tierra y el Sol. Esta es aproximadamente la distancia al planeta Júpiter, aunque no era conocida en tiempos de Euclides y se hubiera conjeturado mucho menor.
3.11. Véase, por ejemplo, Russell (1903), cap. 4.
3.12. Schrödinger (1952), pp. 30-31.
3.13. Véase Stachel (1995).
3.14. Einstein (1955), p. 166.
3.15. Véanse, por ejemplo, Snyder (1947), Schild (1949) y Ahmavaara (1965).
3.16. Véanse Ashtekar (1986), Ashtekar y Lewandowski (2004), Smolin (1998, 2001) y Rovelli (1998, 2003).

Sección 3.4

3.17. La noción de «número ordinal», proporcionada aquí en el caso finito, se extiende también a números ordinales *infinitos*, de los que el menor es el «ω» de Cantor, que es la colección ordenada de todos los ordinales finitos.
3.18. No obstante, esta noción de «constructo» no debería tomarse en un sentido demasiado fuerte. En §16.6 encontraremos que hay ciertos números reales (de hecho, la mayoría de ellos) que son inaccesibles mediante cualquier procedimiento computacional.

Sección 3.5

3.19. El físico irlandés George Johnstone Stoney fue el primero, en 1874, en dar una (cruda) estimación de la carga eléctrica básica y, en 1891, acuñó el término «electrón» para esta unidad fundamental. En 1909, el físico estadounidense Robert Andrews Millikan diseñó su famoso experimento de la «gota de aceite», que demostró de forma precisa que la carga en cuerpos eléctricamente cargados (las gotas de aceite en su experimento) se da en múltiplos enteros de un valor bien definido: la carga del electrón.
3.20. En 1959, R. A. Lyttleton y H. Bondi propusieron que una ligerísima diferencia entre la carga del protón y (menos) la del electrón, del orden de una parte en 10^{18}, podría explicar la expansión del universo (para la cual, véanse §§27.11,13, y el capítulo 28). Véase Lyttleton y Bondi (1959). Por desgracia para esta teoría, semejante discrepancia fue pronto refutada en varios experimentos. De todas formas, esta idea proporcionó un excelente ejemplo de pensamiento creativo.
3.21. Aquí estoy distinguiendo los números cuánticos «aditivos» de los números que los físicos llaman «multiplicativos», a los que llegaremos en §5.5.
3.22. Por ejemplo, en el «efecto Hall cuántico fraccionario» se observa que los números fraccionarios desempeñan un papel clave; véase, por ejemplo, Fröhlich y Pedrini (2000).

4

Los mágicos números complejos

4.1. El mágico número «i»

¿Cómo es posible que -1 tenga una raíz cuadrada? El cuadrado de un número positivo es siempre positivo, y el cuadrado de un número negativo es también positivo (y el cuadrado de 0 es simplemente 0, de modo que de poco nos sirve aquí). Parece imposible que podamos encontrar un número cuyo cuadrado sea realmente negativo. Sin embargo, esta es una situación similar a la que hemos visto antes, cuando hemos establecido que 2 no tiene una raíz cuadrada dentro del sistema de los números racionales. En ese caso hemos resuelto la situación ampliando nuestro sistema de números desde los racionales a un sistema mayor, y hemos establecido el sistema de los reales. Tal vez el mismo truco funcione de nuevo.

Realmente lo hace. De hecho, lo que tenemos que hacer es algo mucho más fácil y menos drástico que el paso de los racionales a los reales. (Raphael Bombelli introdujo el procedimiento en 1572 en su obra *L'Algebra*, siguiendo los encuentros originales de Girolamo Cardano con los números complejos en su *Ars magna* de 1545.) Todo lo que tenemos que hacer es introducir una simple cantidad, llamada «i» cuyo cuadrado sea -1, y añadirla al sistema de los reales, permitiendo combinaciones de i con números reales para formar expresiones de la forma

$$a + \mathrm{i}b,$$

donde a y b son números reales arbitrarios. Cualquiera de estas combinaciones se denomina un *número complejo*. Es fácil ver cómo se suman los números complejos:

$$(a + \mathrm{i}b) + (c + \mathrm{i}d) = (a + c) + \mathrm{i}(b + d),$$

que es de la misma forma que antes (pero ahora los números reales $a + c$ y $b + d$ toman el lugar de los «a» y «b» que teníamos en nuestra expresión original). ¿Qué pasa con la multiplicación? Es casi igual de fácil. Vamos a encontrar el producto de $a + \mathrm{i}b$ por $c + \mathrm{i}d$. En primer lugar, multiplicamos simplemente estos factores, desarrollando la expresión mediante el uso de las reglas ordinarias del álgebra:[1]

$$\begin{aligned}(a + \mathrm{i}b)(c + \mathrm{i}d) &= ac + \mathrm{i}bc + a\mathrm{i}d + \mathrm{i}b\mathrm{i}d \\ &= ac + \mathrm{i}(bc + ad) + \mathrm{i}^2bd.\end{aligned}$$

Pero $\mathrm{i}^2 = -1$, de modo que podemos reescribir esto como

$$(a + \mathrm{i}b)(c + \mathrm{i}d) = (ac - bd) + \mathrm{i}(bc + ad),$$

que de nuevo tiene la misma forma que nuestro $a + \mathrm{i}b$ original, pero donde $ac - bd$ toma el lugar de a y $bc + ad$ toma el lugar de b.

Es bastante fácil restar dos números complejos, pero ¿qué pasa con la división? Recordemos que en la aritmética ordinaria nos está permitido dividir por cualquier número real distinto de cero. Tratemos ahora de dividir el número complejo $a + \mathrm{i}b$ por el número complejo $c + \mathrm{i}d$. Debemos tomar este último distinto de cero, lo que significa que los números reales c y d no pueden ambos ser cero. Entonces, $c^2 + d^2 > 0$, y por consiguiente $c^2 + d^2 \neq 0$, de modo que podemos dividir por $c^2 + d^2$. Es un ejercicio directo[4.1] comprobar (multiplicando ambos miembros de la expresión por $c + \mathrm{i}d$) que

$$\frac{(a + \mathrm{i}b)}{(c + \mathrm{i}d)} = \frac{ac + bd}{c^2 + d^2} + \mathrm{i}\frac{bc - ad}{c^2 + d^2}.$$

Esta expresión es de la misma forma que antes, de modo que es de nuevo un número complejo.

Cuando nos acostumbramos a jugar con estos números complejos, dejamos de pensar en $a + \mathrm{i}b$ como un «par» de objetos, a saber, los dos números reales a y b; ahora pensamos en $a + \mathrm{i}b$ como un «objeto» completo autónomo, de modo que podríamos utilizar una sola letra, diga-

[4.1] Hágalo. Alternativamente, ¿puede comprobarlo multiplicando ambos miembros por $(c - \mathrm{i}d)$?

mos z, para denotar el número complejo global $z = a + ib$. Puede comprobarse que todas las reglas usuales del álgebra son satisfechas por los números complejos.[4.2] De hecho, todo esto es mucho más sencillo que comprobar cada una de estas cosas para los números reales. (Para esa comprobación, suponemos que ya estábamos convencidos de que las reglas del álgebra son satisfechas por las fracciones, y luego tenemos que utilizar «cortaduras» de Dedekind para demostrar que las reglas siguen funcionando para los números reales.) Desde este punto de vista, resulta bastante extraño que los números complejos se vieran con recelo durante mucho tiempo, mientras que la mucho más complicada ampliación de las racionales a los reales había sido aceptada sin problemas después de la época de la Grecia antigua.

Presumiblemente, el motivo de este recelo era que la gente no podía «ver» que los números complejos se les presentasen de un modo obvio en el mundo físico. En el caso de los números reales existía la sensación de que las distancias, los tiempos y otras magnitudes físicas proporcionaban la realidad que tales números requerían; pero los números complejos parecían ser entidades meramente «inventadas», sacadas de la imaginación de los matemáticos que deseaban números con un alcance mayor que los que ya conocían. Pero habría que recordar de §3.3 que la conexión que tienen los números reales matemáticos con aquellos conceptos físicos de longitud o tiempo no es tan clara como habíamos imaginado. No podemos ver directamente los mínimos detalles de una cortadura de Dedekind, ni está claro que realmente existan en la naturaleza longitudes o tiempos arbitrariamente grandes o arbitrariamente pequeños. Se podría decir que los denominados «números reales» son tan producto de la imaginación de los matemáticos como lo son los números complejos. Pese a todo, encontraremos que los números complejos, tanto como los reales, y quizá más incluso, componen una notable unidad con la naturaleza. Es como si la propia naturaleza estuviera tan impresionada por el alcance y consistencia del sistema de los números complejos como lo estamos nosotros, y hubiera confiado a estos números las operaciones detalladas de su mundo en

[4.2] Compruébelo, siendo las reglas relevantes $w + z = z + w$, $w + (u + z) = (w + u) + z$, $wz = zw$, $w(uz) = (wu)z$, $w(u + z) = wu + wz$, $w + 0 = w$, $w1 = w$.

sus escalas más minúsculas. En los capítulos 21-23 veremos en detalle cómo funciona esto.

Además, mencionar solo el alcance y la consistencia de los números complejos no hace justicia a este sistema. Hay algo más que, en mi opinión, solo puede ser calificado de «mágico». En lo que queda de este capítulo y en el siguiente, me propongo transmitir al lector algo del sabor de esta magia. Más adelante, en los capítulos 7-9, seremos testigos de nuevo de esta magia de los números complejos en algunas de sus más sorprendentes e inesperadas manifestaciones.

Durante los cuatro siglos que hace que se conoce el sistema de los números complejos, muchas cualidades mágicas se han ido revelando poco a poco. Pero esta es una magia que se percibía dentro de las matemáticas, y ofrecía realmente una utilidad y una profundidad de intuición matemática que no podía conseguirse solo con el uso de los reales. No había ninguna razón para esperar que el mundo físico estuviera interesado en ello. Y durante los aproximadamente trescientos cincuenta años transcurridos desde la época en que dichos números fueron introducidos en las obras de Cardano y Bombelli, la magia del sistema de los números complejos solo fue percibida a través de su papel matemático. Para todos aquellos que se habían mostrado recelosos de los números complejos habría sido sin duda una sorpresa encontrar que, según la física de los últimos tres cuartos del siglo XX, las leyes que gobiernan el comportamiento del mundo en sus escalas más minúsculas dependen fundamentalmente del sistema de los números complejos.

Estos temas serán capitales para algunas de las secciones posteriores de este libro (especialmente en los capítulos 21-23, 26 y 31-33). Por el momento nos concentraremos en una parte de la magia matemática de los números complejos y dejaremos su magia física para más adelante. Recordemos que todo lo que hemos hecho es exigir que -1 tenga una raíz cuadrada, además de exigir que se mantengan las leyes usuales de la aritmética, y hemos descubierto que dichas exigencias pueden satisfacerse de forma consistente. Esto parece algo bastante simple de hacer. ¡Pero pasemos a la magia!

4.2. Resolviendo ecuaciones con números complejos

En lo que sigue, será necesario introducir una notación algo más matemática que la utilizada hasta ahora. Pido disculpas por ello. Sin embargo, difícilmente pueden transmitirse ideas matemáticas serias sin cierta cantidad de notación. Soy consciente de que habrá muchos lectores que se sientan incómodos con ello. Mi consejo para tales lectores es que tan solo lean el texto y no se preocupen demasiado por tratar de entender las ecuaciones. Cuando menos, lean por encima las diversas fórmulas y sigan adelante. De hecho, habrá numerosas expresiones matemáticas serias desperdigadas a lo largo de este libro, especialmente en algunos de los últimos capítulos. Creo que, con el tiempo, empezarán a calar ciertas dosis de conocimiento si se hace un pequeño intento por entender lo que todas las expresiones significan realmente. Así lo espero, porque la magia de los números complejos es un milagro que vale la pena apreciar. Si usted puede entender la notación matemática, entonces mucho mejor.

Antes de nada, podemos preguntarnos si otros números tienen raíces cuadradas. ¿Qué pasa, por ejemplo, con -2? Esto es fácil. El número complejo $i\sqrt{2}$ ciertamente tiene -2 como cuadrado, y también lo tiene $-i\sqrt{2}$. Más aún, para cualquier número real positivo a, el número complejo $i\sqrt{a}$ tiene como cuadrado $-a$, y lo mismo sucede con $-i\sqrt{a}$. No hay aquí ninguna magia. Pero ¿qué pasa con el número complejo más general $a + ib$ (donde a y b son reales)? Encontramos que el número complejo

$$\sqrt{\frac{1}{2}\left(a+\sqrt{a^2+b^2}\right)}+i\sqrt{\frac{1}{2}\left(-a+\sqrt{a^2+b^2}\right)}$$

tiene como cuadrado $a + ib$ (y lo mismo sucede con su negativo).[4.3] Así pues, vemos que incluso si solo añadimos una raíz cuadrada para una cantidad (a saber -1), ¡encontramos que todos los números del sistema resultante tienen ahora automáticamente una raíz cuadrada! Esto es muy diferente de lo que sucedía al pasar de los racionales a los reales. En este caso, la mera introducción de la cantidad $\sqrt{2}$ en el sistema de los racionales no nos hubiera llevado casi a ninguna parte.

[4.3] Compruébelo.

Pero esto es solo el principio. Podemos preguntarnos sobre raíces cúbicas, raíces quintas, raíces 999-ésimas, raíces π-ésimas —e incluso raíces i-ésimas—. Milagrosamente, encontramos que para cualquier raíz compleja que escojamos y cualquier número complejo al que la apliquemos (excluyendo el 0), hay siempre una solución compleja a este problema. (De hecho, normalmente habrá varias soluciones diferentes al problema, como veremos dentro de poco. Antes he señalado que para raíces cuadradas teníamos dos soluciones, pues el negativo de la raíz cuadrada de un número complejo z es también una raíz cuadrada de z. Para raíces más altas hay más soluciones; véase §5.4.)

Apenas estamos arañando todavía en la superficie de la magia de los números complejos. Lo que acabo de afirmar es realmente muy fácil de establecer (una vez que tengamos la noción del *logaritmo* de un número complejo, como veremos en el capítulo 5). Algo más notable es el denominado «teorema fundamental del álgebra», que afirma que cualquier ecuación polinómica, tal como

$$1 - z + z^4 = 0$$

o

$$\pi + iz - \sqrt{417}z^3 + z^{999} = 0,$$

debe tener soluciones con números complejos. Dicho de forma más explícita, siempre habrá una solución (y normalmente varias diferentes) para cualquier ecuación de la forma

$$a_0 + a_1z + a_2z^2 + a_3z^3 + \ldots + a_nz^n = 0,$$

donde a_0, a_1, a_2, a_3, ..., a_n son números complejos dados, con el a_n tomado no nulo.[2] (Aquí n puede ser cualquier entero positivo que escojamos, tan grande como queramos.) Deberíamos recordar, por comparación, que «i» fue introducida simplemente para proporcionar una solución a la ecuación particular

$$1 + z^2 = 0.$$

¡Todo lo demás lo obtenemos *gratis*!

Antes de seguir adelante vale la pena mencionar el problema que interesaba a Cardano, aproximadamente en 1539, cuando encontró por

primera vez los números complejos y captó un indicio de otro aspecto de sus propiedades mágicas inherentes. Dicho problema era, en efecto, encontrar una expresión para la solución general de una ecuación (real) *cúbica* (i.e., $n = 3$, en el caso anterior). Cardano encontró que la cúbica general podía reducirse mediante una sencilla transformación a la forma

$$x^3 = 3px + 2q$$

Aquí p y q son números reales (y he vuelto a utilizar x en la ecuación, en lugar de z, para indicar que ahora estamos interesados en soluciones reales en lugar de complejas). La solución completa de Cardano (tal como la publicó en 1545 en su libro *Ars magna*) parece haber sido desarrollada a partir de una solución parcial anterior que había aprendido en 1539 de Niccolò Fontana («Tartaglia»), aunque esta solución parcial (y quizá incluso la solución completa) había sido encontrada previamente (antes de 1526) por Scipione del Ferro.[3] La solución de (Del Ferro-)Cardano era en esencia la siguiente (escrita en notación moderna):

$$x = (q + w)^{\frac{1}{3}} + (q - w)^{\frac{1}{3}},$$

donde

$$\mathrm{w} = (q^2 - p^3)^{\frac{1}{2}}.$$

Esta ecuación no plantea ningún problema fundamental dentro del sistema de los números reales si

$$q^2 \geq p^3.$$

En este caso solo hay una solución real x a la ecuación, y viene dada correctamente por la fórmula de (Del Ferro-)Cardano expuesta antes. Pero si

$$q^2 < p^3,$$

el denominado *caso irreducible*, entonces, aunque ahora hay *tres* soluciones reales, la fórmula incluye la raíz cuadrada del número *negativo* $q^2 - p^3$ y por ello no puede ser utilizada sin introducir los números complejos. De hecho, como Bombelli demostró más tarde (en el capítulo 2 de *L'Algebra* de 1572), si admitimos los números complejos, entonces las tres soluciones reales están correctamente expresadas por la fórmula.[4] (Esto tiene sentido porque la expresión nos proporciona dos números complejos sumados, y las partes que contienen «i» se cancelan en la suma para dar una respuesta real.)[5] Lo que hay de misterioso en esto es que, aunque pudiera parecer que el problema no tiene nada que ver con los números complejos —pues la ecuación tiene coeficientes reales y todas sus soluciones son reales (en este caso, «irreducible»)—, tenemos que viajar por este territorio en apariencia ajeno del mundo de los números complejos para que la fórmula nos permita regresar con nuestras soluciones puramente reales. Si nos hubiéramos restringido al recto y estrecho camino «real», habríamos vuelto con las manos vacías. (Resulta irónico que solo puede haber soluciones *complejas* a la ecuación original en aquellos casos en que la fórmula *no* implica necesariamente este viaje por los complejos.)

4.3. Convergencia de series de potencias

A pesar de estos hechos notables, aún no hemos ido muy lejos en la magia de los números complejos. ¡Queda mucho más por venir! Por ejemplo, algo en que los números complejos resultan verdaderamente inestimables es en ofrecer una comprensión del comportamiento de lo que se denominan *series de potencias*. Una serie de potencias es una suma *infinita* de la forma

$$a_0 + a_1x + a_2x^2 + a_3x^3 + \ldots$$

Puesto que esta suma implica un número infinito de términos, puede darse el caso de que la serie *diverja*, lo que significa que no se asienta en un valor particular finito a medida que se suman más términos. Por ejemplo, consideremos la serie

$$1 + x^2 + x^4 + x^6 + x^8 + \ldots$$

(donde he tomado $a_0 = 1$, $a_1 = 0$, $a_2 = 1$, $a_3 = 0$, $a_4 = 1$, $a_5 = 0$, $a_6 = = 1, \ldots$). Si hacemos $x = 1$, entonces, sumando los términos sucesivamente, obtenemos

$$1,\quad 1+1=2,\quad 1+1+1=3,$$
$$1+1+1+1=4,\quad 1+1+1+1+1=5,\text{ etc.},$$

y vemos que la serie no tiene posibilidad de asentarse en un valor particular finito, es decir, es *divergente*. Las cosas son aún peores si probamos con $x = 2$, por ejemplo, puesto que ahora los términos individuales son más grandes, y sumando términos sucesivamente obtenemos

$$1,\quad 1+4=5,\quad 1+4+16=21,\quad 1+4+16+64=85,\text{ etc.},$$

que claramente diverge. Por el contrario, si hacemos $x = 1/2$, obtenemos

$$1,\quad 1+\frac{1}{4}=\frac{5}{4},\quad 1+\frac{1}{4}+\frac{1}{16}=\frac{21}{16},\quad 1+\frac{1}{4}+\frac{1}{16}+\frac{1}{64}=\frac{85}{64},\text{ etc.},$$

y resulta que estos números se acercan cada vez más al valor límite 4/3, de modo que la serie es ahora *convergente*.

Con esta serie no es difícil darse cuenta de que, en cierto sentido, hay una razón subyacente por la que la serie no puede sino divergir para $x = 1$ y $x = 2$, mientras que converge para $x = 1/2$ para dar la respuesta 4/3. En efecto, podemos escribir explícitamente la «respuesta» a la suma de toda la serie, encontrando[4.4]

$$1 + x^2 + x^4 + x^6 + x^8 + \ldots = (1 - x^2)^{-1}.$$

Cuando sustituimos $x = 1$, encontramos que esta respuesta es $(1 - 1^2)^{-1} = 0^{-1}$, que es «infinito»,[6] y esto nos permite entender por qué la serie tiene que divergir para ese valor de x. Cuando sustituimos $x = 1/2$, la respuesta es $(1 - 1/4)^{-1} = 4/3$, y la serie converge realmente a este valor concreto, tal como se ha establecido antes.

Todo esto parece muy razonable. Pero ¿qué pasa con $x = 2$? Ahora hay una «respuesta» dada por la fórmula explícita, a saber, $(1 - 4)^{-1} = -1/3$, aunque no parece que obtengamos dicho valor sumando simplemente los términos de la serie. Difícilmente podríamos obtener esta respuesta porque estamos sumando cantidades positivas, mientras que $-1/3$ es negativo. La razón de que la serie diverja es que, cuando $x = 2$, cada término es mayor que el término correspondiente para $x = 1$, de

[4.4] ¿Puede ver la manera de comprobar esta expresión, de un modo algebraico formal?

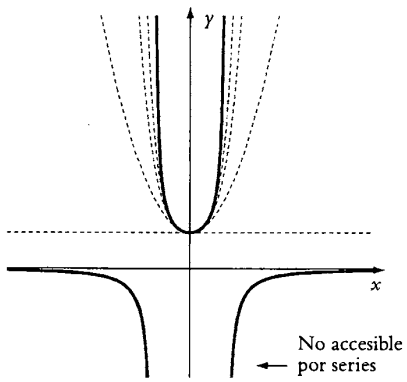


Fig. 4.1. Se representan las respectivas sumas parciales $1, 1 + x^2, 1 + x^2 + x^4, 1 + x^2 + x^4 + x^6$ (líneas de trazos) de la serie para $(1 - x^2)^{-1}$, que ilustran la convergencia de la serie para $(1 - x^2)^{-1}$ para $|x| < 1$ y la divergencia para $|x| > 1$.

modo que la divergencia para $x = 2$ se sigue lógicamente de la divergencia para $x = 1$. En el caso $x = 2$ no se trata de que la «respuesta» sea realmente infinita, sino de que no podemos llegar a esta respuesta intentando sumar la serie directamente. En la Fig. 4.1 he representado las sumas parciales de la serie (i.e., las sumas hasta cierto número finito de términos), sucesivamente hasta cuatro términos, al lado de la «respuesta» $(1 - x^2)^{-1}$ y podemos ver que, con tal de que x esté comprendida estrictamente[7] entre los valores -1 y $+1$, las curvas que muestran estas sumas parciales convergen realmente a esta respuesta, a saber $(1 - x^2)^{-1}$, como cabía esperar. Pero fuera de este intervalo, la serie simplemente diverge y no llega realmente a ningún valor finito.

A modo de pequeña digresión, será útil abordar aquí una cuestión que tendrá importancia para nosotros más adelante. Planteemos la siguiente pregunta: la ecuación que obtenemos haciendo $x = 2$ en la expresión anterior, es decir,

$$1 + 2^2 + 2^4 + 2^6 + 2^8 + \ldots = (1 - 2^2)^{-1} = -\frac{1}{3},$$

¿tiene sentido realmente? El gran matemático del siglo XVIII Leonhard Euler solía escribir ecuaciones como esta, y llegó a ser una moda hacer bromas a su costa por sostener tales absurdos, aunque se le podrían per-

donar sobre la base de que en aquellos primeros días no se entendía adecuadamente nada relativo a la «convergencia» de series y cosas similares. De hecho, lo cierto es que el tratamiento matemático riguroso de las series no llegó hasta finales del siglo XVIII y principios del XIX, con la obra de Augustin Cauchy y otros. Además, según este tratamiento riguroso, la ecuación anterior sería clasificada oficialmente como «sin sentido». Pese a todo, creo que es importante darse cuenta de que, en el sentido adecuado, Euler sabía bien lo que estaba haciendo cuando escribía absurdos aparentes de esta naturaleza, y que hay sentidos en los que la ecuación anterior debe considerarse «correcta».

En matemáticas es realmente imperativo dejar absolutamente claro que las ecuaciones que se obtienen tienen un sentido estricto y preciso. Sin embargo, es asimismo importante no ser insensible a «lo que sucede fuera de escena», que puede, en última instancia, llevar a intuiciones más profundas. Es fácil perder de vista tales cosas si nos atenemos de forma demasiado rígida a lo que parece ser estrictamente lógico, como es el hecho de que la suma de los términos positivos $1 + 4 + 16 + 64 + 256 + \ldots$ no puede ser $-1/3$ de ningún modo. Como ejemplo pertinente, recordemos el absurdo lógico de encontrar una solución real a la ecuación $x^2 + 1 = 0$. No hay solución; pero si nos quedamos en esto, perderemos todas las intuiciones profundas que proporciona la introducción de los números complejos. Un comentario similar es aplicable a lo absurdo de una solución racional a $x^2 = 2$. De hecho, es perfectamente posible dar un sentido matemático a la respuesta «$-1/3$» a la serie infinita anterior, pero hay que tener cuidado con las reglas que nos dicen lo que está permitido y lo que no está permitido. No es mi intención discutir aquí estas cuestiones en detalle,[8] pero habría que señalar que en la física moderna, especialmente en el área de la teoría cuántica de campos, es frecuente encontrar series divergentes de este tipo (véanse especialmente §§26.7,9 y §§31.2,13). Es una cuestión muy delicada la de decidir si las «respuestas» que se obtienen de este modo son realmente significativas y, además, realmente correctas. A veces se obtienen respuestas muy precisas manipulando tales expresiones divergentes, y en ocasiones son sorprendentemente confirmadas al compararlas con experimentos físicos reales. Otras veces, por el contrario, no se tiene tanta suerte. Estas delicadas cuestiones

desempeñan papeles importantes en las teorías físicas actuales y son muy relevantes en nuestros intentos de evaluarlas. El punto de relevancia inmediata para nosotros aquí es que el «sentido» que podemos ser capaces de atribuir a semejantes expresiones aparentemente sin significado depende con frecuencia, y de un modo esencial, de las propiedades de los números complejos.

Volvamos ahora a la cuestión de la convergencia de series, y tratemos de ver cómo encajan los números complejos en la imagen. Para ello, consideremos una función tan solo ligeramente diferente de $(1 - x^2)^{-1}$, a saber $(1 + x^2)^{-1}$, y tratemos de ver si tiene un desarrollo en serie de potencias razonable. Ahora parece haber una mejor oportunidad de convergencia completa, puesto que $(1 + x^2)^{-1}$ se mantiene suave y finita sobre todo el intervalo de los números reales. Existe, de hecho, una serie de potencias de apariencia sencilla para $(1 + x^2)^{-1}$, tan solo ligeramente diferente de la que teníamos antes, a saber

$$1 - x^2 + x^4 - x^6 + x^8 - \ldots = (1 + x^2)^{-1},$$

donde la diferencia es meramente un cambio de signo en términos alternos.[4.5] En la Fig. 4.2 he representado las sumas parciales de la serie, sucesivamente hasta cinco términos, de modo similar a lo que hemos hecho antes, junto con esta respuesta $(1 + x^2)^{-1}$. Lo que parece sorprendente es que las sumas parciales siguen convergiendo a la respuesta solo en el intervalo estricto entre los valores -1 y $+1$. Parece que vamos a obtener una divergencia fuera de este intervalo, incluso si la respuesta no tiende ni mucho menos a infinito, a diferencia de nuestro caso anterior. Podemos comprobar esto de forma explícita utilizando los mismos tres valores $x = 1$, $x = 2$, $x = 1/2$ que hemos utilizado antes, encontrando que, como antes, la convergencia ocurre solo en el caso $x = 1/2$, para el que la respuesta da correctamente el valor límite $4/5$ para la suma de la serie entera:

$$x = 1\text{: } 1, 0, 1, 0, 1, 0, 1, \text{etc.},$$
$$x = 2\text{: } 1, -3, 13, -51, 205, -819, \text{etc.},$$
$$x = \frac{1}{2}\text{: } 1, \frac{3}{4}, \frac{13}{16}, \frac{51}{64}, \frac{205}{256}, \frac{819}{1.024}, \text{etc.}$$

[4.5] ¿Puede ver una razón elemental para esta relación simple entre las dos series?

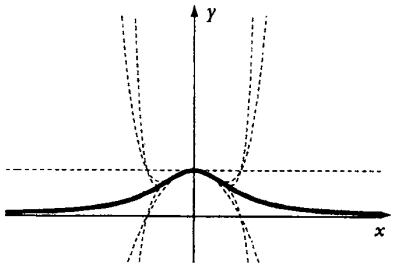


Fig. 4.2. Del mismo modo se representan las sumas parciales $1, 1-x^2, 1-x^2+x^4, 1-x^2+x^4-x^6, 1-x^2+x^4-x^6+x^8$, de la serie para $(1+x^2)^{-1}$, y de nuevo hay convergencia para $|x|<1$ y divergencia para $|x|>1$, pese al hecho de que la función tiene un comportamiento perfectamente bueno en $x=\pm 1$.

Advertimos que la «divergencia» en el primer caso consiste simplemente en que las sumas parciales de la serie no llegan nunca a asentarse, aunque no divergen realmente a infinito.

Así pues, si solo se consideran números reales, hay una discrepancia enigmática entre sumar realmente la serie y pasar directamente a la «respuesta» que se supone que representa la suma infinita de la serie. Las sumas parciales simplemente «despegan» (o más bien oscilan incontroladamente arriba y abajo) precisamente en los mismos lugares (a saber, $x=\pm 1$) donde surgían problemas en el caso anterior, aunque ahora la supuesta respuesta a la suma infinita, a saber $(1+x^2)^{-1}$, no muestra ninguna característica notable en dichos lugares. Encontraremos la solución a este misterio si examinamos valores *complejos* de esta función en lugar de restringir nuestra atención a los reales.

4.4. El plano complejo de Caspar Wessel

Para ver lo que está pasando aquí, será importante utilizar la ahora estándar representación *geométrica* de los números complejos en el plano euclídeo. Caspar Wessel en 1797, Jean Robert Argand en 1806, John Warren en 1828 y Carl Friedrich Gauss mucho antes de 1831 dieron todos ellos, de forma independiente, con la idea del *plano complejo* (véase la Fig. 4.3), en el que ofrecieron interpretaciones geométricas claras de las operaciones de adición y multiplicación de números complejos. En la Fig. 4.3 he utilizado ejes cartesianos estándar, con el eje x apuntando horizontalmente hacia la derecha y el eje y apuntando verticalmente hacia arriba. El número complejo

$$z = x + iy$$

viene representado como el punto del plano con coordenadas cartesianas (x, y).

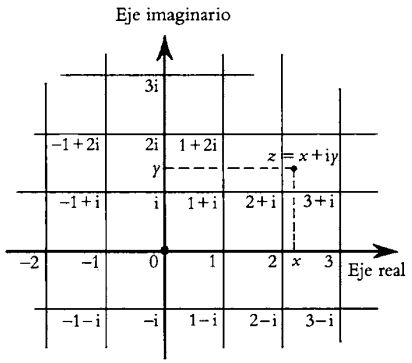


Fig. 4.3. El plano complejo de $z = x + iy$. En coordenadas cartesianas (x, y), el eje x dirigido horizontalmente hacia la derecha es el eje *real*; el eje y dirigido verticalmente hacia arriba es el eje *imaginario*.

Ahora tenemos que considerar un número real x como un caso particular del número complejo $z = x + iy$ en el que $y = 0$. Así pues, tenemos que considerar que el eje x de nuestro diagrama representa la *recta real* (i.e., la totalidad de los números reales, ordenados linealmente a lo largo de una recta). El plano complejo nos da así una representación gráfica directa de cómo se amplía el sistema de los números reales para convertirse en el sistema completo de los números complejos. Esta recta real se suele conocer como el «eje real» en el plano complejo. En correspondencia, el eje y se conoce como el «eje imaginario». Consiste en todos los múltiplos de i por un número real.

Volvamos ahora a nuestras dos funciones que hemos estado tratando de representar en términos de series de potencias. Tomábamos estas como funciones de la variable real x, a saber $(1-x^2)^{-1}$ y $(1+x^2)^{-1}$, pero ahora vamos a extender dichas funciones de modo que se aplican a una variable compleja z. No hay ningún problema en hacer esto:

simplemente escribimos estas funciones ampliadas como $(1 - z^2)^{-1}$ y $(1 + z^2)^{-1}$, respectivamente. En el caso de la primera función real $(1 - x^2)^{-1}$ éramos capaces de reconocer dónde empezaban los problemas de «divergencia», puesto que la función es *singular* (en el sentido de que se hace infinita) en los dos lugares $x = -1$ y $x = +1$; pero con $(1 + x^2)^{-1}$ no veíamos ninguna singularidad en estos lugares y, de hecho, ninguna singularidad real en absoluto. Sin embargo, en términos de la variable compleja z, vemos que estas dos funciones son mucho más parecidas. Hemos advertido las singularidades de $(1 - z^2)^{-1}$ en dos puntos $z = \pm 1$ a distancia unidad del origen a lo largo del eje real; pero ahora vemos que $(1 + z^2)^{-1}$ tiene también singularidades, a saber, en los dos lugares $z = \pm i$ (puesto que entonces $1 + z^2 = 0$), siendo estos los dos puntos situados a distancia unidad del origen en el eje imaginario.

Pero ¿qué tienen que ver estas singularidades complejas con la cuestión de la convergencia o divergencia de la serie de potencias correspondiente? Hay una respuesta sorprendente a esta pregunta. Ahora debemos considerar nuestras series de potencias como funciones de la variable compleja z, en lugar de la variable real x, y podemos preguntar por aquellas localizaciones de z en el plano complejo para las que la serie converge y aquellas para las que diverge. La notable respuesta general[9] para cualquier serie de potencias

$$a_0 + a_1 z + a_2 z^2 + a_3 z^3 + \ldots,$$

es que existe un círculo en el plano complejo, centrado en 0, llamado *círculo de convergencia*, con la propiedad de que si el número complejo z está estrictamente dentro del círculo, entonces la serie converge para dicho valor de z, mientras que si z está estrictamente fuera del círculo, entonces la serie diverge para dicho valor de z. (Si la serie converge o no cuando z está exactamente sobre el círculo es realmente una cuestión algo delicada en la que no entraremos aquí, aunque tiene cierta relevancia para las cuestiones a las que llegaremos en §§9.6,7.) En este enunciado estoy incluyendo las dos situaciones límite: aquella para la que la serie diverge para todos los valores no nulos de z, cuando el círculo de convergencia se ha contraído hasta un radio cero, y aquella en que converge para todo z, en cuyo caso el círculo se ha dilatado hasta un radio infinito. Para encontrar dónde está realmente el círculo

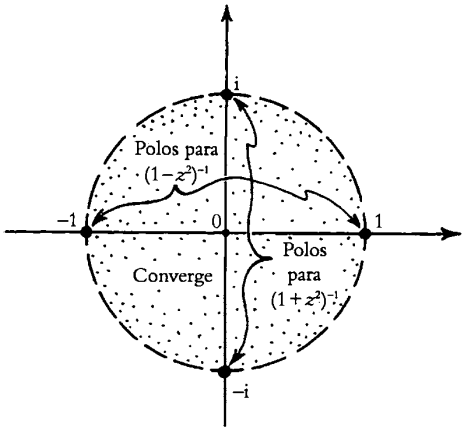


Fig. 4.4. En el plano complejo, las funciones $(1 - z^2)^{-1}$ y $(1 + z^2)^{-1}$ tienen el mismo círculo de convergencia, habiendo polos para la primera en $z = \pm 1$ y polos para la segunda en $z = \pm i$, todos ellos a la misma distancia (unidad) del origen.

de convergencia para una función dada, miramos en qué parte del plano complejo están localizadas las singularidades, y dibujamos el círculo más grande con centro en el origen $z = 0$ y que no contenga ninguna singularidad en su interior (i.e., lo dibujamos de forma que pase por la singularidad más próxima al origen).

En los casos particulares $(1 - z^2)^{-1}$ y $(1 + z^2)^{-1}$ que hemos estado considerando, las singularidades son de un tipo simple denominado *polos* (que surgen donde se anula un polinomio que aparece como divisor en una expresión). Aquí todos estos polos están a distancia unidad del origen, y vemos que el círculo de convergencia es, en ambos casos, el círculo unidad alrededor del origen. Los lugares donde dicho círculo corta al eje real son los mismos en ambos casos, a saber, los dos puntos $z = \pm 1$ (véase la Fig. 4.4). Esto explica por qué las dos funciones convergen y divergen en las mismas regiones, un hecho que no era manifiesto simplemente a partir de sus propiedades como funciones de variables reales. Así pues, los números complejos nos ofrecen ideas profundas sobre el comportamiento de series de potencias, ideas que sencillamente no están disponibles a partir de la consideración de su estructura en variable real.

4.5. Cómo se construye el conjunto de Mandelbrot

Para terminar este capítulo, consideremos otro tipo de cuestión convergencia/divergencia. Es la que subyace en la construcción de esa configuración extraordinaria, mencionada en §1.3 y mostrada en la Fig. 1.2, conocida como *el conjunto de Mandelbrot*. De hecho, este es solo un subconjunto del plano complejo de Wessel que puede definirse de un modo notablemente simple, si lo comparamos con la extrema complicación de dicho conjunto. Todo lo que tenemos que hacer es examinar las aplicaciones repetidas del reemplazamiento

$$z \mapsto z^2 + c,$$

donde c es un número complejo escogido. Consideremos c como un punto en el plano complejo y empecemos con $z = 0$. Entonces *iteramos* esta transformación (i.e., la aplicamos repetidamente una y otra vez) y vemos cómo se comporta el punto z en el plano complejo. Si escapa al infinito, entonces el punto c se marca en color blanco. Si z se mueve por cierta región restringida sin llegar nunca al infinito, entonces c se marca en color negro. ¡La región negra nos da el *conjunto de Mandelbrot*!

Describamos este procedimiento con un poco más de detalle. ¿Cómo procede la iteración? En primer lugar, fijamos c. Luego tomamos algún punto z y aplicamos la transformación, de modo que z se convierte en $z^2 + c$. Después la aplicamos otra vez, de modo que ahora reemplazamos la «z» en $z^2 + c$ por $z^2 + c$ y obtenemos $(z^2 + c)^2 + c$. A continuación reemplazamos la «z» en $z^2 + c$ por $(z^2 + c)^2 + c$, de modo que nuestra expresión se convierte en $((z^2 + c)^2 + c)^2 + c$. Luego continuamos reemplazando la «z» en $z^2 + c$ por $((z^2 + c)^2 + c)^2 + c$, y obtenemos $(((z^2 + c)^2 + c)^2 + c)^2 + c$, y así sucesivamente.

Veamos ahora qué sucede si empezamos en $z = 0$ e iteramos de esta manera. (Simplemente ponemos $z = 0$ en las expresiones anteriores.) Obtenemos la secuencia

$$0, c, c^2 + c, (c^2 + c)^2 + c, ((c^2 + c)^2 + c)^2 + c, \ldots$$

Esto nos da una sucesión de puntos en el plano complejo. (Si trabajáramos con un computador, no utilizaríamos las expresiones algebraicas anteriores, sino que procederíamos de forma puramente numérica para

cada elección individual del número complejo c. Resulta mucho más «barato» computacionalmente hacer la aritmética desde el principio cada vez.) Ahora, para cualquier valor dado de c, puede suceder una de dos cosas: (i) los puntos de la secuencia se alejan a distancias cada vez mayores, i.e., la secuencia *no está acotada*, o (ii) cada uno de los puntos yace a una distancia del origen menor que cierto valor dado (es decir, dentro de cierto círculo alrededor del origen) en el plano complejo, i.e., la secuencia está *acotada*. Las regiones blancas de la Fig. 1.2 son las localizaciones de c que dan una secuencia no acotada (i), mientras que las regiones negras son las localizaciones de c para las que se da el caso (ii); el conjunto de Mandelbrot propiamente dicho es la totalidad de la región negra.[10]

La complejidad del conjunto de Mandelbrot surge del hecho de que la secuencia iterada puede permanecer acotada de muchas maneras diferentes, y a veces muy complicadas. Puede haber combinaciones de ciclos y «casi» ciclos de varios tipos, que puntean el plano de diversas formas complicadas, pero nos llevaría demasiado lejos tratar de entender en detalle cómo surge la extraordinaria complejidad de este conjunto, donde están implicadas cuestiones sutiles de análisis complejo y teoría de números. El lector interesado puede consultar Peitgen y Reichter (1986) y Peitgen y Saupe (1988) para más información e imágenes (véase también Douady y Hubbard, 1985).

Notas

Sección 4.1

4.1. Véase el ejercicio [4.2] para estas reglas.

Sección 4.2

4.2. Una consecuencia directa[4.6] es que cualquier polinomio complejo de la sola variable z factoriza en factores lineales,

$$a_0 + a_1 z + a_2 z^2 + \ldots + a_n z^n = a_n(z - b_1)(z - b_2)\ldots(z - b_n)$$

[4.6] Demuéstrelo. (*Sugerencia*: Demuestre que no queda ningún resto si este polinomio se «divide» por $z - b$ cuando $z = b$ resuelve la ecuación dada.)

y es a *este* enunciado al que normalmente se llama «el teorema fundamental del álgebra».

4.3. La historia es que Tartaglia había revelado su solución parcial a Cardano solo después de que este hubiera jurado guardar el secreto. Por lo tanto, Cardano no podía publicar su solución más general sin romper su juramento. Sin embargo, en un viaje posterior a Bolonia, en 1543, Cardano examinó los papeles póstumos de Del Ferro y se convenció de la prioridad real de Del Ferro. Consideró que esto le dejaba libre para publicar todos estos resultados (con los debidos reconocimientos a Tartaglia y Del Ferro) en *Ars magna* en 1545. Tartaglia discrepó, y la disputa tuvo consecuencias muy amargas (véase Wykes, 1969).

4.4. Para más información, véase Van der Waerden (1985).

4.5. La razón para esto es que estamos sumando dos números que son mutuamente *complejos conjugados* (véase §10.1), y una suma semejante es siempre un número real.

Sección 4.3

4.6. Recordemos de la nota 2.4 que 0^{-1} debería significar $1/0$, i.e., «uno dividido por cero». Una abreviatura conveniente para expresar el «resultado» de esta operación ilegal es «$0^{-1} = \infty$».

4.7. «Estrictamente»: significa que los valores extremos no están incluidos en el intervalo.

4.8. Para más información, véase, por ejemplo, Hardy (1949).

Sección 4.4

4.9. Véanse, por ejemplo, Priestley (2003), p. 71, en lo referente a «radio de convergencia», y Needham (1967), pp. 67 y 264.

Sección 4.5

4.10. En las imágenes del conjunto de Mandelbrot generadas por ordenador (tales como la Fig. 1.2), uno no puede, por supuesto, computar indefinidamente para asegurar que una secuencia en apariencia acotada está *realmente* acotada. Es habitual «cortar» la iteración al cabo de un número apropiado de pasos. Sin embargo, retrasar simplemente el corte no mejora necesariamente la apariencia exacta de una imagen, porque los filamentos tienden a desaparecer.

5

Geometría de logaritmos, potencias y raíces

5.1. La geometría del álgebra compleja

Los aspectos de la magia de los números complejos discutidos al final del capítulo anterior implican muchas sutilezas, de modo que retrocedamos un poco y consideremos algunos otros fragmentos de magia más elementales aunque igualmente enigmáticos e importantes. En primer lugar, veamos cómo se representan geométricamente en el plano complejo las reglas para la suma y la multiplicación que encontramos en §4.1. Podemos presentarlas como la ley del *paralelogramo* y la ley del *triángulo semejante*, respectivamente, que se muestran en la Fig. 5.1a,b. En concreto, para dos números complejos generales w y z, los puntos que representan $w + z$ y wz vienen determinados por las afirmaciones respectivas:

los puntos $0, w, w + z, z$ *son los vértices de un paralelogramo*

y

los triángulos con vértices $0, 1, w$ y $0\ z, wz$ *son semejantes.*

(Aquí se han adoptado los convenios usuales sobre ordenamientos y orientaciones. Por eso, entiendo que recorremos el paralelogramo cíclicamente, de modo que el segmento de recta que va de w a $w + z$ es paralelo al que va de 0 a z, etc.; más aún, la relación de semejanza entre los dos triángulos no incluye «reflexión». Además, hay casos especiales en que los triángulos y el paralelogramo degeneran de varias maneras.)[5.1] El lector interesado puede comprobar estas reglas por tri-

[5.1] Examine las diversas posibilidades.

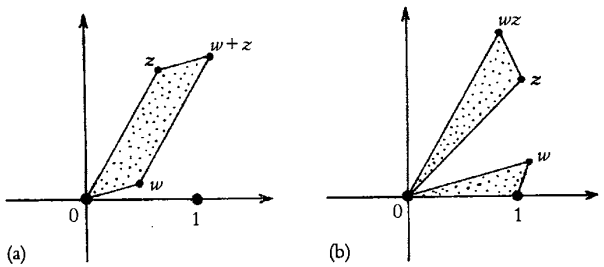


Fig. 5.1. Descripción geométrica de las leyes básicas del álgebra de los números complejos. (a) Ley del paralelogramo de la adición: $0, w, w+z, z$ dan los vértices de un paralelogramo. (b) Ley del triángulo semejante de la multiplicación: los triángulos con vértices $0, 1, w$ y $0, z, wz$ son semejantes.

gonometría y cálculo directo.[5.2] Sin embargo, hay otra manera de considerar estas cosas que evita el cálculo detallado y proporciona intuiciones mayores.

Consideremos la suma y la multiplicación en términos de diferentes *aplicaciones* (o «transformaciones») que aplican el plano complejo entero en sí mismo. Cualquier número complejo w dado define una «aplicación suma» y una «aplicación multiplicación», que son las operaciones que, cuando se aplican a un número complejo arbitrario z, sumarán w a z y harán el producto de w por z, respectivamente, i.e.,

$$z \mapsto w + z \text{ y } z \mapsto wz.$$

Es fácil ver que la aplicación suma desliza simplemente el plano complejo sin rotación ni cambio de tamaño o forma —un ejemplo de una *traslación* (véase §2.1)— desplazando el origen 0 al punto w; véase la Fig. 5.2a. La ley del paralelogramo es básicamente una reformulación de esto. Pero ¿qué pasa con la aplicación multiplicación? Esta proporciona una transformación que deja el origen fijo y conserva las formas —y que envía 1 al punto w—. En el caso general combina una rotación con una dilatación (o contracción) uniforme; véase la Fig. 5.2b.[5.3]

[5.2] Hágalo.

[5.3] Trate de demostrar esto sin cálculo detallado, y sin trigonometría. (*Sugerencia*: Esto es una consecuencia de la «ley distributiva» $w(z_1 + z_2) = wz_1 + wz_2$, que muestra que se conserva la estructura «lineal» del plano complejo, y $w(iz) = i(wz)$, que

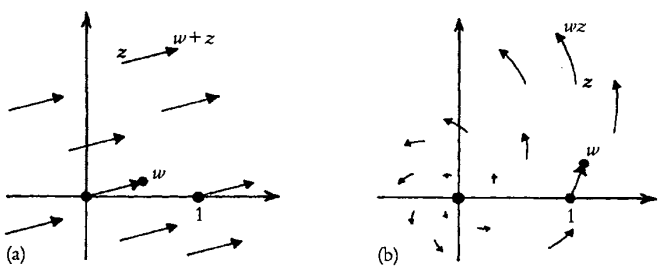


Fig. 5.2. (a) La aplicación suma «$+w$» proporciona una traslación del plano complejo, que envía 0 a w. (b) La aplicación multiplicación «$\times w$» proporciona una rotación y dilatación (o contracción) del plano complejo en torno a 0, que envía 1 a w.

La ley del triángulo semejante muestra esto de modo efectivo. Esta aplicación tendrá una especial importancia para nosotros en §8.2.

En el caso particular $w = i$, la aplicación multiplicación es simplemente una rotación a izquierdas (i.e., en sentido contrario a las agujas del reloj) de un ángulo recto $\left(\frac{1}{2}\pi\right)$. Si aplicamos esta operación dos veces, obtenemos una rotación de π, que es simplemente una reflexión respecto al origen; en otras palabras, es la aplicación multiplicación que hace corresponder a cada número complejo z su negativo. Esto nos proporciona una realización gráfica de la «misteriosa» ecuación $i^2 = -1$ (Fig. 5.3). La operación «multiplicar por i» queda realizada como la transformación geométrica «rotar un ángulo recto».Visto de este modo, no parece tan misterioso que el «cuadrado» de esta operación (i.e., hacerla dos veces) produzca el mismo efecto que la operación de «tomar el negativo». Por supuesto, esto no elimina la magia y el misterio de por qué el álgebra compleja funciona tan bien, ni nos habla de un papel físico claro para estos números. Podríamos preguntarnos, por ejemplo: ¿por qué rotar solo en un plano? ¿Qué pasa en tres dimensiones? Más adelante abordaré diferentes aspectos de esta cuestión, especialmente en §§11.2,3, §18.5, §§21.6,9, §§22.2,3,8-10, §33.2 y §34.8.

En nuestra descripción de un número complejo en el plano utili-

muestra que se conserva la rotación de un ángulo recto: i.e., se conservan los ángulos rectos.)

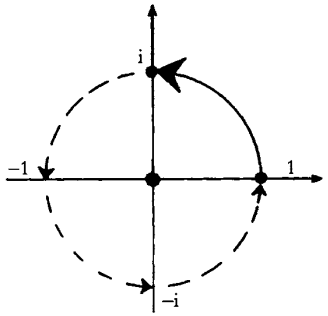


Fig. 5.3. La operación concreta «multiplicar por i» se realiza, en el plano complejo, como la transformación geométrica «rotar un ángulo recto». Se visualiza la «misteriosa» ecuación $i^2 = -1$.

zábamos las coordenadas cartesianas estándar (x, y) para un punto en el plano, pero alternativamente podríamos utilizar las coordenadas *polares* $[r, \theta]$. Aquí, el número real positivo r mide la distancia al origen y θ mide el *ángulo* que forma la recta que une el origen y el punto z con el eje real, medido dicho ángulo en sentido contrario a las agujas del reloj; véase la Fig. 5.4a. La cantidad r se conoce como el *módulo* del número complejo z, que solemos escribir como

$$r = |z|,$$

y θ como su *argumento* (o, en teoría cuántica, se conoce a veces como su *fase*). En el caso $z = 0$, no tenemos que preocuparnos por θ, pero podemos seguir definiendo r como la distancia al origen, que en este caso da simplemente $r = 0$.

Para ser más precisos, podríamos insistir en que θ esté comprendido en un intervalo concreto, tal como $-\pi < \theta \leq \pi$ (que es el convenio estándar). Alternativamente, podemos considerar simplemente el argumento como algo con la ambigüedad de que está permitido añadirle múltiplos enteros de 2π sin que nada se vea afectado. Se trata solo de que, cuando medimos el ángulo, podemos dar tantas vueltas alrededor del origen como queramos y en cualquier sentido (véase la Fig. 5.4b). (Este segundo punto de vista es en realidad el más profundo, y enseguida tendrá algunas implicaciones para nosotros.) A partir de la Fig. 5.5 y la trigonometría elemental, vemos que

$$x = r\cos\theta \text{ e } y = r\,\text{sen}\,\theta,$$

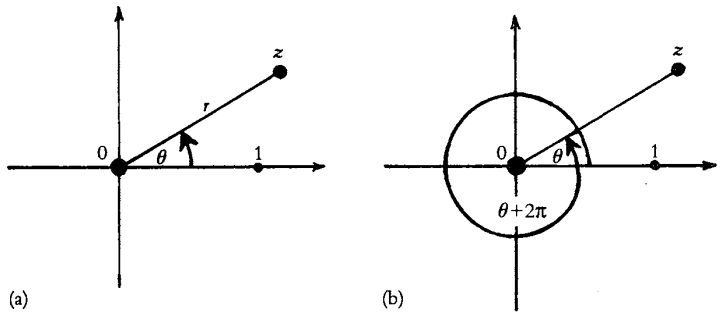


Fig. 5.4. (a) Al pasar de cartesianas (x, y) a polares $[r, \theta]$, tenemos $z = x + iy = re^{i\theta}$, el *módulo* $r = |z|$ es la *distancia* al origen y el *argumento* θ es el ángulo que forma la recta que va del origen a z con el eje real, medido en sentido contrario a las agujas del reloj. (b) Si no insistimos en $-\pi > \theta \leq \pi$, podemos permitir que z dé muchas vueltas alrededor del origen, sumando cualquier múltiplo entero de 2π a θ.

y recíprocamente que

$$r = \sqrt{x^2 + y^2} \text{ y } \theta = \tan^{-1}\frac{y}{x},$$

donde $\theta = \tan^{-1}(y/x)$ significa un valor concreto de la función multivaluada $\tan^{-1}$. (Para aquellos lectores que han olvidado la trigonometría, las dos primeras fórmulas simplemente reexpresan las definiciones del seno y el coseno de un ángulo en términos de un triángulo rectángulo: «el coseno de un ángulo es igual al lado adyacente dividido por la hipotenusa» y «el seno de un ángulo es igual al lado opuesto dividido por la hipotenusa», siendo r la hipotenusa; las dos siguientes expresan el teorema de Pitágoras y, en forma inversa, «la tangente de un ángulo es igual al lado opuesto dividido por el lado adyacente». Asimismo, habría que advertir que «$\tan^{-1}$» es la *función inversa* de «tan», y no el recíproco, de modo que la ecuación anterior $\theta = \tan^{-1}(y/x)$ representa $\tan\theta = y/x$. Finalmente, existe una ambigüedad en $\tan^{-1}$, pues puede añadirse a θ cualquier múltiplo entero de 2π y la relación seguirá siendo válida.)[1]

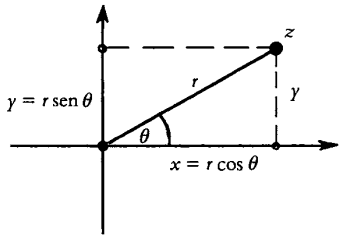


Fig. 5.5. Relación entre las formas cartesiana y polar de un número complejo: $x = r\cos\theta$, e $y = r\,\text{sen}\,\theta$, donde recíprocamente $r = \sqrt{(x^2 + y^2)}$ y $\theta = \tan^{-1}(y/x)$.

5.2. La idea del logaritmo complejo

Ahora, la «ley del triángulo semejante» de la multiplicación de dos números complejos, que se ilustra en la Fig. 5.1b, puede reexpresarse en términos del hecho de que cuando multiplicamos dos números complejos sumamos sus argumentos y multiplicamos sus módulos.[5.4] Adviértase aquí el hecho notable de que, por lo que respecta a la regla para los argumentos, hemos *convertido la multiplicación en suma*. En esto se basa el uso de los *logaritmos* (el logaritmo del producto de dos números es igual a la suma de sus logaritmos: $\log ab = \log a + \log b$), como se manifiesta en la regla de cálculo (Fig. 5.6), y esta propiedad fue de fundamental importancia para la práctica del cálculo en épocas anteriores.[2] Ahora utilizamos calculadoras electrónicas que hacen las multiplicaciones por nosotros. Aunque esto es mucho más rápido y más exacto que el uso de una regla de cálculo o de unas tablas de logaritmos, perdemos algo muy importante para nuestra comprensión si no extraemos ninguna experiencia directa de la bella y muy importante operación logarítmica. Veremos que los logaritmos desempeñan un papel profundo en relación con los números complejos. De hecho, el argumento de un número complejo *es* realmente un logaritmo, en un sentido muy claro. Trataremos de entender cómo se produce esto.

Recordemos también que en §4.2 he afirmado que tomar raíces de números complejos es básicamente una cuestión de comprender los logaritmos complejos. Encontraremos que existen algunas relaciones

[5.4] Explique esto en detalle.

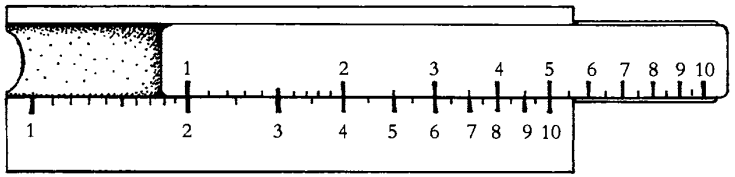


Fig. 5.6. Las reglas de cálculo muestran números en una escala logarítmica, posibilitando con ello que la multiplicación se exprese por la suma de distancias, de acuerdo con la fórmula $\log_b(p \times q) = \log_b p + \log_b q$. (Se ilustra la multiplicación por 2.)

sorprendentes entre logaritmos complejos y trigonometría. Tratemos de ver cómo se combinan todas estas cosas.

En primer lugar, recordemos algo sobre los logaritmos ordinarios. Tomar un logaritmo es la operación inversa de «elevar un número a una potencia», o de la *exponenciación*. «Elevar a una potencia» *es* una operación que convierte suma en multiplicación. ¿Por qué es esto? Tomemos cualquier número b. Notemos entonces la fórmula (que convierte suma en multiplicación)

$$b^{m+n} = b^m \times b^n,$$

que es obvia si m y n son enteros positivos, porque cada miembro representa simplemente $m + n$ copias del número b, todas ellas multiplicadas. Lo que tenemos que hacer es encontrar una manera de generalizar esto de modo que m y n no tengan que ser enteros positivos, sino que puedan ser números complejos cualesquiera. Para ello necesitamos encontrar la definición correcta de «b elevado a la potencia z», para z complejo, y queremos que la misma fórmula anterior, a saber $b^{(w+z)} = b^w \times b^z$, sea válida cuando los exponentes w y z son complejos.

De hecho, el procedimiento para hacer esto refleja, en cierta medida, la historia misma de generalizar, paso a paso, desde los enteros positivos a los números complejos, como se hizo, partiendo de Pitágoras, por la obra de Eudoxo, pasando por Brahmagupta, hasta llegar a la época de Cardano y Bombelli (y más tarde), tal como se ha indicado en §3.4 y §4.1. Primero, la noción de «b^z» es inicialmente entendida cuando z es un número positivo, simplemente como $b \times b \times b \times \ldots \times b$, con z b's multiplicadas; en particular, $b^1 = b$. Luego (siguiendo a Brahmagupta,

permitimos que z sea cero, advirtiendo que para conservar $b^{(w+z)} = b^w \times b^z$ tenemos que definir $b^0 = 1$. A continuación permitimos que z sea negativo, y advertimos, por la misma razón, que para el caso $z = -1$ debemos definir b^{-1} como el recíproco de b (i.e., $1/b$), y que b^{-n} para un número natural n debe ser la n-ésima potencia de b^{-1}. Luego tratamos de generalizar a situaciones en que z es una fracción, empezando con el caso $z = 1/n$, donde z es un entero positivo. La aplicación repetida de $b^w \times b^z = b^{(w+z)}$ nos lleva a concluir que $(b^z)^n = b^{zn}$; así pues, haciendo $z = 1/n$, llegamos a que $b^{1/n}$ es una raíz n-ésima de b.

Podemos hacer esto dentro del dominio de los números reales, con tal de que el número b se haya tomado positivo. Entonces podemos tomar $b^{1/n}$ como la única raíz n-ésima positiva de b (cuando n es un entero positivo) y podemos continuar definiendo unívocamente b^z para cualquier número racional $z = m/n$ como la m-ésima potencia de la raíz n-ésima de b y de aquí pasar (utilizando un proceso de paso al límite) a cualquier número real z. Pero si se permite que b sea negativo, entonces tropezamos con una pega en $z = 1/2$, puesto que $\sqrt{b}$ requiere entonces la introducción de i y caemos en la pendiente resbaladiza hacia los números complejos. En el fondo de dicha pendiente encontramos nuestro mágico mundo complejo, de modo que agarrémonos y sigamos hasta abajo.

Necesitamos una definición de b^p tal que para todos los números complejos p, q y b (con $b \neq 0$), tengamos:

$$b^{p+q} = b^p \times b^q.$$

Podríamos confiar entonces en definir el *logaritmo en base b* (la operación denotada por «$\log_b$») como la inversa de la función definida por $f(z) = b^z$, es decir,

$$z = \log_b w \quad \text{si} \quad w = b^z.$$

Entonces esperaríamos

$$\log_b(p \times q) = \log_b p + \log_b q,$$

de modo que esta noción de logaritmo convertiría realmente la multiplicación en suma.

5.3. Multivaluación, logaritmos naturales

Aunque esto es básicamente correcto, hay ciertas dificultades técnicas en hacerlo (y dentro de poco veremos cómo tratarlas). En primer lugar, b^z es «multivaluada». Es decir, existen, en general, muchas respuestas diferentes para el significado de «b^z». Hay también una multivaluación adicional para $\log_b w$. Hemos visto ya la multivaluación de b^z con valores fraccionarios de z. Por ejemplo, si $z = 1/2$, entonces «b^z» debería significar «cierta cantidad t cuyo cuadrado es b», puesto que exigimos que $t^2 = t \times t = b^{\frac{1}{2}} \times b^{\frac{1}{2}} = b^{\frac{1}{2}+\frac{1}{2}} = b^1 = b$. Si un número t satisface esta propiedad, entonces $-t$ también lo hará (puesto que $(-t) \times (-t) = t^2 = b$). Suponiendo que $b \neq 0$, tenemos dos respuestas distintas para $b^{1/2}$ (normalmente escritas $\pm\sqrt{b}$). Con más generalidad, tenemos n respuestas complejas distintas para $b^{1/n}$, cuando n es un entero positivo: 1, 2, 3, 4, 5, ... De hecho, tenemos un número finito de respuestas cuando quiera que n sea un número racional (no nulo). Si n es irracional, entonces tenemos un número infinito de respuestas, como veremos enseguida.

Tratemos de ver cómo podemos hacer frente a estas ambigüedades. Empezaremos haciendo una elección concreta de b, a saber, el número fundamental «e», conocido como la *base de los logaritmos naturales*. Esto reducirá nuestro problema de ambigüedad. Tenemos, como definición de e:

$$\mathrm{e} = 1 + \frac{1}{1!} + \frac{1}{2!} + \frac{1}{3!} + \frac{1}{4!} + \ldots = 2{,}718\ 281\ 828\ 5\ldots,$$

donde los signos de admiración denotan *factoriales*, i.e.,

$$n! = 1 \times 2 \times 3 \times 4 \times \ldots \times n,$$

de modo que $1! = 1$, $2! = 2$, $3! = 6$, etc. La función definida por $f(z) = \mathrm{e}^z$ se conoce como función *exponencial*, y a veces se escribe «exp»; puede considerarse como «e elevado a la potencia z» y esta potencia se define mediante la siguiente modificación sencilla de la serie anterior para e:

$$\mathrm{e}^z = 1 + \frac{z}{1!} + \frac{z^2}{2!} + \frac{z^3}{3!} + \frac{z^4}{4!} + \ldots$$

Esta importante serie de potencias converge realmente para todos los valores de z (de modo que tiene un círculo de convergencia infini-

to; véase §4.4). La suma infinita hace una elección particular para la ambigüedad en «b^z» cuando $b = e$. Por ejemplo, si $z = 1/2$, entonces la serie nos da la cantidad concreta positiva $+\sqrt{e}$ y no $-\sqrt{e}$. El hecho de que $z = 1/2$ dé realmente una cantidad $e^{1/2}$ cuyo cuadrado es e se sigue del hecho de que e^z, definida por esta serie,[5.5] siempre tiene la propiedad requerida «de suma a multiplicación»

$$e^{a+b} = e^a e^b,$$

de modo que $(e^{\frac{1}{2}})^2 = e^{\frac{1}{2}} e^{\frac{1}{2}} = e^{\frac{1}{2}+\frac{1}{2}} = e^1 = e$.

Tratemos de utilizar esta definición de e^z para darnos un logaritmo inambiguo, definido como la *inversa* de la función exponencial:

$$z = \log w \quad \text{si} \quad w = e^z.$$

Esto se conoce como el logaritmo *natural* (y escribiré la función simplemente como «log», sin un símbolo para la base).[3] De la anterior propiedad «de suma a multiplicación», esperamos una regla «de multiplicación a suma»:

$$\log ab = \log a + \log b.$$

No es inmediatamente obvio que tal inverso de e^z exista necesariamente. Sin embargo, resulta en efecto que para cualquier número complejo w, distinto de 0, siempre existe z tal que $w = e^z$, de modo que podemos definir $\log w = z$. Pero aquí hay una trampa: hay más de una respuesta.

¿Cómo expresamos estas respuestas? Si $[r, \theta]$ es la representación polar de w, entonces podemos escribir su logaritmo z en la forma cartesiana habitual ($z = x + iy$) como

$$z = \log r + i\theta,$$

donde $\log r$ es el logaritmo natural ordinario de un número real positivo —la inversa de la exponencial real—. Resulta intuitivamente claro de la Fig. 5.7 que semejante función logarítmica real existe. En la Fig. 5.7a tenemos la gráfica de $r = e^x$. Simplemente intercambiamos los

[5.5] Compruébelo directamente a partir de la serie. (*Sugerencia*: El «teorema del binomio» para exponentes enteros afirma que el coeficiente de $a^p b^q$ en $(a + b)^n$ es $n!/p!q!$)

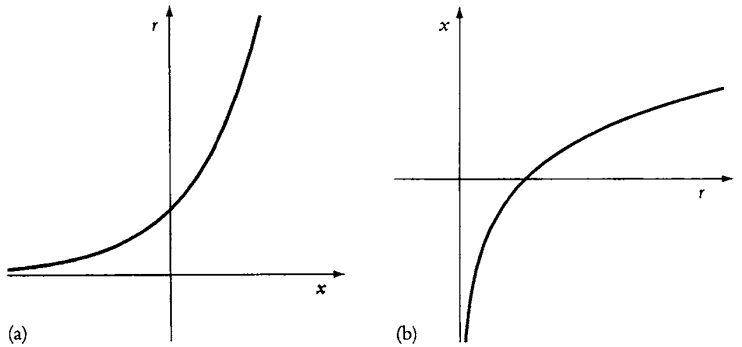


Fig. 5.7. Para obtener el logaritmo de un número real positivo r, consideremos la gráfica (a) de $r = e^x$. Se alcanzan todos los valores positivos de r, de modo que invirtiendo la figura, obtenemos la gráfica (b) de la función inversa $x = \log r$ para r positivo.

ejes para obtener la gráfica de la función inversa $x = \log r$, como en la Fig. 5.7b. No es tan sorprendente que la parte real de $z = \log w$ sea simplemente un logaritmo real ordinario. Lo que es algo más notable[4] es que la parte imaginaria de z es simplemente el ángulo θ, que es el argumento del número complejo w. Este hecho hace explícito mi anterior comentario acerca de que el argumento de un número complejo es realmente una forma de logaritmo.

Recordemos que existe una ambigüedad en la definición del argumento de un número complejo. Podemos añadir a θ cualquier múltiplo entero de 2π, y esto también servirá (recordemos la Fig. 5.4b). En consecuencia, hay muchas soluciones z diferentes para una elección dada de w en la relación $w = e^z$. Si tomamos uno de estos z, entonces $z + 2\pi i n$ es otra posible solución, donde n es cualquier entero que queramos escoger. Así pues, el logaritmo de w es ambiguo en cuanto que está definido salvo la suma de cualquier múltiplo entero de $2\pi i$. Debemos tener esto en cuenta con expresiones tales como $\log ab = \log a + \log b$, asegurándonos de que se hacen las correspondientes y adecuadas elecciones de logaritmo.

En esta etapa, esta característica de los logaritmos complejos parece ser solo una incomodidad. Sin embargo, veremos en §7.2 que es absolutamente capital para algunas de las más poderosas, útiles y mágicas

propiedades de los números complejos. El análisis complejo depende de ello de forma crucial. Por el momento, tratemos solo de apreciar la naturaleza de la ambigüedad.

Otra forma de entender esta ambigüedad en $\log w$ consiste en advertir la sorprendente fórmula

$$e^{2\pi i} = 1,$$

de donde $e^{z+2\pi i} = e^z = w$, etc., lo que demuestra que $z + 2\pi i$ es un logaritmo de w tan bueno como lo es z (y entonces podemos repetir esto tantas veces como queramos). La fórmula anterior está estrechamente relacionada con la famosa *fórmula de Euler*

$$e^{\pi i} + 1 = 0$$

(que relaciona los cinco números fundamentales 0, 1, i, π y e en una expresión casi mística).[5.6]

Podemos entender mejor estas propiedades si tomamos la exponencial de la expresión $z = \log r + i\theta$ para obtener

$$w = e^z = e^{\log r + i\theta} = e^{\log r} e^{i\theta} = re^{i\theta}.$$

Esto demuestra que la forma polar de cualquier número complejo w, que yo había estado denotando hasta ahora por $[r, \theta]$, puede escribirse de forma más reveladora como

$$w = re^{i\theta}.$$

De esta forma, es evidente que si multiplicamos dos números complejos, tomamos el producto de sus módulos y la suma de sus argumentos ($re^{i\theta} se^{i\phi} = rse^{i(\theta+\phi)}$, de modo que r y s se multiplican, mientras que θ y ϕ se suman, teniendo en cuenta que restar $2\pi i$ de $\theta + \phi$ no supone ninguna diferencia), como está implícito en la «ley del triángulo semejante» de la Fig. 5.1b. En lo sucesivo abandonaré la notación $[r, \theta]$ y utilizaré en su lugar la expresión mostrada más arriba. Nótese que si $r = 1$ y $\theta = \pi$, obtenemos -1 y recuperamos la famosa fórmula de Euler anterior $e^{\pi i} + 1 = 0$, utilizando la geometría de la Fig. 5.4a; si $r = 1$ y $\theta = 2\pi$, obtenemos $+1$ y recuperamos $e^{2\pi i} = 1$.

[5.6] Demuestre a partir de esto que $z + \pi i$ es un logaritmo de $-w$.

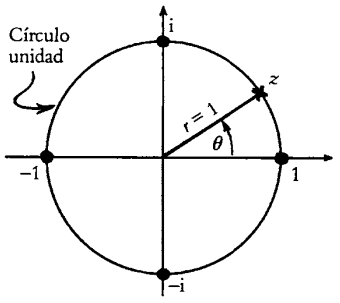


Fig. 5.8. El círculo unidad, que consiste en números complejos de módulo unidad. La fórmula de Cotes-Euler los da como $e^{i\theta} = \cos\theta + i \operatorname{sen}\theta$ para θ real.

El círculo con $r = 1$ se denomina el *círculo unidad* en el plano complejo (véase la Fig. 5.8). Este está dado por $w = e^{i\theta}$ para θ real, según la expresión de arriba. Comparando esta expresión con las anteriores $x = r\cos\theta$ e $y = r \operatorname{sen}\theta$ dadas arriba para las partes real e imaginaria de lo que ahora es la cantidad $w = x + iy$, obtenemos la prolífica «fórmula de (Cotes-)Euler»[5]

$$e^{i\theta} = \cos\theta + i \operatorname{sen}\theta,$$

que básicamente engloba los elementos esenciales de la trigonometría en las propiedades mucho más simples de las funciones exponenciales complejas.

Veamos ahora cómo funciona esto en casos elementales. En particular, la relación básica $e^{a+b} = e^a e^b$, cuando se desarrolla en términos de partes real e imaginaria, da inmediatamente[5.7] las expresiones de apariencia mucho más complicada (sin duda tristemente familiares para algunos lectores):

$$\cos(a + b) = \cos a \cos b - \operatorname{sen} a \operatorname{sen} b,$$
$$\operatorname{sen}(a + b) = \operatorname{sen} a \cos b + \cos a \operatorname{sen} b.$$

Análogamente, el desarrollo de $e^{3i\theta} = (e^{i\theta})^3$, por ejemplo, da rápidamente[6,5.8]

$$\cos 3\theta = \cos^3\theta - 3\cos\theta \operatorname{sen}^2\theta,$$
$$\operatorname{sen} 3\theta = 3 \operatorname{sen}\theta \cos^2\theta - \operatorname{sen}^3\theta.$$

[5.7] Compruébelo.
[5.8] Hágalo.

Hay realmente magia en la forma directa en que tales fórmulas algo complicadas surgen de sencillas expresiones con números complejos.

5.4. Potencias complejas

Volvamos ahora a la cuestión de definir w^z (o b^z, como hemos escrito previamente). Podemos conseguir esto escribiendo

$$w^z = e^{z \log w}$$

(puesto que esperamos que $e^{z \log w} = (e^{\log w})^z$ y $e^{\log w} = w$). Pero notemos que, debido a la ambigüedad en log w, podemos añadir a log z cualquier múltiplo entero de $2\pi i$ para obtener otra respuesta admisible. Esto significa que podemos multiplicar o dividir cualquier elección particular de w^z por $e^{z \cdot 2\pi i}$ cualquier número de veces y seguimos teniendo un «w^z» admisible. Resulta divertido ver la configuración de puntos en el plano complejo que da esto en el caso general. Esto se ilustra en la Fig. 5.9. Los puntos están en las intersecciones de dos espirales equiangulares. (Una espiral equiangular —o *logarítmica*— es una curva en el plano que forma un ángulo constante con las líneas rectas que irradian de un punto en el plano.)[5.9]

Esta ambigüedad nos lleva a todo tipo de problemas si no tenemos cuidado.[5.10] La mejor forma de evitar tales problemas parece consistir en adoptar la regla de que la notación w^z se utilice solo cuando se ha especificado una *elección particular* de log w. (En el caso especial de e^z, el convenio tácito es hacer siempre la elección particular log e = 1. Entonces, la notación estándar e^z es consistente con nuestra más general w^z.) Una vez que se ha hecho esta elección de log w, entonces w^z está definida inequívocamente para todos los valores de z.

Puede comentarse en este punto que también necesitamos una especificación de log b si queremos definir el «logaritmo en base b» mencionado antes en esta sección (la función denotada por «$\log_b$»), porque necesitamos un $w = b^z$ inequívoco para definir $z = \log_b w$. Aun así,

[5.9] Demuéstrelo. ¿Cuántas maneras? Encuentre también todos los casos especiales.
[5.10] Resuelva esta «paradoja»: $e = e^{1+2\pi i}$, de modo que $e = (e^{1+2\pi i})^{1+2\pi i} = e^{1+4\pi i-4\pi^2} = e^{1-4\pi^2}$.

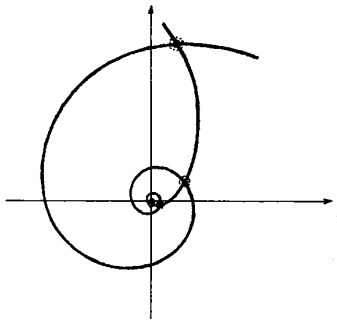


Fig. 5.9. Los diferentes valores de $w^z (= e^{z \log w})$. Puede sumarse cualquier múltiplo entero de $2\pi i$ a log 2, que multiplica o divide w^z por $e^{z2\pi i}$ un número entero de veces. En el caso general, se representan en el plano complejo como las intersecciones de dos espirales equiangulares (cada una de las cuales forma un ángulo constante con líneas rectas que pasan por el origen).

$\log_b w$ será por supuesto multivaluada (como lo era log w), donde podemos sumar a $\log_b w$ cualquier múltiplo entero de $2\pi i/\log b$.[5.11]

Una curiosidad que ha intrigado enormemente a algunos matemáticos en el pasado es la cantidad i^i. Podría parecer que «no puede haber nada más imaginario que esto». Sin embargo, encontramos la respuesta *real*

$$i^i = e^{i \log i} = e^{i \cdot \frac{1}{2}\pi i} = e^{-\pi/2} = 0{,}207\ 879\ 576\ldots,$$

especificando $\log i = \frac{1}{2}\pi i$.[5.12] Hay también muchas otras respuestas, dadas por las otras especificaciones de log i. Estas se obtienen multiplicando la cantidad anterior por $e^{2\pi n}$, donde n es un entero (o, de forma equivalente, elevando la cantidad anterior a cualquier potencia de la forma $4n + 1$, donde n es un entero positivo o negativo).[5.13] Es sorprendente que *todos* los valores de i^i son, de hecho, números reales.

Veamos cómo funciona la notación w^z para $z = 1/2$. Esperamos ser capaces de representar las dos cantidades $\pm\sqrt{w}$ como «$w^{1/2}$» en algún sentido. De hecho, para obtener estas dos cantidades basta con especificar primero un valor para log w y especificar luego otro, donde sumamos $2\pi i$ al primero para obtener el segundo. Esto da lugar a un cambio de signo en $w^{1/2}$ (debido a la fórmula de Euler $e^{\pi i} = -1$). De

[5.11] Demuéstrelo.
[5.12] ¿Por qué es esta una especificación admisible?
[5.13] Demuestre por qué funciona esto.

modo análogo, podemos generar todas las n soluciones $z^n = w$ cuando n es 3, 4, 5, ... como la cantidad $w^{1/n}$, cuando se especifican sucesivamente diferentes valores de log w.[5.14] Con más generalidad, podemos volver a la cuestión de las raíces z-ésimas de un número complejo w no nulo que he mencionado en §4.2. Podemos expresar esta raíz z-ésima como la expresión $w^{1/z}$, y generalmente obtenemos un número infinito de valores alternativos para esto, dependiendo de qué elección de log w se especifique. Con la elección específica correcta para log $(w^{1/z})$, a saber, la dada por $(\log w)/z$, obtenemos realmente $(w^{1/z})^z = w$. Notamos, con más generalidad, que

$$(w^a)^b = w^{ab},$$

donde una vez que hemos hecho una especificación de log w (para el segundo miembro), debemos especificar (para el primer miembro) que log (w^a) sea a log w.[5.15]

Cuando $z = n$ es un entero positivo, entonces las cosas son mucho más simples, y obtenemos precisamente n raíces. En este caso se da una situación de interés particular cuando $w = 1$. Entonces, especificando sucesivamente posibles valores de log 1, a saber, $0, 2\pi i, 4\pi i, 6\pi i, \ldots$ obtenemos $1 = e^0, e^{2\pi i/n}, e^{4\pi i/n}, e^{6\pi i/n}, \ldots$ para los posibles valores de $1^{1/n}$. Podemos escribirlos como $1, \epsilon, \epsilon^2, \epsilon^3, \ldots$ donde $\epsilon = e^{2\pi i/n}$. En términos del plano complejo, obtenemos n puntos igualmente espaciados alrededor del círculo unidad, llamados *raíces* n-ésimas de la unidad. Estos puntos constituyen los vértices de un n-ágono regular (véase la Fig. 5.10). (Nótese que las elecciones $-2\pi i, -4\pi i, -6\pi i$, etc., para log 1 darían meramente las mismas raíces n-ésimas, en orden inverso.)

Resulta interesante observar que, para un n dado, las raíces n-ésimas de la unidad constituyen lo que se denomina un *grupo multiplicativo finito*, más concretamente, *el grupo cíclico* $\mathbb{Z}_n$ (véase §13.1). Tenemos n cantidades con la propiedad de que podemos multiplicar dos cualesquiera de ellas y obtener otra. También podemos dividir una por otra para obtener una tercera. Como ejemplo, consideremos el caso $n = 3$. Ahora obtenemos tres elementos $1, \omega$ y ω^2, donde $\omega = e^{2\pi i/3}$ (de modo que

[5.14] Explique esto en detalle.
[5.15] Demuéstrelo.

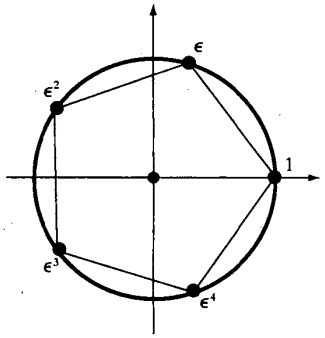


Fig. 5.10. Las raíces n-ésimas de la unidad $e^{2\pi ri/n}$ $(r = 1, 2, \ldots, n)$, igualmente espaciadas alrededor del círculo unidad, proporcionan los vértices de un n-ágono regular. Aquí $n = 5$.

$\omega^3 = 1$ y $\omega^{-1} = \omega^2$). Tenemos las siguientes tablas simples de multiplicación y división para estos números:

×	1	ω	ω^2
1	1	ω	ω^2
ω	ω	ω^2	1
ω^2	ω^2	1	ω

÷	1	ω	ω^2
1	1	ω^2	ω
ω	ω	1	ω^2
ω^2	ω^2	ω	1

En el plano complejo, estos números concretos están representados como los vértices de un triángulo equilátero. La multiplicación por ω rota el triángulo $\frac{2}{3}\pi$ (i.e., 120°) en sentido contrario a las agujas del reloj, y la multiplicación por ω^2 lo rota $\frac{2}{3}\pi$ en el sentido de las agujas; para la división, la rotación es en sentido opuesto (véase la Fig. 5.11).

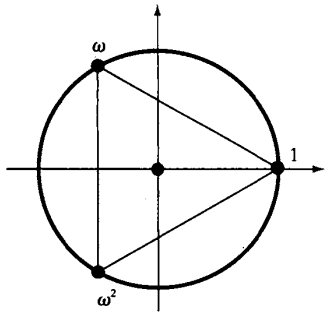


Fig. 5.11. Triángulo equilátero de raíces cúbicas 1, ω y ω^2 de la unidad. La multiplicación por ω rota 120° en sentido contrario a las agujas del reloj, y la multiplicación por ω^2, en el sentido de las agujas del reloj.

5.5. Algunas relaciones con la física de partículas moderna

Números como estos tienen interés en la moderna física de partículas, pues proporcionan los casos posibles de un *número cuántico multiplicativo*. En §3.5 he comentado el hecho de que los números cuánticos (escalares) aditivos de la física de partículas están invariablemente cuantificados, hasta donde conocemos, por enteros. Existen también unos pocos ejemplos de números cuánticos multiplicativos, y estos parecen estar cuantificados en términos de las raíces n-ésimas de la unidad. Solo conozco unos pocos ejemplos de tales cantidades en la física de partículas convencional, y en la mayoría de estos la situación es el caso relativamente sin interés $n = 2$. Hay un caso claro en que $n = 3$ y posiblemente un caso para el que $n = 4$. Por desgracia, en la mayoría de los casos el número cuántico no es universal, i.e., no puede aplicarse de forma consistente a todas las partículas. En tales situaciones, me referiré al número cuántico como solo *aproximado*.

La magnitud física denominada *paridad* es un número cuántico multiplicativo (aproximado) con $n = 2$. (Hay también otras magnitudes aproximadas $n = 2$, similares en muchos aspectos a la paridad, tal como la *g-paridad*. No las discutiré aquí.) La noción de paridad para un sistema compuesto se construye (multiplicativamente) a partir de las paridades de sus partículas constituyentes básicas. Para una de tales partículas constituyentes, su paridad puede ser par, en cuyo caso la reflexión especular de la partícula es igual que la propia partícula (en un sentido apropiado); alternativamente, su paridad puede ser impar, en cuyo caso su reflexión especular es lo que se denomina su antipartícula (véanse §3.5, §§24.1-3,8 y §26.4). Puesto que la noción de reflexión especular, o de tomar la antipartícula, es algo «cuyo cuadrado es la unidad», i.e., hacerla dos veces nos devuelve al punto de partida, el número cuántico —llamémoslo ϵ— tiene que tener la propiedad $\epsilon^2 = 1$, de modo que debe ser una «raíz n-ésima de la unidad» con $n = 2$, i.e., $\epsilon = 1$ o $\epsilon = -1$). Esta noción es solo aproximada porque la paridad es una magnitud que no se conserva en las denominadas «interacciones débiles» y, en realidad, quizá no haya una paridad bien definida para ciertas partículas a causa de ello (véanse §§25.3,4).

Además, la noción de paridad se aplica, en las descripciones normales, solo a la familia de partículas conocidas como *bosones*. Las demás partículas pertenecen a otra familia y se conocen como *fermiones*. La distinción entre bosones y fermiones es muy importante aunque algo sofisticada, y llegaremos a ella más adelante, en §§23.7,8. (En una de sus manifestaciones, tiene que ver con lo que sucede cuando rotamos de forma continua el estado de la partícula hasta completar un giro de 2π (i.e., 360°). Solo los bosones quedan completamente restaurados en sus estados originales tras una rotación semejante. En el caso de los fermiones habría que hacer dos veces esta rotación. (Véanse §11.3 y §22.8.) Hay un sentido en el que «dos fermiones hacen un bosón», mientras que «dos bosones hacen también un bosón» y «un bosón y un fermión hacen un fermión». Así pues, podemos asignar el número cuántico multiplicativo -1 a un fermión y $+1$ a un bosón para describir su naturaleza fermión/bosón, y tenemos otro número cuántico multiplicativo con $n = 2$. Hasta donde se sabe, esta magnitud es un número cuántico multiplicativo *exacto*.

En mi opinión, hay también una noción de paridad que puede aplicarse a fermiones, aunque no parece que esta sea una terminología convencional. Esta debe combinarse con el número cuántico fermión/bosón para dar un número cuántico combinado multiplicativo con $n = 4$. Para un fermión, el valor de la paridad tendría que ser $+i$ o $-i$, y su doble reflexión especular tendría el efecto de una rotación de 2π. Para un bosón, el valor de la paridad sería ± 1, como antes.

El número cuántico multiplicativo con $n = 3$ que he mencionado es a lo que me referiré como *quarkedad*. (Esta no es una terminología estándar, ni es habitual referirse a este concepto como número cuántico, pero recoge un aspecto importante de nuestro conocimiento actual de la física de partículas.) En §3.5 he mencionado el moderno punto de vista según el cual se supone que las partículas que «interaccionan fuertemente» conocidas como *hadrones* (protones, neutrones, mesones-π, etc.) están compuestas de *quarks* (véase §25.6). Estos quarks tienen valores para su carga eléctrica que no son múltiplos enteros de la carga del electrón, sino múltiplos enteros de un tercio de dicha carga. Sin embargo, los quarks no pueden existir como partículas individuales independientes, y sus compuestos solo pueden existir como individuos independientes si la suma de sus cargas es un número entero, en

unidades de la carga del electrón. Sea q el valor de la carga eléctrica medida en unidades negativas de la carga del electrón (de modo que para el propio electrón tenemos $q = -1$, ya que la carga del electrón se cuenta como negativa según el convenio usual). Para los quarks, tenemos $q = 2/3$ o $-1/3$; para los antiquarks, $q = 1/3$ o $-2/3$. Así pues, si tomamos para la quarkedad el número cuántico multiplicativo $e^{-2q\pi i}$, encontramos que toma valores 1, ω y ω^2. Para un quark, la quarkedad es ω, y para un antiquark es ω^2. Una partícula puede existir independientemente solo si su quarkedad es 1. De acuerdo con §5.4, los grados de quarkedad constituyen el grupo cíclico $\mathbb{Z}_3$. (En §16.1 veremos cómo, con un elemento «0» adicional y una noción de suma, este grupo puede extenderse al *campo finito* $\mathbb{F}_4$.)

Tanto en esta sección como en la anterior, he presentado algunos de los aspectos de la magia de los números complejos y he sugerido tan solo algunas de sus aplicaciones. Pero no he mencionado todavía aquellos aspectos de los números complejos (que se verán en el capítulo 7) que a mí personalmente me parecieron los más mágicos de todos cuando los aprendí siendo estudiante de licenciatura en matemáticas. En años posteriores he encontrado aspectos aún más sorprendentes de dicha magia, y uno de estos (descrito al final del capítulo 9) es extrañamente complementario del que más me impresionó siendo estudiante. Estas cosas, sin embargo, dependen de las nociones básicas del *cálculo infinitesimal*, de modo que para transmitir algo de esto al lector será necesario decir algo sobre ciertas nociones básicas del cálculo. Hay, por supuesto, una razón adicional para hacerlo. ¡El cálculo infinitesimal es esencial para un conocimiento adecuado de la física!

Notas

Sección 5.1

5.1. También deberían señalarse las funciones trigonométricas $\cot\theta = \cos\theta/\operatorname{sen}\theta = (\tan\theta)^{-1}$, $\sec\theta = (\cos\theta)^{-1}$ y $\operatorname{cosec}\theta = (\operatorname{sen}\theta)^{-1}$, así como las versiones «hiperbólicas» de las funciones trigonométricas, $\operatorname{senh} t = \frac{1}{2}(e^t + e^{-t})$, $\cosh t = \frac{1}{2}(e^t - e^{-t})$, $\tanh t = \operatorname{senh} t/\cosh t$, etc. Nótese también que las inversas de estas operaciones se denotan por $\cot^{-1}$, senh^{-1}, etc., como sucede con la «$\tan^{-1}(y/x)$» de §5.1.

Sección 5.2

5.2. Los logaritmos fueron introducidos en 1614 por John Neper (Napier) y utilizados con fines prácticos por Henri Briggs en 1624.

Sección 5.3

5.3. Los logaritmos naturales también se escriben habitualmente como «ln».

5.4. De lo que hemos establecido hasta aquí, no podemos inferir que «$i\theta$» en la fórmula $z = \log r + i\theta$ no debería ser un múltiplo real de $i\theta$. Esto requiere cálculo infinitesimal.

5.5. Roger Cotes (1714) tenía la fórmula equivalente $\log(\cos\theta + i\operatorname{sen}\theta) = i\theta$. La $e^{i\theta} = \cos\theta + i\operatorname{sen}\theta$ de Euler parece haber aparecido por primera vez treinta años más tarde (véase Euler, 1748).

5.6. Aquí estoy utilizando la notación conveniente (aunque no muy lógica) $\cos^3\theta$ para $(\cos\theta)^3$, etc. Debería advertirse la inconsistencia notacional con la (más lógica) $\cos^{-1}\theta$, que también se denota habitualmente como $\arccos\theta$. La fórmula $\operatorname{sen} n\theta + i\cos n\theta = (\operatorname{sen}\theta + i\cos\theta)^n$ se suele conocer como «teorema de De Moivre». Parece que también Abraham de Moivre, un contemporáneo de Roger Cotes (véase la nota anterior), ha sido codescubridor de $e^{i\theta} = \cos\theta + i\operatorname{sen}\theta$.

6

Cálculo infinitesimal con números reales

6.1. ¿Qué hace respetable a una función?

El cálculo infinitesimal —o, según su nombre más sofisticado, *análisis matemático*— está construido a partir de dos ingredientes básicos: *diferenciación* e *integración*. La diferenciación está relacionada con velocidades y aceleraciones, con pendientes y curvaturas de curvas y superficies, y cosas similares. Estas son tasas de cambio de las cosas, y son cantidades definidas *localmente*, en términos de estructura o comportamiento en los entornos más minúsculos de puntos simples. La integración, por el contrario, está relacionada con áreas y volúmenes, con centros de gravedad y con muchas otras cosas de esa misma naturaleza general. Estas son cosas que implican de una u otra forma medidas de *totalidad*, y no están definidas meramente por lo que pasa en los entornos locales o infinitesimales de puntos individuales. El hecho notable, conocido como *teorema fundamental del cálculo infinitesimal*, es que cada uno de estos ingredientes es, en esencia, justo el *inverso* del otro. Básicamente es esto lo que posibilita que estos dos importantes dominios de estudio matemático se combinen y proporcionen un poderoso cuerpo de conocimiento y de técnica de cálculo.

Esta disciplina del análisis matemático, tal como se originó en el siglo XVII en la obra de Fermat, Newton y Leibniz, con ideas que se remontan hasta Arquímedes en el siglo III a.C., se denomina «cálculo infinitesimal» porque proporciona un cuerpo de técnica de cálculo mediante el que algunos problemas que, de otro modo, serían conceptualmente difíciles de manejar, pueden ser resueltos con frecuencia «de forma automática», siguiendo unas pocas reglas relativamente sencillas

que a menudo pueden aplicarse sin necesidad de hacer un gran esfuerzo mental. Pese a todo, en este cálculo hay un contraste sorprendente entre las operaciones de diferenciación e integración atendiendo a lo que es lo «fácil» y lo que es lo «difícil». Cuando se trata de aplicar las operaciones a fórmulas explícitas que implican funciones conocidas, la diferenciación es la «fácil» y la integración la «difícil», y en muchos casos la última puede ser imposible de llevar a cabo de una manera explícita. Por el contrario, cuando las funciones no están dadas en términos de fórmulas, sino que vienen dadas en forma de listas tabuladas de datos numéricos, entonces la integración es la «fácil» y la diferenciación la «difícil» y la que, estrictamente hablando, quizá no sea posible en la forma usual. Las técnicas numéricas están relacionadas generalmente con aproximaciones, pero también en la teoría exacta hay una analogía muy estrecha con este aspecto de las cosas, y de nuevo es la integración la que puede realizarse en circunstancias en las que la diferenciación no puede hacerse. Tratemos de entender algo de esto. Estas cuestiones tienen que ver, de hecho, con lo que realmente se entiende por una «función».

Para Euler y los demás matemáticos de los siglos XVII y XVIII, una «función» significaría algo que se podría escribir explícitamente, como x^2 o sen x o $\log(3 - x + e^x)$, o quizá algo definido por cierta fórmula que incluye una integración o tal vez por una serie de potencias dada explícitamente. En la actualidad, se prefiere pensar en términos de «aplicaciones», mediante las que cierta colección A de números (o entes más generales) llamada *dominio* de la función se «aplica» en cierta colección B, llamada *imagen* de la función (véase la Fig. 6.1). El punto esencial de esto es que la función asignará un miembro de la imagen B

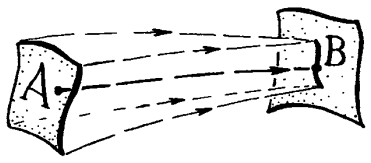


Fig. 6.1. Una *función* como «aplicación», en donde su dominio (una colección A de números o de otras entidades) se «aplica» en su *imagen* (otra colección B). A cada elemento de A se le asigna un valor particular en B, aunque diferentes elementos de A pueden alcanzar el mismo valor y algunos valores de B pueden no ser alcanzados.

a cada miembro del dominio A. (Podemos pensar que la función «examina» un número que pertenece a A y, entonces, dependiendo solo de qué número encuentre, produce un número definido perteneciente a B.) Una función de este tipo puede ser simplemente una «tabla de búsqueda». No es necesario que haya una «fórmula» de apariencia razonable que exprese la acción de la función de una manera manifiestamente explícita.

Consideremos algunos ejemplos. En la Fig. 6.2 he dibujado las gráficas de tres funciones[1] sencillas, a saber, las dadas por x^2, $|x|$ y $\theta(x)$. En cada caso, los espacios dominio e imagen son ambos la totalidad de los *números reales*, que suelen representarse mediante el símbolo $\mathbb{R}$. La función que estoy denotando por «x^2» da simplemente el cuadrado del número real que está examinando. La función denotada por «$|x|$» (llamada *valor absoluto*) da simplemente x si x es no negativo, pero da $-x$ si x es negativo; así pues, el propio valor $|x|$ no es nunca negativo. La función $\theta(x)$ es 0 si x es negativo y 1 si x es positivo; también es habitual definir $\theta(0) = 1/2$. (Esta función se denomina *función escalón* de Heaviside; véase §21.1 para otra importante influencia matemática de Oliver Heaviside, quizá más conocido por haber postulado por primera vez la «capa de Heaviside» en la atmósfera terrestre, tan importante para las transmisiones por radio.) Cada una de estas es una función perfectamente buena en el sentido moderno del término, pero Euler[2] habría tenido dificultad en aceptar que $|x|$ o $\theta(x)$ fueran realmente una «función» en *su* sentido del término.

¿Por qué podría ser así? Una posibilidad es pensar que la dificultad

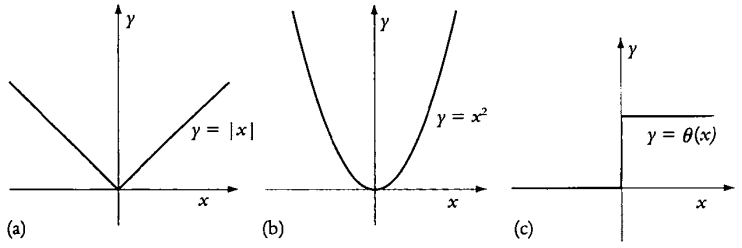


Fig. 6.2. Gráficas de (a) $|x|$, (b) x^2, y (c) $\theta(x)$; en cada caso el dominio y la imagen son el sistema de los números reales.

con $|x|$ y $\theta(x)$ es que hay demasiadas cosas del tipo: «si x es tal y cual, entonces tomemos esto y aquello, mientras que si x es ...», y no hay una «fórmula bonita» para la función. Sin embargo, esto es un poco vago, y en cualquier caso podríamos preguntarnos qué hay realmente erróneo en que $|x|$ cuente como una «fórmula». Además, una vez que hemos aceptado $|x|$, podríamos escribir[6.1] una *fórmula* para $\theta(x)$

$$\theta(x) = \frac{|x| + x}{2x}$$

(aunque podríamos preguntarnos si hay un sentido aceptable en el que esto da el valor correcto para $\theta(0)$, puesto que la fórmula da simplemente $0/0$). Una objeción más pertinente es decir que la dificultad con $|x|$ consiste en que no es «suave», y no en que su expresión explícita no sea «bonita». Vemos esto en el «ángulo» en el centro de la Fig. 6.2a. La presencia de este ángulo es lo que impide que $|x|$ tenga una *pendiente* bien definida en $x = 0$. Tratemos ahora de entender esta idea.

6.2. Pendientes de funciones

Como se ha comentado antes, una de las cosas con las que tiene que ver el cálculo diferencial es con hallar «pendientes». Vemos claramente en la gráfica de $|x|$, tal como se muestra en la Fig. 6.2a, que no tiene una pendiente única en el origen, donde está nuestro molesto ángulo. En cualquier otro lugar, la pendiente está bien definida, pero no en el origen. Debido a esta dificultad en el origen, decimos que $|x|$ *no es diferenciable* en el origen o, de forma equivalente, que no es *suave* allí. Por el contrario, la función x^2 tiene una pendiente perfectamente buena definida unívocamente en todo lugar, como se ilustra en la Fig. 6.2b. De hecho, la función x^2 es diferenciable en todo lugar.

La situación con $\theta(x)$, tal como se ilustra en la Fig. 6.2c, es incluso peor que para $|x|$. Nótese que $\theta(x)$ da un «salto» desagradable en el origen ($x = 0$). Decimos que $\theta(x)$ es *discontinua* en el origen. Por el contrario, las dos funciones x^2 y $|x|$ son *continuas* en todo lugar. La di-

[6.1] Demuéstrelo (ignorando $x = 0$).

ficultad de $|x|$ en el origen no es una solución de continuidad, sino de diferenciabilidad. (Aunque la solución de continuidad y de suavidad son cosas diferentes, son realmente conceptos interrelacionados, como veremos en breve.)

Ninguno de estos fallos hubiera agradado, presumiblemente, a Euler, y parecen proporcionar razones por las que $|x|$ y $\theta(x)$ podrían no ser consideradas como funciones «propiamente dichas». Pero consideremos ahora las dos funciones ilustradas en la Fig. 6.3. La primera, x^3, sería aceptable para el criterio de cualquiera; pero ¿qué pasa con la segunda, que puede definirse mediante la expresión $x|x|$, y que ilustra la función que es x^2 cuando x es no negativo y $-x^2$ cuando x es negativo? A simple vista, las dos gráficas parecen muy similares y ciertamente «suaves». De hecho, ambas tienen un valor perfectamente bueno para la «pendiente» en el origen, a saber, «cero» (lo que significa que las curvas tienen allí una pendiente horizontal) y son en realidad «diferenciables» en todo lugar, en el sentido más directo de dicha palabra. Pese a todo, $x|x|$ no parece ser ciertamente el tipo «bonito» de función que hubiera satisfecho a Euler.

Lo que «falla» con $x|x|$ es que no tiene una *curvatura* bien definida en el origen, y la noción de curvatura es ciertamente algo que concierne al cálculo diferencial. De hecho, la «curvatura» es algo que incluye lo que se denominan «derivadas segundas», que significa hacer la «diferenciación» dos veces. En efecto, decimos que la función $x|x|$ *no es dos veces diferenciable* en el origen. Llegaremos a las derivadas segundas y superiores en §6.3.

Para empezar a entender todo esto, necesitaremos ver qué es lo que

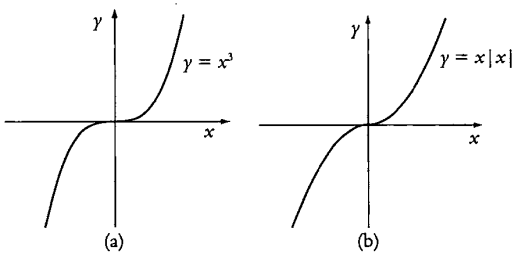


Fig. 6.3. Gráficas de (a) x^3 y (b) $x|x|$ (i.e., x^2 si $x \geq 0$ y $-x^2$ si $x < 0$).

hace realmente la operación de diferenciación. Para ello tenemos que saber cómo se mide una *pendiente*. Esto se ilustra en la Fig. 6.4. He mostrado una función de apariencia bastante representativa que llamaré $f(x)$. La curva de la Fig. 6.4a muestra la relación $y = f(x)$, donde el valor de la coordenada y mide la altura, mientras que x mide el desplazamiento horizontal, como es habitual en una descripción cartesiana. He indicado la pendiente de la curva en un punto particular p, como el incremento en la coordenada y dividido por el incremento en la coordenada x, cuando procedemos a lo largo de la *línea tangente* a la curva en el punto p. (La definición técnica de «línea tangente» depende de los procedimientos adecuados de paso al límite, pero no es mi propósito ahora el discutir estos tecnicismos. Espero que el lector encontrará mis descripciones intuitivas adecuadas para nuestros propósitos inmediatos.)[3] La notación estándar para el valor de dicha pendiente es dy/dx. Podemos considerar «dy» como un incremento muy pequeño en el valor de y a lo largo de la curva y dx como el correspondiente incremento minúsculo en el valor de x. (Aquí, la corrección técnica nos exigiría un paso al «límite», a medida que cada uno de estos minúsculos incrementos se reducen a cero.)

Ahora podemos considerar otra curva, que representa (frente a x) dicha pendiente en cada punto p, para las diversas elecciones posibles de la coordenada x; véase la Fig. 6.4b. De nuevo, estoy utilizando una descripción cartesiana, pero ahora es dy/dx, en lugar de y, lo que se representa en vertical. El desplazamiento horizontal sigue estando medido por x. La función que se está representando aquí se denomina normalmente $f'(x)$, y podemos escribir $dy/dx = f'(x)$. Llamamos a dy/dx la *derivada de y con respecto a x*, y decimos que la función $f'(x)$ es la *derivada*[4] de $f(x)$.

6.3. Derivadas de orden superior; funciones C^∞-suaves

Veamos ahora lo que sucede cuando tomamos una derivada *segunda*. Esto significa que ahora estamos considerando la función-pendiente para la nueva curva de la Fig. 6.4b, que muestra $u = f'(x)$, donde ahora u representa a dy/dx. En la Fig. 6.4c he mostrado esta función-pen-

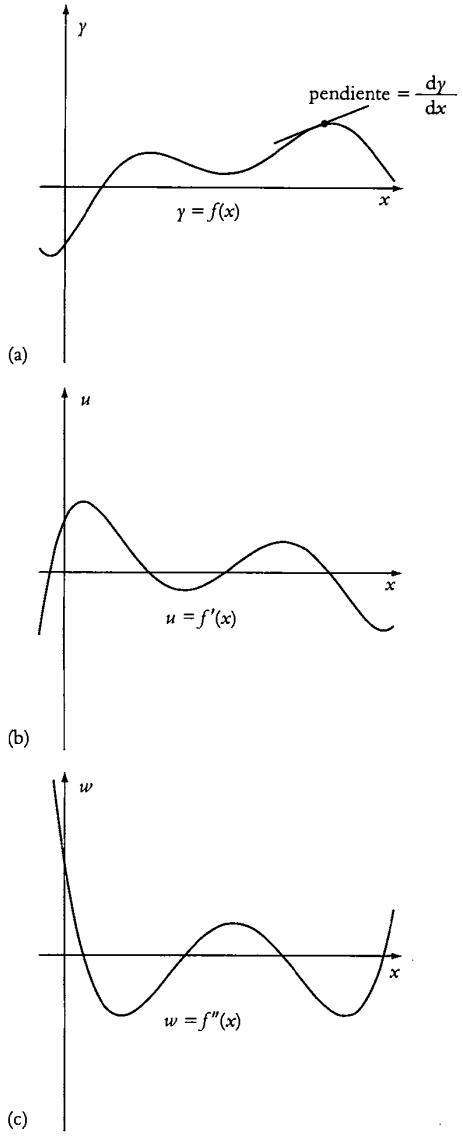


Fig. 6.4. Representación cartesiana de (a) $y = f(x)$; (b) la derivada $u = f'(x)$ $(= dy/dx)$, y (c) la segunda derivada $f''(x) = d^2y/dx^2$. (Nótese que $f(x)$ tiene pendiente horizontal solamente donde $f'(x)$ corta al eje x, y tiene un punto de inflexión donde $f''(x)$ corta al eje x.)

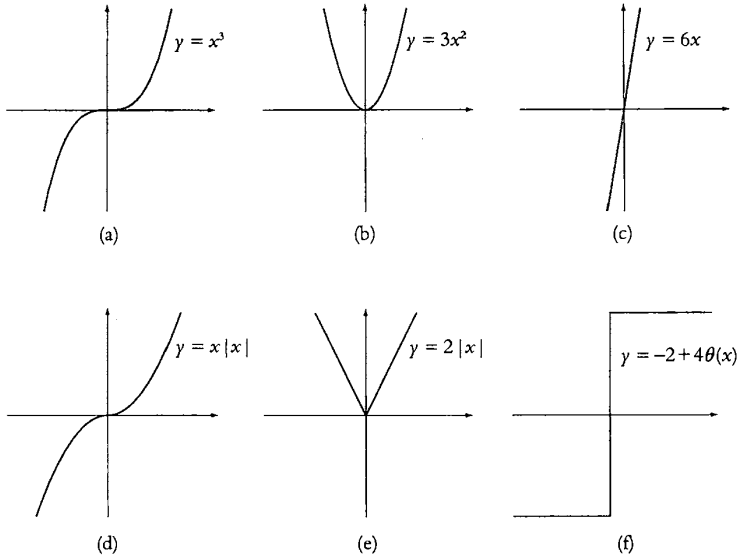


Fig. 6.5. (a) (b), (c) Representaciones de x^3, su primera derivada $3x^2$ y su segunda derivada $6x$, respectivamente. (d), (e), (f) Representaciones de $x|x|$, su primera derivada $2|x|$ y la segunda derivada $-2 + 4\theta(x)$, respectivamente.

diente de «segundo orden», que es la gráfica de du/dx frente a x, de la misma forma que lo hemos hecho antes para dy/dx, de modo que el valor du/dx nos da ahora la pendiente de la segunda curva $u = f'(x)$. Esto nos da lo que se denomina derivada segunda de la función original $f(x)$, y se escribe normalmente $f''(x)$. Cuando sustituimos u por dy/dx en la cantidad du/dx obtenemos la *derivada segunda* de y con respecto a x, que se escribe (con no mucha lógica) d^2y/dx^2.

Nótese que los valores de x donde la función original $f(x)$ tiene pendiente horizontal son precisamente los valores de x donde $f'(x)$ corta al eje x (de modo que dy/dx se anula para dichos valores de x). Los lugares donde $f(x)$ alcanza un máximo o un mínimo (local) se dan en tales posiciones, lo que es importante cuando estamos interesados en encontrar los valores máximos y mínimos (localmente) de una función. ¿Qué pasa con los lugares donde la derivada segunda $f''(x)$ corta al eje x? Estos se dan donde la *curvatura* de $f(x)$ se anula. En general, estos

puntos son aquellos donde la dirección en que se «dobla» la curva $y = f(x)$ cambia de un lado de la curva a otro, en un lugar llamado *punto de inflexión*. (De hecho, no sería correcto decir que $f''(x)$ «mide» realmente la curvatura de la curva definida por $y = f(x)$, en general; la curvatura real viene dada por una expresión más complicada[5] que $f''(x)$, pero que incluye a $f''(x)$, y la curvatura se anula cuando $f''(x)$ se anula.

Consideremos a continuación nuestras dos funciones x^3 y $|x|x$ de apariencia (superficialmente) similar consideradas más arriba. En la Fig. 6.5a,b,c, he representado x^3 y sus derivadas primera y segunda, como he hecho con la función $f(x)$ en la Fig. 6.4, y en la Fig. 6.5d,e,f he hecho lo mismo con $|x|x$. En el caso de x^3 vemos que no hay ningún problema con la continuidad o la suavidad ni en la primera ni en la segunda derivada. De hecho, la primera derivada es $3x^2$ y la segunda es $6x$, ninguna de las cuales hubiese dado ningún momento de preocupación a Euler. (Veremos dentro de poco cómo obtener estas expresiones explícitas.) Sin embargo, en el caso de $|x|x$ encontramos algo muy parecido al «ángulo» de la Fig. 6.2a para la derivada primera, y un comportamiento de «función-escalón» para la derivada segunda, muy similar a la Fig. 6.2c. Tenemos una solución de suavidad para la derivada primera y una solución de continuidad para la segunda. A Euler no le hubiera gustado nada. La derivada primera es realmente $2|x|$ y la derivada segunda es $-2 + 4\theta|x|$. (Mis lectores más escrupulosos podrían quejarse y aducir que yo no debería escribir tan alegremente una «derivada» para $2|x|$, que no es realmente diferenciable en el origen. Cierto, pero no vamos a discutir por eso: puede conseguirse una justificación completa para esto utilizando las nociones que se introducirán al final del capítulo 9.)

Es fácil imaginar que se pueden construir funciones para las que tales soluciones de suavidad o de continuidad no se manifiestan hasta que no se han llevado a cabo un gran número de derivadas. Lo cierto es que bastará con funciones de la forma $x^n|x|$, donde podemos tomar n como un número entero positivo tan grande como queramos. La terminología matemática para este tipo de cosas es decir que la función $f(x)$ es C^n-*suave* si puede ser diferenciada n veces (en cada punto de su dominio) y la n-ésima derivada es continua.[6] La función $x^n|x|$ es, de hecho, C^n-suave pero no es C^{n+1}-suave en el origen.

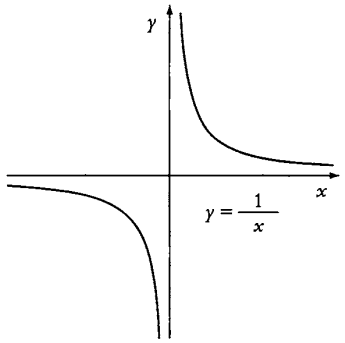


Fig. 6.6. Representación de $1/x$.

¿Qué valor de n satisfaría a Euler? Parece claro que él no se hubiera contentado con detenerse en *cualquier* valor particular de n. Con seguridad, el tipo de función autorrespetable que Euler hubiera aprobado sería una que se pudiera diferenciar *tantas veces como queramos*. Para cubrir esta situación, los matemáticos denominan a una función C^∞-suave si es C^n-suave para *todo* entero positivo n. Para decirlo de otra forma, una función C^∞-suave debe ser diferenciable tantas veces como queramos.

Podemos presumir que la idea de Euler de una función hubiera exigido algo parecido a la C^∞-suavidad. Al menos podríamos imaginar que él hubiera esperado que su función fuera C^∞-suave en la mayoría de los lugares de su dominio. Pero ¿qué pasa con la función $1/x$? (véase la Fig. 6.6.) Ciertamente, esta no es C^∞-suave en el origen. Ni siquiera está *definida* en el origen en el sentido moderno de una función. A pesar de este problema, nuestro Euler hubiera aceptado ciertamente $1/x$ como una «función» decente. Después de todo, existe para ella una sencilla fórmula de apariencia natural. Cabe imaginar que Euler no se hubiera interesado tanto en que sus funciones fueran C^∞-suaves en *todos* los puntos de su dominio (suponiendo que él se hubiera preocupado siquiera de «dominios».) Quizá el que las cosas no funcionasen en el punto singular no importaría. Pero $|x|$ y $\theta(x)$ solo fallaban en el mismo «punto singular» en que lo hacía $1/x$. Parece que, a pesar de nuestros esfuerzos, aún no hemos captado la noción «euleriana» de función por la que nos estamos esforzando.

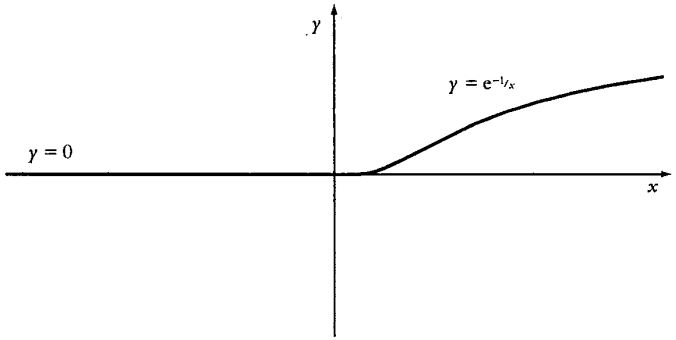


Fig. 6.7. Representación de $y = h(x)$ (= 0 si $x \leq 0$ y $= e^{-1/x}$ si $x > 0$), que es C^∞-suave.

Tomemos otro ejemplo. Consideremos la función $h(x)$ definida por las reglas

$$h(x) = \begin{cases} 0 & \text{si } x \leq 0, \\ e^{-1/x} & \text{si } x > 0. \end{cases}$$

La gráfica de esta función se muestra en la Fig. 6.7. Ciertamente, tiene el aspecto de una función suave. De hecho, es *muy* suave. Es C^∞-suave sobre todo el dominio de los números reales. (Probar esto es el tipo de cosas que uno hace en un curso de la licenciatura en matemáticas. Recuerdo cómo tuve que enfrentarme personalmente a ello cuando era estudiante.)[6.2] A pesar de su absoluta suavidad, uno puede imaginar a Euler mirando por encima del hombro a una función definida de esta manera. Claramente no es «una sola función» en el sentido de Euler. Son «dos funciones pegadas», con independencia de lo suave que sea el trabajo de pegado hecho para empalmar la «falla» en el origen. En contraste, para Euler $1/x$ *es* una sola función, pese al hecho de que está separada en dos piezas por una «punta» muy desagradable en el origen, donde no es ni siquiera continua, y mucho menos suave (Fig. 6.6). Para nuestro Euler, la función $h(x)$ no es realmente mejor que $|x|$ o $\theta(x)$. En estos casos teníamos claramente «dos funciones pegadas», aunque con trabajos de pegado mucho más chapuceros (y con $\theta(x)$, el pegado parece haberse separado por completo).

[6.2] Intente demostrar esto si tiene la formación necesaria.

6.4. ¿La noción «euleriana» de función?

¿Cómo hay que entender esta noción «euleriana» de tener solo una única función y no una colección de funciones separadas? Como el ejemplo de $h(x)$ muestra con claridad, la C^∞-suavidad no es suficiente. Resulta que hay dos enfoques de apariencia completamente diferente para resolver esta cuestión. Uno de ellos utiliza los números complejos, y es engañosamente sencillo de enunciar, aunque trascendental en sus consecuencias. Exigimos que nuestra función $f(x)$ se pueda extender a una función $f(z)$ de la variable compleja z, de modo que $f(z)$ sea suave en el sentido de que meramente se requiere que sea *una vez* diferenciable con respecto a la variable compleja z. (Así pues, $f(z)$ es, en el sentido complejo, un tipo de C^1-función.) El hecho de que sea esto todo lo que necesitamos es una muestra extraordinaria de auténtica magia. Si $f(z)$ puede ser diferenciada una vez con respecto a la variable compleja z, ¡entonces puede ser diferenciada tantas veces como queramos!

Volveré al tema del cálculo infinitesimal complejo en la próxima sección. Pero hay otro enfoque para la solución de este problema de la «noción euleriana de función» que utiliza solo números reales, y este enfoque implica el concepto de serie de potencias que encontramos en §2.5. (Una de las cosas en las que Euler era un auténtico maestro en manipular series de potencias.) Será útil considerar la cuestión de las series de potencias, en esta sección, antes de volver al tema de la diferenciabilidad compleja. El hecho de que, localmente, la diferenciabilidad compleja resulta ser equivalente a la validez del desarrollo en serie de potencias es una de las piezas verdaderamente grandes de la magia de los números complejos.

Llegaré a todo esto a su debido tiempo, pero por el momento quedémonos en las funciones reales. Supongamos que cierta función $f(x)$ tiene realmente una representación en serie de potencias:

$$f(x) = a_0 + a_1x + a_2x^2 + a_3x^3 + a_4x^4 + \dots$$

Ahora bien, existen métodos para encontrar, a partir de $f(x)$, cuáles deben ser los coeficientes $a_0, a_1, a_2, a_3, a_4, \dots$ Para que exista tal desarrollo, es necesario (aunque no suficiente, como veremos de inmediato) que $f(x)$ sea C^∞-suave, de modo que tendremos nuevas funciones $f'(x)$,

$f''(x), f'''(x), f''''(x), \ldots$, etc., que son las derivadas primera, segunda, tercera, cuarta, etc., de $f(x)$, respectivamente. De hecho, nos interesaremos en los valores de dichas funciones solo en el origen ($x = 0$), y solo allí necesitamos la C^∞-suavidad de $f(x)$. El resultado (a veces denominado *serie de Maclaurin*)[7] es que si $f(x)$ tiene un desarrollo tal en serie de potencias, entonces[6.3]

$$a_0 = f(0),\ a_1 = \frac{f'(0)}{1!},\ a_2 = \frac{f''(0)}{2!},\ a_3 = \frac{f'''(0)}{3!},\ a_4 = \frac{f''''(0)}{4!},\ \ldots$$

(Recordemos, de §5.3, que $n! = 1 \times 2 \times \ldots \times n$.) Pero ¿qué pasa a la inversa? Si se dan las a^n de este modo, ¿se sigue de ello que la suma realmente nos da $f(x)$ (en un intervalo que incluye al origen)?

Volvamos a nuestra $h(x)$ aparentemente sin fisuras. Quizá podamos detectar un fallo en el punto de unión ($x = 0$) utilizando esta idea. Tratemos de ver si $h(x)$ tiene realmente un desarrollo en serie de potencias. Haciendo $f(x) = h(x)$ en lo que precede, consideremos los diferentes coeficientes $a_0, a_1, a_2, a_3, a_4, \ldots$ notando que todos ellos tienen que anularse siempre que x esté a la izquierda del origen, puesto que la serie tiene que coincidir con el valor $h(x) = 0$. De hecho, encontramos que todos ellos se anulan también para $e^{-1/x}$, que es básicamente la razón por la que $h(x)$ es C^∞-suave en el origen, pues empalman todas las derivadas que vienen de los dos lados. Pero esto nos dice también que no hay forma de que la serie de potencias pueda funcionar, puesto que todos los términos son cero (véase el ejercicio [6.1]) y por consiguiente no pueden sumar $e^{-1/x}$. Así pues, *hay* un fallo en la unión en $x = 0$; la función $h(x)$ no puede expresarse como una serie de potencias. Decimos que $h(x)$ no es *analítica* en $x = 0$.

En la exposición anterior me he estado refiriendo realmente a lo que se denominaría un desarrollo en serie de potencias *en torno al origen*. Un análisis similar se aplicaría a cualquier otro punto del dominio real de la función. La diferencia está entonces en que tenemos que «desplazar el origen» a algún otro punto particular, definido por el número real p en el dominio, lo que significa reemplazar x por $x - p$ en el anterior desarrollo en serie de potencias, para obtener:

[6.3] Demuéstrelo utilizando las reglas dadas al final de §6.5.

$$f(x) = a_0 + a_1(x - p) + a_2(x - p)^2 + a_3(x - p)^3 + \ldots,$$

donde ahora

$$a_0 = f(p),\ a_1 = \frac{f'(p)}{1!},\ a_2 = \frac{f''(p)}{2!},\ a_3 = \frac{f'''(p)}{3!}\ \ldots,$$

Esto se denomina un desarrollo en serie de potencias *en torno a p*. La función $f(x)$ se llama *analítica en p* si puede expresarse como un desarrollo semejante en serie de potencias en algún intervalo que englobe a $x = p$. Si $f(x)$ es analítica en *todos* los puntos de su dominio, la llamamos simplemente *función analítica* o, lo que es equivalente, una función C^ω-suave. Las funciones analíticas son, en un sentido claro, incluso «más suaves» que las funciones C^∞-suaves. Además, tienen la propiedad de que no es posible salir del paso pegando dos funciones analíticas «diferentes», a la manera de los ejemplos $\theta(x)$, $|x|$, $x|x|$, $x^n|x|$, o $h(x)$, que se han dado antes. Euler se habría mostrado feliz con las funciones analíticas. ¡Son funciones realmente «respetables»!

Sin embargo, resulta complicado seguir estas series de potencias, aunque solo sea en la imaginación. La forma «compleja» de considerarlas es más económica, e incluso nos ofrece una comprensión más profunda. Por ejemplo, la función $1/x$ no es analítica en $x = 0$, pero sigue siendo «una función».[6.4] La «filosofía de la serie de potencias» no nos dice esto directamente. Pero desde el punto de vista de los números complejos, $1/x$ es claramente una sola función, como veremos.

6.5. Las reglas de diferenciación

Antes de examinar estas materias será útil decir algo acerca de las maravillosas reglas que nos proporciona el cálculo diferencial; reglas que nos permiten diferenciar funciones casi sin pensar, ¡aunque, por supuesto, solo tras meses de práctica! Estas reglas nos permiten ver la forma de escribir directamente la derivada de muchas funciones, en especial cuando se representan como series de potencias.

[6.4] Consideremos la «función única» e^{-1/x^2}. Demuestre que es C^∞, pero no analítica en el origen.

Recordemos, de paso, que antes he comentado que la derivada de x^3 es $3x^2$. Este es un caso particular de una fórmula sencilla pero importante: la derivada de x^n es nx^{n-1}, que podemos escribir

$$\frac{\mathrm{d}(x^n)}{\mathrm{d}x} = nx^{n-1}.$$

(Nos distraeríamos demasiado si intentara explicar por qué es válida esta fórmula. No es realmente difícil demostrarlo, y el lector interesado puede encontrar todo lo que se necesita en cualquier libro de texto elemental sobre cálculo infinitesimal.[8] Dicho sea de paso, n no tiene por qué ser entero.) También podemos expresar[9] esta ecuación («multiplicando por dx») mediante la fórmula conveniente

$$\mathrm{d}(x^n) = nx^{n-1}\mathrm{d}x.$$

No necesitamos saber mucho más por lo que respecta a la diferenciación de series de potencias. Hay básicamente otras dos cosas. En primer lugar, la derivada de una suma de funciones es la suma de las derivadas de las funciones:

$$\mathrm{d}[f(x) + g(x)] = \mathrm{d}f(x) + \mathrm{d}g(x).$$

Esto se extiende a una suma de cualquier número finito de funciones.[10] En segundo lugar, la derivada del producto de una función por una constante es el producto de la constante por la derivada de la función:

$$\mathrm{d}\{a\,f(x)\} = a\,\mathrm{d}f(x).$$

Por una «constante» entiendo un número que no varía con x. Los *coeficientes* $a_0, a_1, a_2, a_3, \ldots$ en la serie de potencias son constantes. Con estas reglas podemos diferenciar directamente *cualquier* serie de potencias.[6.5]

Otra manera de expresar la constancia de a es

$$\mathrm{d}a = 0.$$

[6.5] Utilizando la serie de potencias para e^x dada en §5.3, demuestre que $\mathrm{d}e^x = e^x\mathrm{d}x$.

Teniendo esto en cuenta, encontramos que la regla inmediatamente anterior a esta es realmente un caso especial (con $g(x) = a$) de la «ley de Leibniz»:

$$\mathrm{d}\{f(x)\,g(x)\} = f(x)\,\mathrm{d}g(x) + g(x)\,\mathrm{d}f(x)$$

(y $\mathrm{d}(x^n)/\mathrm{d}x = nx^{n-1}$, para cualquier número natural n, también puede obtenerse de la ley de Leibniz).[6.6] Otra ley útil es

$$\mathrm{d}\{f(g(x))\} = f'(g(x))g'(x)\mathrm{d}x.$$

A partir de las dos últimas y de la primera, poniendo $f(x)[g(x)]^{-1}$ en la ley de Leibniz, podemos deducir[6.7]

$$\mathrm{d}\left(\frac{f(x)}{g(x)}\right) = \frac{g(x)\,\mathrm{d}f(x) - f(x)\,\mathrm{d}g(x)}{g(x)^2}.$$

Provisto de estas pocas reglas (y muchísima práctica), uno puede convertirse en un «experto en diferenciación» sin necesidad de tener mucho *conocimiento* real de por qué funcionan las reglas. Esta es la potencia que tiene un buen cálculo.[6.8] Además, con el conocimiento de las derivadas de tan solo algunas funciones especiales,[6.9] uno puede convertirse en un experto aún *mayor*. Solo para que el lector no iniciado pueda convertirse en un «miembro al instante» del club de diferenciadores expertos, permítanme dar los ejemplos principales:[11,6.10]

$$\mathrm{d}(e^x) = e^x\,\mathrm{d}x,$$

$$\mathrm{d}(\log x) = \frac{\mathrm{d}x}{x},$$

$$\mathrm{d}(\mathrm{sen}\,x) = \cos x\,\mathrm{d}x,$$

$$\mathrm{d}(\cos x) = -\mathrm{sen}\,x\,\mathrm{d}x,$$

$$\mathrm{d}(\tan x) = \frac{\mathrm{d}x}{\cos^2 x},$$

[6.6] Establezca esto.
[6.7] Dedúzcalo.
[6.8] Calcule $\mathrm{d}y/\mathrm{d}x$ para $y = (1 - x^2)^4$, $y = (1 + x)/(1 - x)$.
[6.9] Con a constante, calcule $\mathrm{d}(\log_a x)$, $\mathrm{d}(\log_x a)$, $\mathrm{d}(x^x)$.
[6.10] Para el primero, véase el ejercicio [6.5]; deducir el segundo a partir de $\mathrm{d}(e^{\log x})$; el tercero y el cuarto a partir de $\mathrm{d}e^{ix}$, suponiendo que las cantidades complejas funcionan como las reales; y el resto de las anteriores, utilizando $\mathrm{d}(\mathrm{sen}\,(\mathrm{sen}^{-1}\,x))$, etc., y notando que $\cos^2 x + \mathrm{sen}^2 x = 1$.

$$\mathrm{d}(\sin^{-1}x) = \frac{\mathrm{d}x}{\sqrt{1-x^2}},$$

$$\mathrm{d}(\cos^{-1}x) = \frac{-\mathrm{d}x}{\sqrt{1-x^2}},$$

$$\mathrm{d}(\tan^{-1}x) = \frac{\mathrm{d}x}{1+x^2}.$$

Esto ilustra el punto mencionado al principio de esta sección según el cual, cuando se nos dan fórmulas explícitas, la operación de diferenciación es «fácil». Por supuesto, con ello no quiero decir que sea algo que uno pueda hacer mientras duerme. De hecho, en ejemplos concretos, puede darse el caso de que se obtengan expresiones realmente muy complicadas. Cuando digo «fácil», quiero decir que hay un procedimiento computacional explícito para llevar a cabo la diferenciación. Si sabemos cómo diferenciar cada uno de los ingredientes de una expresión, entonces los procedimientos del cálculo, como los dados antes, nos dicen cómo llegar a diferenciar la expresión completa. «Fácil», aquí significa en realidad algo que podría ser programado directamente en un computador. Pero las cosas son muy diferentes si tratamos de ir en dirección inversa.

6.6. Integración

Como se ha afirmado al principio de esta sección, la *integración* es la operación inversa de la diferenciación. Esto equivale a tratar de encontrar una función $g(x)$ para la que $g'(x) = f(x)$, i.e., encontrar una solución $y = g(x)$ a la ecuación $\mathrm{d}y/\mathrm{d}x = f(x)$. Otra forma de decirlo es que en lugar de bajar de una imagen a otra en la Fig. 6.4 (o 6.5), tratamos de hacer nuestro camino hacia arriba. La belleza del «teorema fundamental del cálculo infinitesimal» reside en que este procedimiento nos está diciendo cómo calcular las áreas bajo cada curva sucesiva. Echemos una ojeada a la Fig. 6.8. Recordemos que la curva inferior $u = f(x)$ puede obtenerse a partir de la curva superior $y = g(x)$ porque representa las pendientes de esta curva, al ser $f(x)$ la derivada de $g(x)$. Esto es precisamente lo que teníamos antes. Pero empecemos ahora con la curva in-

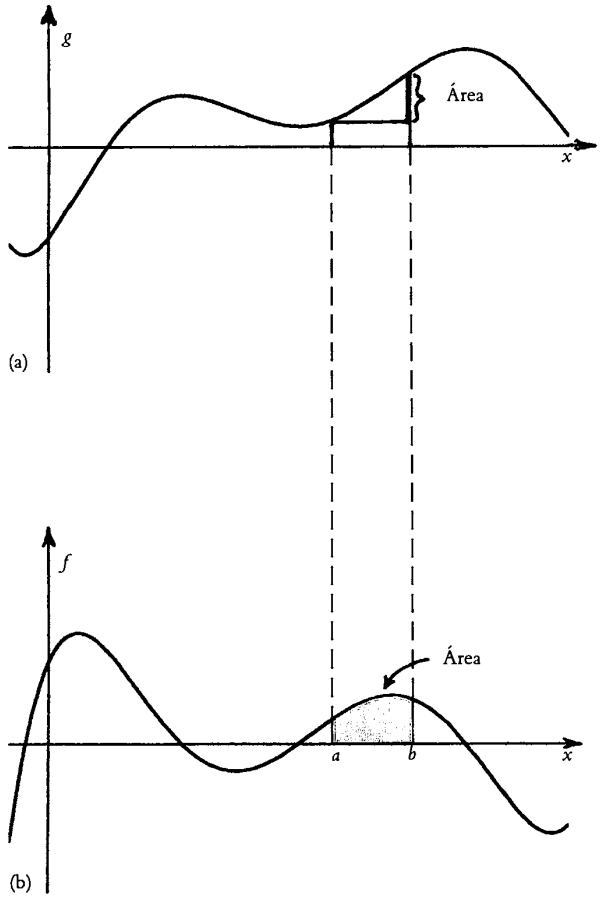


Fig. 6.8. Teorema fundamental del cálculo: reinterpretamos la Fig. 6.4a,b, procediendo hacia arriba y no hacia abajo. La curva superior (a) representa las áreas bajo la curva inferior (b), donde el área limitada por dos líneas verticales $x = a$ y $x = b$, el eje x y la curva inferior es la diferencia $g(b) - g(a)$ de alturas de la curva superior en esos dos valores de x (teniendo en cuenta los signos).

ferior. Encontramos que la curva superior simplemente expresa las áreas que hay bajo la curva inferior. De forma un poco más explícita: si en la imagen inferior tomamos dos rectas verticales dadas por $x = a$ y $x = b$, respectivamente, entonces el área limitada por estas dos rectas,

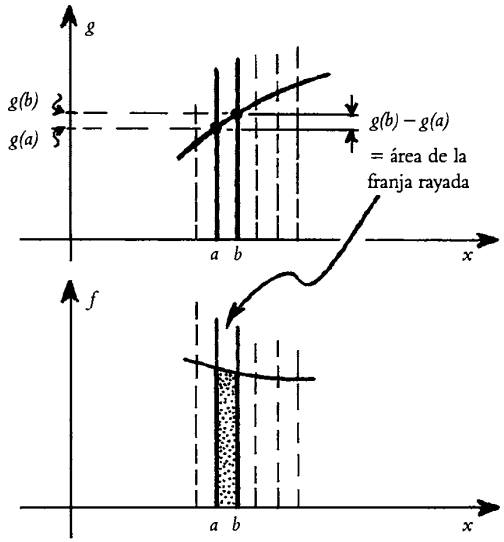


Fig. 6.9. Tómese $b > a$ en una cantidad minúscula. En la imagen inferior, el área de una franja muy estrecha entre líneas vecinas $x = a$, $x = b$ es esencialmente el producto de la anchura de la franja $b - a$ por su altura (del eje x a la curva). Esta altura es la pendiente de la curva superior, de donde el área de la franja es esta pendiente × la anchura de la franja, que es la cantidad en que crece la curva superior de a a b, i.e., $g(b) - g(a)$. Sumando muchas franjas estrechas, encontramos que el área de una franja ancha bajo la curva inferior es la cantidad correspondiente en que aumenta la curva superior.

el eje x y la propia curva, será la diferencia entre las alturas de la curva superior en esos dos valores de x. Por supuesto, en cuestiones como estas debemos tener cuidado con los «signos». En las regiones donde la curva inferior está por debajo del eje x, las áreas cuentan negativamente. Además, en la imagen he tomado $a < b$ y la «diferencia entre las alturas» de la curva superior en la forma $g(b) - g(a)$. Los signos se invertirían si $a > b$.

En la Fig. 6.9 he tratado de hacer intuitivamente creíble por qué existe esta relación inversa entre pendientes y áreas. Imaginemos que b sea mayor que a tan solo en una minúscula cantidad. Entonces, el área a considerar, en la imagen inferior, es el área de la franja muy estrecha limitada por las rectas vecinas $x = a$ y $x = b$. La medida de esta área es

esencialmente el producto de la anchura minúscula de la franja (i.e., $b - a$) por su altura (desde el eje x a la curva). Pero se supone que la altura de la franja está midiendo la pendiente de la curva superior en dicho punto. Por lo tanto, el área de la franja es esta pendiente multiplicada por la anchura de la franja. Pero la pendiente de la curva superior multiplicada por la anchura de la franja es la cantidad en que aumenta la curva superior desde a a b, es decir, la diferencia $g(b) - g(a)$. Así pues, para franjas muy estrechas, el área está medida realmente por esta diferencia. Las franjas más anchas se consideran construidas a partir de grandes números de franjas estrechas, y obtenemos el área total midiendo cuánto aumenta la curva superior en el intervalo completo.

Aquí hay un punto importante que debería sacar a colación. En el paso desde la curva inferior a la curva superior hay una no unicidad en lo que respecta a la altura a que debe situarse globalmente la curva superior. Solo estamos interesados en *diferencias* entre alturas en la curva superior, de modo que desplazar la curva hacia arriba o hacia abajo en una cantidad constante no supone ninguna diferencia. Esto también queda claro por la interpretación de la «pendiente», puesto que la pendiente en puntos diferentes de la curva superior será exactamente la misma si la desplazamos hacia arriba o hacia abajo. Esto equivale, en nuestro cálculo, a que si sumamos una constante C a $g(x)$, entonces la diferenciación de la función resultante sigue dando $f(x)$:

$$\mathrm{d}(g(x) + C) = \mathrm{d}g(x) + \mathrm{d}C = f(x)\,\mathrm{d}x + 0 = f(x)\,\mathrm{d}x.$$

Una función semejante $g(x)$, o equivalentemente $g(x) + C$ para una constante arbitraria C, se denomina *integral indefinida* de $f(x)$, y escribimos

$$\int f(x)\,\mathrm{d}x = g(x) + \text{const.}$$

Esta es solo otra forma de expresar la relación $\mathrm{d}[g(x) + \text{const.}] = f(x)\,\mathrm{d}x$, de modo que simplemente consideramos el signo «$\int$» como el inverso del símbolo «d». Si queremos especificar el área entre $x = a$ y $x = b$, entonces queremos lo que se denomina la *integral definida*, y escribimos

$$\int_a^b f(x)\,\mathrm{d}x = g(b) - g(a).$$

Si conocemos la función $f(x)$ y queremos obtener su integral $g(x)$, apenas tenemos reglas tan directas para obtenerla como las teníamos para la diferenciación. Se conocen muchos trucos, varios de los cuales pueden hallarse en los libros de texto estándar y en los paquetes informáticos, pero no bastan para tratar todos los casos. De hecho, con frecuencia encontramos que hay que ampliar la familia de funciones estándar explícitas que habíamos estado utilizando previamente, e «inventar» nuevas funciones para expresar los resultados de la integración. En efecto, ya hemos visto esto en los ejemplos especiales que se han dado antes. Supongamos que estuviéramos familiarizados solo con funciones construidas a partir de combinaciones de potencias de x. En el caso de una potencia general x^n, podemos integrarla para obtener $x^{n+1}/(n+1)$. (Esto es simplemente utilizar nuestra fórmula anterior, en §6.5, con $n+1$ en lugar de n: $\mathrm{d}(x^{n+1})/\mathrm{d}x = (n+1)x^n$.) Todo va bien hasta que nos preocupamos por lo que hay que hacer en el caso $n = -1$. Entonces, la supuesta respuesta $x^{n+1}/(n+1)$ tiene un cero en el denominador, de modo que esto no funciona. ¿Cómo integramos entonces x^{-1}? Bien, notemos que, por la mayor de las fortunas, en nuestra lista existe la fórmula $\mathrm{d}(\log x) = x^{-1}\,\mathrm{d}x$. De modo que la respuesta es $\log x$ + const.

¡Esta vez hemos tenido suerte! Sencillamente, ya habíamos estudiado antes la función logaritmo por una razón diferente, y conocíamos algunas de sus propiedades. Pero en otras ocasiones bien podríamos encontrarnos con que no hay ninguna función previamente conocida y en cuyos términos pudiéramos expresar nuestra respuesta. En realidad, a menudo las integrales proporcionan los medios adecuados por los que se definen nuevas funciones. En este sentido, dicha integración explícita es «difícil».

Por el contrario, si no estamos tan interesados en expresiones explícitas sino más bien en cuestiones de *existencia* de funciones que son las derivadas o integrales de funciones dadas, entonces las cosas cambian. La integración es ahora la operación que funciona suavemente y la diferenciación la que causa problemas. Lo mismo es válido cuando se realizan estas operaciones con datos numéricos. Básicamente, el problema con la diferenciación reside en que depende de forma muy crítica de los detalles finos de la función que va a ser diferenciada. Esto puede presentar un problema si no tenemos una expresión explícita

para la función que tiene que ser diferenciada. La integración, por el contrario, es relativamente insensible a tales cuestiones, al estar interesada en la naturaleza global de la función que va a integrar. De hecho, cualquier función continua (una C^0-función) cuyo dominio es un intervalo «cerrado» $a \leq x \leq b$ puede ser integrada,[12] y el resultado es una función C^1 (i.e., C^1-suave). Esta puede ser integrada de nuevo, siendo el resultado C^2, y luego una vez más, dando una función C^3-suave, y así sucesivamente. La integración hace las funciones cada vez más suaves, y podemos continuar así indefinidamente. La diferenciación, por el contrario, solo empeora las cosas, y puede llegar a un final en cierto punto, donde la función se hace «no diferenciable».

Pese a todo, existen enfoques de estas cuestiones que hacen posible que el proceso de diferenciación se continúe también indefinidamente. Ya he insinuado esto cuando me he permitido diferenciar la función $|x|$ para obtener $\theta(x)$, incluso si $|x|$ es «no diferenciable». Podríamos intentar ir más lejos y diferenciar también $\theta(x)$, pese al hecho de que tiene una pendiente infinita en el origen. La «respuesta» es lo que se denomina una *función delta* de Dirac:[13] una entidad de enorme importancia en las matemáticas de la mecánica cuántica. En realidad, la función delta no es una función en absoluto, en el sentido ordinario (moderno) de «función» que aplica espacios dominio en espacios imagen. No hay ningún «valor» para la función delta en el origen (donde solo podría haber sido *infinita*) y es cero en los demás lugares. Pese a todo, la función delta encuentra una definición matemática clara dentro de varias clases más amplias de entidades matemáticas, de las que las *distribuciones* son las mejor conocidas.

Para esto tenemos que extender nuestra noción de C^n-funciones a casos en los que n puede ser un entero negativo. La función $\theta(x)$ es entonces una C^{-1}-función y la función delta es C^{-2}. Cada vez que diferenciamos, debemos disminuir en una unidad la clase de diferenciabilidad (i.e., la clase se hace más negativa en una unidad). Parecería que con todo esto nos estamos alejando cada vez más de la noción de Euler de una «función decente», y que él se negaría a tener ningún trato con cosas semejantes si no fuera por el hecho de que parecen ser útiles. Pese a todo, encontraremos, a su debido tiempo, que es aquí donde los números complejos nos sorprenden con una ironía, ¡una ironía que se ex-

presa en una de sus mayores hazañas mágicas! Tendremos que esperar hasta el final del capítulo 9 para ser testigos de esta hazaña, pues no es algo que pueda describirse ahora adecuadamente. El lector debe seguir conmigo durante un rato, pues primero hay que preparar el terreno, pavimentado con otros ingredientes de soberbia magia.

Notas

Sección 6.1

6.1. Aquí estoy cometiendo un pequeño «abuso de notación», pues técnicamente x^2, por ejemplo, denota el *valor* de la función más que la función. La propia función aplica x en x^2 y podría denotarse por $x \mapsto x^2$, o por $\lambda x[x^2]$, según el *cálculo lambda* de Alonzo Church (1941); véase el capítulo 2 de Penrose (1989).

6.2. En esta sección me referiré a menudo a las opiniones que Euler podría haber tenido con respecto a la idea de función. Sin embargo, debería dejar claro que el «Euler» al que me estoy refiriendo es en realidad un individuo hipotético o idealizado. No tengo ninguna información directa acerca de cuáles podrían haber sido los puntos de vista del Leonhard Euler real en cualquier caso concreto. Pero las opiniones que estoy atribuyendo a mi «Euler» no parecen estar muy alejadas del tipo de opiniones que el Euler real podría haber expresado. Para más información sobre Euler, véanse Boyer (1968), Thiele (1982) y Dunham (1999).

Sección 6.2

6.3. Para más detalles, véase Burkill (1962).

6.4. Estrictamente hablando, es la función f' la que es la derivada de la función f; no podemos obtener el valor de f' en x simplemente a partir del valor de f en x. Véase la nota 6.1.

Sección 6.3

6.5. Viz., $f''(x)/[1 + f'(x)^2]^{3/2}$.

6.6. De hecho, esto implica que todas las derivadas hasta la n-ésima, esta incluida, deben ser continuas, porque la definición técnica de diferenciabilidad requiere continuidad.

Sección 6.4

6.7. Tradicionalmente, este desarrollo en serie de potencias en torno al origen se conoce (con poca justificación histórica) como serie de Maclaurin; el resultado más general en torno al punto p (véase más adelante en la sección) se atribuye a Brook Taylor (1685-1731).

Sección 6.5

6.8. Véase Edwards y Penney (2002).

6.9. Por el momento, trate solo formalmente las siguientes expresiones, o si no «divida por dx» mentalmente si eso le hace más feliz. La notación que estoy utilizando aquí es compatible con la de las formas diferenciales, que se discutirán en §§12.3-6.

6.10. Sin embargo, hay una sutileza técnica en la aplicación de esta ley a la suma del número infinito de términos que necesitamos para una serie de potencias. Esta sutileza puede ignorarse para valores de x estrictamente dentro del círculo de convergencia; véase §4.4. Véase también Priestley (2003).

6.11. Recordemos de §5.1 que sen^{-1}, $\cos^{-1}$ y $\tan^{-1}$ son las funciones inversas de sen, cos y tan, respectivamente. Así, $\text{sen}\,(\text{sen}^{-1}\,x) = x$, etc. No obstante, debemos tener en cuenta que estas funciones inversas son «funciones multivaluadas», y es habitual seleccionar los valores para los cuales

$$-\frac{\pi}{2} \leq \text{sen}^{-1}\,x \leq \frac{\pi}{2},\ 0 \leq \cos^{-1}\,x \leq \pi,\ \text{y}\ -\frac{\pi}{2} < \tan^{-1}\,x < \frac{\pi}{2}.$$

Sección 6.6

6.12. El requisito importante sobre el dominio es que sea lo que se denomina *compacto*; véase §12.6. Los intervalos finitos de la recta real que incluyen sus puntos extremos son, de hecho, compactos.

6.13. Al parecer, Oliver Heaviside también había concebido la «función delta» muchos años antes que Dirac.

7

Cálculo infinitesimal con números complejos

7.1. Suavidad compleja; funciones holomorfas

¿Cómo debemos entender la noción de diferenciación cuando se aplica a una función *compleja* $f(z)$? Ciertamente, no sería oportuno que intentara abordar tal cuestión con todo detalle en este libro.[1] Ni siquiera he abordado de manera adecuada en §6.2 todos estos detalles para una función real. Pero al menos puedo intentar transmitir las ideas básicas. Lo que sigue es un rápido esbozo del argumento esencial para mostrar qué es lo que consigue la diferenciabilidad compleja. Después de esto, seré un poco más explícito acerca de alguno de sus sorprendentes ingredientes.

Básicamente, para la diferenciación compleja exigimos que exista una noción de «pendiente» de la curva compleja $w = f(z)$ en cualquier punto z en el dominio de la función. (Ahora se permite que tanto la función $f(z)$ como la variable z tomen valores complejos.) Para que esta noción de «pendiente» tenga un sentido consistente, cuando movemos ligeramente la variable z en diferentes direcciones en el plano complejo de la z, es necesario que $f(z)$ satisfaga un par de ecuaciones llamadas ecuaciones de *Cauchy-Riemann*[2] (que incluyen las derivadas de las partes real e imaginaria de $f(z)$, tomadas con respecto a las partes real e imaginaria de z; véase §10.5). Estas ecuaciones nos dicen algo bastante notable acerca de la integración compleja —algo que hace posible definir una nueva noción de integración, llamada *integración de contorno*—. Entonces puede darse una bonita fórmula para la n-ésima derivada de $f(z)$ en términos de dicha integración de contorno. Así pues, una vez que tenemos la derivada primera, obtenemos *gratis* todas las derivadas superiores.

A continuación utilizamos esta fórmula para que nos dé los coeficientes de una serie de Taylor que se propone para $f(z)$, y que tenemos que demostrar que converge realmente a $f(z)$. Al haber conseguido esto, tenemos una expresión en serie de Taylor para $f(z)$ que funciona dentro de cualquier círculo en el z-plano complejo en el que $f(z)$ está definida y es diferenciable. Se da así el hecho mágico de que cualquier función compleja que sea suave-compleja ¡es necesariamente analítica!

En consecuencia, en el análisis complejo no hay ningún problema en reconocer las limitaciones de los «trabajos de pegado» en ciertas C^∞-funciones tales como la «$h(x)$» definida en el capítulo anterior. Con seguridad, a Euler le hubiera encantado la potencia de la suavidad compleja. (Por desgracia para el Leonhard Euler real, la sorprendente potencia de esta «suavidad compleja» fue apreciada demasiado tarde para él, ya que fue descubierta por primera vez por Augustin Cauchy en 1821, unos treinta y ocho años después de la muerte de Euler.) Vemos así que la suavidad compleja proporciona una vía para expresar lo que se necesita para nuestra noción «euleriana» de función de una forma mucho más económica que la que ofrece la existencia de un desarrollo en serie de potencias. Pero hay también otra ventaja en considerar tales funciones desde el punto de vista complejo. Recordemos nuestra incómoda «$1/x$» que parecía ser «una sola función», pese al hecho de que la curva real $y = 1/x$ consta de dos fragmentos separados que no están empalmados «analíticamente» a través de valores reales de x. Desde la perspectiva compleja, vemos con claridad que $1/z$ es realmente una sola función. El único lugar donde la función «no funciona» en el plano complejo es el origen $z = 0$. Si eliminamos este punto del plano complejo, seguimos teniendo una región conexa. La parte de la recta real para la que $x < 0$ está conectada a la parte para la que $x > 0$ a través del plano complejo. Así pues, $1/z$ es realmente una función compleja conexa, siendo esto completamente diferente de la situación que teníamos con números reales.

Las funciones que son suave-complejas (analítico-complejas) en este sentido se denominan *holomorfas*. Las funciones holomorfas desempeñarán un papel importante en muchas de nuestras discusiones posteriores. Veremos su importancia en relación con las aplicaciones conformes y las superficies de Riemann en el capítulo 8, y con las se-

ries de Fourier (fundamentales para la teoría de vibraciones) en el capítulo 9. Tienen papeles importantes que desempeñar en la teoría cuántica y en la teoría cuántica de campos (como veremos en §24.3 y §26.3). Son asimismo fundamentales en diversos enfoques para el desarrollo de nuevas teorías físicas (en particular, la teoría de twistores —véase el capítulo 33—, y también desempeñan un papel importante en la teoría de cuerdas; véanse §§31.5,13,14).

7.2. Integración de contorno

Aunque no es este el lugar para desgranar los detalles de los argumentos matemáticos indicados en §7.1, en cualquier caso será ilustrativo desarrollar el esbozo anterior. En particular, será conveniente dar aquí una explicación de la integración de contorno que proporcione al lector una idea de cómo puede utilizarse la integración de contorno para establecer lo que se necesita para los requisitos de §7.1. Recordemos, en primer lugar, la notación que se ha introducido en el capítulo anterior para una integral definida con una variable real x, y consideremos ahora que se aplica a una variable compleja z:

$$\int_a^b f(z)\,\mathrm{d}z = g(b) - g(a),$$

donde $g'(z) = f(z)$. En el caso real, la integral se toma entre un punto a de la recta real y otro punto b en dicha recta. Hay solo un camino para ir de a a b a lo largo de la recta real. Ahora la consideraremos como una fórmula compleja. Aquí tenemos a y b como dos puntos en el plano complejo. Ahora ya no solo tenemos un camino para ir de a a b, sino que podemos dibujar montones de caminos diferentes que conectan a y b. Lo que nos dicen las ecuaciones de Cauchy-Riemann es que si hacemos nuestra integración a lo largo de uno de estos caminos,[3] obtenemos la misma respuesta que si la hacemos a lo largo de cualquier otro camino que pueda obtenerse a partir del primero por una deformación continua dentro del dominio de la función. (Véase la Fig. 7.1.) Esta propiedad es una consecuencia de un simple caso del «teorema fundamental del cálculo exterior», descrito en §12.6. Para algunas funciones, entre las que se cuenta $1/z$, el dominio tiene un «agujero» (en

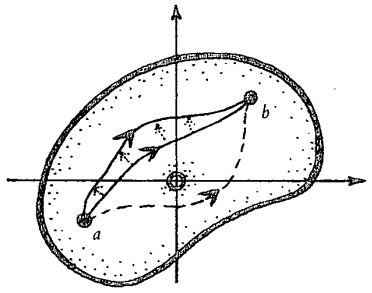


Fig. 7.1. Diferentes caminos de a a b. Integrando una función holomorfa f a lo largo de un camino se obtiene la misma respuesta que a lo largo de cualquier otro camino que pueda obtenerse a partir del primero por una deformación continua dentro del dominio de la f. Para algunas funciones, el dominio tiene un «agujero» dentro (por ejemplo, $z = 0$, para $1/z$) que impide ciertas deformaciones, de modo que pueden obtenerse respuestas diferentes.

el caso $1/z$, el agujero está en $z = 0$), de modo que puede haber varias formas esencialmente diferentes para ir de a a b. Aquí, «esencialmente diferente» se refiere al hecho de que un camino no puede ser deformado de forma continua hasta el otro sin dejar de permanecer en el dominio de la función. En tales casos, el valor de la integral entre a y b puede dar una respuesta diferente para los dos caminos.

En este punto habría que introducir una nota de aclaración (o más bien de corrección). Cuando hablo de un camino que es deformado de forma continua hasta otro, me refiero a lo que los matemáticos llaman deformaciones *homólogas*, y no a las *homotópicas*. En el caso de una deformación homóloga, es legítimo que partes de los caminos se cancelen mutuamente, con tal de que dichas porciones se recorran en sentidos opuestos. Véase la Fig. 7.2 para un ejemplo de este tipo de deformaciones permitido. De dos caminos que son deformables uno hasta otro en este sentido se dice que *pertenecen a la misma clase de homología*. Por el contrario, las deformaciones homotópicas no permiten este tipo de cancelación. Caminos deformables hasta coincidir uno con otro donde no está permitida tal cancelación, pertenecen a la misma *clase de homotopía*. Las curvas homotópicas son siempre homólogas, pero la inversa no es necesariamente cierta. Tanto la homotopía como la homología tienen que ver con la equivalencia bajo movimientos continuos. Así pues, forman parte de la disciplina de la *topología*. Más tarde veremos diferentes aspectos de la topología que desempeñan papeles importantes en otras áreas.

La función $f(z) = 1/z$ es, de hecho, una función para la que se *ob-*

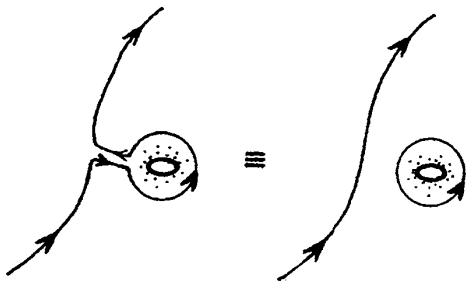


Fig. 7.2. Con una deformación homóloga, partes de los caminos se cancelan mutuamente, si se recorren en sentidos opuestos. A veces esto da lugar a lazos separados.

tienen diferentes respuestas cuando los caminos no son homólogos. Podemos ver por qué debe ser así a partir de lo que ya sabemos sobre los logaritmos. Hacia el final de la sección anterior se advirtió que $\log z$ es una integral indefinida de $1/z$. (De hecho, esto solo se estableció para una variable real x, pero el mismo razonamiento por el que se llega a la respuesta real sirve, asimismo, para llegar a la correspondiente respuesta compleja. Este es básicamente un principio general que se aplica también a cualesquiera otras fórmulas explícitas.) Tenemos así

$$\int_a^b \frac{dz}{z} = \log b - \log a.$$

Pero recordemos de §5.3 que existen diferentes «respuestas» alternativas a un logaritmo complejo. Más pertinente es que podemos pasar de forma continua de una respuesta a otra. Para ilustrarlo, mantengamos a fijo y dejemos que b varíe. De hecho, vamos a permitir que b describa de forma continua un círculo alrededor del origen en sentido positivo (i.e., contrario a las agujas del reloj), que lo lleva de nuevo a su posición original (véase la Fig. 7.3a). Recordemos de §5.3 que la parte imaginaria de $\log b$ es simplemente su *argumento* (i.e., el ángulo que forma b con el eje real positivo, medido en el sentido positivo; véase la Fig. 5.4b). Este argumento se incrementa precisamente en 2π en el curso de este movimiento, de modo que encontramos que $\log b$ ha aumentado en $2\pi i$ (véase la Fig. 7.3b). Así pues, el valor de nuestra integral aumenta en $2\pi i$ por cada nueva vuelta alrededor del origen (en sentido positivo) que da el camino a lo largo del cual se realiza la integral.

Podemos reexpresar este resultado en términos de *contornos cerrados*,

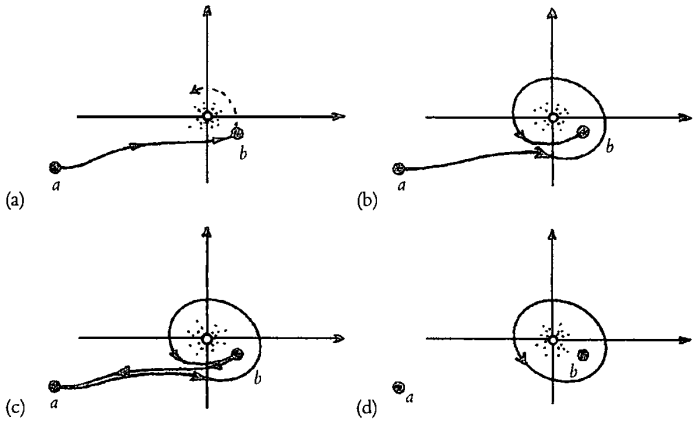


Fig. 7.3. (a) La integración de $z^{-1}\,dz$ de a a b da $\log b - \log a$. (b) Mantener a fijo, y permitir que b dé una vuelta en el sentido contrario a las agujas del reloj alrededor del origen, incrementa $\log b$ en la respuesta en $2\pi i$. (c) Luego volvemos a a hacia atrás a lo largo de la ruta original. (d) Cuando la parte del camino se cancela desde a, nos quedamos con una integral de contorno cerrada en sentido contrario a las agujas del reloj $\oint z^{-1}\,dz = 2\pi i$.

cuya existencia es un rasgo característico y poderoso del análisis complejo. Consideremos la diferencia entre el segundo y el primero de nuestros dos caminos, es decir, recorramos primero el segundo camino y luego recorramos el primero en sentido contrario (Fig. 7.3c). Consideramos esta diferencia en el sentido homólogo, de modo que podemos cancelar porciones que «vuelven sobre sus pasos» y resolver las demás de una forma continua. El resultado es un camino —o *contorno*— cerrado que forma un solo lazo alrededor del origen (véase la Fig. 7.3d), y no se interesa por la elección de a ni de b. Esto proporciona un ejemplo de una integral de contorno (cerrado), que normalmente se escribe con el símbolo «$\oint$», y encontramos, en este ejemplo,[7.1]

$$\oint \frac{dz}{z} = 2\pi i.$$

Por supuesto, cuando utilizamos este símbolo debemos tener cuidado en dejar claro qué contorno real se está utilizando, o más bien

[7.1] Explique por qué $\oint z^n\,dz = 0$ cuando n es un entero distinto de -1.

qué clase de homología de contorno se está utilizando. Si nuestro contorno hubiera dado dos vueltas (en el sentido positivo), entonces habríamos obtenido la respuesta $4\pi i$. Si hubiera dado una vuelta alrededor del origen en el sentido opuesto (i.e., en el de las agujas del reloj), entonces la respuesta habría sido $-2\pi i$.

Resulta interesante que esta propiedad de obtener una respuesta no trivial con semejante contorno cerrado dependa críticamente de la multivaluación del logaritmo complejo, una característica que podría haber parecido ser solo una complicación en la definición de un logaritmo. Más adelante veremos que esto no es una simple curiosidad. De hecho, la potencia del análisis complejo depende de ello de forma crucial. En las dos secciones siguientes esbozaré algunas de las consecuencias de este tipo de cosas. Espero que los lectores no matemáticos puedan sacar algún provecho de la exposición. Creo que transmite algo a la vez genuino y sorprendente en la naturaleza del razonamiento matemático.

7.3. Series de potencias a partir de la suavidad compleja

La expresión mostrada antes es un caso particular (para la función constante $f(z) = 2\pi i$) de la famosa *fórmula de Cauchy* que expresa el valor de una función holomorfa en el origen en términos de una integral de contorno alrededor del origen:[4]

$$\frac{1}{2\pi i}\oint \frac{f(z)}{z}\,dz = f(0).$$

Aquí $f(z)$ es holomorfa en el origen (i.e., suave-compleja en una región que contiene el origen), y el contorno es cierto lazo que precisamente rodea el origen, o podría ser cualquier lazo homólogo a aquel, en el dominio de la función del que se ha eliminado el origen. Así pues, tenemos el hecho notable de que lo que la función está haciendo *en* el origen está completamente determinado por lo que está haciendo en un conjunto de puntos *alrededor* del origen. (La fórmula de Cauchy es básicamente una consecuencia de las ecuaciones de Cauchy-Riemann, junto con la expresión anterior $\oint z^{-1}\,dz = 2\pi i$, to-

mada en el límite de lazos pequeños; pero no sería oportuno que entrara aquí en estos detalles.)

Si en lugar de utilizar $1/z$ en la fórmula de Cauchy utilizamos $1/z^{n+1}$, donde n es algún entero positivo, obtenemos una versión de «orden superior» de la fórmula de Cauchy que da lo que resulta ser la n-ésima derivada $f^{(n)}(z)$ de $f(z)$ en el origen:

$$\frac{n!}{2\pi i}\oint \frac{f(z)}{z^{n+1}}\,dz = f^{(n)}(0).$$

(Recordemos $n!$ de §5.3.) Podemos ver que esta fórmula «tiene que ser la respuesta correcta» examinando la serie de potencias para $f(z)$,[7.2] pero *utilizar* este hecho sería una petición de principio, puesto que aún no sabemos si existe el desarrollo en serie de potencias, y ni siquiera si existe la n-ésima derivada de f. Todo lo que sabemos por ahora es que $f(z)$ es suave-compleja, sin saber si puede diferenciarse más de una vez. Sin embargo, utilizamos simplemente esta fórmula como algo que nos proporciona la *definición* de la derivada n-ésima en el origen. Podemos entonces incorporar esta «definición» en la fórmula de Taylor $a_n = f^{(n)}(0)/n!$ para los coeficientes en la serie de potencias (véase §6.4)

$$a_0 + a_1 z + a_2 z^2 + a_3 z^3 + a_4 z^4 + \ldots,$$

y con un poco de trabajo podemos demostrar que la suma de esta serie es realmente $f(z)$ en cierta región que comprende el origen. En consecuencia, la función tiene una derivada n-ésima en el origen dada por la fórmula.[7.3] Esto contiene la esencia del argumento que demuestra que la suavidad compleja en una región que rodea el origen implica, de hecho, que la función es realmente *analítica* (-compleja) en el origen (i.e., holomorfa).

Por supuesto, en todo esto no hay nada de especial en el origen. Igualmente podemos hablar de series de potencias en torno a cualquier

[7.2] Demuéstrelo simplemente sustituyendo en la integral la serie de Maclaurin para $f(z)$.

[7.3] Demuestre todo esto al menos en el nivel de las expresiones formales; no se preocupe por una justificación rigurosa. *Sugerencia*: Examine la fórmula de Cauchy con el origen desplazado.

otro punto p en el plano complejo y utilizar la serie de Taylor, como hemos hecho en §6.4. Para esto, simplemente desplazamos el origen al punto p para obtener la fórmula de Cauchy en la forma con «origen desplazado»

$$\frac{1}{2\pi i}\oint \frac{f(z)}{(z-p)}\,dz = f(p),$$

y también la expresión para la derivada n-ésima

$$\frac{n!}{2\pi i}\oint \frac{f(z)}{(z-p)^{n+1}}\,dz = f^{(n)}(p),$$

donde ahora el contorno rodea al punto p en el plano complejo. Así pues, la suavidad compleja implica la analiticidad (holomorficidad) en *todo* punto del dominio.

He preferido mostrar la esencia del argumento por el que, localmente, la suavidad compleja implica analiticidad, en lugar de exigir que el lector acepte el resultado como cierto, porque es un ejemplo maravilloso del modo en que los matemáticos pueden obtener a veces sus resultados. Ni la premisa ($f(z)$ es suave-compleja) ni la conclusión ($f(z)$ es analítica) contienen ningún indicio de la noción de integración de contorno o de la multivaluación de un logaritmo complejo. Pese a todo, estos ingredientes proporcionan las claves esenciales hacia la verdadera ruta para encontrar la respuesta. Es difícil ver cómo podría haberlo logrado un argumento «directo» (cualquiera que pudiera ser). La clave es el juego matemático. La naturaleza tentadora del propio logaritmo complejo es lo que nos seduce para estudiar sus propiedades. Este atractivo intrínseco es en apariencia independiente de cualquier aplicación que pudiera tener el logaritmo en otras áreas. Lo mismo se puede decir, incluso en mayor medida, de la integración de contorno. Hay una extraordinaria elegancia en la idea base, donde la libertad topológica se combina con las expresiones explícitas con exquisita precisión.[7.4] Pero no se trata meramente de elegancia; la integración de

[7.4] La función $f(z)$ es holomorfa en cualquier punto de un contorno cerrado Γ, y también dentro de Γ, excepto en un conjunto finito de puntos donde f tiene polos. Recordemos de §4.4 que $f(z)$ tiene un *polo* de orden n en $z = \alpha$ si $f(z)$ es de la forma $h(z)/(z-\alpha)^n$, donde $h(z)$ es regular en α. Demuestre que $\oint_\Gamma f(z)dz = 2\pi i \times$ {suma de los residuos en dichos polos}, donde el residuo en el polo α es $h^{(n-1)}(\alpha)/(n-1)!$

contorno proporciona también una técnica matemática muy poderosa y útil en muchas áreas diferentes, que contiene bastante de la magia de los números complejos. En particular, conduce a formas sorprendentes de evaluar integrales definidas y sumar explícitamente diversas series infinitas.[7.5],[7.6] Asimismo, encuentra muchas otras aplicaciones en física e ingeniería, así como en otras áreas de las matemáticas. ¡Euler hubiera disfrutado con todo esto!

7.4. Prolongación analítica

Tenemos ahora el extraordinario resultado de que la suavidad-compleja en cierta región es equivalente a la existencia de un desarrollo en serie de potencias en torno a cualquier punto en dicha región. Sin embargo, debería aclarar un poco más lo que significa «región» en este contexto. En términos técnicos, entiendo por ello lo que los matemáticos llaman una región *abierta*. Podemos expresarlo diciendo que si un punto a está en la región, entonces hay un círculo centrado en a cuyo interior está también contenido en la región. Quizá esto no sea muy intuitivo, así que permítanme poner algunos ejemplos. Un único punto no es una región abierta, ni lo es una curva ordinaria. Pero el *interior* del círculo unidad en el plano complejo, es decir, el conjunto de puntos cuya distancia al origen es estrictamente menor que la unidad, es una región abierta. Esto se debe a que cualquier punto que esté estrictamente en el interior del círculo, por muy próximo que se halle a la circunferencia que lo limita, está rodeado por un círculo mucho más

[7.5] Demuestre que $\int_0^\infty x^{-1}\,\text{sen}\,x\,dx = \pi/2$ integrando ze^{iz} alrededor de un contorno cerrado Γ que consiste en dos porciones del eje real, desde $-R$ a $-\epsilon$ y de ϵ a R (con $R > \epsilon > 0$) conectadas por dos arcos semicirculares en el semiplano superior, de radios respectivos ϵ y R. Tomar entonces $\epsilon \to 0$ y $R \to \infty$.

[7.6] Demuestre que $1 + \frac{1}{2^2} + \frac{1}{3^2} + \frac{1}{4^2} + \ldots = \frac{\pi}{6}$, integrando $f(z) = z^{-2}\cot \pi z$ (véase la nota 5.1) alrededor de un contorno grande, por ejemplo un cuadrado de lado $2N+1$ centrado en el origen (siendo N un entero grande) y hacer luego $N \to \infty$. (*Sugerencia*: Utilice el ejercicio [7.4], encontrando los polos de $f(z)$ y sus residuos. Trate de demostrar por qué la integral de $f(z)$ alrededor de Γ tiende al valor límite 0 cuando $N \to \infty$.)

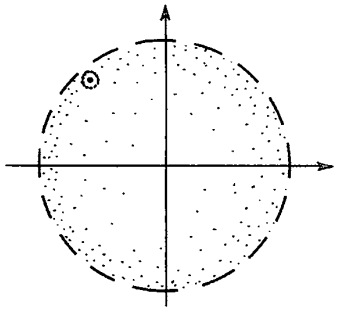


Fig. 7.4. El disco unidad abierto $|x| < 1$. Cualquier punto estrictamente interior, por muy próximo que esté a la circunferencia, está rodeado por muchos círculos más pequeños cuyo interior sigue yaciendo estrictamente dentro del círculo cerrado $|x| \leq 1$, lo que no ocurre para puntos en la frontera.

pequeño que aún yace estrictamente dentro del círculo unidad (véase la Fig. 7.4). Por el contrario, el *disco cerrado*, que consiste en los puntos cuya distancia al origen es menor o *igual* que la unidad, *no* es una región abierta porque ahora está incluida la circunferencia, y un punto de la circunferencia no tiene la propiedad de que existe un círculo centrado en dicho punto que esté contenido dentro de la región.

Consideremos ahora el *dominio*[5] D de cierta función holomorfa $f(z)$, donde tomamos D como una región abierta. En todo punto de D la función $f(z)$ será suave-compleja. Así pues, de acuerdo con lo anterior, si seleccionamos un punto cualquiera p en D, tenemos una serie de potencias convergente en torno a p que representa a $f(z)$ en una región apropiada que contiene a p. ¿Qué tamaño tiene esta «región apropiada»? Tenderá a darse el caso de que para un p concreto, la serie de potencias no funcionará para la totalidad de D. Recordemos el *círculo de convergencia* descrito en §4.4. Este sería algún círculo centrado en p (se admite un radio infinito) tal que para puntos estrictamente dentro de ese círculo la serie de potencias convergerá, pero no lo hará para puntos z estrictamente fuera del círculo. Supongamos que $f(z)$ tiene una *singularidad* en algún punto q, a saber, un punto en el que la función $f(z)$ no puede extenderse y seguir siendo suave-compleja. (Por ejemplo, el origen $q = 0$ es una singularidad de la función $f(z) = 1/z$; véase §7.1. Una singularidad se conoce a veces como un «punto singular» de la función. Un punto *regular* es simplemente un lugar donde la función es no singular, y con ello holomorfa.) Entonces, el círculo de convergencia no puede ser tan grande que contenga a q en su interior.

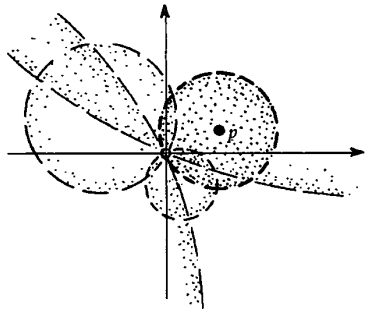


Fig. 7.5. Para $f(z) = 1/z$, el dominio D es el plano complejo con el origen eliminado. El círculo de convergencia alrededor de cualquier punto p en D está centrado en p y pasa por el origen. Para cubrir D por entero, necesitamos un atlas (infinito) de círculos semejantes.

Tenemos así un mosaico de círculos de convergencia (normalmente en número infinito) que, en conjunto, recubren la totalidad de D, aunque en general uno solo de dichos círculos no lo hará. El caso $f(z) = 1/z$ ilustra este punto (véase la Fig. 7.5). Aquí, el dominio D es el plano complejo del que se ha eliminado el origen. Si seleccionamos un punto p en D, encontramos que el círculo de convergencia es el círculo centrado en p y que pasa por el origen.[7.7] Necesitamos un número infinito de tales círculos para recubrir la región entera D.

Esto nos lleva a la importante cuestión de la *prolongación analítica*. Supongamos que se nos da cierta función $f(z)$, holomorfa en cierto dominio D, y consideremos la pregunta: «¿Podemos extender D a una región mayor D', de modo que $f(z)$ también se extienda de forma holomorfa a D'?». Por ejemplo, se nos podría haber dado $f(z)$ en forma de una serie de potencias, convergente dentro de su círculo de convergencia específico, y podríamos desear extender $f(z)$ fuera de este círculo. Con frecuencia esto es posible. En §4.4 hemos considerado la serie $1 - z^2 + z^4 - z^6 + \ldots$, que tiene al círculo unidad como círculo de convergencia; pese a todo, tiene la extensión natural a la función $(1 + z^2)^{-1}$, que es holomorfa en todo el plano complejo del que solo se han eliminado los dos puntos $+i$ y $-i$. Así pues, en este caso la función puede ser extendida analíticamente mucho más allá del dominio sobre el que estaba dada inicialmente.

Aquí hemos sido capaces de escribir una fórmula explícita para la función, pero en otros casos puede no ser tan fácil. De todas formas,

[7.7] ¿Cuál es la serie de potencias, tomada en torno al punto p, para $f(z) = 1/z$?

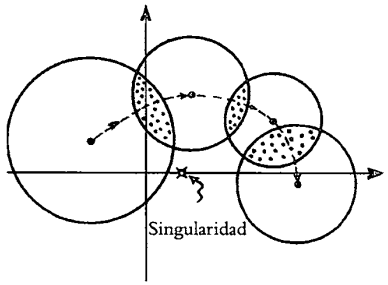


Fig. 7.6. Una función holomorfa puede ser prolongada analíticamente, utilizando una sucesión de expresiones en serie de potencias en torno a una secuencia de puntos. Esto procede unívocamente a lo largo del camino que los conecta, suponiendo que se solapan los sucesivos círculos de convergencia.

existe un procedimiento general según el cual puede llevarse a cabo frecuentemente la prolongación analítica. Podemos imaginar que partimos de alguna pequeña región en la que se conoce una expresión en serie de potencias válida localmente para la función $f(z)$. Entonces podemos movernos a lo largo de cierto camino, prolongando la función mediante el uso repetido de series de potencias basadas en puntos diferentes. Para esto utilizaríamos una secuencia de puntos a lo largo del camino y tomaríamos una sucesión de desarrollos en serie de potencias en torno a cada uno de estos puntos. Esto determinará la prolongación con tal de que puedan solaparse los interiores de los sucesivos círculos de convergencia (véase la Fig. 7.6). Cuando puede llevarse a cabo este procedimiento de forma consistente, la función resultante está unívocamente determinada por el valor de la función en la región inicial y en el camino a lo largo del cual esta siendo prolongada.

Existe así una notable «rigidez» en las funciones holomorfas, como se manifiesta en este proceso de prolongación analítica. En el caso de C^∞-funciones reales, por el contrario, era posible «ir cambiando de idea» acerca de lo que tenía que hacer la función (como con la $h(x)$ suavemente parcheada de §6.3, Fig. 6.7, que repentinamente «despega» después de haber sido nula para todos los valores negativos de x). Esto no puede suceder con las funciones holomorfas. Una vez que la función está determinada en su región original y se fija el camino, no hay elección sobre la forma de extender la función. De hecho, lo mismo es cierto para funciones analítico-reales de una variable real. También estas tienen una «rigidez» similar, pero ahora tampoco hay mucha elección sobre el camino. Solo puede ser en un sentido o en el otro a lo

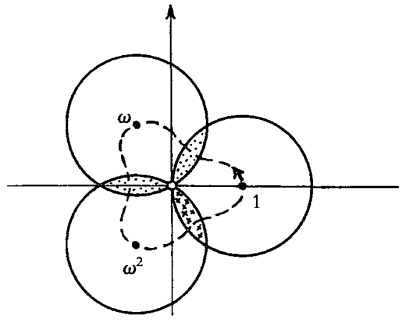


Fig. 7.7. Partimos de $z = 1$, prolongando analíticamente $f(z) = \log z$ a lo largo de un camino que rodea al origen en el sentido contrario a las agujas del reloj (desarrollando en torno a los puntos sucesivos $1, \omega, \omega^2, 1$; $\omega = e^{2\pi i/3}$. Encontramos que $2\pi i$ se suma a f.

largo de la recta real. Con las funciones complejas, la prolongación analítica puede ser más interesante debido a esta libertad del camino dentro de un plano bidimensional.

Para ilustrarlo, consideremos a nuestra vieja amiga $\log z$. Ciertamente, no tiene un desarrollo en serie de potencias en torno al origen, puesto que tiene una singularidad allí. Pero, si queremos, podemos desarrollarla en torno al punto $p = 1$, por ejemplo, para obtener la serie[7.8]

$$\log z = (z-1) - \frac{1}{2}(z-1)^2 + \frac{1}{3}(z-1)^3 - \frac{1}{4}(z-1)^4 + \dots$$

El círculo de convergencia es el círculo de radio unidad centrado en $z = 1$. Imaginemos que se realiza una prolongación analítica a lo largo de un camino que describe un círculo alrededor del origen en el sentido contrario a las agujas del reloj. Si quisiéramos, podríamos utilizar, series de potencias tomadas en torno a los puntos sucesivos 1, ω, ω^2, y vuelta a 1, regresando así a nuestro punto de partida después de haber dado una vuelta alrededor del origen (Fig. 7.7). Aquí he utilizado las tres raíces cúbicas de la unidad, situadas regularmente a lo largo del círculo unidad, a saber, 1, $\omega = e^{2\pi i/3}$ y $\omega^2 = e^{4\pi i/3}$, como se ha expuesto al final de la sección §5.4, y la ruta alrededor del origen es un triángulo equilátero. Alternativamente, podría haber utilizado $1, i, -1, -i, 1$, que es ligeramente menos económico. En cualquier caso, no hay necesidad de calcular la serie de potencias, puesto que ya conocemos la respuesta explícita para la propia función, a saber, $\log z$. El problema, por su-

[7.8] Deduzca esta serie.

puesto, es que cuando hemos dado una vuelta alrededor del origen, siguiendo unívocamente la función, encontramos que la hemos extendido unívocamente a un valor diferente de aquel del que hemos partido. De algún modo, se ha sumado $2\pi i$ a la función cuando giramos una vuelta. Si hubiésemos decidido proceder alrededor del origen en sentido contrario, entonces habríamos encontrado que se había restado $2\pi i$ de la función de la que hemos partido. Así pues, la unicidad de la prolongación analítica puede ser algo muy sutil, y puede depender decididamente del camino tomado. Para funciones «multivaluadas» más complicadas que $\log z$, podemos obtener algo mucho más elaborado que el mero añadido de una constante (como $2\pi i$) a la función.

Como un aparte, vale la pena señalar que la noción de prolongación analítica no necesita referirse concretamente a series de potencias, pese al hecho de que he encontrado útil emplearlas en algunas de mis descripciones. Por ejemplo, hay otra clase de series que tienen gran importancia en la teoría de números, a saber, las denominadas *series de Dirichlet*. La más importante de estas es la *función zeta de (Euler-)Riemann*,[6] definida por la suma infinita[7]

$$\zeta(z) = 1^{-z} + 2^{-z} + 3^{-z} + 4^{-z} + 5^{-z} + \ldots,$$

que converge con la función holomorfa denotada por $\zeta(z)$ cuando la parte real de z es mayor que 1. La prolongación analítica de esta función la define unívocamente (y «univaluadamente») en la totalidad del plano complejo, pero del que se ha eliminado el punto $z = 1$. Quizá el más importante de los problemas matemáticos no resueltos hoy es la *hipótesis de Riemann*, que concierne a los ceros de dicha función zeta extendida analíticamente, es decir, las soluciones de $\zeta(z) = 0$. Es relativamente fácil demostrar que $\zeta(z)$ se anula para $z = -2, -4, -6, \ldots$ La hipótesis de Riemann afirma que todos los ceros restantes yacen en la recta $\mathrm{Re}(z) = 1/2$, es decir, $\zeta(z)$ se anula (a menos que z sea un número entero par) solo cuando la parte real de z es igual a $1/2$. Toda la evidencia numérica hasta la fecha apoya esta hipótesis, pero su verdad real es una incógnita. Tiene implicaciones fundamentales para la teoría de los números primos.[8]

Notas

Sección 7.1

7.1. A aquellos lectores que quieran explorar estas fascinantes cuestiones con un mayor detalle geométrico, les recomiendo Needham (1997).

7.2. Las daré en §10.5, cuando se haya introducido la idea de *derivada parcial*.

Sección 7.2

7.3. De forma más explícita, la integración de f «a lo largo» de un camino dado por $z = p(t)$ (donde p es una función suave de valor complejo de un parámetro real t) puede expresarse como la integral definida $\int_u^v f(p(t))p'(t)\,dt = \int_a^b f(z)dz$, donde $p(u)$ es el punto inicial a del camino y $p(v)$ es el punto final b.

Sección 7.3

7.4. Una «razón» por la que la fórmula de Cauchy debe ser verdadera es que para un lazo pequeño alrededor del origen, $f(z)$ debe ser tratada realmente como el *valor constante* $f(0)$, y entonces la situación se reduce a la estudiada en §7.2.

Sección 7.4

7.5. Una de las cosas irritantes de la terminología de este tema es que el término «dominio» tiene dos significados distintos. Uno, que *no* es el que se pretende aquí, es simplemente cualquier «región abierta conexa en el plano complejo». Aquí, como antes (véase §6.1), entiendo por región la que se da en el plano complejo en la que está definida la función f, que en general no tiene por qué ser necesariamente abierta no conexa (aunque aquí se toma abierta).

7.6. La función zeta fue considerada por primera vez por Euler, pero normalmente se le da el nombre de Riemann, en vista de su fundamental trabajo, que incluye la extensión de esta función al plano complejo.

7.7. Nótese la curiosa relación «al revés» entre esta serie y una serie de potencias ordinaria, a saber, $(-z) + (-z)^2 + (-z)^3 + \ldots = -z(1 + z)^{-1}$.

7.8. Para más información sobre la función ζ y la hipótesis de Riemann, véanse Apostol (1976) y Priestley (2003). Para exposiciones de divulgación, véanse Derbyshire (2003), Du Sautoy (2003), Sabbagh (2002) y Devlin (1988, 2002).

8

Superficies de Riemann y aplicaciones complejas

8.1. La idea de una superficie de Riemann

Hay un modo de entender qué está pasando con esta prolongación analítica de la función logaritmo —o de cualquier otra «función multivaluada»— en términos de lo que se denominan *superficies de Riemann*. La idea de Riemann consistía en considerar que tales funciones están definidas en un dominio que no es simplemente un subconjunto del plano complejo, sino que es una región con muchas hojas. En el caso de $\log z$ podemos representarla como una especie de rampa espiral aplastada verticalmente sobre el plano complejo. He tratado de mostrarlo en la Fig. 8.1. La función logaritmo es univaluada en esta versión de muchas hojas enrolladas del plano complejo, porque cada vez que damos una vuelta alrededor del origen, y tenemos que sumar $2\pi i$ al logaritmo, nos encontramos en otra hoja del dominio. Ahora no hay conflicto entre los diferentes valores del logaritmo, puesto que su

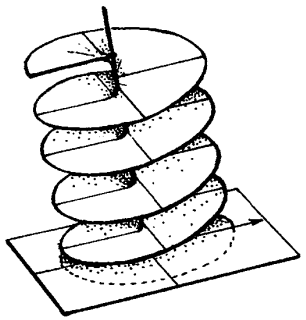


Fig. 8.1. La superficie de Riemann para $\log z$, representada como una rampa espiral aplastada en dirección vertical.

dominio es este espacio enrollado más amplio —un ejemplo de superficie de Riemann—, un espacio sutilmente diferente del propio plano complejo.

Bernhard Riemann, que introdujo esta idea, fue uno de los más grandes matemáticos, y en su corta vida (1826-1866) propuso muchísimas ideas matemáticas que cambiaron profundamente el curso del pensamiento matemático en este planeta. Más adelante encontraremos algunas contribuciones más de él, tal como la que subyace en la teoría de la relatividad general de Einstein (y una contribución muy importante de Riemann, de un tipo muy diferente, se ha mencionado al final del capítulo 7). Antes de que Riemann introdujera la idea de lo que ahora se denomina una «superficie de Riemann», los matemáticos estaban divididos acerca de la forma de tratar las denominadas «funciones multivaluadas», de las que el logaritmo es uno de los ejemplos más sencillos. Para ser rigurosos, muchos habían sentido la necesidad de considerar estas funciones de una forma que personalmente me resultaría bastante desagradable. (Dicho sea de paso, así es como me enseñaron a considerarlas cuando estaba en la universidad, pese a que ya habían transcurrido casi cien años desde la publicación del histórico artículo de Riemman sobre el tema). En particular, el dominio de la función logaritmo estaría «cortado» de una forma arbitraria por una recta que iría desde el origen al infinito. Según mi opinión, esto era una mutilación brutal de una estructura matemática sublime. Riemann nos enseñó que debemos considerar las cosas de un modo diferente. Las funciones holomorfas no encajan fácilmente en la ahora habitual noción de una «función» que aplica un dominio determinado en un espacio imagen definido. Como hemos visto con la prolongación analítica, una función holomorfa «tiene su propia opinión» y decide por sí misma cuál debería ser su dominio, con independencia de la región del plano complejo que nosotros podamos haberle asignado inicialmente. Aunque podamos considerar que el dominio de la función está determinado por la superficie de Riemann asociada a la función, el dominio no está dado por adelantado; es la forma explícita de la propia función la que nos dice qué superficie de Riemann es realmente el dominio.

Pronto encontraremos otros muchos tipos de superficies de Riemann. Este bello concepto desempeña un papel importante en algunos

de los intentos modernos de encontrar una nueva base para la física matemática —muy especialmente en la teoría de cuerdas (§§31.5,13), pero también en la teoría de twistores (§§33.2,10). De hecho, la superficie de Riemann para $\log z$ es una de las más simples de dichas superficies. Nos da simplemente un indicio de lo que nos aguarda. La función z^a es quizá ligeramente más interesante que $\log z$ con respecto a su superficie de Riemann, pero solo cuando el número complejo a es un número racional. Cuando a es irracional, la superficie de Riemann para z^a tiene precisamente la misma estructura que para $\log z$, pero para un a racional, cuya expresión en términos más simples es $a = m/n$, las hojas en espiral vuelven a empalmarse al cabo de n vueltas.[8.1] En todos estos ejemplos, el origen $z = 0$ se denomina un *punto rama*. Si se vuelven a empalmar las hojas al cabo de un número finito n de vueltas (como en el caso $z^{m/n}$, cuando m y n no tienen ningún factor común), diremos que el punto rama tiene *orden finito*, o que es *de orden n*. Cuando no se empalman al cabo de ningún número finito de vueltas (como en el caso $\log z$), decimos que el punto rama tiene *orden infinito*.

Expresiones como $(1 - z^3)^{1/2}$ nos ofrecen más materia de reflexión. Aquí la función tiene tres puntos rama en $z = 1$, $z = \omega$ y $z = \omega^2$ (donde $\omega = e^{2\pi i/3}$; véanse §5.4 y §7.4), de modo que $1 - z^3 = 0$, y hay otro «punto rama en el infinito». Cuando damos una vuelta completa alrededor de cada punto rama individual, permaneciendo en su inmediata vecindad (y en el caso del «infinito» esto significa simplemente describir un círculo muy grande), encontramos que la función cambia de signo, y al dar una vuelta más, la función vuelve a su valor original. Vemos así que todos los puntos rama tienen orden 2. Tenemos dos hojas para la superficie de Riemann, cosidas de la forma que he tratado de mostrar en la Fig. 8.2a. En la Fig. 8.2b he intentado mostrar, utilizando algunas contorsiones topológicas, que la superficie de Riemann tiene en realidad la topología de un *toro*, que topológicamente es la superficie de una rosquilla (o de un donut), pero con cuatro agujeros minúsculos en ella que corresponden a los puntos rama. De hecho, los agujeros pueden ser rellenados sin ambigüedad (con cuatro simples

[8.1] Explique por qué.

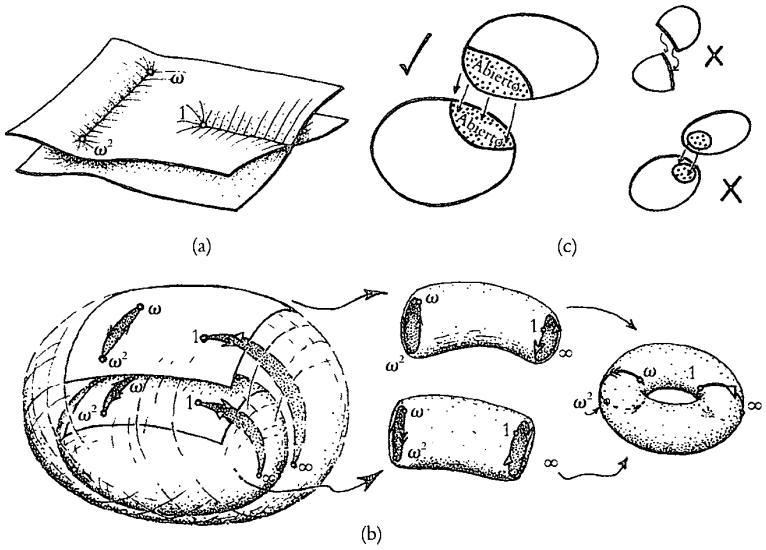


Fig. 8.2. (a) Construcción de la superficie de Riemann para $(1 - z^3)^{1/2}$ a partir de dos hojas, con puntos rama de orden dos en 1, ω, ω^2 (y también ∞). (b) Para ver que la superficie de Riemann para $(1 - z^3)^{1/2}$ es topológicamente un toro, imaginemos los planos de (a) como dos esferas de Riemann con ranuras cortadas desde ω a ω^2 y desde 1 a ∞, identificadas a lo largo de las flechas de empalme. Hay cilindros topológicos pegados en correspondencia, que dan un toro. (c) Para construir una superficie de Riemann (o una variedad en general), podemos empalmar cartas del espacio de coordenadas —aquí porciones abiertas del plano complejo—. Debe haber solapamientos (conjuntos abiertos) entre cartas (y cuando se unen, no debe haber «ramificación no Hausdorff», como en el último caso de arriba; véanse la Fig. 12.5b y §12.2).

puntos), y la superficie de Riemann tiene entonces exactamente la topología de un toro.[8.2]

Las superficies de Riemann proporcionaron los primeros ejemplos de la noción general de *variedad*, que es un espacio que puede pensarse «curvado» de diversas maneras, pero que, *localmente* (i.e., en un entorno suficientemente pequeño de cualquiera de sus puntos), parece un fragmento de espacio euclídeo ordinario. Encontraremos variedades de una forma más seria en los capítulos 10 y 12. La noción de va-

[8.2] Intente ahora $(1 - z^4)^{1/2}$.

riedad es crucial en muchas áreas diferentes de la física moderna. Y lo que es más sorprendente, constituye una parte esencial de la teoría de la relatividad general de Einstein. Puede considerarse que las variedades están construidas pegando varias cartas diferentes, con un trabajo de pegado realmente sin fisuras, a diferencia de la situación con la función $h(x)$ al final de §6.3. La naturaleza sin fisuras del mosaico de cartas se consigue asegurando que siempre existe un solapamiento (conjunto abierto) apropiado entre una carta y la siguiente (véanse la Fig. 8.2c y también §12.2 y la Fig. 12.5).

En el caso de las superficies de Riemann, la variedad (i.e., la propia superficie de Riemann) está construida pegando varias cartas del plano complejo correspondientes a las diferentes «hojas» que van a formar la superficie entera. Como antes, podemos terminar con unos pocos «agujeros» en forma de algunos puntos individuales que faltan, que proceden de los puntos rama de orden finito, pero estos puntos que faltan siempre pueden ser reemplazados sin ambigüedad, como antes. Por el contrario, para puntos rama de orden infinito, las cosas pueden ser más complicadas, y no puede hacerse ningún enunciado general de este tipo.

A modo de ejemplo, consideremos la superficie de Riemann «rampa espiral» de la función logaritmo. Una forma de construirla, en un modelo de papel, sería tomar, sucesivamente, cartas alternadas que son copias de (a) el plano complejo del que se han eliminado los números reales no negativos, y (b) el plano complejo del que se han eliminado los números reales no positivos. La mitad superior de cada carta-(a) estaría pegada a la mitad superior de la carta-(b) siguiente, y la mitad inferior de cada carta-(b) estaría pegada a la mitad inferior de la carta-(a) siguiente (a); véase la Fig. 8.3. Existe un punto rama de orden infinito en el origen y también en el infinito, pero, curiosamente, encontramos que la rampa espiral entera es equivalente a una esfera en la que falta un único punto, y este punto puede ser reemplazado sin ambigüedad de modo que se tiene simplemente una esfera.[8.3]

[8.3] ¿Puede ver cómo se obtiene esto? (*Sugerencia*: Piense en la esfera de Riemann de la variable $w(= \log z)$; véase §8.3.)

8.2. Aplicaciones conformes

Cuando construimos una variedad, tenemos que considerar qué estructura local debe conservarse de una carta a la siguiente. Normalmente, uno trata con variedades reales, y las diferentes cartas son fragmentos de espacio euclídeo (de una dimensión determinada) que están pegadas a lo largo de varias regiones (abiertas) que se solapan. La estructura local que hay que empalmar de una carta a la siguiente se reduce normalmente a conservar la continuidad o la suavidad. Esta cuestión se discutirá en §10.2. Sin embargo, en el caso de superficies de Riemann, estamos interesados en la suavidad *compleja*, y recordamos, de §7.1, que esta es una cuestión más sofisticada, que implica lo que se denominan *ecuaciones de Cauchy-Riemann*. Aunque todavía no las hemos visto explícitamente (llegaremos a ellas en §10.5), ahora será oportuno entender el significado geométrico de la estructura que está codificado en dichas ecuaciones. Es una estructura de elegancia, flexibilidad y potencia notables, que conduce a conceptos matemáticos con un gran ámbito de aplicación.

La noción es la de *geometría conforme*. Hablando en términos generales, en la geometría conforme estamos interesados en la forma pero no en el tamaño, entendiendo forma en la escala infinitesimal. En una aplicación conforme de una región (abierta) del plano en otra, las formas de tamaño finito quedan en general distorsionadas, pero las formas *infinitesimales* se conservan. Podemos considerar que esto se aplica a

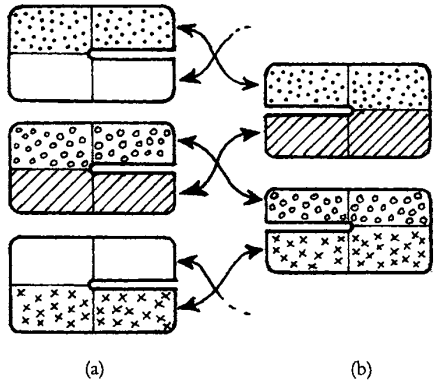


Fig. 8.3. Podemos construir la superficie de Riemann para $\log z$ tomando cartas alternas de (a) el plano complejo del que se ha eliminado el eje real no negativo y (b) el plano complejo del que se ha eliminado el eje real no positivo. La mitad superior de cada (a)-carta se pega a la mitad superior de la siguiente (b)-carta, y la mitad inferior de cada (b)-carta se pega a la mitad inferior de la siguiente (a)-carta.

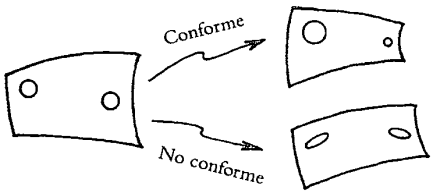


Fig. 8.4. En el caso de un mapa conforme, los círculos pequeños (infinitesimales) pueden ampliarse o contraerse, pero no distorsionarse en pequeñas elipses.

círculos pequeños (infinitesimales) dibujados en el plano. En una aplicación conforme, estos pequeños círculos pueden dilatarse o contraerse, pero no se distorsionan dando pequeñas elipses. Véase la Fig 8.4.

Para hacernos una idea de a qué puede parecerse una transformación conforme, consideremos el grabado de M.C. Escher, dado en la Fig. 2.11, que proporciona una representación conforme del plano hiperbólico en el plano euclídeo, como se ha descrito en §2.4 («disco de Poincaré» de Beltrami). El plano hiperbólico es muy simétrico. En particular, existen transformaciones que llevan las figuras de la región central del grabado de Escher a las correspondientes figuras minúsculas que yacen en el interior inmediato al círculo frontera. Podemos representar una transformación semejante como un movimiento conforme del plano euclídeo que aplica el interior del círculo frontera sobre sí mismo. Evidentemente una transformación semejante no conservaría en general los tamaños de las figuras individuales (puesto que las que están en el centro son mucho más grandes que las que están cerca del borde), pero las formas se conservan aproximadamente. Esta conservación de la forma se hace más precisa cuanto menor es el detalle de cada figura que se está examinando, de modo que las formas *infinitesimales* quedarían completamente inalteradas. Quizá el lector encontraría más útil una caracterización ligeramente diferente: los *ángulos* entre curvas no son alterados por una transformación conforme. Esto caracteriza la naturaleza conforme de una transformación.

¿Qué tiene que ver esta propiedad conforme con la «suavidad compleja» (holomorficidad) de cierta función $f(z)$? Trataremos de obtener una idea intuitiva del contenido geométrico de la suavidad compleja. Volvamos al punto de vista de una función f como «aplicación» y consideremos que la relación $w = f(z)$ proporciona una aplicación de cierta región del plano complejo de z (el dominio de la función f) en

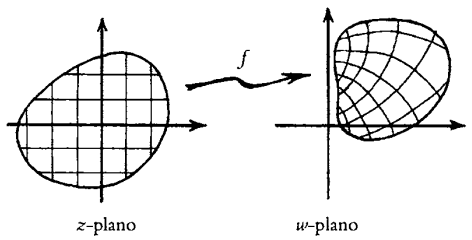


Fig. 8.5. La aplicación $w = f(z)$ tiene como dominio una región abierta en el z-plano complejo y como imagen una región abierta en el w-plano complejo. La holomorficidad de f es equivalente a que sea conforme y no reflexiva.

el plano complejo de w (la imagen); véase la Fig. 8.5. Preguntemos: ¿qué propiedad geométrica local caracteriza esta aplicación como holomorfa? Hay una respuesta sorprendente. La holomorficidad de f es equivalente a que la aplicación sea *conforme* y *no reflexiva* (donde no reflexiva —o que *conserva la orientación*— significa que las formas pequeñas conservadas en la transformación no están reflejadas, i.e., «no están invertidas»; véase la parte final de §12.6).

La noción de «suavidad» en nuestra transformación $w = f(z)$ se refiere a cómo actúa la transformación en el límite infinitesimal. Consideremos primero el caso real, y reexaminemos nuestra función real $f(x)$ de §6.2, donde la gráfica de $y = f(x)$ se ilustra en la Fig. 6.4. La función f es *suave* en un punto si la gráfica tiene una tangente bien definida en dicho punto. Podemos representar la tangente imaginando que se hace una ampliación cada vez mayor de la curva en dicho punto; si la función es suave, a medida que la ampliación aumenta la curva se parece cada vez más a una recta que pasa por dicho punto, confundiéndose con la tangente en el límite de ampliación infinita. La situación con la suavidad compleja es similar, pero ahora aplicamos la idea a la aplicación del z-plano en el w-plano. Para examinar la naturaleza infinitesimal de esta aplicación, tratemos de representar el entorno inmediato de un punto z, en un plano, aplicándolo en el entorno inmediato de w en el otro plano. Para examinar el entorno inmediato del punto, imaginemos que ampliamos el entorno de z en un factor enorme y el correspondiente entorno de w en el mismo factor enorme. En el límite, la aplicación del entorno ampliado de z en el entorno ampliado de w será simplemente una transformación lineal del plano, pero si va a ser holomorfa, debe ser básicamente una de las transformaciones estudiadas en §5.1. De esto se sigue que, en el caso general, la transformación

del entorno de z en el entorno de w combina simplemente una rotación con una dilatación (o contracción) uniforme; véase la Fig. 5.2b. Es decir, formas (o ángulos) pequeñas se conservan, sin reflexión, lo que demuestra que la aplicación es realmente conforme y no reflexiva.

Veamos algunos ejemplos sencillos. Las circunstancias muy concretas de las aplicaciones dadas por la suma de una constante b a z o la multiplicación de z por una constante a, que se han considerado ya en §5.1 (véase la Fig. 5.2), son obviamente holomorfas (al ser $a + b$ y az claramente diferenciables) y son también obviamente conformes. Estos son casos particulares de la transformación general combinada (lineal inhomogénea)

$$w = az + b.$$

Tales transformaciones proporcionan los movimientos euclídeos del plano (sin reflexión), combinados con dilataciones (o contracciones) uniformes. De hecho, son las únicas aplicaciones conformes (no reflexivas) del z-plano complejo entero en el w-plano complejo entero. Además, tienen la propiedad muy especial de que los círculos reales —no solo los círculos infinitesimales— se aplican en círculos reales, y también las líneas rectas se aplican en líneas rectas.

Otra función holomorfa sencilla es la función *recíproca*

$$w = z^{-1},$$

que aplica el plano complejo, con el origen eliminado, en el plano complejo, con el origen eliminado. Sorprendentemente, esta transformación también aplica círculos reales en círculos reales[8.4] (donde consideramos las líneas rectas como casos particulares de círculos —de radio infinito). Esta transformación, junto con una reflexión en el eje real, es lo que se denomina una *inversión*. Combinándola con las aplicaciones lineales inhomogéneas que acabamos de considerar, obtenemos la transformación más general[8.5]

$$w = \frac{az + b}{cz + d},$$

[8.4] Demuéstrelo.

[8.5] Verifique que la secuencia de transformaciones $z \mapsto Az + B$, $z \mapsto z^{-1}$, $z \mapsto Cz + D$ conduce realmente a una aplicación bilineal.

llamada transformación *bilineal* o de *Moebius*. Por lo que se ha dicho antes, estas transformaciones también deben aplicar círculos en círculos (considerando de nuevo las líneas rectas como círculos especiales). La transformación de Moebius aplica en realidad el plano complejo entero, con el punto $-d/c$ eliminado, en el plano complejo entero, con el punto a/c eliminado, donde, para que la transformación proporcione una aplicación no trivial, debemos tener $ad \neq bc$ (de modo que el numerador no sea un múltiplo determinado del denominador).

Nótese que el punto eliminado del z-plano es el valor ($z = -d/c$) que daría «$w = \infty$»; en correspondencia, el «punto eliminado» del w-plano es el valor ($w = a/c$) que se alcanzaría en «$z = \infty$». De hecho, la transformación completa tendría más sentido global si incorporáramos una cantidad «∞» tanto en el dominio como en la imagen. Esta es una forma de considerar la superficie de Riemann (cerrada) más simple de todas: la *esfera de Riemann*, a la que llegamos ahora.

8.3. La esfera de Riemann

Añadir simplemente un punto extra llamado «∞» al plano completo no deja del todo claro que la estructura sin fisuras requerida se mantenga en el entorno de ∞, igual que en cualquier otro lugar. La forma de abordar esta cuestión es considerar que la esfera está construida a partir de dos «cartas de coordenadas», una de las cuales es el z-plano y la otra el w-plano. A todos los puntos de la esfera, salvo a dos, se les asigna una z-coordenada y una w-coordenada (relacionadas por la transformación de Moebius anterior). Pero un punto tiene solo una z-coordenada (donde w sería «infinito») y otro tiene solo una w-coordenada (donde z sería «infinito»). Utilizamos o bien z o bien w para definir la estructura conforme necesaria y, allí donde utilizamos las dos, obtenemos la misma estructura conforme utilizando una u otra, puesto que la relación entre ambas coordenadas es holomorfa.

De hecho, para esto no necesitamos una transformación tan complicada entre z y w como la transformación general de Moebius. Basta con considerar la transformación de Moebius, particularmente simple, dada por

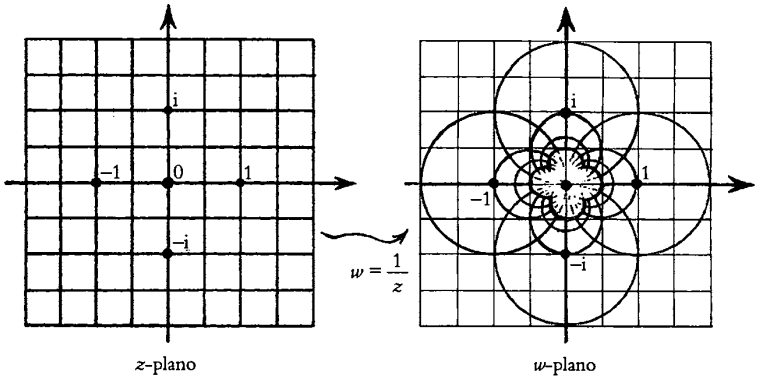


Fig. 8.6. Haciendo cartas de la esfera de Riemann a partir de los z- y w-planos complejos, vía $w = 1/z$, $z = 1/w$. (Aquí las líneas de mallado z se muestran también en el w-plano). Las regiones que se solapan excluyen solo los orígenes, $z = 0$ y $w = 0$, cada una de las cuales da «∞» en la carta opuesta.

$$w = \frac{1}{z}, \quad z = \frac{1}{w},$$

en la que los dos valores $z = 0$ y $w = 0$ darían «∞» en la carta opuesta. En la Fig. 8.6 he indicado cómo aplica esta transformación las rectas coordenadas real e imaginaria de z.

Todo esto define la esfera de Riemann de una forma más bien abstracta. Podemos ver más claramente la razón de que la esfera de Riemann sea llamada una «esfera» utilizando la geometría ilustrada en la Fig. 8.7a. El plano z representa aquí el *plano ecuatorial* de esta esfera geométrica. Los puntos de la esfera se aplican en puntos del plano mediante lo que se denomina *proyección estereográfica* desde el polo sur. Esto solo significa que trazo una línea recta en el 3-espacio euclídeo (en el que suponemos que todo tiene lugar) que sale del polo sur y pasa por el punto z del plano. El punto donde la recta corta de nuevo a la esfera es el punto de la esfera que representa el número complejo z. Hay un punto adicional en la esfera, a saber, el propio polo sur, y este representa a $z = \infty$. Para ver cómo encaja w en esta imagen, imaginemos que su plano complejo se inserta *boca abajo* (con $w = 1, \mathrm{i}, -1$ y $-\mathrm{i}$ empalmando con $z = 1, -\mathrm{i}, -1, \mathrm{i}$, respectivamente), y ahora proyectemos

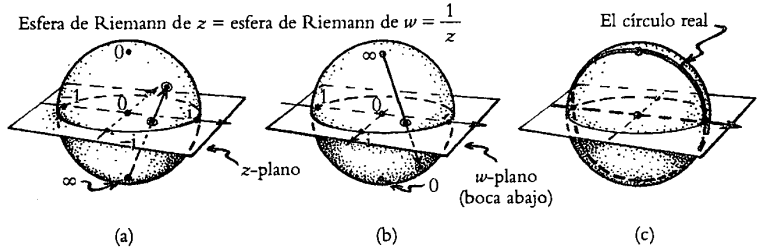


Fig. 8.7. (a) La esfera de Riemann como esfera unidad cuyo ecuador coincide con el círculo unidad en el plano complejo (horizontal) de z. La esfera se proyecta (estereográficamente) en el z-plano a lo largo de líneas rectas que pasan por su polo sur; este da $z = \infty$. (b) Reinterpretando el plano ecuatorial como el w-plano, mostrado al revés pero con el mismo eje real, la proyección estereográfica se hace ahora desde el polo norte ($w = \infty$), donde $w = 1/z$. (c) El eje real es un círculo máximo en esta esfera de Riemann, como el círculo unidad pero dibujado verticalmente en lugar de horizontalmente.

estereográficamente desde el *polo norte* (Fig. 8.7b).[8.6] Una importante y bella propiedad de la proyección estereográfica es que aplica círculos en la esfera en círculos (o líneas rectas) en el plano.[1] De ahí que las transformaciones (de Moebius) bilineales envíen círculos a círculos en la esfera de Riemann. Este hecho notable tiene importancia para la teoría de la relatividad a la que llegaremos en §18.5 (y tiene profunda relevancia para la teoría de espinores y la de twistores; véanse §22.8, §24.7 y §§33.2,4).

Advertimos que, desde el punto de vista de la esfera de Riemann, el eje real es «solo otro círculo», no esencialmente diferente del círculo unidad, aunque dibujado vertical en lugar de horizontalmente (Fig. 8.7c). Uno se obtiene a partir del otro por rotación. Una rotación es ciertamente conforme, de modo que está dada por una aplicación holomorfa de la esfera sobre sí misma. De hecho, toda aplicación conforme (no reflexiva) que aplica la esfera de Riemann entera sobre sí misma se consigue mediante una transformación bilineal (i.e., de Moebius). La rotación particular en la que estamos interesados puede mostrarse ex-

[8.6] Compruebe que estas dos proyecciones estereográficas están relacionadas por $w = z^{-1}$.

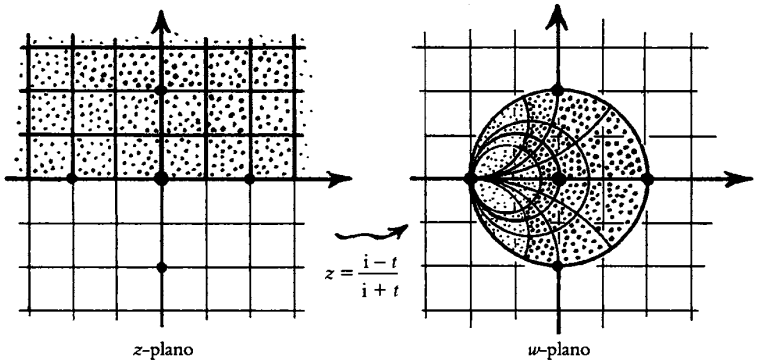


Fig. 8.8. La correspondencia $t = (z - 1)/(iz + i)$, $z = (-t + i)/(t + i)$ en términos de los planos complejos de t y z. El semiplano superior de t, limitado por su eje real, se aplica en el disco unidad de z, limitado por su círculo unidad.

plícitamente como una relación entre las esferas de Riemann de los parámetros complejos z y t dada por la correspondencia bilineal[8.7]

$$t = \frac{z - 1}{iz + i}, \quad z = \frac{-t + i}{t + i}.$$

En la Fig. 8.8 he representado esta correspondencia en términos de los planos complejos de t y z, donde he marcado específicamente cómo el semiplano superior de t, limitado por su eje real, se aplica en el disco unidad de z, limitado por su círculo unidad. Esta transformación particular tendrá importancia para nosotros en el capítulo siguiente.

La esfera de Riemann es la más simple de las superficies de Riemann *compactas*, o «*cerradas*».[2] Véase §12.6 para la noción de «compacto». Por el contrario, la superficie de Riemann «rampa espiral» de la función logaritmo, tal como la he descrito, es *no compacta*. En el caso de la superficie de Riemann de $(1 - z^3)^{1/2}$, necesitamos rellenar los cuatro agujeros que aparecen en los puntos rama para hacerla compacta (y es no compacta si no lo hacemos), pero esta «compactificación» es la que se hace normalmente. Como se ha comentado antes, este «rellenado de agujeros» es siempre posible con un punto rama de orden finito. Como

[8.7] Demuéstrelo.

hemos visto al final de §8.1, en el caso del logaritmo podemos llenar realmente los puntos rama en el origen y en el infinito, ambos simultáneamente, con un solo punto, para obtener la esfera de Riemann como compactificación. De hecho, existe una clasificación completa de las superficies de Riemann compactas (obtenida por el propio Riemann), que es importante en muchas áreas (incluida la teoría de cuerdas). A continuación daré un breve esbozo de dicha clasificación.

8.4. El género de una superficie de Riemann compacta

La primera etapa consiste en clasificar las superficies de acuerdo con su *topología*, es decir, de acuerdo con ese aspecto de las cosas que se conserva en las transformaciones continuas. La clasificación topológica de superficies orientables 2-dimensionales compactas (véase el final de §12.6) es realmente muy sencilla. Viene dada por un simple número natural denominado el *género* de la superficie. Hablando en general, todo lo que tenemos que hacer es contar el número de «asas» que tiene la superficie. En el caso de la esfera, el género es 0, mientras que para el toro es 1. La superficie de una taza de té ordinaria también tiene género 1 (¡un asa!), de modo que es topológicamente igual a un toro. La superficie de un pretzel normal tiene género 3. Véase la Fig. 8.9 para varios ejemplos.

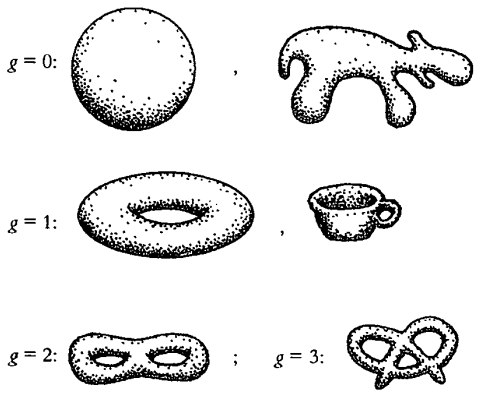


Fig. 8.9. El género de una superficie de Riemann es su número de «asas». El género de la esfera es 0, el del toro, o superficie de la taza de té, es 1. La superficie de un pretzel tiene género 3.

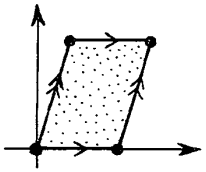


Fig. 8.10. Para construir una superficie de Riemann de género 1, se toma una región del plano complejo limitada por un paralelogramo de vértices $0, 1, 1 + p, p$ (cíclicamente), con los bordes opuestos identificados. La cantidad p proporciona un módulo para la superficie de Riemann.

Sin embargo, el género no determina por sí solo la superficie de Riemann (excepto para el caso de género 0). También necesitamos conocer ciertos parámetros complejos conocidos como *módulos*. A continuación ilustraré este punto en el caso del toro (género 1). Una manera fácil de construir una esfera de Riemann de género 1 es tomar una región del plano complejo limitada por un paralelogramo con vértices, pongamos por caso, $0, 1, 1 + p, p$ (descritos cíclicamente). Véase la Fig. 8.10. Ahora debemos imaginar que los lados opuestos del paralelogramo se empalman, es decir, el lado que va de 0 a 1 p se empalma con el que va de p a $1 + p$, y el lado que va de 0 a p se empalma con el que va de 1 a $p + 1$. (Si quisiéramos, siempre podríamos encontrar otras cartas para cubrir las costuras.) La superficie de Riemann resultante es topológicamente un toro. Ahora bien, resulta que, para diferentes valores de p, las superficies resultantes son generalmente *no equivalentes* entre sí; es decir, no es posible transformar una en otra por medio de una aplicación holomorfa. (Sin embargo, existen ciertas equivalencias discretas, tales como la que se obtiene reemplazando p por $1 + p$, por $-p$ o por $1/p$.)[8.8] Puede hacerse intuitivamente plausible que no todas las superficies de Riemann con la misma topología pueden ser equivalentes si consideramos los dos casos ilustrados en la Fig. 8.11. En un caso he escogido un valor muy pequeño para p, y tenemos un toro muy delgado, y en el otro caso he escogido p próximo a i, para el que el toro es bonito y grueso. Intuitivamente, parece muy claro que no puede haber equivalencia conforme entre los dos, y de hecho no la hay.

En el caso de género 1 solo existe este único módulo complejo p, pero para género 2 encontramos que hay tres. Para construir una superficie de Riemann de género 2 pegando una forma, a la manera del paralelogramo que hemos utilizado para género 1, podríamos construir dicha forma a partir de un fragmento del plano hiperbólico; véase la Fig. 8.12. Lo mismo sería válido para cualquier género más alto. El número m de módulos complejos para un género g, cuando $g \geq 2$, es $m = 3g - 3$.

Podría considerarse un poco extraño que la fórmula $3g - 3$ para el número de módulos funcione para todos los valores del género $g = 2, 3, 4, 5, \ldots$ pero falle para $g = 0$ o 1. En realidad, hay una «razón» para ello, que tiene que ver con el número s de parámetros complejos que son necesarios para especificar las diferentes *autotransformaciones* continuas (holomorfas) de la superficie de Riemann. Para $g \geq 2$ no existen tales autotransformaciones continuas (aunque puede haberlas discretas), de modo que $s = 0$. Sin embargo, para $g = 1$, el plano complejo del paralelogramo de la Fig. 8.10 puede ser desplazado (movido rígidamente sin rotación) en cualquier dirección del plano. La magnitud (y dirección) de este desplazamiento puede especificarse por un solo parámetro complejo a, al estar la traslación definida por $z \mapsto z + a$, de modo que $s = 1$ cuando $g = 1$. En el caso de la esfera (género 0), las autotransformaciones se obtienen mediante las transformaciones bilineales descritas antes, a saber, $z \mapsto (az + b)/(cz + d)$. Aquí la libertad

[8.8] Demuestre que tales reemplazamientos dan espacios holomórficamente equivalentes. Encuentre todos los valores especiales de p donde estas equivalencias llevan a simetrías discretas adicionales de la superficie de Riemann.

Fig. 8.11. Dos superficies de Riemann no equivalentes con topología de toro.

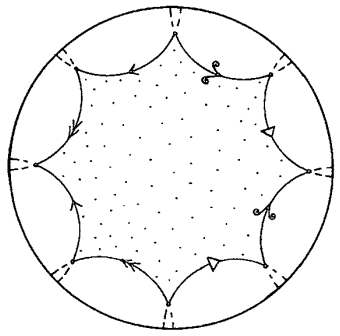


Fig. 8.12. Una región octogonal del plano hiperbólico, en la representación conforme de la Fig. 2.12, con identificaciones para dar una superficie de Riemann de género 2.

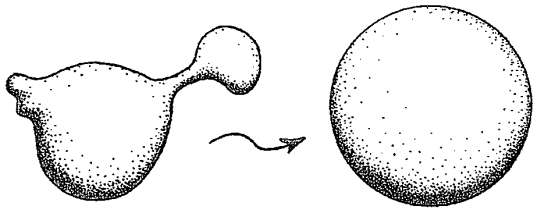


Fig. 8.13. Toda métrica $g = 0$ es conformemente idéntica a la de la esfera unidad estándar («redonda»).

está dada por las tres[3] razones independientes $a : b : c : d$. Así pues, en el caso $g = 0$ tenemos $s = 3$. De ahí que, en todos los casos, la diferencia $m - s$ entre el número de módulos complejos y el número de parámetros complejos requeridos para especificar una autotransformación satisface

$$m - s = 3g - 3.$$

(Esta fórmula está relacionada con algunas cuestiones más profundas que van más allá del alcance de este libro.)[4]

Es evidente que, dentro de la familia de las transformaciones conformes (holomorfas), hay una libertad considerable para alterar la «forma» aparente de una superficie de Riemann manteniendo inalterada su estructura como superficie de Riemann. En el caso de una topología esférica, por ejemplo, son posibles muchas métricas diferentes (como se ilustra en la Fig. 8.13); no obstante, todas ellas son conformemente idénticas a la esfera unidad estándar («redonda»). (Seré más explícito acerca de la noción de «métrica» en §14.7.) Más aún, para géneros más altos, la en apariencia gran cantidad de libertad en la «forma» de la superficie puede reducirse al número finito de módulos complejos dados por las fórmulas anteriores. Pero sigue habiendo una información global en la forma de la superficie que no puede eliminarse mediante el uso de esta libertad conforme, a saber, la que está definida por los propios módulos. Cuánto se puede conseguir con exactitud globalmente con el uso de tal libertad, es una cuestión muy sutil.

8.5. El teorema de la aplicación de Riemann

No obstante, una estimación de la considerable libertad implícita en las transformaciones holomorfas se puede obtener a partir de un famoso resultado conocido como *teorema de la aplicación de Riemann*. Se afirma que si tenemos cierta región en el plano complejo (véase la nota 8.2), acotada por un lazo cerrado no autointersecante, entonces existe una aplicación holomorfa que aplica esta región en el disco unidad cerrado (véase la Fig. 8.14). (Existen ligeras restricciones sobre la «docilidad» del lazo, pero estas no excluyen que el lazo tenga esquinas o lugares aún peores donde el lazo pueda no ser diferenciable, como se ilustra en el ejemplo particular de la Fig. 8.14.) Podemos ir más allá de esto y seleccionar, de una manera completamente arbitraria, tres puntos distintos a, b, c en el lazo e insistir en que sean llevados por la aplicación a tres puntos especificados a', b', c' en el círculo unidad (digamos $a' = 1$, $b' = \omega, c' = \omega^2$), con la única restricción de que el orden cíclico de los puntos a, b, c, alrededor del lazo coincida con el de a', b', c' alrededor del círculo unidad. Además, la aplicación queda entonces determinada unívocamente. Otra manera de especificar unívocamente la aplicación sería escoger solo un punto a en el lazo y un punto adicional j en su interior, e insistir entonces en que a se aplique en un punto específico a' en el círculo unidad (digamos $a' = 1$) y j se aplique en un punto específico j' en el interior del círculo unidad (digamos $j' = 0$).

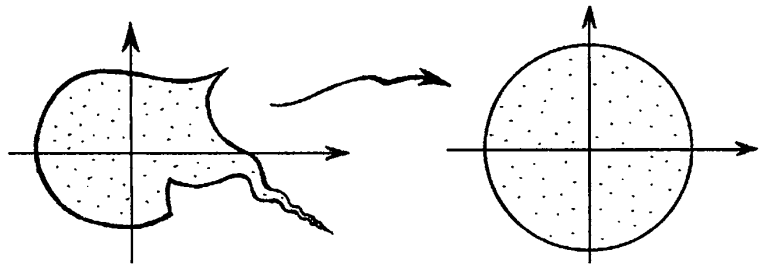


Fig. 8.14. El teorema de la aplicación de Riemann afirma que cualquier región abierta en el plano complejo, limitada por un lazo cerrado simple (no necesariamente suave), puede aplicarse de forma holomorfa en el interior del círculo unidad, siendo también aplicada en consecuencia la frontera.

Ahora imaginemos que estamos aplicando el teorema de la aplicación de Riemann en la *esfera de Riemann*, en lugar de en el plano complejo. Desde el punto de vista de la esfera de Riemann, el «interior» de un lazo cerrado está en pie de igualdad con el «exterior» del lazo (basta con mirar la esfera desde el otro lado), de modo que el teorema puede aplicarse igual al exterior que al interior del lazo. Así pues, existe una forma «invertida» del teorema de la aplicación de Riemann que afirma que el exterior de un lazo en el plano complejo puede ser aplicado en el exterior del círculo unidad y la unicidad está ahora asegurada por el simple requisito de que un punto especificado a en el lazo se aplica en un punto especificado a' en el círculo unidad (digamos $a' = 1$), donde ahora ∞ toma el papel de j y j' en la descripción ofrecida al final del párrafo anterior.[5]

Con frecuencia tales aplicaciones deseadas se pueden conseguir explícitamente, y una de las razones de que tales aplicaciones pudieran «desearse» realmente es que pueden proporcionar soluciones a problemas físicos de interés, por ejemplo, al flujo de aire que pasa junto a una forma de ala (en la situación ideal en la que el fluido es lo que se denomina «no viscoso», «incompresible» e «irrotacional»). Recuerdo cómo me sorprendieron estos aspectos cuando era un estudiante de licenciatura en matemáticas, muy especialmente la que se conoce como la transformación del alerón de Zhoukowski (o Joukowski), ilustrada en la Fig. 8.15, que puede darse explícitamente por el efecto de la transformación

$$w = \frac{1}{2}\left(z + \frac{1}{z}\right),$$

en un círculo adecuado que pasa por el punto $z = -1$. Esta forma se parece mucho a la sección del ala de un avión de los años treinta, de modo que el flujo de aire (idealizado) a su alrededor puede obtenerse directamente a partir del flujo alrededor de un «ala» de sección circular, que, a su vez, se obtiene mediante otra transformación holomorfa semejante. (En cierta ocasión, me contaron que la razón de que esta forma fuera utilizada a menudo para alas de aeroplanos era que uno podía estudiarla matemáticamente con solo emplear la transformación de Zhoukowski. ¡Espero que esto no sea verdad!)

Por supuesto, hay hipótesis específicas y simplificaciones implícitas

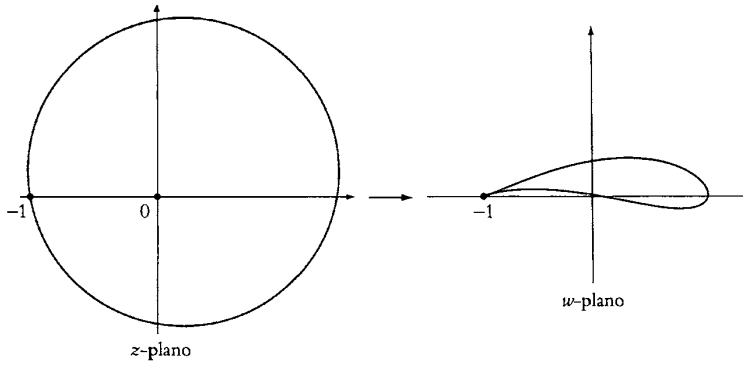


Fig. 8.15. La transformación de Zhoukowski $w = 1/2(z + 1/z)$ lleva el exterior de un círculo que pasa por $z = -1$ a la sección transversal de un ala, haciendo posible el cálculo del flujo de aire alrededor de esta.

en aplicaciones como estas. No solo las hipótesis de viscosidad nula y de flujo incompresible e irrotacional son meras simplificaciones convenientes, sino que existe también una simplificación muy drástica en que el flujo pueda considerarse igual a lo largo del ala, de modo que un problema esencialmente tridimensional puede reducirse a un problema en solo dos dimensiones. Está claro que, para un cálculo realista del flujo alrededor de un ala de aeroplano, sería necesario un tratamiento matemático mucho más complejo. No hay ninguna razón para esperar que en un tratamiento más realista pudiéramos salir con algo que se acerque a un uso tan directo y elegante de las funciones holomorfas como el que tenemos con la transformación de Zhoukowski. Podría argumentarse que hay una buena dosis de buena suerte en el hallazgo de una aplicación tan atractiva de los números complejos a un problema que tenía una importancia patente en el mundo real. El aire consiste, por supuesto, en un enorme número de partículas fundamentales individuales (de hecho, del orden de 10^{20} de ellas por centímetro cúbico), de modo que el flujo de aire es algo cuya descripción macroscópica implica muchos promedios y aproximaciones. No hay razón para esperar que las ecuaciones matemáticas de la aerodinámica reflejen mucho de las matemáticas que están profundamente implicadas en las leyes físicas por las que se rigen dichas partículas individuales.

En §4.1 he mencionado el «papel extraordinario y muy básico» que desempeñan realmente los números complejos en las «escalas más minúsculas» de la acción física, y de hecho hay una ecuación holomorfa que rige el comportamiento de las partículas (véase §21.2). Sin embargo, en el caso de los sistemas macroscópicos esta «estructura compleja» está en general completamente enterrada, y podría parecer que solo en circunstancias excepcionales (tales como el problema del flujo de aire considerado antes) encontrarían una utilidad natural los números complejos y la geometría holomorfa. Pese a todo, hay circunstancias en las que una estructura compleja subyacente se manifiesta incluso en el nivel macroscópico. Esto puede verse a veces en la teoría electromagnética de Maxwell y otros fenómenos ondulatorios. Hay también un ejemplo particularmente llamativo en la teoría de la relatividad (véase §18.5). En el capítulo siguiente veremos algo acerca de la forma notable en que los números complejos y las funciones holomorfas ejercen su magia entre bastidores.

Notas

Sección 8.3

8.1. Véase el ejercicio [2.5].

8.2. Hay lugar para la confusión terminológica en el uso de la palabra «cerrado» en el contexto de superficies, o de las variedades (n-superficies) más generales que se considerarán en el capítulo 12. En el caso de una variedad semejante, «cerrado» significa «compacto sin frontera», antes que meramente «cerrado» en el sentido topológico, que es la noción complementaria de «abierto», tal como se ha expuesto en §7.4. (Topológicamente, un conjunto *cerrado* es uno que contiene todos sus puntos límite. El *complemento* de un conjunto cerrado es uno abierto, y viceversa, donde el «complemento» de un conjunto $\mathcal{S}$ dentro de un espacio topológico ambiente $\mathcal{V}$ es el conjunto de miembros de $\mathcal{V}$ que no están en $\mathcal{S}$.) Hay confusión adicional en que el término «frontera», que aparece más arriba, se refiere a una noción de «variedad con frontera», que no discuto en este libro. En el caso de las variedades ordinarias mencionadas en el capítulo 12 (i.e., «variedades sin frontera»), la noción de «cerrado» (en oposición al sentido topológico) es equivalente a «compacto». Para evitar confusiones, normalmente utilizaré solo el término «compacto» en este libro, antes que «cerrado». Excepciones son el uso de «curva cerrada» para una 1-variedad real que es topológicamente un círculo S^1 y «universo cerrado» para un modelo de universo que es *espacialmente* compacto, esto es, que contiene una hipersuperficie de tipo espacio compacta; véase la nota 27.36.

Sección 8.4

8.3. La transformación no se ve afectada si multiplicamos (reescalamos) cada uno de los a, b, c, d por el mismo número complejo no nulo, pero cambia si alteramos cada uno de ellos por separado. Esta libertad de reescalamiento global reduce en uno el número de parámetros independientes implicados en la transformación, de cuatro a tres.

8.4. Esto puede considerarse el principio de una larga historia cuyo clímax es el muy general y potente teorema de Atiyah-Singer (1963).

Sección 8.5

8.5. Habría que señalar que, solo en el caso de un lazo que sea un círculo exacto, la combinación de ambas versiones del teorema de la aplicación de Riemann nos dará una esfera de Riemann completa y suave.

9

Descomposición de Fourier e hiperfunciones

9.1. Series de Fourier

Volvamos a la cuestión, planteada en §6.1, acerca de lo que Euler y sus contemporáneos podrían haber considerado una idea aceptable de «función decente». En §7.1 nos hemos decidido por las funciones holomorfas (analítico-complejas) como las que mejor se ajustan a lo que Euler podría haber tenido en mente. Pese a todo, la mayoría de los matemáticos actuales considerarían que semejante idea de «función» es muy irrazonablemente restrictiva. ¿Quién tiene razón? Llegaremos a una respuesta muy notable a esta pregunta al final de este capítulo. Pero antes tratemos de entender cuáles son las cuestiones.

En la aplicación de las matemáticas a problemas del mundo físico, a menudo se requiere una flexibilidad que ni las funciones holomorfas ni sus contrapartidas en el campo real, las funciones analíticas (i.e., las C^ω-funciones), parecen poseer. Debido a la unicidad de la prolongación analítica, tal como se ha descrito en §7.4, el comportamiento global de una función holomorfa definida en una región abierta y conexa $\mathcal{D}$ del plano complejo está completamente determinado una vez que se conoce en una pequeña subregión abierta de $\mathcal{D}$. Análogamente, una función analítica de una variable real, definida en cierto segmento conexo $\mathcal{R}$ de la recta real $\mathbb{R}$, está también completamente determinada una vez que se conoce la función en una pequeña subregión abierta de $\mathcal{R}$. Esta rigidez parece poco razonable para la modelización de sistemas físicos realistas.

Sería particularmente incómoda cuando se considera la propagación de *ondas*. Esta propagación, que incluye el envío de señales mediante las vibraciones electromagnéticas de ondas de radio o luz, es de

gran utilidad por el hecho de que puede transmitirse información por este medio. Después de todo, el punto crucial de la señalización consiste en que debe existir la capacidad para enviar un mensaje que pueda resultar inesperado para el receptor. Si la forma de la señal tiene que estar dada por una función analítica, entonces no existe la posibilidad de «cambiar de idea» a mitad de mensaje. Cualquier pequeña parte de la señal determinaría la señal en su totalidad para cualquier instante. De hecho, una forma habitual de estudiar la propagación consiste en ver cómo se propagarán realmente las discontinuidades u otras desviaciones de la analiticidad.

Consideremos ondas y preguntemos cómo se describen matemáticamente estos objetos. Una de las formas más efectivas de estudiar formas de onda es mediante el procedimiento conocido como *análisis de Fourier*. Joseph Fourier fue un matemático francés que vivió entre 1768 y 1830. Estaba interesado en la descomposición de vibraciones periódicas en componentes sinusoidales. En el caso de la música, esto es básicamente lo que está implícito en la representación de cierto sonido musical en términos de sus «tonos puros» constituyentes. El término «periódico» significa que la pauta (digamos de los desplazamientos físicos del objeto que está vibrando) se repite con exactitud al cabo de cierto período de tiempo, o podría referirse a periodicidad en el espacio, como las pautas repetitivas en un cristal o en un papel pintado o en las olas del mar abierto. En términos matemáticos decimos que una función f (de una variable real χ, por ejemplo[1]) es *periódica*, para todo χ, si satisface

$$f(\chi + l) = f(\chi),$$

donde l es algún número fijo que se conoce como *período*. Así, si «deslizamos» la gráfica de $y = f(\chi)$ a lo largo del eje χ en una cantidad l, sigue teniendo el mismo aspecto que tenía antes. (Fig. 9.1a). (La forma en que Fourier manejaba funciones que *no* eran necesariamente periódicas —mediante el uso de la *transformada* de Fourier— se describirá en §9.4.)

Los «tonos puros» son cosas como sen χ o cos χ (Fig 9.1b). Estos tienen período 2π, puesto que

$$\text{sen}\,(\chi + 2\pi) = \text{sen}\,\chi, \quad \cos(\chi + 2\pi) = \cos\chi,$$

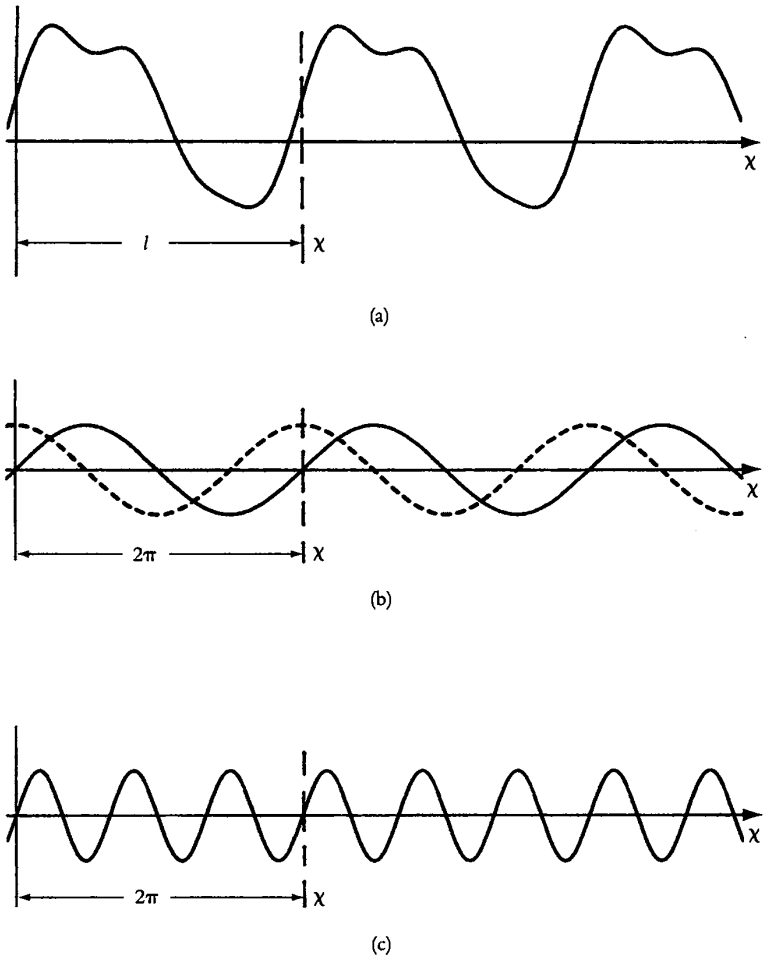


Fig. 9.1. Funciones periódicas. (a) $f(\chi)$ tiene período l si $f(\chi) = f(\chi + l)$, lo que significa que si deslizamos la gráfica $y = f(\chi)$ a lo largo del eje-χ en l, sigue pareciendo la misma que antes. (b) Los «tonos puros» básicos sen χ o cos χ (mostrados en trazos) tienen período $l = 2\pi$. (c) Los tonos puros «armónicos superiores» oscilan varias veces en el período l; siguen teniendo período l, aunque también tienen un período más corto (se ilustra sen 3χ, que tiene período $l = 2\pi$ y también el período más corto $2\pi/3$).

siendo estas relaciones manifestaciones de la periodicidad de una única cantidad compleja $e^{i\chi} = \cos\chi + i \operatorname{sen}\chi$,

$$e^{i(\chi + 2\pi)} = e^{i\chi},$$

que hemos encontrado ya en §5.3. Si queremos periodicidad l, en lugar de 2π, podemos «reescalar» la χ que aparece en la función, y tomar $e^{i2\pi\chi/l}$ en lugar de $e^{i\chi}$. Por consiguiente, las partes real e imaginaria $\cos(2\pi\chi/l)$ y $\operatorname{sen}(2\pi\chi/l)$ tendrán también período l. Pero esta no es la única posibilidad. En lugar de oscilar solo una vez en el período l, la función podría oscilar dos veces, tres veces, o de hecho n veces, donde n es cualquier entero positivo (véase la Fig. 9.1c), de modo que encontramos que cada una de las

$$e^{i\cdot 2\pi n\chi/l}, \quad \operatorname{sen}\left(\frac{2\pi n\chi}{l}\right), \quad \cos\left(\frac{2\pi n\chi}{l}\right)$$

tiene período l (además de tener también un período más pequeño l/n). En música, estas expresiones, para $n = 2, 3, 4, \ldots$ se conocen como *armónicos superiores*.

Un problema que abordó (y resolvió) Fourier era encontrar el modo de expresar una función periódica general $f(\chi)$, de período l, como suma de tonos puros. La contribución del correspondiente tono puro al total tendrá en general un valor diferente para cada n, que dependerá de la forma de onda (i.e., de la forma de la gráfica $y = f(\chi)$). Algunos ejemplos sencillos se ilustran en la Fig. 9.2. No obstante, lo normal es que el número de tonos puros diferentes que contribuyen a $f(\chi)$ sea infinito. Más concretamente, lo que Fourier necesitaba eran los coeficientes $c, a_1, b_1, a_2, b_2, a_3, b_3, a_4, \ldots$ de la descomposición de $f(\chi)$ en sus tonos puros constituyentes, tal como viene dada en la expresión

$$f(\chi) = c + a_1 \cos\omega\chi + b_1 \operatorname{sen}\omega\chi + a_2 \cos 2\omega\chi + b_2 \operatorname{sen} 2\omega\chi + \\ + a_3 \cos 3\omega\chi + b_3 \operatorname{sen} 3\omega\chi + \ldots,$$

donde, para hacer las expresiones más simples, las he escrito en términos de la *frecuencia angular* ω (nada que ver con la «ω» de §§5.4,5 y §8.1) dada por $\omega = 2\pi/l$.

Quizá para algunos lectores esta expresión sigue pareciendo excesivamente complicada —¡y esos lectores están realmente en lo cierto!—. La fórmula tiene una apariencia mucho más ordenada si agrupamos

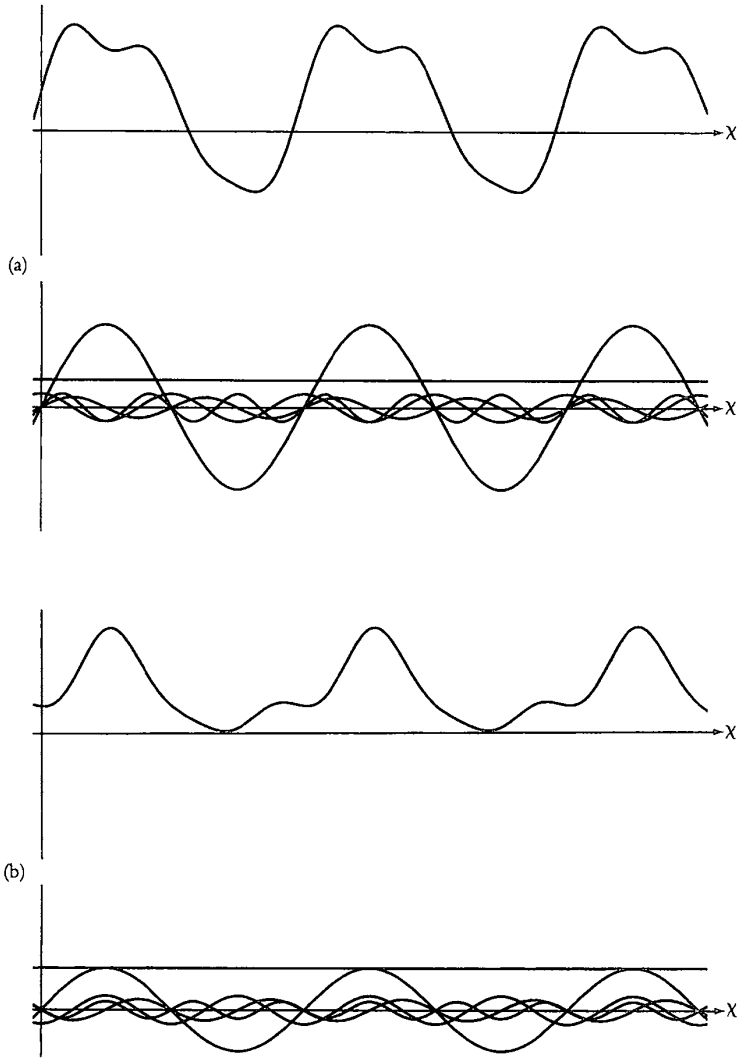


Fig. 9.2. Ejemplos de descomposición de Fourier de funciones periódicas. La forma de onda (forma de la gráfica) está determinada por los coeficientes de Fourier. Las funciones y bajo ellas sus componentes individuales de Fourier. (a) $f(\chi) = \frac{2}{3} + 2 \operatorname{sen} \chi + \frac{1}{3} \cos 2\chi + \frac{1}{4} \operatorname{sen} 2\chi + \frac{1}{3} \operatorname{sen} 3\chi$. (b) $f(\chi) = \frac{1}{2} + \operatorname{sen} \chi - \frac{1}{3} \cos 2\chi - \frac{1}{4} \operatorname{sen} 2\chi - \frac{1}{5} \operatorname{sen} 3\chi$.

los términos cos y sen como exponenciales complejas ($e^{iA\chi} = \cos A\chi + i \operatorname{sen} A\chi$), de modo que

$$f(\chi) = \ldots + \alpha_{-2}e^{-2i\omega\chi} + \alpha_{-1}e^{-i\omega\chi} + \alpha_0 + \alpha_1 e^{i\omega\chi} + \alpha_2 e^{2i\omega\chi} + \alpha_3 e^{3i\omega\chi} + \ldots,$$

donde[2,[9.1]]

$$a_n = \alpha_n + \alpha_{-n}, \quad b_n = i\alpha_n - i\alpha_{-n}, \quad c = \alpha_0$$

para $n = 1, 2, 3, 4, \ldots$ La expresión parece todavía más ordenada si ponemos $z = e^{i\omega\chi}$, y definimos la función $F(z)$ como la misma cantidad que $f(\chi)$, pero ahora expresada en términos de la nueva variable compleja z; pues entonces obtenemos

$$F(z) = \ldots + \alpha_{-2}z^{-2} + \alpha_{-1}z^{-1} + \alpha_0 z^0 + \alpha_1 z^1 + \alpha_2 z^2 + \alpha_3 z^3 + \ldots,$$

donde

$$F(z) = F(e^{i\omega\chi}) = f(\chi).$$

Y podemos hacer que parezca aún más ordenada utilizando el signo de suma Σ, que significa «sumar todos los términos, para todos los valores enteros de r':

$$F(z) = \sum \alpha_r z^r.$$

Esto se parece a una serie de potencias (véase §4.3), salvo que existen potencias *negativas* tanto como positivas. Se denomina una serie de *Laurent*. En la próxima sección veremos la importancia de esta expresión.[9.2]

9.2. Funciones sobre un círculo

La serie de Laurent nos proporciona un modo muy económico de representar las series de Fourier. Pero esta expresión también sugiere una

[9.1] Demuéstrelo.
[9.2] Demuestre que cuando F es analítica en el círculo unidad los coeficientes α_n, y con ello los a_n, b_n y c, pueden obtenerse mediante el uso de la fórmula $\alpha_n = (2\pi i)^{-1} \oint z^{-n-1} F(z)dz$.

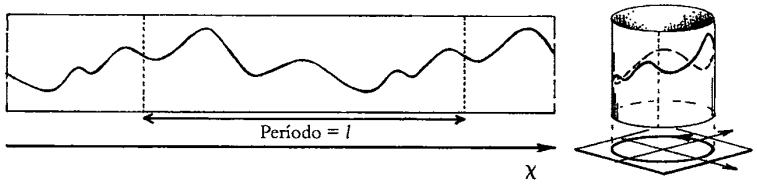


Fig. 9.3. Una función periódica de una variable real χ puede considerarse definida en un círculo de circunferencia l donde «enrollamos» el eje real de χ en el círculo. Con $l = 2\pi$, podemos tomar este círculo como el círculo unidad en el plano complejo.

interesante forma alternativa de considerar nuestra descomposición de Fourier. Puesto que una función periódica se repite a sí misma indefinidamente, podemos considerar que semejante función (de una variable real χ) está definida en un *círculo* (Fig. 9.3), donde el período l de la función es la longitud de la circunferencia del círculo y χ mide la distancia a lo largo de esta. Más que ir simplemente en línea recta, estas distancias dan vueltas ahora alrededor del círculo, de modo que la periodicidad es tenida en cuenta automáticamente.

Por conveniencia (al menos por ahora), tomo este círculo como el círculo unidad en el plano complejo, cuya circunferencia es 2π, y tomo el período l como 2π. Por consiguiente,

$$\omega = 1, \quad \text{de modo que } z = e^{i\chi}.$$

(Para cualquier otro valor del período, todo lo que tenemos que hacer es readmitir ω reescalando convenientemente la variable χ.) Los diferentes términos cos y sen que representan los diversos «tonos puros» de la descomposición de Fourier están ahora representados simplemente como potencias positivas o negativas de z, a saber, $z^{\pm n}$ para los n-ésimos armónicos. En el círculo unidad, estas potencias nos dan precisamente los términos cos y sen oscilatorios que necesitamos; véase la Fig. 9.4.

Ahora tenemos esta forma muy ordenada de representar la descomposición de Fourier de cierta función periódica $f(\chi)$. Consideramos que $f(\chi) = F(z)$ está definida en el círculo unidad en el z-plano, con $z = e^{i\chi}$, y entonces la descomposición de Fourier es simplemente

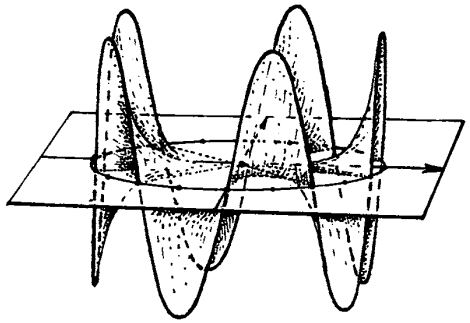


Fig. 9.4. En el círculo unidad, las partes real e imaginaria de la función z^n aparecen como ondas cos y sen (las partes real e imaginaria de $e^{in\chi}$, respectivamente, donde $z = e^{i\chi}$), que son n-ésimos armónicos. Aquí, para $n = 5$, se representa la parte real de z^5.

la descripción de dicha función como serie de Laurent, en términos de una variable compleja z. Pero la ventaja no está solo en una cuestión de pulcritud. Esta representación también nos proporciona ideas más profundas acerca de la naturaleza de las series de Fourier y del tipo de funciones que pueden representar. Más importante para el objetivo final de este libro es que tiene fuertes conexiones con la mecánica cuántica y, por consiguiente, es importante para nuestra comprensión profunda de la naturaleza. Esto es resultado de la magia de los números complejos, pues también podemos utilizar nuestra expresión en serie de Laurent cuando z yace lejos del círculo unidad. Resulta que esta serie nos dice algo importante sobre $F(z)$ cuando z está *sobre* el círculo unidad, en términos de lo que hace la serie cuando z está *fuera* del círculo unidad.

Recordemos ahora (de §4.4) la noción de círculo de convergencia, dentro del cual converge una serie de potencias y fuera del cual diverge. Hay un análogo de esto para una serie de Laurent: el *anillo de convergencia*. Este es la región comprendida estrictamente entre dos círculos en el plano complejo, ambos centrados en el origen (véase la Fig. 9.5a). Resulta sencillo de comprender una vez que tenemos la noción de círculo de convergencia para una serie de potencias ordinaria. La parte de la serie con potencias positivas,[3]

$$F^- = \alpha_1 z^1 + \alpha_2 z^2 + \alpha_3 z^3 + \ldots,$$

tendrá un círculo de convergencia ordinario, de radio A, pongamos por caso, y dicha parte de la serie converge para todos los valores de z

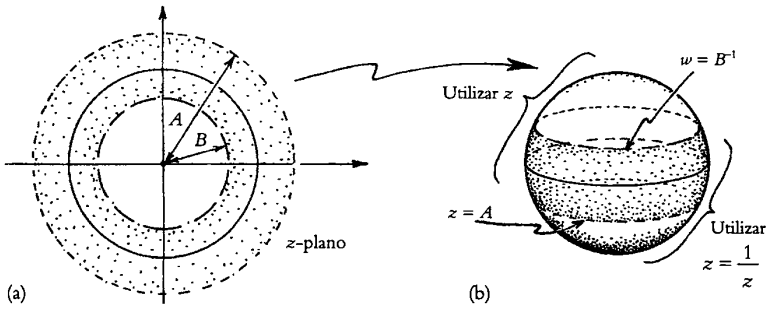


Fig 9.5. (a) El anillo de convergencia para una serie de Laurent $F(z) = F^+ + \alpha_0 + F^-$, donde $F^+ = \ldots + \alpha_{-2}z^{-2} + \alpha_{-1}z^{-1}$, $F^- = \alpha_1 z^1 + \alpha_2 z^2 + \ldots$ El radio de convergencia para F^+ es A y, en términos de $w = z^{-1}$, para F^- es B^{-1}. (b) Lo mismo, en la esfera de Riemann (véase la Fig. 8.7, donde z se refiere al hemisferio norte ampliado y w $(= z^{-1})$ al hemisferio sur ampliado.

cuyos módulos sean menores que A. Con respecto a la parte de la serie con potencias negativas, esto es,

$$F^+ = \ldots + \alpha_{-3}z^{-3} + \alpha_{-2}z^{-2} + \alpha_{-1}z^{-1},$$

podemos entenderla como una serie de potencias ordinaria en la variable recíproca $w = 1/z$. Habrá un círculo de convergencia en el w-plano, de radio $1/B$, pongamos por caso, y esa parte de la serie convergerá para valores de w cuyo módulo sea menor que $1/B$. (Aquí estamos hablando realmente de la esfera de Riemann, tal como se ha descrito en el capítulo 8; véase la Fig. 8.7, con la z-coordenada referida a un hemisferio y la w-coordenada referida al otro. Véase la Fig. 9.5b. Exploraremos el aspecto de esfera de Riemann de esto en la próxima sección.) Por consiguiente, para valores de z cuyos módulos sean mayores que B, la parte de la serie con potencias negativas convergerá. Con tal de que $B < A$, estas dos regiones de convergencia se solaparán, y obtendremos el anillo de convergencia para la serie de Laurent completa. Nótese que la serie entera de Laurent o de Fourier para la función $f(\chi) = F(e^{i\chi}) = F(z)$ viene dada por

$$F(z) = F^+ + \alpha_0 + F^-,$$

donde debe incluirse el término constante adicional α_0.

En la situación presente, buscamos convergencia *en* el círculo unidad, puesto que es en este donde podemos tener $z = e^{i\chi}$ para valores reales de χ, y la cuestión de la convergencia de nuestra serie de Fourier para $f(\chi)$ es precisamente la cuestión de la convergencia de la serie de Laurent para $F(z)$ cuando z está sobre el círculo unidad. Así pues, parece que necesitamos $B < 1 < A$, lo que asegura que el círculo unidad esté dentro del anillo de convergencia. ¿Significa esto que, para la convergencia de las series de Fourier, necesitamos que el círculo unidad esté dentro del anillo de convergencia?

Este sería el caso realmente si $f(\chi)$ fuera *analítica* (i.e., C^ω); pues entonces la función $f(\chi)$ puede ampliarse a una función $F(z)$ que es holomorfa en alguna región abierta que incluye al círculo unidad.[4] Pero si $f(\chi)$ no es analítica, surge una pregunta interesante. En este caso, o bien el anillo de convergencia se contrae hasta convertirse en el propio círculo unidad —lo que, estrictamente hablando, no está permitido para un anillo de convergencia genuino, puesto que el anillo de convergencia debería ser una región abierta, algo que el círculo unidad no es—, o, de lo contrario, el círculo unidad se convierte en la frontera exterior o interior del anillo de convergencia. Estas cuestiones tendrán importancia para nosotros en §§9.6,7.

Por el momento, no nos preocupemos por lo que sucede cuando $f(\chi)$ no es analítica y consideremos la situación más simple que aparece cuando $f(\chi)$ sí es analítica. Entonces tenemos el círculo unidad en el z-plano estrictamente contenido dentro de un anillo de convergencia genuino para $F(z)$, que está acotado por círculos (centrados en el origen) de radios A y B, con $B < 1 < A$. La parte de la serie de Laurent con potencias positivas, F^-, converge para puntos en el z-plano cuyos módulos son menores que A, y la parte con potencias negativas, F^+, lo hace para puntos en el z-plano cuyos módulos son mayores que B, de modo que ambas convergen dentro del propio anillo (y, en un sentido muy trivial, el término constante α_0 «converge» obviamente para todo z). Esto nos proporciona una «separación» de la función $F(z)$ en dos partes, una holomorfa dentro del círculo exterior y la otra holomorfa fuera del círculo interior, que están definidos, respectivamente, por las expresiones de las series para F^- y F^+.

Hay una (ligera) ambigüedad acerca de si el término constante α_0

debe ser incluido en F^- o en F^+ en esta separación. De hecho, es mejor vivir con esta ambigüedad. Existe una simetría entre F^- y F^+ que se hace más clara si adoptamos la imagen de la esfera de Riemann a la que nos hemos referido antes (véase la Fig. 9.5b). Esto nos da una imagen más completa de la situación, de modo que vamos a explorarla a continuación.

9.3. Separación de frecuencias sobre la esfera de Riemann

Las coordenadas z y w $(= 1/z)$ nos dan dos cartas que recubren la esfera de Riemann. El círculo unidad se convierte en el ecuador de la esfera y el anillo es ahora simplemente un «collar» del ecuador. Consideramos que nuestra separación de $F(z)$ la expresa como suma de dos partes: una de ellas —llamada la parte de *frecuencia positiva* de $F(z)$— se extiende de forma holomorfa al hemisferio sur y está definida por $F^+(z)$, junto con cualquier porción del término constante que queramos incluir; y la otra —llamada parte de *frecuencia negativa* de $F(z)$— se extiende de forma holomorfa al hemisferio norte y está definida por $F^-(z)$ y la porción restante del término constante. Si ignoramos el término constante, esta separación está unívocamente determinada por este requisito de holomorficidad para la extensión a uno u otro de los dos hemisferios.[9.3]

En ocasiones, será práctico referirse al «interior» y al «exterior» de un círculo (u otro lazo cerrado) trazado sobre la esfera de Riemann apelando a una *orientación* que debe asignarse al círculo. La orientación estándar del círculo unidad en el z-plano está dada por la dirección en que aumenta la coordenada θ estándar, i.e., contraria a las agujas del reloj. Si invertimos esta orientación (por ejemplo, reemplazando θ por $-\theta$), entonces intercambiamos frecuencias positivas y negativas. Nuestra convención para un lazo cerrado general debe ser consistente con esto. La orientación es contraria a las agujas del reloj si la «esfera del reloj» está en el interior del lazo, por así decirlo, mientras que sería en el sentido de las agujas si hubiera que colocar la «esfera del reloj» en el exterior del lazo. Esto sirve para definir el «interior» y el «exterior» de un lazo cerrado orientado. La Fig. 9.6 debería dejar clara esta cuestión.

[9.3] ¿Puede ver por qué?

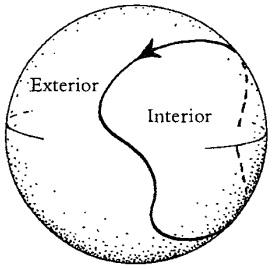


Fig. 9.6. Una orientación asignada a un lazo cerrado en la esfera de Riemann define su «interior» y su «exterior» como se indica: esta orientación es en sentido contrario a las agujas del reloj para una «esfera de reloj» dentro del lazo (y en el sentido de las agujas si está fuera).

Esta separación de una función en sus partes de frecuencia positiva y negativa es un ingrediente crucial de la teoría cuántica, y muy especialmente de la teoría cuántica de campos, como veremos en §24.3 y §§26.2-4. La formulación concreta que he dado aquí no es ni mucho menos la forma más habitual en que se expresa esta separación, pero tiene algunas ventajas considerables en varios contextos diferentes (especialmente en la teoría de twistores, por ejemplo; véase §33.10). La formulación habitual no se interesa tanto por las extensiones holomorfas como por el desarrollo de Fourier directamente. Las componentes de frecuencia positiva son aquellas dadas por múltiplos de e^{-inx}, donde n es positiva, en oposición a aquellas dadas por múltiplos de e^{inx}, que son componentes de frecuencia negativa. Una función de frecuencia positiva es aquella que está compuesta enteramente de componentes de frecuencia positiva.

Sin embargo, esta descripción no revela la completa generalidad de lo que está implícito en esta separación. Existen muchas aplicaciones holomorfas de la esfera de Riemann en sí misma que envían cada hemisferio sobre sí mismo, pero que no conservan los polos norte o sur (i.e., los puntos $z = 0$ o $z = \infty$).[9.4] Estas conservan la separación de frecuencias positivas/negativas, pero no conservan las componentes de Fourier individuales e^{-inx} o e^{inx}. Así, la cuestión de la separación en frecuencias positivas y negativas (que resulta crucial en teoría cuántica) es una noción más general que la selección de las componentes de Fourier individuales.

[9.4] ¿Cuáles son, explícitamente, estas aplicaciones?

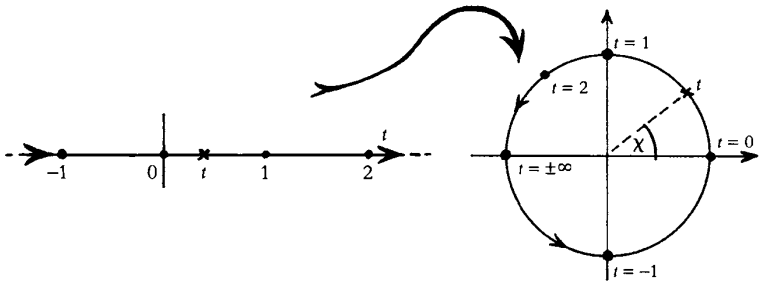


Fig. 9.7. En mecánica cuántica, la separación en frecuencias positiva/negativa se refiere a funciones del tiempo t, que no se suponen periódicas. La separación de la Fig. 9.5 aún puede aplicarse, para el intervalo completo de t (desde $-\infty$ a $+\infty$), si utilizamos la transformación que relaciona t con $z(= e^{i\chi})$, donde rodeamos el círculo unidad en sentido contrario a las agujas del reloj, desde $z = -1$ y vuelta a $z = -1$, de modo que χ va de $-\pi$ a π.

En las discusiones habituales de la mecánica cuántica, la separación en frecuencias positivas/negativas se refiere a funciones del *tiempo t*, y normalmente no consideramos que el tiempo esté dando vueltas en un círculo. Pero podemos utilizar una transformación sencilla para obtener el intervalo completo de t, desde el «límite pasado» $t = -\infty$ al «límite futuro» $t = \infty$, a partir de una χ que recorre una vez el círculo; aquí considero que χ corre entre los límites $\chi = -\pi$ y $\chi = \pi$ (de modo que $z = e^{i\chi}$ recorre el círculo unidad en el plano complejo en sentido contrario a las agujas del reloj, partiendo del punto $z = -1$ y volviendo de nuevo a $z = -1$; véase la Fig. 9.7). Esta transformación viene dada por

$$t = \tan \frac{1}{2}\chi.$$

La gráfica de esta relación se da en la Fig. 9.8 y una descripción geométrica simple se ofrece en la Fig. 9.9.

Una ventaja de esta transformación particular es que se extiende de manera holomorfa a toda la esfera de Riemann; es una transformación que ya consideramos en §8.3, que lleva el círculo unidad (z-plano) a la recta real (t-plano):[9.5]

[9.5] Demuestre que esto da la misma t que antes.

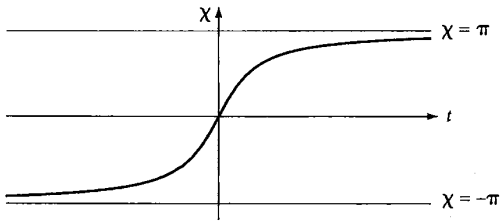


Fig. 9.8. Gráfica de $t = \tan \chi/2$.

$$t = \frac{z-1}{iz+i}, \quad z = \frac{-t+i}{t+i}.$$

El interior del círculo unidad en el z-plano corresponde al semi-t-plano superior y el exterior del círculo unidad corresponde al semi-t-plano inferior. Así, las funciones de t de frecuencia positiva son aquellas que se extienden de forma holomorfa al semiplano inferior de t. (Hay, sin embargo, un tecnicismo adicional con el que debemos tener cuidado cuando trabajamos con el punto «∞» del t-plano; pero este se maneja adecuadamente si pensamos siempre en términos de la esfera de Riemann, en lugar de simplemente el t-plano complejo.)

Sin embargo, en las presentaciones estándar, la noción de «frecuencia positiva» en términos de una coordenada temporal t no se enuncia normalmente de la forma concreta que yo acabo de hacer, sino más bien en términos de lo que se denomina *transformada de Fourier* de $f(\chi)$. La respuesta es en realidad la misma[5] que la que he dado, pero puesto

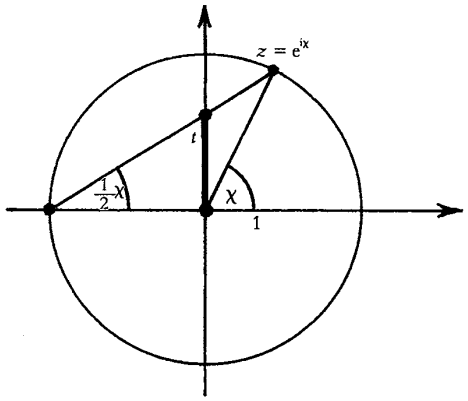


Fig. 9.9. Geometría de $t = \tan \frac{\chi}{2}$.

que las transformadas de Fourier son de importancia crucial para la mecánica cuántica (y también en muchas otras áreas), será importante explicar aquí qué es realmente esta transformada.

9.4. La transformada de Fourier

Básicamente, una transformada de Fourier es el límite de una serie de Fourier cuando el período l de nuestra función periódica $f(\chi)$ se hace cada vez mayor hasta que se hace infinito. En este límite infinito no hay ninguna restricción de periodicidad sobre $f(\chi)$; es solo una función ordinaria.[6] Esto tiene ventajas considerables cuando estamos estudiando la propagación de ondas y la capacidad de enviar señales «inesperadas», pues entonces no queremos insistir en que la forma de la señal sea periódica. La transformada de Fourier nos permite considerar tales señales «excepcionales» mientras seguimos analizándolas en términos de «tonos puros» periódicos. De hecho, lo consigue considerando que nuestra función $f(\chi)$ tiene período $l \to \infty$. A medida que el período l se hace mayor, los tonos puros armónicos, que tienen períodos l/n para algún entero positivo n, se harán cada vez más próximos a cualquier número real positivo que escojamos. (Recordemos que cualquier número real puede aproximarse por racionales con la precisión que queramos.) Lo que esto nos dice es que cualquier tono puro de cualquier frecuencia está ahora permitido como componente de Fourier. En lugar de tener $f(\chi)$ expresada como suma discreta de componentes de Fourier, ahora tenemos $f(\chi)$ expresada como una suma continua sobre todas las frecuencias, lo que significa que $f(\chi)$ está ahora expresada como una *integral* (véase §6.6) sobre la frecuencia.

Veamos, a grandes rasgos, cómo funciona esto. En primer lugar, recordemos nuestra expresión «más ordenada» para la descomposición de Fourier de una función periódica $f(\chi)$, de período l, dada más arriba:

$$F(z) = \sum \alpha_r z^r, \quad \text{donde } z = e^{i\omega\chi}$$

(la frecuencia angular ω está dada por $\omega = 2\pi/l$). Consideremos que el período es inicialmente 2π, de modo que $\omega = 1$. Ahora vamos a tratar de aumentar el período en un factor grande N (así que ahora $l = 2\pi N$),

de modo que la frecuencia se reduce en el mismo factor (i.e., $\omega = N^{-1}$). La onda oscilatoria que acostumbraba a ser el tono puro fundamental se convierte ahora en el N-ésimo armónico con respecto a esta nueva frecuencia más baja. Un tono puro que antes era un n-ésimo armónico será ahora un (nN)-ésimo armónico. Cuando tomamos el límite de N tendiendo a infinito, no es conveniente identificar una componente oscilatoria particular etiquetándola por su «número armónico» (i.e., por el número n), puesto que este número sigue cambiando. Es decir, no es conveniente etiquetar esta componente oscilatoria por el entero r, en la suma anterior, puesto que un valor fijo de r etiqueta un armónico particular ($r = \pm n$, para el n-ésimo armónico), en lugar de identificar un tono de frecuencia particular. En su lugar, es r/N el que identifica esta frecuencia, y necesitamos una nueva variable para etiquetarlo. Teniendo en cuenta el uso importante que haremos de la transformada de Fourier en capítulos posteriores (véase especialmente §21.11), llamaré a esta variable «p», que, en el límite cuando N tiende a infinito, representa el *momento*[7] de una partícula mecanocuántica cuya posición viene medida por χ.

Para N finito, escribo

$$p = \frac{r}{N}.$$

En el límite $N \to \infty$, el parámetro p se convierte en una variable continua y, puesto que los «coeficientes α_r» en nuestra suma dependerán entonces del parámetro continuo con valores reales p en lugar del parámetro discreto con valores enteros r, es mejor escribir la dependencia de los coeficientes α_r de r utilizando el tipo estándar de notación funcional, digamos $g(p)$, en lugar de utilizar solo un sufijo (por ejemplo, «g_p»), como en α_r. En efecto, haremos el reemplazamiento

$$\alpha_r \mapsto g(p)$$

en nuestra suma $\Sigma\alpha_r z^r$, pero debemos tener en cuenta que a medida que N crece, el número de términos que yacen dentro de cierto intervalo pequeño de valores de p se hace mayor (básicamente, en proporción a N, puesto que estamos considerando fracciones n/N que yacen en dicho rango). En consecuencia, la cantidad $g(p)$ es realmente una medida de densidad, y debe ir acompañada por la cantidad diferencial

dp en el límite, cuando la suma Σ se convierte en una integral $\int$. Finalmente, consideremos el término z^r en nuestra suma $\Sigma\alpha_r z^r$. Tenemos $z = e^{i\omega\chi}$, con $\omega = N^{-1}$, de modo que $z = e^{i\chi/N}$. Así pues, $z^r = e^{i\chi r/N} = e^{i\chi p}$, de modo que, juntando todo esto, obtenemos en el límite $N \to \infty$ la expresión

$$\sum \alpha_r z^r \to \int_{-\infty}^{\infty} g(p) e^{i\chi p} dp,$$

que representa nuestra función $f(\chi)$. De hecho, es habitual incluir un factor de escala de $(2\pi)^{-1/2}$ con la integral, pues con ello se da una notable simetría que consiste en que la relación *inversa*, que expresa $g(p)$ en términos de $f(\chi)$, tiene exactamente la misma forma (aparte de un signo menos) que la que expresa $f(\chi)$ en términos de $g(p)$:

$$f(\chi) = (2\pi)^{-1/2} \int_{-\infty}^{\infty} g(p) e^{i\chi p} dp, \quad g(p) = (2\pi)^{-1/2} \int_{-\infty}^{\infty} f(\chi) e^{-i\chi p} d\chi.$$

Las funciones $f(\chi)$ y $g(p)$ se denominan *transformadas de Fourier* una de otra.[9.6]

9.5. Separación de frecuencias a partir de la transformada de Fourier

Una función (compleja) $f(\chi)$, definida en toda la recta real, se dice de *frecuencia positiva* si su transformada de Fourier $g(p)$ es *cero* para todo $p \geq 0$. Así pues, $f(\chi)$ está compuesta solo de componentes de la forma $e^{i\chi p}$ con $p < 0$. (Euler podría muy bien haberse preocupado —véase §6.1— con una $g(p)$ semejante, que parece ser un descarado «trabajo de pegado» entre una función no nula para $p < 0$ y simplemente cero para $p > 0$. Pese a todo, esto parece estar representando una propiedad «holomorfa» perfectamente respetable de $f(\chi)$.) Otra forma de expresar esta condición de «frecuencia positiva» es mediante la extensibilidad holomorfa de $f(\chi)$, como se ha hecho antes para la serie de Fourier. Ahora consideramos que la variable χ etiqueta los puntos del eje real

[9.6] Demuestre (a grandes rasgos) cómo obtener la expresión para $g(p)$ en términos de $f(\chi)$ utilizando una forma límite de la expresión para la integral de contorno $\alpha_n = (2\pi i)^{-1} \oint z^{-n-1} F(z) dz$ del ejercicio [9.2].

(de modo que podemos tomar $\chi = x$ sobre este eje), donde en la esfera de Riemann este «eje real» (incluyendo el punto «$\chi = \infty$») es ahora el *círculo real* (véase la Fig. 8.7c). Este círculo divide la esfera en dos hemisferios, de los que el «exterior» corresponde al semiplano inferior en la imagen estándar del plano complejo. La condición de que $f(\chi)$ sea de frecuencia positiva equivale ahora a que se extienda de manera holomorfa a este hemisferio exterior.

No obstante, hay una cuestión que requiere cierto cuidado cuando comparamos estas dos definiciones de «frecuencia positiva». Está relacionada con la cuestión de cómo tratamos el punto $z = \infty$, puesto que la función $f(\chi)$ tendrá, en general, algún tipo de singularidad allí. De hecho, con tal de que adoptemos el punto de vista «hiperfuncional» que describiré en breve (en §9.7), esta singularidad en $z = \infty$ no nos presenta ninguna dificultad esencial. Con el punto de vista adecuado

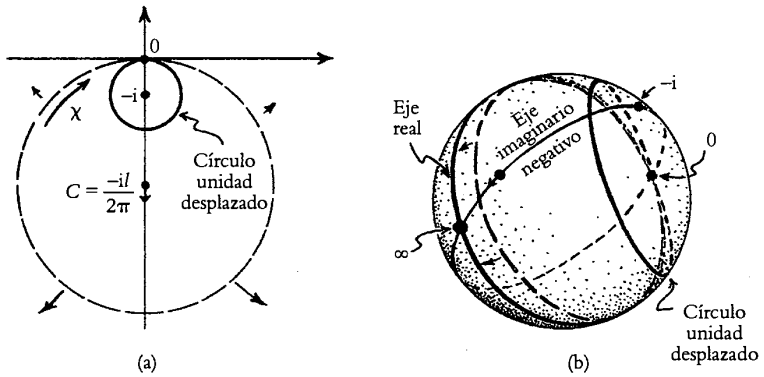


Fig. 9.10. Condición de frecuencia positiva, cuando $l \to \infty$, donde l es el período de $f(\chi)$. (a) Empezamos con $l = 2\pi$, con f definida en el círculo unidad desplazado para que su centro esté en $z = -i$. Para l creciente, el círculo tiene radio l y centro en $C = -il/2\pi$. En cada caso χ mide la longitud de arco *en el sentido de las agujas del reloj*. La frecuencia positiva se expresa en que f se extiende de forma holomorfa al interior del círculo, y en el límite $l = \infty$, al semiplano inferior. (b) Lo mismo, en la esfera de Riemann. Para l finito, la serie de Fourier se obtiene a partir de una serie de Laurent en torno a $z = -il/2\pi$, pero en la esfera este punto no es el centro del círculo, convirtiéndose en el punto ∞ (que yace en él) en el límite $l = \infty$, donde la serie de Fourier se convierte en la transformada de Fourier.

con respecto a «$f(\infty)$», resulta que las dos definiciones de frecuencia positiva que he dado en el párrafo anterior están en mutuo acuerdo.[8]

Para el lector interesado, puede ser útil examinar, en términos de la esfera de Riemann, algo de la geometría que está implícita en nuestro límite de §9.4, que nos lleva de la serie de Fourier a la transformada de Fourier. Volvamos a la descripción en el z-plano que hemos estado considerando anteriormente para una función $f(\chi)$ de período 2π, donde χ mide la longitud del arco a lo largo de un círculo de radio unidad. Supongamos que queremos cambiar el período a valores mayores que 2π, en pasos sucesivamente crecientes, aunque manteniendo la interpretación de χ como una distancia a lo largo de un círculo. Podemos hacerlo considerando una secuencia de círculos cada vez mayores, pero para que el procedimiento de paso al límite tenga sentido geométrico supondremos que todos los círculos se tocan en el punto de partida $\chi = 0$ (véase la Fig. 9.10a). Por simplicidad en lo que sigue, escojamos este punto en el origen $z = 0$ (en lugar de $z = 1$), con todos los círculos yaciendo en el semiplano inferior. Esto hace que nuestro círculo inicial, para período $l = 2\pi$, sea el círculo unidad centrado en $z = -i$ en lugar de estar centrado en el origen. Para un período $l > 2\pi$, el círculo está centrado en el punto $C = -il/2\pi$, y en el límite $l \to \infty$ obtenemos el propio eje real (de modo que $\chi = x$), al haber movido el «centro» hasta el infinito a lo largo del eje imaginario negativo. En cada caso, tomamos χ como una medida de la longitud de arco *en el sentido de las agujas del reloj* alrededor del círculo (o, en el caso límite, la distancia positiva a lo largo del eje real), con $\chi = 0$ en el origen. Puesto que nuestros círculos tienen ahora una orientación no estándar (i.e., en el sentido de las agujas del reloj), sus «exteriores» son sus *interiores* (véanse §9.3 y la Fig. 9.6), de modo que nuestra condición de frecuencia positiva se refiere a su interior. Tenemos ahora la relación entre χ y z expresada en la forma[9.7]

$$z = \frac{il}{2\pi}(e^{-i\chi} - 1).$$

Para l finito, podemos expresar $f(\chi)$ como una serie de Fourier remitiendo a una serie de Laurent en torno al punto $C = -il/2\pi$. Obte-

[9.7] Deduzca esta expresión.

nemos la transformada de Fourier tomando el límite $l \to \infty$. Para l finito, obtenemos la condición de frecuencia positiva como la extensibilidad holomorfa de $f(\chi)$ al *interior* del círculo relevante; en el límite $l \to \infty$, esto se convierte en la extensibilidad holomorfa al semiplano inferior, de acuerdo con lo que se ha afirmado antes.

¿Qué sucede con la serie de Laurent en el límite $l \to \infty$? Tendremos que considerar la esfera de Riemann para entender lo que sucede en este límite. Para cada valor finito de l, el punto $C(= il/2\pi)$ es el centro del χ-círculo, pero, en la esfera de Riemann, el punto C no tiene por qué ser nada como el centro del círculo. Cuando l aumenta, C se desplaza a lo largo del círculo en la esfera de Riemann que representa al eje imaginario (véase la Fig. 9.10b), y el punto $C(= il/2\pi)$ se parece cada vez menos al centro del círculo. Finalmente, cuando se alcanza el límite $l \to \infty$, C se convierte en el punto $z = \infty$ en la esfera de Riemann. Pero cuando $C = \infty$, ¡encontramos que en realidad *yace sobre* el círculo del que se supone que es el centro! (Este círculo es ahora, por supuesto, el eje real.) Así pues, hay algo peculiar (o «singular») en el hecho de tomar una serie de potencias en torno a dicho punto, lo que es de esperar, por supuesto, porque ya no tenemos una suma de términos individuales sino una integral continua.

9.6. ¿Qué tipo de función es apropiada?

Volvamos a la pregunta planteada al comienzo de este capítulo, concerniente al tipo de «función» que es apropiado utilizar. Plantearemos la siguiente pregunta: ¿qué tipo de funciones podemos representar como transformadas de Fourier? Parecería inadecuado restringir la atención únicamente a las funciones analíticas (i.e., a las C^ω-funciones) porque, como hemos visto antes, la transformada de Fourier $g(p)$ de una función $f(\chi)$ de frecuencia positiva —que ciertamente puede ser analítica— es un «trabajo de pegado» característicamente no analítico de una función no nula con la función cero. La relación entre una función y su transformada de Fourier es simétrica, de modo que no parece razonable adoptar criterios diferentes para cada una de ellas. Como cuestión adicional, se advirtió antes que el comportamiento de $f(\chi)$ en el

punto $\chi = \infty$ es relevante para la cuestión de su separación en frecuencias positivas/negativas, pero solo en circunstancias muy especiales sería $f(\chi)$ realmente analítica (C^ω) en ∞ (puesto que esto requeriría un empalme preciso entre el comportamiento de $f(\chi)$ cuando $\chi \to +\infty$ y cuando $\chi \to -\infty$). Además de todo esto, está nuestra motivación *física* inicial, mencionada antes, para estudiar las transformadas de Fourier, a saber: que nos permiten tratar señales que pueden transmitir mensajes «inesperados» (no analíticos). Así pues, debemos volver a la pregunta a la que nos enfrentábamos al comienzo de esta sección: ¿qué tipo de objeto deberíamos aceptar como función «respetable»?

Recordemos que, por una parte, Euler y sus contemporáneos hubieran admitido probablemente que una función holomorfa (o analítica) era el tipo de objeto que tenían en mente como una «función» respetable; pero, por otra parte, tales funciones parecen innecesariamente restrictivas para muchos tipos de problemas matemáticos y físicos, incluyendo aquellos relativos a la propagación de ondas, de modo que se necesita una noción más general. ¿Es uno de estos puntos de vista más «correcto» que el otro? Hay tal vez una opinión dominante que afirma que los defensores del primer punto de vista están «pasados de moda», y que los conceptos modernos se inclinan bastante hacia el segundo, siendo las funciones holomorfas o analíticas tan solo casos muy especiales de la noción general de una «función». Pero ¿es esta necesariamente la actitud correcta que se debe tomar? Tratemos de ponernos en el marco mental del siglo XVIII.

Entra Joseph Fourier a comienzos del siglo XIX. Aquellos que pertenecían a la escuela «analítica» («euleriana») de pensamiento habrían recibido un golpe bajo cuando Fourier demostró que ciertas «funciones» periódicas, tales como la onda cuadrada o el diente de sierra mostradas en la Fig. 9.11, ¡tienen representaciones de Fourier de apariencia perfectamente razonable! Fourier encontró mucha oposición por parte de la comunidad matemática de su época. Muchos se mostraron reacios a creer sus conclusiones. ¿Cómo podría haber una «fórmula» para la función onda-cuadrada, por ejemplo? Pese a todo, como Fourier demostró, la suma de la serie

$$s(\chi) = \operatorname{sen}\chi + \frac{1}{3}\operatorname{sen} 3\chi + \frac{1}{5}\operatorname{sen} 5\chi + \frac{1}{7}\operatorname{sen} 7\chi + \ldots$$

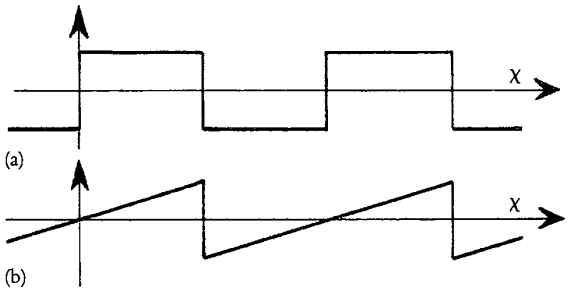


Fig. 9.11. Funciones periódicas discontinuas (con representaciones de Fourier de apariencia perfectamente razonable): (a) Onda cuadrada, (b) Diente de sierra.

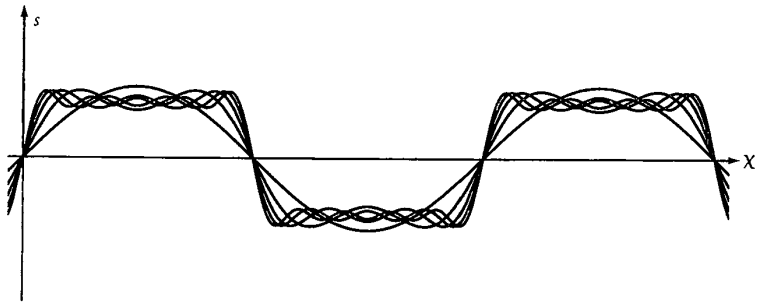


Fig. 9.12. Sumas parciales de la serie de Fourier $s(\chi) = \operatorname{sen}\chi + \frac{1}{3}\operatorname{sen} 3\chi + \frac{1}{5}\operatorname{sen} 5\chi + \frac{1}{7}\operatorname{sen} 7\chi + \frac{1}{9}\operatorname{sen} 9\chi + \ldots$, que converge a una onda cuadrada (como la de Fig. 9.11a).

es realmente una onda cuadrada que oscila entre los valores constantes $\frac{1}{4}\pi$ y $-\frac{1}{4}\pi$ en el semiperíodo π (véase la Fig. 9.12).

Consideremos la descripción en serie de Laurent para esto, tal como se ha dado antes. Tenemos la expresión de apariencia bastante elegante[9.8]

$$2is(\chi) = \ldots -\frac{1}{5}z^{-5} - \frac{1}{3}z^{-3} - z^{-1} + z + \frac{1}{3}z^3 + \frac{1}{5}z^5 + \ldots,$$

[9.8] Demuéstrelo.

donde $z = e^{iX}$. De hecho, este es un ejemplo donde el anillo de convergencia se contrae hasta el círculo unidad, sin que quede ninguna región abierta. Sin embargo, aún podemos dar sentido a las cosas en términos de funciones holomorfas si separamos la serie de Laurent en dos mitades, una con las potencias positivas, que da una serie de potencias ordinaria en z, y otra con potencias negativas, que da una serie de potencias en z^{-1}. En efecto, estas son series bien conocidas, y pueden ser sumadas explícitamente:[9.9]

$$S^- = z + \frac{1}{3}z^3 + \frac{1}{5}z^5 + \ldots = \frac{1}{2}\log\left(\frac{1+z}{1-z}\right)$$

y

$$S^+ = \ldots -\frac{1}{5}z^{-5} - \frac{1}{3}z^{-3} - z^{-1} = -\frac{1}{2}\log\left(\frac{1+z^{-1}}{1-z^{-1}}\right),$$

lo que da $2is(\chi) = S^- + S^+$. Un pequeño reordenamiento de estas expresiones lleva a la conclusión de que S^- y $-S^+$ difieren solo en $\pm\frac{1}{2}i\pi$, lo que nos dice que $s(\chi) = \pm\frac{1}{4}\pi$.[9.10] Pero necesitamos mirar un poco más de cerca para ver por qué obtenemos realmente una onda cuadrada que oscila entre estos valores alternos.

Es un poco más facil apreciar lo que está pasando si aplicamos la transformación $t = (z-1)/(iz+i)$, dada en §8.3, que lleva el interior del círculo unidad en el z-plano al semi-t-plano superior (como se ilustra en la Fig. 8.8). En términos de t, la cantidad S^- se refiere ahora a este semiplano superior y S^+ al semiplano inferior, y encontramos (con posibles ambigüedades $2\pi i$ en los logaritmos)

$$S^- = -\frac{1}{2}\log t + \frac{1}{2}\log i, \quad S^+ = \frac{1}{2}\log t + \frac{1}{2}\log i.$$

Siguiendo los logaritmos de forma continua desde los respectivos puntos de partida $t = i$ (donde $S^- = 0$) y $t = -i$ (donde $S^+ = 0$), encontramos que a lo largo del eje t real positivo $S^- + S^+ = +\frac{1}{2}i\pi$, mientras

[9.9] Hágalo, aprovechando un desarrollo en serie de potencias para $\log z$ en torno a $z = 1$, dado al final de §7.4.

[9.10] Demuéstrelo (suponiendo que $|s(\chi)| < 3\pi/2$).

que a lo largo del eje t real negativo tenemos $S^- + S^+ = -\frac{1}{2}i\pi$.[9.11] De ello deducimos que a lo largo de la mitad superior del círculo unidad en el z-plano tenemos $s(\chi) = +\frac{1}{4}\pi$, mientras que a lo largo de la mitad inferior tenemos $s(\chi) = -\frac{1}{4}\pi$. Esto demuestra que la suma de la serie de Fourier da realmente la onda cuadrada, precisamente lo que Fourier había afirmado.

¿Cuál es la moraleja de este ejemplo? Hemos visto que una función (periódica) particular que ni siquiera es continua, y mucho menos diferenciable (siendo en este caso una C^{-1}-función), puede representarse por una serie de Fourier de apariencia perfectamente razonable. De forma equivalente, cuando consideramos la función definida en el círculo unidad, esta puede representarse por una serie de Laurent de apariencia razonable, aunque una para la que el anillo de convergencia se ha contraído hasta el propio círculo unidad. La suma de cada una de las mitades positiva y negativa de esta serie de Laurent da sendas funciones holomorfas perfectamente buenas en la esfera de Riemann. Una está definida en un lado del círculo unidad y la otra está definida en el otro lado. Podemos considerar que la «suma» de estas dos funciones da la onda cuadrada requerida en el propio círculo unidad. La existencia de singularidades rama en los dos puntos $z = \pm 1$ en el círculo unidad es lo que hace que la suma pueda «saltar» de un lado a otro, dando la onda cuadrada que aparece en la suma. Estas singularidades rama impiden también que las series de potencias en los dos lados converjan más allá del círculo unidad.

9.7. Hiperfunciones

Este ejemplo es tan solo un caso muy especial, pero ilustra lo que debemos hacer en general. Preguntemos cuál es el tipo más general de función que puede ser definida en el círculo unidad (en la esfera de Riemann) y representada como una «suma» de una función holomorfa F^+

[9.11] Demuéstrelo.

en la región abierta que yace a un lado del círculo y de otra función F^- en la región abierta que yace al otro lado, igual que en el ejemplo que hemos estado considerando. Encontraremos que la respuesta a esta pregunta nos lleva directamente a una idea exótica pero importante conocida como una «hiperfunción».

De hecho, resulta más ilustrativo considerar f como la «diferencia» entre F^- y $-F^+$. Una razón para ello es que, en los casos más generales, puede no haber extensión analítica de F^- o de F^+ al círculo unidad, de modo que no está claro lo que una «suma» semejante podría significar sobre el propio círculo. Sin embargo, podemos considerar que la *diferencia* entre F^- y $-F^+$ representa el «salto» entre estas dos funciones cuando sus regiones de definición se juntan en el círculo unidad.

Esta idea de un «salto» entre una función holomorfa en un lado de una curva en el plano complejo y otra función holomorfa en el otro lado —donde ninguna de las dos funciones holomorfas necesita extenderse de forma holomorfa sobre la propia curva— nos proporciona realmente un nuevo concepto de una «función» definida sobre la curva. Esta es, en efecto, la definición de una *hiperfunción* sobre una curva (analítica). Es una noción maravillosa propuesta por el matemático japonés Mikio Sato en 1958,[9] aunque, como veremos enseguida, la definición real de Sato es considerablemente más elegante.[10]

No necesitamos considerar una curva cerrada, como el círculo unidad entero, para la definición de una hiperfunción, pero podemos considerar una parte de la curva. En realidad, es más habitual considerar hiperfunciones definidas en algún segmento γ de la recta real. Tomaremos γ como el segmento de la recta real entre a y b, donde a y b son números reales con $a < b$. Una hiperfunción definida en γ es entonces el *salto* a través de γ, partiendo de una función holomorfa f en un conjunto abierto $\mathcal{R}^-$ (que tiene γ como frontera superior) hasta una función holomorfa g en un conjunto abierto $\mathcal{R}^+$ (que tiene γ como su frontera inferior); véase la Fig. 9.13.

Hablar de un «salto» de esta forma no nos da mucha idea sobre qué hacer con tal objeto (y tampoco es muy preciso en matemáticas). La elegante resolución de Sato para estas cuestiones consiste en proceder de una manera más bien formalmente algebraica, que en realidad es bastante sencilla. Simplemente representamos este salto como el *par* $(f,$

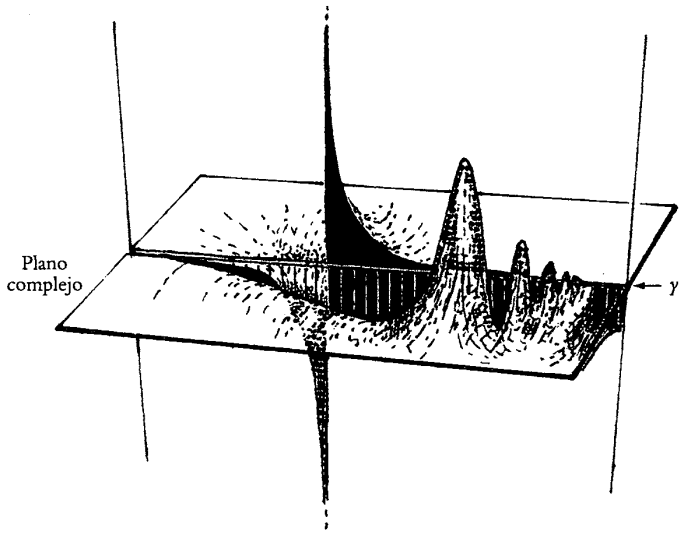


Fig. 9.13. Una hiperfunción en un segmento γ del eje real expresa el «salto» desde una función holomorfa en un lado de γ a una en el otro lado.

$g)$ de estas funciones holomorfas, pero donde decimos que un par (f, g) es *equivalente* a otro par (f_0, g_0) si el segundo se obtiene del primero añadiendo a ambas f y g la misma función holomorfa h, donde h está definida en la región (abierta) combinada $\mathcal{R}$, que consiste en $\mathcal{R}^-$ y $\mathcal{R}^+$ unidas a lo largo del segmento γ; véase la Fig. 9.14. Podemos decir

$$(f, g) \quad \text{es equivalente a} \quad (f + h, g + h),$$

donde las funciones holomorfas f y g están definidas en $\mathcal{R}^-$ y $\mathcal{R}^+$, respectivamente, y donde h es una función holomorfa arbitraria en la región combinada $\mathcal{R}$. Cualquiera de las expresiones mostradas antes puede ser utilizada para representar la misma hiperfunción. La propia hiperfunción se conocería matemáticamente como la *clase de equivalencia* de tales pares, «reducidos módulo»[11] las funciones holomorfas h definidas en $\mathcal{R}$. El lector puede recordar la noción de clase de equivalencia mencionada en el prefacio, en relación con la definición de una fracción. Esta es la misma idea general, y no menos confusa. El punto esencial aquí es que sumar h no afecta al «salto» entre f y g, pero h pue-

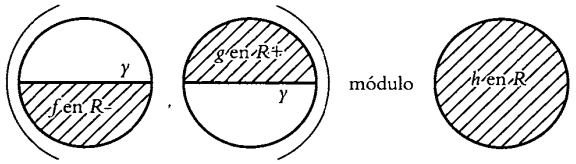


Fig. 9.14. Una hiperfunción, en un segmento γ del eje real, está dada por un par de funciones holomorfas (f, g), con f definida en una región abierta $\mathcal{R}^-$, que se extiende hacia abajo desde γ y g en una región abierta $\mathcal{R}^+$, que se extiende hacia arriba desde γ. La hiperfunción h, en γ, es (f, g) *módulo* cantidades $(f + h, g + h)$, donde h es holomorfa en la unión $\mathcal{R}$ de $\mathcal{R}^-$, γ y $\mathcal{R}^+$.

de cambiar f y g de formas que son irrelevantes para este salto. (Por ejemplo, h puede cambiar la forma de estas funciones lejos de γ en las regiones abiertas $\mathcal{R}^-$ y $\mathcal{R}^+$.) Así pues, el propio salto está impecablemente representado como esta clase de equivalencia.

El lector puede sentirse genuinamente contrariado porque esta pulcra definición parece depender crucialmente de nuestras elecciones arbitrarias de las regiones $\mathcal{R}^-$ y $\mathcal{R}^+$, solo restringidas por el hecho de estar unidas a lo largo de su frontera común γ. Lo extraordinario, sin embargo, es que la definición de hiperfunción *no* depende de esta elección. Según un sorprendente teorema, conocido como *teorema de extirpación*, esta noción de hiperfunción es en realidad completamente independiente de las elecciones particulares de $\mathcal{R}^-$ y $\mathcal{R}^+$; véanse los tres ejemplos superiores de la Fig. 9.15.

De hecho, el teorema de extirpación nos da más que esto. No exigimos que nuestra región abierta $\mathcal{R}$ quede dividida en dos (a saber, en $\mathcal{R}^-$ y $\mathcal{R}^+$) cuando eliminamos γ. Todo lo que necesitamos es que la región abierta $\mathcal{R}$, en el plano complejo, debe contener al segmento abierto[12] γ. Puede ser que $\mathcal{R} - \gamma$ (i.e., lo que queda de $\mathcal{R}$ cuando se elimina γ de ella[13]) consista en dos piezas separadas, igual que habíamos considerado hasta ahora, pero lo más general es que la eliminación de γ de $\mathcal{R}$ nos deje con una única región conexa, como se ilustra en los tres ejemplos inferiores de la Fig. 9.15. En estos casos, también debemos eliminar cualquier punto extremo interno a o b de γ, de modo que nos quedamos con un conjunto abierto, que llamaré $\mathcal{R} - \bar{\gamma}'$. En este caso más general, nuestras hiperfunciones se definen como funciones holo-

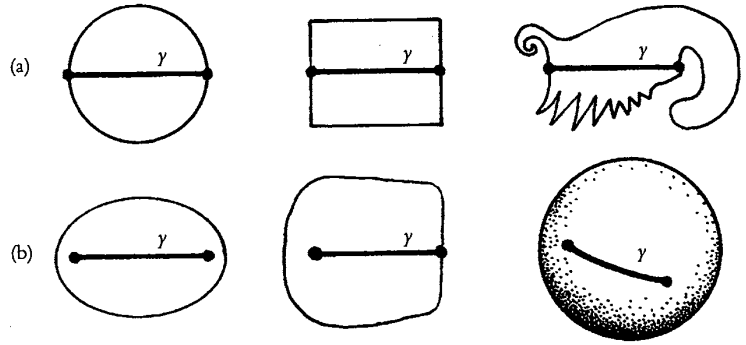


Fig. 9.15. Según el teorema de extirpación, la noción de una hiperfunción es independiente de la elección de región abierta $\mathcal{R}$, con tal de que $\mathcal{R}$ contenga a la curva dada γ. (a) La región $\mathcal{R} - \bar{\gamma}$ puede consistir en dos piezas separadas (de modo que obtenemos dos funciones holomorfas distintas f y g, como en la Fig. 9.14) o (b) la región $\mathcal{R} - \bar{\gamma}$ puede ser una pieza simplemente conexa, en cuyo caso f y g son simplemente dos partes de la misma función holomorfa.

morfas en $\mathcal{R}$, reducidas módulo funciones holomorfas en $\mathcal{R} - \bar{\gamma}'$. Es bastante notable que esta elección muy libérrima de $\mathcal{R}$ no suponga diferencia para la clase de «hiperfunciones» que se define con ello.[9.12] El caso en que tanto a como b yacen dentro de $\mathcal{R}$ es útil para integrales de hiperfunciones, puesto que entonces puede utilizarse un contorno cerrado en $\mathcal{R} - \bar{\gamma}$.

Todo esto se aplica también a nuestro caso anterior de un círculo en la esfera de Riemann. Aquí hay cierta ventaja en tomar $\mathcal{R}$ como la esfera de Riemann entera, porque entonces las funciones que tenemos que tomar «módulo por» son las funciones holomorfas que son globales en la esfera de Riemann entera, y hay un teorema según el cual dichas funciones son precisamente constantes. (Estas son realmente las «constantes» α_0 de las que hemos decidido no preocuparnos en §9.2.) Así pues, módulo constantes, una hiperfunción definida en un círculo en la esfera de Riemann está especificada simplemente por una fun-

[9.12] ¿Por qué las «funciones holomorfas en $\mathcal{R}$, reducidas módulo funciones holomorfas en $\mathcal{R} - \bar{\gamma}'$ se convierten en la definición de una hiperfunción que teníamos previamente, cuando $\mathcal{R} - \bar{\gamma}$ se separa en $\mathcal{R}^-$ y $\mathcal{R}^+$?

ción holomorfa en toda la región a un lado del círculo y otra función holomorfa al otro lado. Esto da la separación de una hiperfunción arbitraria en el círculo de forma unívoca (módulo constantes) en sus partes de frecuencia positiva/negativa.

Terminemos considerando algunas propiedades básicas de las hiperfunciones. Utilizaré la notación $\langle f, g\rangle$ para denotar la hiperfunción especificada por el par f y g definidas holomórficamente en $\mathcal{R}^-$ y $\mathcal{R}^+$, respectivamente (donde estoy volviendo al caso en que γ divide a $\mathcal{R}$ en $\mathcal{R}^-$ y $\mathcal{R}^+$. Así, si tenemos dos representaciones diferentes $\langle f, g\rangle$ y $\langle f_0, g_0\rangle$ de la misma hiperfunción, es decir, $\langle f, g\rangle = \langle f_0, g_0\rangle$, entonces $f - f_0$ y $g - g_0$ son ambas la *misma* función holomorfa h definida en $\mathcal{R}$, pero restringida a $\mathcal{R}^-$ y $\mathcal{R}^+$, respectivamente. Es sencillo, pues, expresar la *suma* de dos hiperfunciones, la *derivada* de una hiperfunción, y el *producto* de una hiperfunción por una función analítica q definida en γ:

$$\langle f, g\rangle + \langle f_1, g_1\rangle = \langle f + f_1, g + g_1\rangle,$$

$$\frac{\mathrm{d}\langle f, g\rangle}{\mathrm{d}z} = \left\langle \frac{\mathrm{d}f}{\mathrm{d}z}\, \frac{\mathrm{d}g}{\mathrm{d}z}\right\rangle,$$

$$q\langle f, g\rangle = \langle qf, qg\rangle,$$

donde en la última expresión, la función analítica q está extendida de manera holomorfa en un entorno de[14] γ.[9.13] Podemos representar la propia q como una hiperfunción por $q = \langle q, 0\rangle = \langle 0, -q\rangle$, pero no hay un producto general definido entre dos hiperfunciones. La carencia de un producto no es culpa de la aproximación de hiperfunción a las funciones generalizadas. Sucede con todas las aproximaciones.[15] El hecho de que la función delta de Dirac (mencionada en §6.6; véase también *infra*) no pueda ser elevada al cuadrado, por ejemplo, causa un sinfín de dificultades a muchos teóricos de los campos cuánticos.

Algunos ejemplos sencillos de representaciones hiperfuncionales, en el caso en que $\gamma = \mathbb{R}$, y $\mathcal{R}^-$ y $\mathcal{R}^+$ son los semiplanos complejos abiertos superior e inferior, la función escalón de Heaviside $\theta(x)$ y la función delta de Dirac(-Heaviside) $\delta(x) (= \mathrm{d}\theta(x)/\mathrm{d}x)$; (véanse §§6.1,6):

[9.13] Hay aquí una pequeña sutileza. Resuélvala. (*Sugerencia*: Piense cuidadosamente en los dominios de definición.)

$$\theta(x) = \left\langle \frac{1}{2\pi\mathrm{i}} \log z, \frac{1}{2\pi\mathrm{i}} \log z - 1\right\rangle,$$

$$\delta(x) = \left\langle \frac{1}{2\pi\mathrm{i}}, \frac{1}{2\pi\mathrm{i}}\right\rangle,$$

donde tomamos la rama de los logaritmos para la que $\log 1 = 0$. La integral de la hiperfunción $\langle f, g\rangle$ sobre toda la recta real puede expresarse como la integral de f a lo largo de un contorno inmediatamente por debajo de la recta real menos la integral de g a lo largo de un contorno inmediatamente por encima de la recta real (suponiendo que estas convergen), ambas de izquierda a derecha.[9.14] Nótese que la hiperfunción puede ser no trivial incluso cuando f y g son prolongaciones analíticas de la misma función.

¿Hasta qué punto son generales las hiperfunciones? Ciertamente, incluyen a todas las funciones analíticas. Incluyen también funciones discontinuas como $\theta(x)$ y la onda cuadrada (como muestran nuestras discusiones anteriores), u otras C^{-1}-funciones obtenidas sumando estos objetos. De hecho, todas las C^{-1}-funciones son ejemplos de hiperfunciones. Además, puesto que podemos diferenciar una hiperfunción para obtener otra hiperfunción, y cualquier C^{-2}-función puede obtenerse como la derivada de alguna C^{-1}-función, se sigue que todas las C^{-2}-funciones son también hiperfunciones. Hemos visto que esto incluye a la función delta de Dirac. Podemos derivar otra vez, y luego otra vez más. En realidad, cualquier C^{-n}-función es una hiperfunción, para cualquier n entero. ¿Qué pasa con las $C^{-\infty}$-funciones, conocidas como *distribuciones*? (Véase §6.6). Sí, también todas estas son hiperfunciones.

Normalmente se define una distribución como un elemento de lo que se denomina el espacio *dual* de las funciones C^∞-suaves. La noción de un «espacio dual» será discutida en §12.3 (y §13.6). De hecho, el dual (en un sentido apropiado) del espacio de las C^n-funciones es el espacio de las C^{-2-n}-funciones para cualquier entero n, y esto se aplica también a $n = \infty$ si escribimos $-2 - \infty = -\infty$ y $-2 + \infty = \infty$. Por consiguiente, las $C^{-\infty}$-funciones son duales de las C^∞-funciones. ¿Qué sucede con el dual ($C^{-\omega}$) de las C^ω-funciones? En realidad, con la definición

[9.14] Compruebe la propiedad estándar de la función delta por la que $\int q(x)\delta(x)\mathrm{d}x = q(0)$, en el caso en que $q(x)$ es analítica.

adecuada de «dual», ¡estas $C^{-\omega}$-funciones son precisamente las hiperfunciones!

Hemos cerrado el círculo. Al tratar de generalizar la noción de «función», alejándonos todo lo posible de la noción aparentemente muy restrictiva de función «analítica» u «holomorfa» —el tipo de función que hubiera hecho feliz a Euler—, hemos llegado a la noción extraordinariamente general y flexible de una *hiperfunción*. Pero las propias hiperfunciones están definidas, de una forma básicamente muy simple, en términos de estas mismas funciones holomorfas «eulerianas» que pensábamos que habíamos abandonado con reticencia. En mi opinión, este es uno de los logros mágicos supremos de los números complejos.[16] ¡Ojalá Euler hubiera vivido para apreciar este hecho maravilloso!

Notas

Sección 9.1

9.1. Aquí utilizo la letra griega χ («chi»), en lugar de una x ordinaria, que podría parecer más natural, solo porque necesitamos distinguir esta variable de la parte real x del número complejo z, que desempeñará un papel importante en lo que sigue.

9.2. No se requiere que $f(\chi)$ sea real para valores reales de χ, esto es, que a_n, b_n y c sean números reales. Es perfectamente legítimo tener funciones complejas de variables reales. La condición para que $f(\chi)$ sea real es que α_{-n} sea el complejo conjugado de α_n. Los complejos conjugados se examinarán en §10.1.

Sección 9.2

9.3. La anomalía notacional de apariencia extraña que supone el utilizar «F^-» para la parte de la serie con potencias positivas y «F^+» para la parte con potencias negativas, surge en definitiva de un convenio de signos desafortunado que ha llegado a hacerse casi universal en la literatura mecanocuántica (véanse §§21.2,3 y §24.3). Pido disculpas por esto, ¡pero no hay nada que pueda hacer de forma razonable!

9.4. Es un principio general que para cualquier C^ω-función f, definida en un dominio real $\mathcal{R}$, es posible «complejificar» $\mathcal{R}$ hasta un dominio complejo $\mathbb{C}\mathcal{R}$ ligeramente ampliado, llamado un «engrosamiento complejo» de $\mathcal{R}$, que contiene a $\mathcal{R}$ en su interior, tal que f se extiende unívocamente a una función holomorfa definida en $\mathbb{C}\mathcal{R}$.

Sección 9.3

9.5. Véase, por ejemplo, Bailey *et al.* (1982).

Sección 9.4

9.6. Por otra parte, es habitual imponer algún requisito de que $f(\chi)$ se comporte «razonablemente» cuando χ tiende a infinito positivo o negativo. Esto no tendrá un interés especial para nosotros aquí y, en cualquier caso, con el enfoque que estoy adoptando, los requisitos habituales serían innecesariamente restrictivos.

9.7. En mecánica cuántica hay también una cantidad constante $\hbar$ introducida para fijar adecuadamente el escalamiento de p con respecto a x (véanse §§21.2,11), pero por el momento estoy simplificando las cosas tomando $\hbar = 1$. De hecho, $\hbar$ es la forma de Dirac de la constante de Planck (i.e., $h/2\pi$, donde h es el «cuanto de acción» original de Planck). La elección $\hbar = 1$ puede hacerse siempre definiendo nuestras unidades básicas de la forma apropiada. Véase §27.10.

Sección 9.5

9.8. Véase Bailey *et al.* (1982).

Sección 9.7

9.9. Véase Sato (1958, 1959, 1960).

9.10. Véase también Bremermann (1965), aunque el término «hiperfunción» no se utiliza explícitamente en esta obra.

9.11. Este aspecto de la noción «módulo» será examinado en §16.1 (y se puede comparar con la nota 3.17).

9.12. Aquí «segmento abierto» se refiere simplemente al hecho de que los verdaderos puntos extremos a y b no están incluidos en γ, de modo que «que contiene» a γ no implica que a y b estén contenidos dentro de $\mathcal{R}$.

9.13. Esta «diferencia» entre conjuntos $\mathcal{R}, \gamma$ también se escribe habitualmente $\mathcal{R}\gamma$.

9.14. La definición técnica de «entorno de» es «que contiene un conjunto abierto».

9.15. Para la aproximación más estándar («distribución») a la idea de «función generalizada», véanse Schwartz (1966), Friedlander (1982), Gel'fand y

Shilov (1964), y Trèves (1967); para una propuesta alternativa, útil en contextos «no lineales», y que traslada el «problema de la no existencia del producto» a un problema de no unicidad, véanse Colombeau (1983, 1985) y Grosser *et al.* (2001).

9.16. Hay también interconexiones importantes entre hiperfunciones y la cohomología de haces holomorfa, que se discutirá en §33.9. Tales ideas desempeñan papeles importantes en la teoría de hiperfunciones en superficies de dimensiones más altas; véanse Sato (1959, 1960) y Harvey (1966).

10

Superficies

10.1. Dimensiones complejas y dimensiones reales

Uno de los logros más impresionantes de las matemáticas en los dos últimos siglos es el desarrollo de varias técnicas notables que pueden tratar con espacios no planos de varias dimensiones. Será importante para nuestros objetivos que transmita al lector algo de estas ideas, pues la física moderna depende de ellas de forma vital.

Hasta ahora hemos considerado espacios de una sola dimensión. El lector podría sentirse intrigado por este comentario, ya que el plano complejo, la esfera de Riemann y otras superficies de Riemann han sido protagonistas en varios de los capítulos anteriores. Sin embargo, en el contexto de las funciones holomorfas hay que pensar que estas superficies son, en esencia, de una sola dimensión, siendo esta una dimensión *compleja*, como se ha comentado en §8.2. Los puntos de un espacio semejante se distinguen (localmente) unos de otros por un único parámetro, aunque este parámetro resulta ser un número complejo. Así pues, hay que considerar estas «superficies» realmente como *curvas*, a saber, curvas *complejas*. Por supuesto, podríamos separar un número complejo z en sus partes real e imaginaria (x, y), donde $z = x + \mathrm{i}y$, y considerar x e y como dos parámetros reales independientes. Pero el proceso de separar de este modo un número complejo no es algo que pertenezca al ámbito de las operaciones holomorfas. Mientras estemos interesados solo en estructuras holomorfas, como lo hemos estado hasta ahora cuando considerábamos nuestros espacios complejos, debemos considerar que un único parámetro complejo proporciona una única dimensión. Esta, al menos, es la actitud mental que yo recomendaría adoptar.

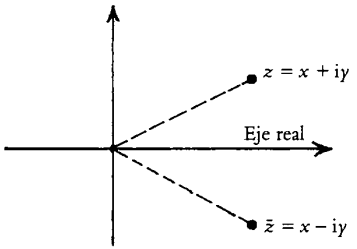


Fig. 10.1. El complejo conjugado de $z = x + iy$ (x, y reales) es $\bar{z} = x - iy$, obtenido como reflexión del z-plano en el eje real.

Por el contrario, se puede adoptar una posición opuesta, a saber, que las operaciones holomorfas constituyen meramente ejemplos particulares de operaciones más generales mediante las que x e y pueden separarse, si así lo deseamos, para ser consideradas como parámetros independientes. La forma adecuada de lograrlo es mediante la noción de *conjugación compleja*, que es una operación *no* holomorfa. El conjugado complejo del número $z = x + iy$, donde x e y son números reales, es el número complejo $\bar{z}$ dado por

$$\bar{z} = x - iy.$$

En el z-plano complejo, la operación de formar el conjugado complejo de un número complejo corresponde a una *reflexión* del plano respecto a la recta real (véase la Fig. 10.1). Recordemos de la exposición de §8.2 que las operaciones holomorfas conservan siempre la orientación del plano complejo. Si queremos considerar una aplicación conforme de (una parte del) plano complejo que invierte la orientación (tal como poner del revés el propio plano complejo), entonces necesitamos incluir la operación de conjugación compleja. Pero cuando se incluye junto a las otras operaciones estándar (sumar, multiplicar, tomar un límite), la conjugación compleja nos permite también generalizar nuestras aplicaciones, de modo que ya no necesitan ser conformes en absoluto. De hecho, cualquier aplicación de una porción del plano complejo en otra porción del plano complejo (mediante una transformación continua, pongamos por caso) puede lograrse añadiendo la operación de conjugación compleja a las otras operaciones.

Permítanme desarrollar este comentario. Podemos considerar que las funciones holomorfas son aquellas construidas a partir de las operaciones de suma y multiplicación, tal como se aplican a los números

complejos, junto con el procedimiento de tomar un límite (puesto que estas operaciones bastan para construir series de potencias, ya que una suma infinita es un límite de sumas parciales sucesivas).[10.1] Si incorporamos también la operación de conjugación compleja, entonces podemos generar funciones generales (digamos continuas) de x e y, ya que podemos expresar x e y individualmente por

$$x = \frac{z + \bar{z}}{2}, \quad y = \frac{z + \bar{z}}{2i}.$$

(Cualquier función continua de x e y puede construirse a partir de números reales mediante sumas, productos y límites.) Utilizaré la notación $F(z, \bar{z})$, con $\bar{z}$ explícitamente mencionado, cuando se esté considerando una función no holomorfa de z. Esto sirve para destacar el hecho de que en cuanto salimos del ámbito holomorfo, debemos considerar que nuestras funciones están definidas en un espacio 2-real-dimensional, en lugar de un espacio de una sola dimensión compleja. Podemos considerar nuestra función $F(z, \bar{z})$ expresada tambien en términos de las partes imaginarias x e y de z, y escribir esta función como $f(x, y)$. Entonces tenemos $f(x, y) = F(z, \bar{z})$, aunque, por supuesto, la forma funcional de f será en general completamente diferente de la de F. Por ejemplo, si $F(z, \bar{z}) = z^2 + \bar{z}^2$, entonces $f(x, y) = 2x^2 - 2y^2$. Como ejemplo adicional, podríamos considerar $F(z, \bar{z}) = z\bar{z}$; entonces $f(x, y) = x^2 + y^2$, que es el cuadrado del *módulo* $|z|$ de z, esto es,[10.2]

$$z\bar{z} = |z|^2$$

10.2. Suavidad, derivadas parciales

Puesto que al considerar funciones de más de una variable empezamos a aventurarnos en espacios de dimensiones más altas, aquí es necesario hacer algunos comentarios concernientes al «cálculo infinitesimal» en tales espacios. Como veremos explícitamente dentro de dos capítulos, los espacios —conocidos como *variedades*— pueden ser de cualquier dimensión n, donde n es un entero positivo. (Una variedad n-dimen-

[10.1] Explique por qué la resta y la división pueden construirse a partir de estas.
[10.2] Deduzca ambos resultados.

sional se suele conocer simplemente como una n-variedad.) La teoría de la relatividad general de Einstein utiliza una 4-variedad para describir el espaciotiempo, y muchas teorías modernas utilizan variedades de dimensiones aún más altas. Exploraremos las n-variedades generales en el capítulo 12, pero, por simplicidad, en este capítulo consideraremos solo la situación de una 2-variedad (o superficie) real $\mathcal{S}$. Entonces podemos utilizar las coordenadas locales x e y (reales) para etiquetar los diferentes puntos de $\mathcal{S}$ (en una región local de $\mathcal{S}$). De hecho, la discusión es muy representativa del caso general n-dimensional.

Una superficie 2-dimensional podría ser, por ejemplo, un plano ordinario o una esfera ordinaria. Pero la superficie no debe ser considerada como un «plano complejo» o una «esfera de Riemann» porque no estamos interesados en asignarle una estructura como espacio complejo (i.e., con la noción acompañante de «función holomorfa» definida en la superficie). Su estructura solo tiene que ser la de una *variedad suave*. Geométricamente, esto significa que no necesitamos seguir la pista de nada parecido a una estructura local conforme, como hacíamos con nuestras superficies de Riemann en §8.2, sino que necesitamos poder decir cuándo una función definida en el espacio (i.e., una función cuyo dominio es el espacio) debe considerarse «suave».

Para hacernos una idea intuitiva de lo que es una variedad «suave», pensemos en una esfera en contraposición a un cubo (donde, por supuesto, en cada caso me estoy refiriendo a la superficie y no al interior). Como ejemplo de una función suave sobre la esfera podríamos considerar una «función altura», por ejemplo la distancia por encima del plano ecuatorial (estando la esfera representada en un 3-espacio euclídeo ordinario en la forma normal, y contando negativamente las distancias por debajo del plano).Véase la Fig. 10.2a. Por el contrario, si nuestra función es el *módulo* de esta función altura (véanse §6.1 y la Fig 10.2b), de modo que las distancias por debajo del ecuador también cuentan positivamente, entonces esta función no es suave a lo largo del ecuador. Pero si consideramos el *cuadrado* de la función altura, entonces esta función *es* suave sobre la esfera (Fig. 10.2c). Es instructivo notar que en todos estos casos la función es suave en los polos norte y sur, pese a la apariencia «singular» que tienen en los polos las líneas de nivel de altura constante. El único caso de no suavidad ocurre en nuestro segundo ejemplo, en el ecuador.

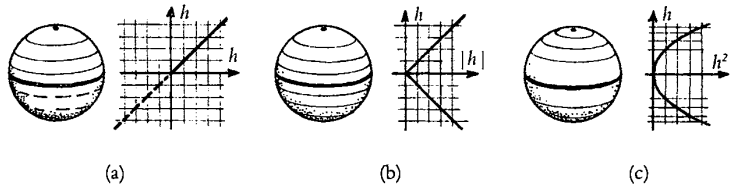


Fig. 10.2. Funciones en una esfera $\mathcal{S}$, representadas como si estuvieran en un 3-espacio euclídeo, donde h mide la distancia por encima del plano ecuatorial. (a) La propia función h es suave en $\mathcal{S}$ (valores negativos indicados por líneas discontinuas). (b) El módulo $|h|$ (véase la Fig. 6.2b) no es suave a lo largo del ecuador. (c) El cuadrado h^2 es suave sobre toda la $\mathcal{S}$.

Para entender lo que esto significa de manera un poco más precisa, introduzcamos un sistema de *coordenadas* en nuestra superficie $\mathcal{S}$. Estas coordenadas solo tienen que aplicarse localmente, y podemos imaginar $\mathcal{S}$ construida a partir de fragmentos locales —*cartas de coordenadas*— «pegados», de forma análoga a nuestro procedimiento para las esferas de Riemann en §8.1. (En el caso de la esfera, por ejemplo, necesitamos más de una carta.) Dentro de una carta, los diferentes puntos están etiquetados por coordenadas suaves; véase la Fig. 10.3. Nuestras coordenadas deben tomar valores reales, y las llamaremos x e y (sin que esto pretenda sugerir que deberían combinarse en forma de un número complejo). Supongamos ahora que tenemos una función suave Φ definida en $\mathcal{S}$. En terminología matemática moderna, Φ es una *aplicación* suave de $\mathcal{S}$ en el espacio de los números reales $\mathbb{R}$ (o de los números complejos $\mathbb{C}$, en el caso de que Φ sea una función de valor complejo sobre $\mathcal{S}$) porque Φ asigna a cada punto de $\mathcal{S}$ un número real (o com-

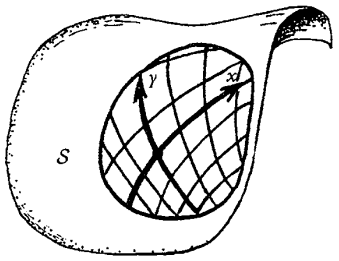


Fig. 10.3. Dentro de una carta local, coordenadas suaves (x, y) (con números reales) etiquetan los puntos.

plejo) i.e., Φ *aplica* $\mathcal{S}$ en los números reales (o complejos). Dicha función se denomina a veces un *campo escalar* sobre $\mathcal{S}$. En una carta de coordenadas particular, la cantidad Φ puede representarse como una función de las dos coordenadas, digamos

$$\Phi = f(x, y),$$

donde la *suavidad* de la cantidad Φ se expresa como la *diferenciabilidad* de la función $f(x, y)$.

No he explicado todavía lo que va a entenderse por «diferenciabilidad» para una función de más de una variable. Aunque intuitivamente clara, la definición precisa es un poco técnica para que entre aquí en ella con todo detalle.[1] En cualquier caso, haré algunos comentarios clarificadores.

En primer lugar, para que f sea diferenciable, como una función del par de variables (x, y), es necesario que si consideramos $f(x, y)$ en su calidad como función de la variable x, mientras y se mantiene en un valor constante, entonces esta función debe ser suave (al menos C^1) como función de x, en el sentido de las funciones de una *única* variable (véase §6.3); además, si consideramos $f(x, y)$ como una función de la variable y solamente, mientras que ahora es x la que se mantiene constante, entonces debe ser suave (C^1) como función de y. Sin embargo, esto no es en absoluto suficiente. Existen muchas funciones $f(x, y)$ que son suaves en x e y por separado, pero para las que sería irrazonable llamarlas suaves en el *par* (x, y).[10.3] Un requisito adicional suficiente para la suavidad es que cada una de las derivadas con respecto a x e y por separado sean funciones *continuas* del par (x, y). Enunciados similares (de particular relevancia para §4.3) serían válidos si consideramos funciones de más de dos variables. Utilizamos el símbolo ∂ de «derivada parcial» para denotar la diferenciación con respecto a una variable, manteniendo la otra (u otras) fija. Las derivadas parciales de $f(x, y)$ con respecto a x y con respecto a y, respectivamente, se escriben

$$\frac{\partial f}{\partial x} \text{ y } \frac{\partial f}{\partial y}.$$

(A modo de ejemplo, advertimos que si $f(x, y) = x^2 + xy^2 + y^3$, entonces $\partial f/\partial x = 2x + y^2$, $\partial f/\partial y = 2xy + 3y^2$.) Si estas cantidades existen y son continuas, entonces decimos que Φ es una función (C^1-)suave en la superficie.

También podemos considerar derivadas de órdenes superiores, denotando la derivada parcial *segunda* de f con respecto a x e y, respectivamente, por

$$\frac{\partial^2 f}{\partial x^2} \text{ y } \frac{\partial^2 f}{\partial y^2}.$$

(Ahora necesitamos C^2-suavidad, por supuesto.) Hay también una derivada segunda «cruzada» $\partial^2 f/\partial x \partial y$, que significa $\partial(\partial f/\partial x)/\partial y$, es decir, la derivada parcial con respecto a y de la derivada parcial de f con respecto a x. Podemos tomar esta derivada cruzada en el orden inverso para obtener la derivada $\partial^2 f/\partial y \partial x$. De hecho, una consecuencia de la diferenciabilidad (segunda) de f es que estas dos cantidades son iguales:[10.4]

$$\frac{\partial^2 f}{\partial x \partial y} = \frac{\partial^2 f}{\partial y \partial x}.$$

(La definición completa de C^2-suavidad, para una función de dos variables, requiere esto.)[10.5] Para derivadas más altas (y suavidades más altas), tenemos las correspondientes cantidades

$$\frac{\partial^3 f}{\partial x^3}, \quad \frac{\partial^3 f}{\partial x^2 \partial y} = \frac{\partial^3 f}{\partial x \partial y \partial x} = \frac{\partial^3 f}{\partial y \partial x^2}, \text{ etc.}$$

[10.3] Consideremos la función real $f(x, y) = xy(x^2 + y^2)^{-N}$, en los casos respectivos $N = 2$, 1 y 1/2. Demuestre que en cada caso la función es diferenciable (C^∞) con respecto a x, para cualquier valor fijo de y (y que lo mismo es válido cuando se invierten los papeles de x e y). De todas formas, f no es suave como función del par (x, y). Demuéstrelo en el caso $N = 2$ probando que la función no está siquiera acotada en el entorno del origen (0,0) (i.e., toma valores arbitrariamente grandes allí), en el caso $N = 1$ demostrando que la función aunque acotada, no es realmente continua como una función de (x, y), y en el caso $N = 1/2$ demostrando que aunque la función es ahora continua, no es suave a lo largo de la recta $x = y$. (*Sugerencia*: Examine los valores de cada función a lo largo de líneas rectas que pasan por el origen en el plano (x, y).) Algunos lectores pueden encontrar iluminador utilizar un adecuado programa de ordenador para representaciones gráficas 3-dimensionales, si dispone de uno, pero esto no es en absoluto necesario.

[10.4] Demuestre que las segundas derivadas cruzadas $\partial^2 f/\partial x \partial y$ y $\partial^2 f/\partial y \partial x$ son siempre iguales si $f(x, y)$ es un polinomio. (Un *polinomio* en x e y es una expresión formada a partir de x, y y constantes utilizando solamente la adición y la multiplicación.)

[10.5] Demuestre que las segundas derivadas cruzadas de la función $f = xy(x^2 - y^2)/(x^2 + y^2)$ son desiguales en el origen. Establezca directamente la falta de continuidad en sus segundas derivadas parciales en el origen.

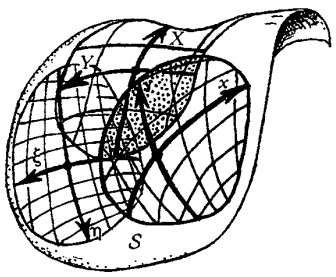


Fig. 10.4. Para cubrir la totalidad de $\mathcal{S}$, quizá tenemos que «pegar» varias cartas de coordenadas. Una función suave Φ en $\mathcal{S}$ tendría una expresión en coordenadas $\Phi = f(x, y)$ en una carta y $\Phi = F(X, Y)$ en otra (con coordenadas locales respectivas (x, y), (X, Y). En una región de solapamiento $f(x, y) = F(X, Y)$, donde X e Y son funciones suaves de x e y.

Una razón importante por la que aquí he tenido cuidado en distinguir f de Φ, utilizando letras diferentes (y quizá sea mucho menos «cuidadoso» sobre este aspecto más adelante) es que quizá queramos considerar una cantidad Φ, definida en la superficie, pero expresada con respecto a varios sistemas de coordenadas diferentes. La expresión matemática para la función $f(x, y)$ puede muy bien cambiar de una carta a otra, incluso si el valor de la cantidad Φ en cualquier punto concreto de la superficie «recubierta» por dichas cartas no cambia. Más concretamente, esto puede ocurrir cuando consideremos una región de solapamiento entre las diferentes cartas de coordenadas (véase la Fig. 10.4). Si un segundo conjunto de coordenadas se denota por (X, Y), entonces tenemos una nueva expresión

$$\Phi = F(X, Y),$$

para los valores de Φ en la nueva carta de coordenadas. En una región de solapamiento entre dos cartas, tendremos así

$$F(X, Y) = f(x, y).$$

Pero, como se ha indicado antes, la expresión particular que representa «F», en términos de cantidades X e Y, será en general muy diferente de la expresión que representa f en términos de x e y. En realidad, X podría ser una función complicada de x e y en la región de solapamiento, como también podría serlo Y, y estas funciones tendrían que ser incorporadas en el paso de f a F.[10.6] Tales funciones, que representan las coordenadas de un sistema en términos de las coordenadas del otro,

$$X = X(x, y) \quad \text{e} \quad Y = Y(x, y)$$

y sus inversas

$$x = x(X, Y) \quad \text{e} \quad y = y(X, Y)$$

son las denominadas *funciones de transición*, que expresan el cambio de coordenadas de una carta a la otra. Estas funciones de transición deben ser suaves —digamos, por simplicidad, C^∞-suaves— y esto tiene como consecuencia que la noción de «suavidad» para la cantidad Φ es independiente de la elección de coordenadas que se utilicen en un solapamiento de cartas.

[10.6] Encuentre explícitamente la forma de $F(X, Y)$ cuando $f(x, y) = x^3 - y^3$, donde $X = x - y$, $Y = xy$. *Sugerencia*: ¿Qué es $x^2 + xy + y^2$ en términos de X e Y?, ¿qué tiene que ver con f?

10.3. Campos vectoriales y 1-formas

Existe una noción de «derivada» de una función que es independiente de la elección de coordenadas. Una notación estándar para esto, tal como se aplica a la función Φ definida en $\mathcal{S}$, es $\mathrm{d}\Phi$, donde

$$\mathrm{d}\Phi = \frac{\partial f}{\partial x}\mathrm{d}x + \frac{\partial f}{\partial y}\mathrm{d}y.$$

Aquí empezamos a tropezar con algunas de las confusiones que hay en el tema, y necesitamos algún tiempo para acostumbrarnos. En primer lugar, una cantidad tal como «$\mathrm{d}\Phi$» o «$\mathrm{d}x$» tiende a ser considerada inicialmente como una cantidad «infinitesimalmente pequeña», que aparece al aplicar el procedimiento de paso al límite implícito en el cálculo cuando se formula la derivada «$\mathrm{d}y/\mathrm{d}x$» (véase §6.2). En algunas de las expresiones en §6.5 consideré también cosas como $\mathrm{d}(\log x) = \mathrm{d}x/x$. En esta etapa, estas expresiones se consideraban meramente formales,[2] tomando esta última expresión como una forma conveniente («multiplicar por $\mathrm{d}x$») de representar la expresión «más correcta» $\mathrm{d}(\log x)/\mathrm{d}x = 1/x$. Por el contrario, cuando escribo «$\mathrm{d}\Phi$» en la fórmula mostrada más arriba entiendo por ello cierto tipo de entidad geométrica que se denomina una *1-forma* (aunque este no es el tipo más general de 1-forma; véanse §10.4 y §12.6), y esto funciona también para $\mathrm{d}(\log x) =$

$= dx/x$). Una 1-forma no es un «infinitesimal»; tiene un tipo de interpretación algo diferente, un tipo de interpretación cuya importancia ha aumentado con los años, y a la que llegaré en un momento. Sin embargo, hay que destacar que, pese a este significativo cambio de interpretación de «d», las expresiones matemáticas formales (tales como las de §6.5) —siempre que no tratemos de dividir por cosas como dx— no son alteradas en absoluto.

Hay también otra potencial fuente de confusión en la fórmula mostrada antes, que surge del hecho de que he utilizado «Φ» en el miembro izquierdo y «f» en el derecho. He hecho esto sobre todo a causa de las advertencias sobre la distinción entre «Φ» y «f» que he planteado antes. La cantidad Φ es una función cuyo dominio es la variedad $\mathcal{S}$, mientras que el dominio de f es una región (abierta) en el plano (x, y) que remite a una carta de coordenadas particular. Si voy a aplicar la noción de «derivada parcial de f con respecto a x», entonces tengo que saber lo que significa «mantener constante la variable y restante». Por esta razón se utiliza f a la derecha, y no Φ, porque f «sabe» cuáles son las coordenadas x e y, mientras que Φ no lo sabe. Incluso así, hay una confusión en la fórmula que se ha mostrado, puesto que no se mencionan los argumentos de las funciones. La Φ de la izquierda se aplica a un punto particular p de la 2-variedad $\mathcal{S}$, mientras que f se aplica a los valores particulares (x, y) que el sistema de coordenadas asigna al punto p. Estrictamente hablando, esto tendría que hacerse explícito para que la expresión tuviera sentido. Sin embargo, es un fastidio tener que seguir diciendo este tipo de cosas, y sería mucho más conveniente poder escribir esta fórmula como

$$d\Phi = \frac{\partial \Phi}{\partial x} dx + \frac{\partial \Phi}{\partial y} dy,$$

o, en forma de operador «abstracto»

$$d = dx \frac{\partial}{\partial x} + dy \frac{\partial}{\partial y}.$$

En realidad, estoy tratando de dar sentido a estas cosas. Estas fórmulas son casos de algo que se conoce como la *regla de la cadena*. Tal como se establece, requiere que se asignen significados a cosas como «$\partial \Phi / dx$'», cuando Φ es una función definida en $\mathcal{S}$.

¿Cómo hay que considerar un operador, tal como $\partial/\partial x$, de modo

que pueda aplicarse a una función, tal como Φ, definida en la variedad $\mathcal{S}$, en lugar de aplicarse solo a una función de las variables x e y? Tratemos primero de ver qué significa $\partial/\partial x$ cuando referimos las cosas a algún otro sistema de coordenadas (X, Y). La «regla de la cadena» apropiada resulta ser ahora

$$\frac{\partial}{\partial x} = \frac{\partial X}{\partial x} \frac{\partial}{\partial X} + \frac{\partial Y}{\partial x} \frac{\partial}{\partial Y}.$$

Así pues, en términos del sistema (X, Y), tenemos ahora la expresión de apariencia más complicada $(\partial X/\partial x)\partial/\partial X + (\partial Y/\partial x)\partial/\partial Y$ para representar con exactitud la misma operación que en apariencia $\partial/\partial x$ representa en el sistema (x, y). Esta expresión más complicada es una cantidad ξ de la forma

$$\xi = A\frac{\partial}{\partial X} + B\frac{\partial}{\partial Y},$$

donde A y B son funciones (C^∞-)suaves de X e Y. En el caso particular que acabamos de ver, donde ξ representa $\partial/\partial x$ en el sistema (x, y), tenemos $A = \partial X/\partial x$ y $B = \partial Y/\partial x$. Pero podemos considerar cantidades ξ más generales para las que A y B no tienen estas formas particulares. Una cantidad semejante ξ se denomina un *campo vectorial* en $\mathcal{S}$ (en la (X, Y)-carta de coordenadas). Podemos reescribir ξ en el sistema (x, y) original, y encontramos que ξ tiene precisamente la misma forma general que en el sistema (X, Y):

$$\xi = a\frac{\partial}{\partial x} + b\frac{\partial}{\partial y}$$

(aunque las funciones a y b son en general completamente diferentes de A y B).[10.7] Esto nos permite extender el campo vectorial desde la (X, Y)-carta a una (x, y)-carta que se solapa con ella. De esta manera, podemos imaginar que el campo vectorial ξ se extiende a la totalidad de $\mathcal{S}$.

¡Es probable que todo esto haya causado gran confusión en el lector! Sin embargo, mi propósito no es confundir sino encontrar la forma analítica correcta de una noción geométrica muy básica. El operador diferencial ξ, que he denominado un «campo vectorial», con su

[10.7] Encuentre A y B en términos de a y b; por analogía, escriba a y b en términos de A y B.

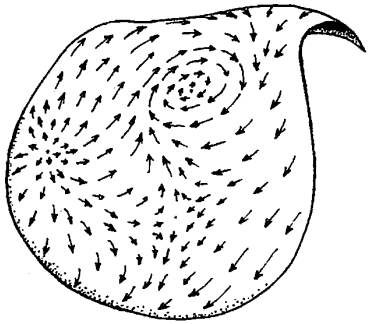


Fig. 10.5. La interpretación geométrica de un campo vectorial ξ como un «campo de flechas» dibujado en $\mathcal{S}$.

(consiguiente) forma muy específica de transformar, cuando pasamos de una carta a otra, tiene una clara interpretación geométrica, como se ilustra en la Fig. 10.5. Tenemos que visualizar ξ como un «campo de pequeñas flechas» dibujado en $\mathcal{S}$ (aunque en algunos lugares de $\mathcal{S}$, una flecha puede reducirse a un punto, siendo estos los lugares donde ξ toma el valor cero). (Para tener una buena imagen de un campo vectorial, pensemos en los mapas de vientos que se muestran en los programas de información meteorológica en la televisión.) Las flechas representan las direcciones en las que debe ser diferenciada la cantidad sobre la que ξ actúa. Considerando que esta cantidad es Φ, la acción de ξ sobre Φ, a saber $\xi(\Phi) = a\,\partial\Phi/\partial x + b\partial\Phi/\partial y$, mide la tasa de incremento de Φ en la dirección de las flechas; véase la Fig. 10.6. Además, la magnitud («longitud») de la flecha tiene importancia para determinar la «escala» en cuyos términos debe medirse este incremento. Una flecha más larga da una medida correspondientemente mayor de la tasa de incremento. Sería más apropiado pensar en todas las flechas como si fueran infinitesimales, conectando cada una de ellas un punto p de $\mathcal{S}$ (la «cola» de la flecha) con un punto «vecino» p' de $\mathcal{S}$ (la «cabeza» de la flecha). Para hacer esto un poco más explícito, escojamos un pequeño número positivo ϵ como una medida de la separación. Entonces la diferencia $\Phi(p') - \Phi(p)$, dividida por ϵ, nos da una aproximación a la magnitud $\xi(\Phi)$. Finalmente, en el límite en que p' se aproxima a p (de modo que $\epsilon \to 0$), obtenemos realmente $\xi(\Phi)$, que se suele denominar *gradiente* (o pendiente) de Φ en la dirección de ξ.

En el caso particular del campo vectorial $\partial/\partial x$, todas las flechas

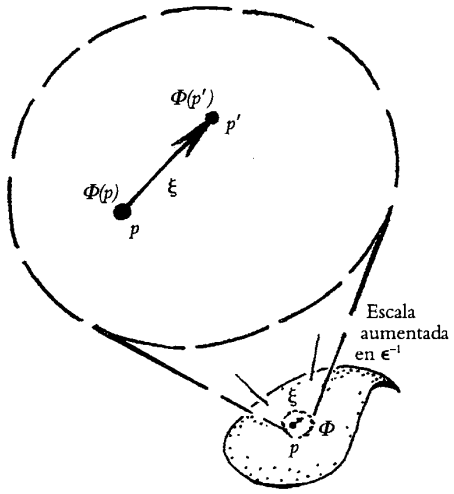


Fig. 10.6. La acción de ξ sobre un campo escalar Φ da su tasa de incremento a lo largo de las ξ-flechas. Considérense las flechas *infinitesimales*, cada una de ellas conectando un punto p de $\mathcal{S}$ («cola» de la flecha) con un punto «vecino» p' de $\mathcal{S}$ («cabeza» de la flecha), representada aplicando un gran aumento (por un factor ϵ^{-1}, donde ϵ es pequeño) al entorno de p. La diferencia $\Phi(p') - \Phi(p)$, dividida por ϵ, es (en el límite $\epsilon \to 0$) el gradiente $\xi(\Phi)$ de Φ a lo largo de ξ.

apuntan a lo largo de las líneas coordenadas de y constante. Esto ilustra un punto que frecuentemente lleva a confusión con la notación matemática estándar «$\partial/\partial x$» para la derivada parcial. Podríamos haber pensado que la expresión «$\partial/\partial x$» se refería más específicamente a la cantidad x. Sin embargo, en un sentido claro, tiene más que ver con la(s) variable(s) que no se menciona(n) explícitamente, aquí la variable y, que con x. La notación es particularmente traicionera cuando se considera un cambio de variables coordenadas, digamos de (x, y) a (X, Y), en el que una de las coordenadas sigue siendo la misma. Consideremos, por ejemplo, el cambio de coordenadas muy simple

$$X = x, \quad Y = y + x.$$

Entonces encontramos[10.8]

$$\frac{\partial}{\partial X} = \frac{\partial}{\partial x} - \frac{\partial}{\partial y}, \quad \frac{\partial}{\partial Y} = \frac{\partial}{\partial y}.$$

Así pues, vemos que $\partial/\partial X$ es diferente de $\partial/\partial x$, pese al hecho de que X es la misma que x —mientras que, en este caso, $\partial/\partial Y$ es la mis-

[10.8] Deduzca esto explícitamente. *Sugerencia*: Puede utilizar expresiones de «regla de la cadena» para $\partial/\partial X$ y $\partial/\partial Y$ que son las analogías exactas de la expresión para $\partial/\partial x$ que se ha mostrado antes.

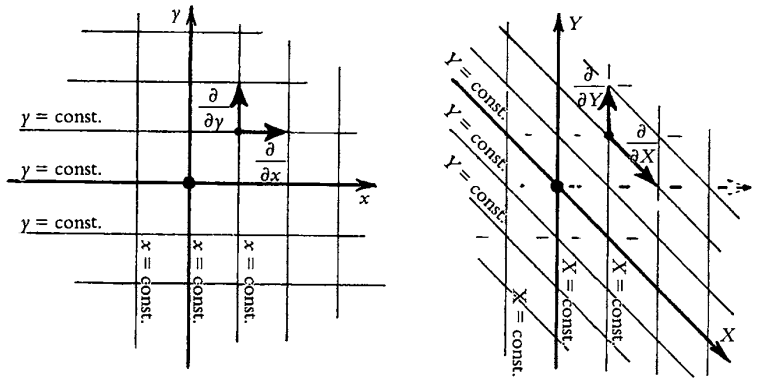


Fig. 10.7. Se ilustra la segunda confusión fundamental del cálculo: $\partial/\partial X \neq \partial/\partial x$ pese a que $X = x$, y $\partial/\partial Y = \partial/\partial y$ pese a que $Y \neq y$, para el cambio de coordenadas $X = x$, $Y = y + x$. La interpretación de los operadores de derivada parcial como «flechas» que apuntan a lo largo de líneas coordenadas clarifica la geometría (x = constante está de acuerdo con X = constante, pero y = constante está en desacuerdo con Y = constante).

ma que $\partial/\partial y$ incluso si Y difiere de y—. Este es un ejemplo de lo que mi colega Nick Woodhouse llama ¡«la segunda confusión fundamental del cálculo infinitesimal»![3] Por otra parte, *geométricamente* está claro por qué $\partial/\partial X \neq \partial/\partial x$, puesto que las «flechas» correspondientes apuntan a lo largo de líneas coordenadas diferentes (Fig. 10.7).

Ahora estamos en situación de interpretar la cantidad $d\Phi$. Se denomina *gradiente* (o derivada exterior) de Φ, y lleva la información de cómo varía Φ en todas las direcciones posibles en $\mathcal{S}$. Una buena forma geométrica de considerar $d\Phi$ es hacerlo en términos de un sistema de *curvas de nivel* en $\mathcal{S}$. Véase la Fig. 10.8a. Podemos considerar que $\mathcal{S}$ es similar a un mapa ordinario (donde por «mapa» entiendo esa cosa hecha de papel rígido que lleva usted cuando va de excursión, no la noción matemática de «mapa» o aplicación), que podría ser un globo terrestre si queremos tener en cuenta que $\mathcal{S}$ podría ser una variedad curva. La función Φ podría representar la altura del terreno sobre el nivel del mar. Entonces $d\Phi$ representa la pendiente del terreno con relación a la horizontal. Las curvas de nivel recorren los puntos de la misma altura.

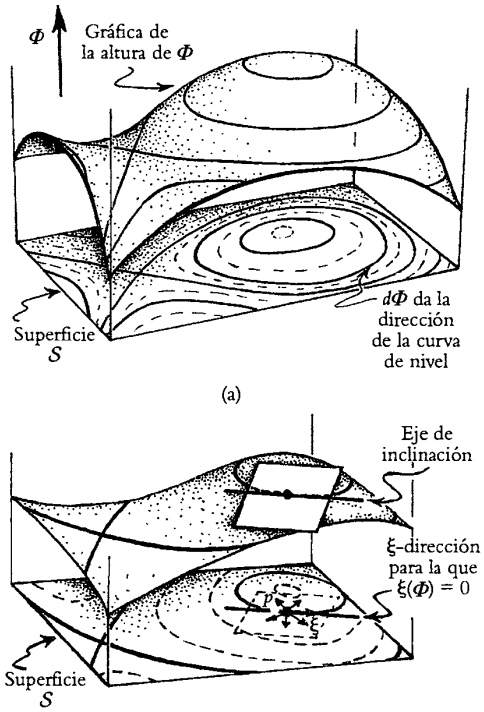


Fig. 10.8. Podemos representar geométricamente el pleno gradiente (derivada exterior) $d\Phi$ de un escalar Φ en términos de un sistema de líneas de nivel en $\mathcal{S}$. (a) El valor Φ se representa aquí verticalmente por encima de $\mathcal{S}$, de modo que las líneas de nivel en $\mathcal{S}$ (Φ constante) describen altura constante. (b) En cualquier punto p de $\mathcal{S}$, la dirección de la línea de nivel nos da la dirección a lo largo de la cual se anula el gradiente (el «eje de inclinación» de la pendiente de la colina), i.e., la dirección de las flechas ξ en p para la que $\xi(\Phi) = 0$. El corte a lo largo de líneas de nivel da un aumento o disminución en Φ, que mide el apiñamiento de las líneas en la dirección de ξ.

En cualquier punto p de $\mathcal{S}$, la dirección de la curva de nivel nos da la dirección a lo largo de la cual se anula el gradiente (el «eje de inclinación» de la pendiente del terreno), de modo que es la dirección de las flechas ξ en p para las que $\xi(\Phi) = 0$. Cuando seguimos una curva de nivel no ascendemos ni descendemos. Pero si cruzamos curvas de nivel, entonces habrá un aumento o disminución en Φ, y la tasa a la que esto ocurre, a saber $\xi(\Phi)$, vendrá medida por el apiñamiento de las líneas en la dirección en que las cruzamos. Véase la Fig. 10.8b.

10.4. Componentes, productos escalares

De acuerdo con la expresión

$$\xi = a\frac{\partial}{\partial x} + b\frac{\partial}{\partial y},$$

el campo vectorial ξ puede considerarse compuesto de dos partes, una de las cuales es proporcional a $\partial/\partial x$, que apunta a lo largo de las curvas de y constante, y la otra, proporcional a $\partial/\partial y$, que apunta a lo largo de las curvas de x constante. Así pues, en el sistema de coordenadas (x,y), el par de factores de ponderación respectivos (a, b) puede utilizarse para etiquetar ξ. Los números a y b se conocen como las *componentes* de ξ en este sistema de coordenadas; véase la Fig. 10.9. (Estrictamente hablando, las dos «componentes» de ξ serían realmente los dos campos vectoriales $a\,\partial/\partial x$ y $b\,\partial/\partial y$, de los que está compuesto el campo vectorial ξ, como se muestra en la Fig. 10.9 —y un comentario similar se aplicaría a las componentes de $d\Phi$—. Sin embargo, el término «componente» ha adquirido ahora este significado de «etiqueta coordenada» en buena parte de la literatura matemática, especialmente en relación con el cálculo tensorial; véase §12.8.)

Análogamente, la cantidad $d\Phi$ (una «1-forma») está compuesta de las dos partes dx y dy, de acuerdo con la expresión

$$d\Phi = u\,dx + v\,dy$$

y así puede utilizarse (u, v) para etiquetar $d\Phi$, y los números u y v son las *componentes* de $d\Phi$ en este mismo sistema de coordenadas. (De hecho, aquí tenemos $u = \partial\Phi/\partial x$ y $v = \partial\Phi/\partial y$.) La relación entre las componentes (u, v) de la 1-forma $d\Phi$ y las componentes (a, b) del campo vectorial ξ se obtiene a través de la cantidad $\xi(\Phi)$, que, como hemos visto antes, mide la tasa de incremento de Φ en la dirección de ξ. Encontramos[10.9] que el valor de $\xi(\Phi)$ viene dado por

$$\xi(\Phi) = au + bv.$$

Llamamos a $au + bv$ *producto escalar* (o producto *interno*) entre ξ, representado por (a, b), y $d\Phi$, representado por (u, v). Este producto escalar se escribirá a veces $d\Phi \bullet \xi$, si queremos expresarlo en abstracto sin referencia a ningún sistema de coordenadas concreto, y tenemos

$$d\Phi \bullet \xi = \xi(\Phi).$$

La razón de tener dos notaciones diferentes para la misma cosa es que la operación expresada en «$d\Phi \bullet \xi$» se aplica también a tipos de 1-formas más generales que aquellos que pueden expresarse como «$d\Phi$». Si $\boldsymbol{\eta}$ es una 1-forma semejante, entonces tiene un producto escalar con cualquier campo vectorial $\boldsymbol{\xi}$, que se escribe como «$\boldsymbol{\eta} \bullet \boldsymbol{\xi}$».

De hecho, la definición de una 1-forma consiste, en esencia, en que es una cantidad que puede combinarse con un campo vectorial para formar de esta manera un «producto escalar». Así pues, el hecho de que la cantidad $d\Phi$ forme de manera natural un producto escalar con un campo vectorial es realmente lo que la caracteriza como una «1-forma». (A veces a las 1-formas se las denomina *covectores*, dependiendo del contexto.) Técnicamente, las 1-formas (covectores) son objetos *duales* de los campos vectoriales en este sentido. Esta noción de un objeto «dual» será explorada con más detalle en §12.3, donde veremos que estas ideas se aplican también con toda generalidad a una «superficie» de dimensión más alta (i.e., a una n-variedad). El significado geométrico de una 1-forma se clarificará también en §12.3-5, en el contexto de dimensiones más altas. Por el momento, la propia familia de curvas de nivel servirá, pues dichas curvas representan las direccio-

[10.9] Demuéstrelo explícitamente, utilizando expresiones de «regla de la cadena» que hemos visto antes.

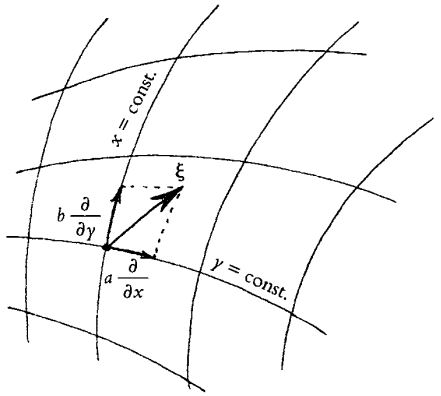


Fig. 10.9. Observamos que el vector $\xi = a\partial/\partial x + b\partial/\partial y$ puede considerarse compuesto de dos partes, una proporcional a $\partial/\partial x$, que apunta a lo largo de y = constante, y la otra, proporcional a $\partial/\partial y$, que apunta a lo largo de x = constante. El par de factores de peso respectivos (a, b) se denominan *componentes* de ξ en el sistema de coordenadas (x, y).

nes a lo largo de las cuales debe apuntar una ξ-flecha si $d\Phi \bullet \xi = 0$ (i.e., si $\xi(\Phi) = 0$).

10.5. Las ecuaciones de Cauchy-Riemann

Pero antes de dar este salto a dimensiones más altas, para el que nos prepararemos en el próximo capítulo, volvamos a la cuestión con la que hemos empezado este capítulo: qué propiedad es necesaria para que una superficie 2-dimensional pueda interpretarse como una 1-variedad compleja. En esencia, lo que necesitamos es un medio de caracterizar aquellas funciones de valor complejo Φ que son holomorfas. La condición de holomorficidad es local, de modo que podemos reconocerla como algo válido en cada carta, y compatible en los solapamientos entre cartas. En la (x, y)-carta exigimos que Φ sea holomorfa en el número complejo $z = x + iy$; en una (X, Y)-carta que se solapa, exigimos que sea holomorfa en $Z = X + iY$. La compatibilidad entre las dos está asegurada por el requisito de que Z sea una función holomorfa de z en el solapamiento, y viceversa. (Si Φ es holomorfa en z, y z es holomorfa en Z, entonces Φ debe ser holomorfa en Z, puesto que una función holomorfa de una función holomorfa es, asimismo, una función holomorfa.)[10.10]

Ahora bien, ¿cómo expresamos la condición de que Φ es holomorfa en z, en términos de las partes reales e imaginarias de Φ y z? Estas son las famosas ecuaciones de *Cauchy-Riemann* mencionadas en §7.1. Pero ¿cuáles son explícitamente estas ecuaciones? Podemos imaginar que Φ se expresa como una función de z y $\bar{z}$ (puesto que, como hemos visto al principio de este capítulo, las partes real e imaginaria de z, a saber, x e y, pueden reexpresarse en términos de z y $\bar{z}$ utilizando las expresiones $x = (z + \bar{z})/2$ e $y = (z - \bar{z})/2i$). Estamos obligados a expresar la condición de que, en efecto, Φ «depende solo de z» (i.e., que es «independiente de $\bar{z}$»).

[10.10] Explíquelo desde tres puntos de vista diferentes: (a) intuitivamente, a partir de principios generales (¿cómo podía aparecer $\bar{z}$?); (b) utilizando la geometría de las aplicaciones holomorfas descrita en §8.2, y (c) explícitamente, utilizando la regla de la cadena y las ecuaciones de Cauchy-Riemann a las que llegamos ahora.

¿Qué significa esto? Imaginemos que, en lugar del par de variables complejas conjugadas z y $\bar{z}$, tuviéramos un par de variables reales independientes u y v, pongamos por caso, y quisiéramos expresar el hecho de que cierta cantidad Ψ, que es una función de u y v, es de hecho independiente de v. Esta independencia puede establecerse en la forma

$$\frac{\partial \Psi}{\partial v} = 0$$

(porque esta ecuación nos dice que, para cada valor de u, la cantidad Ψ es constante en v; de modo que Ψ depende solo de u).[4] En consecuencia, el que Φ sea «independiente de $\bar{z}$» debería expresarse como

$$\frac{\partial \Phi}{\partial \bar{z}} = 0,$$

y esto expresa en realidad la holomorficidad de Φ (aunque el «argumento por analogía» que acabo de dar no debería tomarse como una demostración de este hecho).[5] Utilizando la regla de la cadena, podemos reexpresar esta ecuación[10.11] en términos de derivadas parciales en el sistema (x, y):

$$\frac{\partial \Phi}{\partial x} + i\frac{\partial \Phi}{\partial y} = 0.$$

Escribiendo Φ en términos de sus partes real e imaginaria

$$\Phi = \alpha + i\beta$$

(siendo α y β reales), obtenemos las *ecuaciones de Cauchy-Riemann*[6],[10.12]

$$\frac{\partial \alpha}{\partial x} = \frac{\partial \beta}{\partial y}, \quad \frac{\partial \alpha}{\partial y} = \frac{\partial \beta}{\partial x}.$$

Puesto que, como se ha señalado antes, en un solapamiento entre una (x, y)-carta y una (X, Y)-carta exigimos que $Z = X + iY$ sea holomorfa en $z = x + iy$, tenemos también las ecuaciones de Cauchy-Riemann válidas entre (x, y) y (X, Y):

$$\frac{\partial X}{\partial x} = \frac{\partial Y}{\partial y}, \quad \frac{\partial X}{\partial y} = -\frac{\partial Y}{\partial x}.$$

Si esta condición se satisface entre cualquier par de cartas, entonces hemos ensamblado una superficie de Riemann $\mathcal{S}$. (Estas son las

[10.11] Hágalo.

[10.12] Dé una deducción más directa de las ecuaciones de Cauchy-Riemann, a partir de la definición de derivada.

condiciones analíticas exigidas sobre las que he pasado por encima en §7.1.) Recordemos que una superficie semejante puede considerase también como una 1-variedad compleja. Pero, de acuerdo con la presente forma «Cauchy-Riemann» de considerar las cosas, pensamos en $\mathcal{S}$ como una 2-variedad real con un tipo concreto de estructura (a saber, la determinada por las ecuaciones de Cauchy-Riemann).

Mientras que hay cierta «pureza» en tratar de atenerse por completo a operaciones holomorfas (una perspectiva filosófica que tendrá importancia para nosotros más adelante, en el capítulo 33 y en §34.8) y en considerar $\mathcal{S}$ como una «curva», este punto de vista «Cauchy-Riemann» alternativo es potente en otros contextos. Por ejemplo, nos permite demostrar resultados apelando a muchas técnicas útiles en la *teoría de existencia* de ecuaciones en derivadas parciales. A continuación daré una idea de esto apelando a un (importante) ejemplo.

Si se satisfacen las ecuaciones de Cauchy-Riemann $\partial\alpha/\partial x = \partial\beta/\partial y$ y $\partial\alpha/\partial y = -\partial\beta/\partial x$, entonces cada una de las cantidades α y β por separado resulta satisfacer una ecuación particular (la ecuación de Laplace). En efecto, tenemos[10.13]

$$\nabla^2\alpha = 0, \quad \nabla^2\beta = 0,$$

donde el operador diferencial de segundo orden ∇^2, llamado el *laplaciano* (2-dimensional), está definido por

$$\nabla^2 = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}.$$

El laplaciano es importante en muchas situaciones físicas (véanse §21.2, §22.11, §24.3-6). Por ejemplo, si tenemos una película de jabón que llena un lazo de alambre y la película oscila ligeramente a uno y otro lado de un plano horizontal, entonces la altura de la película sobre la horizontal será una solución de la ecuación de Laplace (con una buena aproximación que se hace cada vez mejor cuanto más pequeña es la desviación vertical).[7] Véase la Fig.10.10. La ecuación de Laplace (en tres dimensiones) desempeña también un papel fundamental en la teoría newtoniana de la gravitación (y en electrostática; véanse los capítulos 17 y 19), puesto que es la ecuación que satisface una función

[10.13] Demuéstrelo.

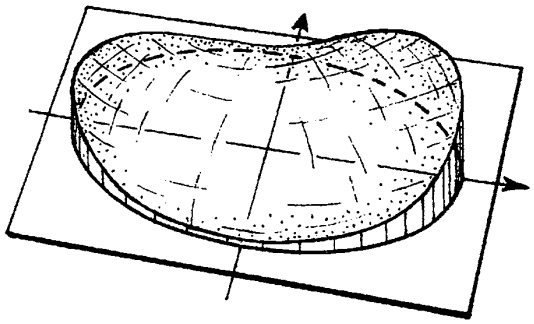


Fig. 10.10. Una película de jabón que llena un lazo de alambre que se desvía muy ligeramente hacia arriba y hacia abajo respecto a un plano horizontal. La altura de la película sobre la horizontal da una solución de la ecuación de Laplace (con una aproximación que se hace mejor cuanto menor es la desviación vertical).

potencial que determina el campo gravitatorio (o eléctrico) en el espacio libre.

Las soluciones de las ecuaciones de Cauchy-Riemann pueden obtenerse a partir de las soluciones de la ecuación de Laplace 2-dimensional de una forma bastante directa. Si tenemos una α que satisface $\nabla^2\alpha = 0$, podemos construir β mediante $\beta = \int(\partial\alpha/\partial x)\mathrm{d}y$; encontramos entonces que ambas ecuaciones de Cauchy-Riemann son satisfechas en consecuencia.[10.14] Este hecho puede utilizarse para demostrar e ilustrar algunas de las afirmaciones que se han hecho al final del capítulo precedente.

Consideremos, en particular, el hecho notable, enunciado al final de §9.7, de que cualquier función continua f, definida en el círculo unidad en el plano complejo, puede representarse como una hiperfunción. Esta afirmación establece efectivamente que cualquier f continua es la suma de dos partes, una de las cuales se extiende de manera holomorfa al interior del círculo unidad y la otra se extiende de manera holomorfa al exterior, donde ahora consideramos el plano complejo completado a la esfera de Riemann. Esta afirmación es equivalente, en efecto (según la discusión del §9.2), a la existencia de una representación de f como serie de Fourier, donde se considera que f es una fun-

[10.14] Demuéstrelo.

ción periódica de una variable real. Supongamos, por simplicidad, que f toma valores reales. (El caso complejo se sigue de este separando f en partes real e imaginaria.) Ahora bien, existen teoremas que nos dicen que podemos extender f de manera continua al interior del círculo, mientras que f satisface $\nabla^2 f = 0$ dentro del círculo. (Este hecho es muy plausible intuitivamente, por el argumento de la película de jabón dado antes; véase la Fig. 10.10. Reescalando f de manera apropiada a una nueva función ϵf, para un ϵ pequeño determinado, podemos imaginar que nuestro lazo de alambre yace en el círculo unidad en el plano complejo, desviándose ligeramente[8] de este en vertical, hacia arriba o hacia abajo, en los valores de ϵf sobre el círculo unidad. La altura de una película de jabón proporciona ϵf y por lo tanto f en el interior.) Mediante la prescripción anterior ($g = \int(\partial f/\partial x)\mathrm{d}y$), podemos suministrar a f una parte imaginaria g, de modo que $f + ig$ es holomorfa en todo el interior del círculo unidad. Este procedimiento también proporciona a f una parte imaginaria g *sobre* el círculo unidad, de modo que $f + ig$ es de frecuencia negativa. Ahora repetimos el procedimiento, aplicándolo al exterior del círculo unidad (que consideramos yace en la esfera de Riemann), y encontramos que $f - ig$ se extiende allí y es de frecuencia positiva. La separación $f = 1/2((f + ig) + 1/2(f - ig))$ consigue lo que se requiere.

Notas

Sección 10.2

10.1. Para una discusión detallada de la diferenciabilidad, en el caso de funciones de varias variables, véase Marsden y Tromba (1996).

Sección 10.3

10.2. Aunque la notación «dx» que introdujo originalmente Leibniz (a finales del siglo XVII) muestra gran potencia y flexibilidad, como se ilustra por el hecho de que cantidades como dx pueden ser tratadas como entidades algebraicas por sí mismas, esto no se extiende a su notación «d^2x» para las derivadas segundas. Si hubiera utilizado una modificación de esta notación en la que la segunda derivada de y con respecto a x se escribiera $(\mathrm{d}^2y - \mathrm{d}^2x\ \mathrm{d}y/\mathrm{d}x)/\mathrm{d}x^2$, entonces la cantidad «$\mathrm{d}^2x$» se comportaría de una manera algebraicamente consistente (donde «$\mathrm{d}x^2$» denota dxdx, etc.). Sin embargo, no está claro que esto hubiera sido muy práctico, debido a la complicación de esta expresión.

10.3. La «primera confusión fundamental» tiene que ver con la confusión entre el uso de f y de Φ que encontramos en §10.2, particularmente en relación con la toma de derivadas parciales. Véase Woodhouse (1991).

Sección 10.5

10.4. Debemos tomar esta condición solamente en un sentido local. Por ejemplo, podemos tener una función suave $\Phi(u, v)$ definida en una región con forma de riñón en el (u, v)-plano, dentro de la cual $\partial\Phi/\partial v = 0$, pero para la que Φ no sea completamente consistente como función de u.[10.15]

10.5. Aunque no es la ruta más rigurosa a las ecuaciones de Cauchy-Riemann, este argumento proporciona la *razón* subyacente para su forma.

10.6. De hecho, Jean LeRond D'Alembert encontró estas ecuaciones en 1752, mucho antes que Cauchy o Riemann (véase Struik, 1954, p. 219).

10.7. Resulta que la ecuación de la película de jabón real (a la que la ecuación de Laplace es una aproximación) tiene una notable solución general, encontrada por Weierstrass (1866), en términos de funciones holomorfas libres.

10.8. Puesto que f es continua en el círculo, debe ser acotada (i.e., sus valores están comprendidos entre un valor inferior determinado y un valor superior determinado). Esto se sigue de teoremas estándar, al ser el círculo un espacio compacto. (Véanse §21.6 para la noción de compacto, y Kahn, 1995, y Frankel, 2001.) Podemos entonces reescalar f (multiplicándola por una pequeña constante ϵ), de modo que las cotas superior e inferior sean ambas minúsculas. La analogía con la película de jabón proporciona entonces un argumento de plausibilidad razonable a favor de la existencia de ϵf extendida al interior del círculo, que satisface la ecuación de Laplace. Por supuesto, no es una demostración; véase Strauss (1992) o Brown y Churchill (2004) para una solución más rigurosa a este denominado «problema de Dirichlet para un disco».

[10.15] Explicar esto en el caso $\Phi(u, v) = \theta(u)h(v)$, donde las funciones θ y h están definidas como en §§6.1,3. La región en forma de riñón debe evitar el eje u no negativo.

11

Números hipercomplejos

11.1. El álgebra de los cuaterniones

¿Cómo se generaliza todo esto a dimensiones más altas? En el próximo capítulo describiré el procedimiento estándar (moderno) para estudiar n-variedades, pero, por diversas razones, será ilustrativo que antes familiarice al lector con algunas ideas más antiguas orientadas al estudio de dimensiones más altas. Estas ideas más antiguas han adquirido una relevancia directa para algunas actividades actuales en la física teórica.

La belleza y potencia del análisis complejo, tal como se muestra en la propiedad antes mencionada por la que las soluciones de la ecuación de Laplace 2-dimensional —una ecuación de importancia física considerable— pueden representarse de forma muy simple en términos de funciones holomorfas, condujo a los matemáticos del siglo XIX a buscar «números complejos generalizados» que pudieran aplicarse de un modo natural a un espacio 3-dimensional. Uno de los que reflexionó largo y tendido sobre este tema fue el reputado matemático irlandés William Rowan Hamilton (1805-1865). Finalmente, la respuesta le vino el 16 de octubre de 1843, mientras paseaba con su esposa por el Canal Real en Dublín, y quedó tan excitado por su descubrimiento que inmediatamente grabó sus ecuaciones fundamentales

$$\mathbf{i}^2 = \mathbf{j}^2 = \mathbf{k}^2 = \mathbf{ijk} = -1$$

en una piedra del Puente Brougham de Dublín.

Cada una de las tres cantidades $\mathbf{i}, \mathbf{j}$ y $\mathbf{k}$ es una «raíz cuadrada de -1» independiente (como la «i» única de los números complejos) y la combinación general

$$q = t + u\mathbf{i} + v\mathbf{j} + w\mathbf{k},$$

donde t, u, v y w son números reales, define el *cuaternión* general. Estas cantidades satisfacen todas las reglas usuales del álgebra salvo una. La excepción —y esta era la auténtica novedad[1] de las entidades de Hamilton— era la violación de la ley *conmutativa* de la multiplicación. En efecto, Hamilton encontró que[11.1]

$$\mathbf{ij} = -\mathbf{ji}, \quad \mathbf{jk} = -\mathbf{kj}, \quad \mathbf{ki} = -\mathbf{ik},$$

lo que constituye una flagrante violación de la ley conmutativa estándar $\boldsymbol{ab} = \boldsymbol{ba}$.

Los cuaterniones siguen satisfaciendo las leyes conmutativa y asociativa de la suma, la ley asociativa de la multiplicación, y las leyes distributivas de la multiplicación respecto a la suma,[11.2] a saber,

$$\begin{gathered} \boldsymbol{a} + \boldsymbol{b} = \boldsymbol{b} + \boldsymbol{a}, \\ \boldsymbol{a} + (\boldsymbol{b} + \boldsymbol{c}) = (\boldsymbol{a} + \boldsymbol{b}) + \boldsymbol{c}, \\ \boldsymbol{a}(\boldsymbol{bc}) = (\boldsymbol{ab})\boldsymbol{c}, \\ \boldsymbol{a}(\boldsymbol{b} + \boldsymbol{c}) = \boldsymbol{ab} + \boldsymbol{ac}, \\ (\boldsymbol{a} + \boldsymbol{b})\boldsymbol{c} = \boldsymbol{ab} + \boldsymbol{ac}, \end{gathered}$$

junto con la existencia de «elementos identidad» 0 y 1 para la suma y la multiplicación, tales que

$$\boldsymbol{a} + 0 = \boldsymbol{a}, \quad 1\boldsymbol{a} = \boldsymbol{a}1 = \boldsymbol{a}.$$

Estas relaciones, si excluimos la última, definen lo que los algebristas llaman un *anillo*. (En mi opinión, el término «anillo» es totalmente contraintuitivo —como lo son muchos términos en la disciplina del álgebra— y no tengo idea alguna de sus orígenes.) Si incluimos la última relación, obtenemos lo que se denomina un *anillo con identidad*.

Los cuaterniones proporcionan también un ejemplo de lo que se denomina un *espacio vectorial* sobre los números reales. En un espacio

[11.1] Demuéstrelas directamente a partir de las «ecuaciones del Puente Brougham» de Hamilton, suponiendo solo la ley asociativa $a(bc) = (ab)c$.

[11.2] Exprese la suma y el producto de dos cuaterniones generales de modo que se cumplan todas estas.

vectorial podemos sumar dos elementos (vectores),[2] $\boldsymbol{\xi}$ y $\boldsymbol{\eta}$, para formar su suma $\boldsymbol{\xi} + \boldsymbol{\eta}$, que está sujeta a conmutatividad y asociatividad

$$\boldsymbol{\xi} + \boldsymbol{\eta} = \boldsymbol{\eta} + \boldsymbol{\xi},$$
$$(\boldsymbol{\xi} + \boldsymbol{\eta}) + \boldsymbol{\zeta} = \boldsymbol{\xi} + (\boldsymbol{\eta} + \boldsymbol{\zeta}),$$

y podemos multiplicar vectores por «escalares» (en este caso, simplemente los números reales f, g), siendo válidas las siguientes propiedades asociativa, distributiva, etc.:

$$(f + g)\boldsymbol{\xi} = f\boldsymbol{\xi} + g\boldsymbol{\xi},$$
$$f(\boldsymbol{\xi} + \boldsymbol{\eta}) = f\boldsymbol{\xi} + f\boldsymbol{\eta},$$
$$f(g\boldsymbol{\xi}) = (fg)\boldsymbol{\xi},$$
$$1\boldsymbol{\xi} = \boldsymbol{\xi}.$$

Los cuaterniones forman un espacio vectorial *4-dimensional* sobre los reales, porque hay precisamente cuatro cantidades «base» independientes $1, \mathbf{i}, \mathbf{j}, \mathbf{k}$, que generan el espacio entero de los cuaterniones. Más adelante veremos otros muchos ejemplos de espacios vectoriales.

Los cuaterniones nos proporcionan también un ejemplo de lo que se denomina un *álgebra* sobre los números reales, debido a la existencia de una ley de multiplicación, tal como se ha descrito antes. Pero lo que es notable en los cuaterniones de Hamilton es que, además, tenemos una operación de *división* o, lo que es lo mismo, un *inverso* (multiplicativo) q^{-1} para cada cuaternión q distinto de cero. Este inverso satisface

$$q^{-1}q = qq^{-1} = 1,$$

lo que da a los cuaterniones la estructura de lo que se denomina un *anillo de división*, y se escribe explícitamente

$$q^{-1} = \bar{q}(q\bar{q})^{-1}$$

donde el *conjugado* (cuaterniónico) $\bar{q}$ de q se define como

$$\bar{q} = t - u\mathbf{i} - v\mathbf{j} - w\mathbf{k},$$

con $q = t + u\mathbf{i} + v\mathbf{j} + w\mathbf{k}$, como antes. Encontramos que

$$q\bar{q} = t^2 + u^2 + v^2 + w^2,$$

de modo que el *número real* $q\bar{q}$ no puede anularse, a menos que $q = 0$ (i.e., $t = u = v = w = 0$), de modo que $(q\bar{q})^{-1}$ existe, de donde se sigue que q^{-1} está bien definido siempre que $q \neq 0$.[11.3]

11.2. ¿Hay un papel físico para los cuaterniones?

Esto nos da una estructura algebraica muy bella y, aparentemente, nos abre un cálculo maravilloso ajustado a la perfección al tratamiento de la física y la geometría de nuestro espacio físico 3-dimensional. De hecho, el propio Hamilton dedicó los veintidós años restantes de su vida a desarrollar dicho cálculo. Sin embargo, cuando miramos retrospectivamente a los siglos XIX y XX desde nuestra perspectiva actual, debemos tener en cuenta que esos esfuerzos heroicos terminaron en un fracaso relativo. Esto no quiere decir que los cuaterniones no sean matemáticamente (o incluso físicamente) importantes. Lo cierto es que desempeñan un papel importante y en un sentido ligeramente indirecto su influencia ha sido enorme, a través de varias generalizaciones. Pero los «cuaterniones puros» originales no han llegado a hacer realidad lo que en un principio parecía ser una extraordinaria promesa.

¿Por qué no lo han hecho? ¿Se puede extraer alguna lección para los intentos modernos de encontrar las matemáticas «correctas» para el mundo físico? En primer lugar, hay un punto obvio. Si vamos a considerar los cuaterniones como algo análogo en dimensiones más altas a los números complejos, la analogía no se basa en que la dimensión ha subido de 2 a 3 dimensiones, sino en que ha pasado de 2 a 4; pues en cada caso una de las dimensiones es el «eje real», que aquí corresponde a la componente t en la representación anterior de q en términos de $\mathbf{i}, \mathbf{j}, \mathbf{k}$. Hay una fuerte tentación a considerar que esta t representa el *tiempo*,[3] de modo que nuestros cuaterniones describirían un *espaciotiempo* tetradimensional, en lugar de un solo espacio. Desde nuestra perspectiva del siglo XX, podríamos pensar que esto sería muy apropiado, puesto que un espaciotiempo tetradimensional es fundamental en la moderna teoría de la relatividad, como veremos en el capítulo 17. Pero

[11.3] Compruebe que esta definición de q^{-1} funciona realmente.

sucede que los cuaterniones no son apropiados para la descripción del espaciotiempo, básicamente por la razón de que la *forma cuadrática* «cuaterniónicamente natural» $q\bar{q} = t^2 + u^2 + v^2 + w^2$ tiene la «signatura incorrecta» para la teoría de la relatividad (una cuestión a la que llegaremos más adelante; véanse §13.8 y §18.1). Por supuesto, Hamilton no conocía la relatividad, puesto que vivió en el siglo equivocado para ello. En cualquier caso, aquí hay un cajón de problemas en los que no deseo involucrarme por el momento. ¡Lo abriré lentamente más tarde! (véanse §§13.8, §18.1-4, final de §22.11, §28.9, §31.13, §32.2).

Hay otra razón, quizá más fundamental, para que los cuaterniones no sean tan matemáticamente «bonitos» como parecen a primera vista. Son «magos» relativamente pobres; no soportan ninguna comparación con los números complejos a este respecto. La razón es que, al parecer, no hay ningún análogo cuaterniónico satisfactorio[4] de la noción de una función holomorfa. La razón básica para esto es sencilla. En el capítulo anterior hemos visto que una función holomorfa de una variable compleja z se caracteriza por ser holomórficamente «independiente» del conjugado complejo $\bar{z}$. Pero con los cuaterniones encontramos que es posible expresar *algebraicamente* el conjugado cuaterniónico $\bar{q}$ de q en términos de q y las cantidades constantes $\mathbf{i}, \mathbf{j}$ y $\mathbf{k}$ mediante el uso de la expresión[11.4]

$$\bar{q} = -\frac{1}{2}(q + \mathbf{i}q\mathbf{i} + \mathbf{j}q\mathbf{j} + \mathbf{k}q\mathbf{k}).$$

Si «cuaterniónica-holomorfa» tuviera que significar «construida a partir de cuaterniones por medio de suma, multiplicación y toma de límites», entonces $\bar{q}$ tendría que contar como una función cuaterniónica-holomorfa de q, lo que echa a perder la idea general.

¿Es posible encontrar modificaciones de los cuaterniones que pudieran tener más relevancia directa para el mundo físico? Encontraremos que sí lo es, pero todas ellas sacrifican la propiedad clave de los cuaterniones, antes mostrada, de que siempre se puede dividir por ellos (si son no nulos). ¿Qué pasa con las generalizaciones a dimensiones más altas? Enseguida veremos cómo lo consiguió Clifford, y qué importancia tiene para la física este tipo de generalización. Pero todos

[11.4] Compruébelo.

estos cambios llevan al abandono de la propiedad de álgebra de división.

¿Existen generalizaciones de los cuaterniones que conserven la propiedad de división? De hecho, sí existen; pero lo primero que hay que señalar es que existen teoremas que nos dicen que eso no es posible a menos que relajemos las reglas del álgebra incluso más allá de nuestro abandono de la ley conmutativa de la multiplicación. Aproximadamente dos meses después de recibir una carta de Hamilton anunciando el descubrimiento de los cuaterniones, en 1843, John Graves descubrió que existe un tipo de cuaternión «doble», entidades ahora conocidas como *octoniones*. Estos fueron redescubiertos por Arthur Cayley en 1845. En el caso de los octoniones, se abandona la ley asociativa $a(bc) = (ab)c$ —aunque se mantiene un vestigio de dicha ley en la forma $a(ab) = a^2b$ y $(ab)b = ab^2$—. La belleza de esta estructura reside en que sigue siendo un álgebra de división, aunque no asociativa. (Para cada a distinto de cero, existe un a^{-1} tal que $a^{-1}(ab) = b = (ba)a^{-1}$.) Los octoniones forman un álgebra de división no asociativa octodimensional. Existen siete análogos de los $\mathbf{i}, \mathbf{j}$ y $\mathbf{k}$, del álgebra de cuaterniones que, junto con 1, llenan las ocho dimensiones del álgebra de octoniones. Las leyes de multiplicación individual para estos elementos (análogas a $\mathbf{ij} = \mathbf{k} = -\mathbf{ji}$, etc.) son un poco complicadas y es mejor que las posponga hasta §16.2, donde se dará una elegante descripción, ilustrada en la Fig. 16.3. Por desgracia, no hay una generalización plenamente satisfactoria de los octoniones a dimensiones aún más altas si debe conservarse la propiedad de álgebra de división, como se sigue de un resultado algebraico de Hurwitz (1898) que demostraba que la identidad cuaterniónica (y octoniónica) «$q\bar{q}$ = suma de cuadrados» no funciona para dimensiones distintas de 1, 2, 4, 8. De hecho, aparte de estas dimensiones concretas, no puede haber *ningún* álgebra en la que la división (salvo por cero) sea siempre posible. Esto se sigue de un notable teorema topológico[5] que encontraremos en §15.4. Las únicas álgebras de división son, de hecho, los números reales, los números complejos, los cuaterniones y los octoniones.

Si estamos dispuestos a abandonar la propiedad de división, entonces *existe* una generalización importante de la noción de cuaterniones a dimensiones más altas, y es una generalización que tiene poderosas

implicaciones en la física moderna. Esta es la noción de *álgebra de Clifford* que fue introducida[6] en 1878 por el brillante (pese a su prematura muerte) matemático inglés William Kingdon Clifford (1845-1879). Se puede considerar que el álgebra de Clifford ha brotado de dos fuentes, cada una de las cuales estaba orientada a la comprensión de espacios de dimensiones mayores que las dos que están descritas directamente por los números complejos. Una de estas fuentes era el álgebra de cuaterniones de Hamilton en la que nos hemos interesado; la otra era un importante desarrollo anterior, propuesto originalmente[7] en 1844 y 1852 por un maestro de escuela alemán poco reconocido, Hermann Grassmann (1809-1877). Las álgebras de Grassmann también desempeñan un papel directo en la física teórica moderna. (En particular, la noción moderna de *supersimetría* —véase §31.3— depende crucialmente de ellas, al ser la supersimetría casi ubicua en los intentos modernos de desarrollar los fundamentos de la física más allá del marco de su modelo estándar.) Será importante que aquí nos familiaricemos con las álgebras de Grassmann y las de Clifford, y lo haremos en §11.6 y §11.5, respectivamente.

Las álgebras de Clifford (y Grassmann) incluyen un nuevo ingrediente que resulta de la dimensionalidad más alta del espacio en consideración. Antes de que podamos apreciar adecuadamente este punto, es mejor que consideremos de nuevo los cuaterniones, aunque desde una perspectiva algo diferente —una perspectiva geométrica—. Esto nos llevará también a algunas otras consideraciones que tienen una importancia fundamental en la física moderna.

11.3. Geometría de cuaterniones

Consideremos que las cantidades cuaterniónicas básicas **i**, **j**, **k** se refieren a tres ejes (dextrógiros) mutuamente perpendiculares en el 3-espacio euclídeo ordinario (véase la Fig. 11.1). Ahora recordemos de §5.1 que la cantidad i en la teoría de los números complejos ordinarios puede interpretarse en términos de la operación «multiplicar por i», que en su acción sobre el plano complejo significa «rotar un ángulo recto alrededor del origen en sentido positivo». Podríamos imaginar que es

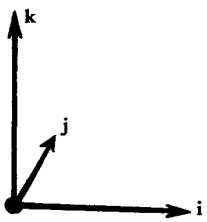


Fig. 11.1. Los cuaterniones básicos **i**, **j**, **k** remiten a tres ejes mutuamente perpendiculares (y dextrógiros) en el 3-espacio euclídeo ordinario.

posible interpretar el cuaternión **i** de un modo similar, aunque ahora como una rotación en 3 dimensiones, en sentido positivo (i.e., dextrógiro) alrededor del eje-**i** (de modo que el plano-**jk** desempeña el papel de plano complejo); en correspondencia, consideraríamos que **j** representa una rotación (en sentido positivo) alrededor del eje-**j** y **k** representa una rotación alrededor del eje-**k**. Sin embargo, si estas rotaciones son realmente rotaciones de un ángulo recto, como sucedía en el caso de los números complejos, entonces las relaciones producto no funcionarán, porque si hacemos la **i**-rotación seguida de la **j**-rotación, no obtenemos (ni siquiera un múltiplo de) la **k**-rotación.

Es muy fácil ver esto explícitamente tomando algún objeto ordinario y rotándolo físicamente. Sugiero hacerlo con un libro. Deje el libro apoyado sobre una mesa horizontal delante de usted de la forma normal, con el libro cerrado, como si estuviera a punto de abrirlo para leer. Imagine que el eje-**k** es vertical y pasa por el centro del libro, de modo que el eje-**i** sale hacia la derecha y el eje-**j** se aleja directamente de usted, pasando ambos también por el centro. Si rotamos el libro un ángulo recto (en sentido dextrógiro) alrededor de **i** y luego lo rotamos (en sentido dextrógiro) alrededor de **j**, encontramos que termina en una configuración (con su lomo hacia arriba) que no puede restituirse a su estado original mediante ninguna rotación simple alrededor de **k** (véase la Fig. 11.2).

Lo que tenemos que hacer para que las cosas funcionen es rotar *dos* ángulos rectos (i.e., 180° o π). Esto parece algo extraño, pues no hay una analogía directa con la forma en que hasta ahora entendíamos la acción del número complejo i. Parece que la dificultad principal reside en que, si aplicamos esta operación dos veces alrededor del mismo eje, obtenemos una rotación de 360° (o 2π), que simplemente devuel-

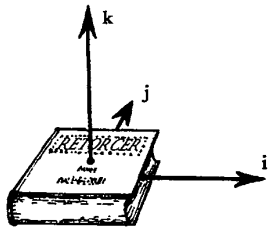


Fig. 11.2. Podemos considerar que los operadores cuaterniónicos **i**, **j**, **k** se refieren a rotaciones (de 180°, i.e., π) de un objeto, que aquí es un libro.

ve el objeto (en este caso nuestro libro) a su estado original, lo que aparentemente representa $\mathbf{i}^2 = 1$, en lugar de $\mathbf{i}^2 = -1$. Pero aquí interviene una nueva y maravillosa idea. Es una idea de una sutileza e importancia considerables, una importancia matemática que es fundamental en la física cuántica de las partículas elementales tales como electrones, protones y neutrones. Como veremos en §23.7, la materia sólida ordinaria no podría existir sin sus consecuencias. La noción matemática esencial es la de *espinor*.[8]

¿Qué es un espinor? En esencia, es un objeto que se transforma en su negativo cuando sufre una rotación completa de 2π. Esto puede parecer un absurdo, puesto que cualquier objeto clásico de la experiencia ordinaria vuelve siempre a su estado original tras una rotación semejante, y no a otra cosa. Para entender esta curiosa propiedad de los espinores —o de lo que llamaré *objetos espinoriales*—, volvamos a nuestro libro que reposa sobre la mesa ante nosotros. Necesitaremos algún medio para seguir la pista de cómo ha sido rotado. Podemos hacerlo colocando un extremo de un cinturón largo sujetado entre las páginas del libro y atando la hebilla rígidamente a alguna estructura fija (una pila de otros libros, por ejemplo; véase la Fig. 11.3a). Una rotación del libro en un ángulo de 2π retuerce el cinturón de una forma que no puede deshacerse sin una rotación posterior del libro (Fig. 11.3b). Pero si rotamos el libro un ángulo adicional de 2π, lo que da una rotación total de 4π, entonces encontramos, de forma bastante sorprendente, que la torsión del cinturón puede ser eliminada por completo haciéndolo pasar por encima del libro mientras este se mantiene en la misma posición durante la maniobra (Fig. 11.3c). Así pues, más que registrar el número total de rotaciones 2π que sufre el libro, lo que hace el cinturón es conservar la huella de la *paridad* de dicho número. Es decir, si rotamos el libro en un número par de rotaciones 2π, entonces podemos hacer que la torsión del cinturón desaparezca por completo, mientras que si rotamos el libro un número impar de rotaciones 2π, el cinturón queda inevitablemente retorcido. Esto se aplica cualquiera que sea el eje de rotación, o una sucesión de ejes de rotación diferentes, que decidamos utilizar.

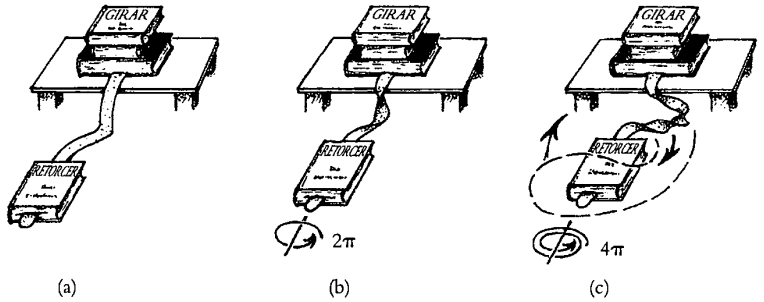


Fig. 11.3. Un objeto espinorial, representado por el libro de la Fig. 11.2. Un número par de rotaciones 2π es equivalente a ausencia de rotación, mientras que un número impar de rotaciones 2π no lo es. (a) Seguimos la pista de la paridad del número de rotaciones 2π del libro uniéndolo, mediante un cinturón largo, a un objeto fijo (aquí, a una pila de libros). (b) Una rotación de nuestro libro en 2π retuerce el cinturón de modo que no puede deshacerse sin una rotación adicional. (c) Una rotación del libro de 4π da una torsión que puede eliminarse por completo pasando el cinturón por encima del libro.

Así pues, para representar un objeto espinorial podemos pensar en un objeto ordinario en el espacio, pero en el que hay una imaginaria atadura flexible a cierta estructura externa, atadura que está representada por el cinturón que acabamos de considerar. La atadura puede moverse de una forma continua, pero sus extremos deben permanecer fijos, uno en el propio objeto y el otro en la estructura externa fija. Hay que pensar que la configuración de nuestro «libro espinorial», así imaginado, tiene una atadura imaginaria de este tipo a cierta estructura externa fija, y se estima que dos configuraciones son equivalentes solo si la atadura imaginaria de una puede deformarse de forma continua hasta coincidir con la atadura imaginaria de la otra. Para cada configuración ordinaria del libro habrá exactamente dos configuraciones

no equivalentes del libro espinorial, y decimos que una es la negativa de la otra.

Veamos ahora si esto nos proporciona las leyes de multiplicación correctas para los cuaterniones. Deje el libro sobre la mesa delante de usted, igual que antes, pero con el cinturón sujeto entre sus páginas. Rótelo, ahora, un ángulo π alrededor de $\mathbf{i}$ y hágalo seguir de una rotación de π alrededor de $\mathbf{j}$. Obtenemos una configuración que es equivalente a una rotación de π alrededor de $\mathbf{k}$, precisamente como debería ser de acuerdo con la $\mathbf{ij} = \mathbf{k}$ de Hamilton.

¿O no es así? Hay solo un pequeño motivo de irritación. Si insistimos cuidadosamente en que todas estas rotaciones sean en el sentido dextrógiro, siguiendo entonces apropiadamente la pista de las torsiones del cinturón, parece que lo que obtenemos en su lugar es $\mathbf{ij} = -\mathbf{k}$. Sin embargo, este no es un punto importante y hay diferentes maneras de corregirlo. Podemos o bien representar nuestros cuaterniones por rotaciones levógiras de 2π en lugar de dextrógiras (en cuyo caso recuperamos $\mathbf{ij} = \mathbf{k}$), o bien tomar nuestros ejes $\mathbf{i}, \mathbf{j}, \mathbf{k}$ con una orientación levógira en lugar de dextrógira. O, mejor, podemos adoptar un convenio sobre el orden de multiplicación de los operadores que es bastante habitual en matemáticas, a saber, que el «producto $\boldsymbol{pq}$» representa $\boldsymbol{q}$ seguido de $\boldsymbol{p}$, en lugar de $\boldsymbol{p}$ seguido de $\boldsymbol{q}$.

De hecho, existe una buena razón para este convenio aparentemente extraño. Esta tiene que ver con el hecho de que en general se entiende que los operadores —objetos tales como $\partial/\partial x$— actúan sobre los objetos que están escritos a su derecha. Así, el operador $\boldsymbol{P}$ actuando sobre $\boldsymbol{\Phi}$ se escribirá $\boldsymbol{P}(\boldsymbol{\Phi})$, o simplemente $\boldsymbol{P\Phi}$. En consecuencia, si aplicamos primero $\boldsymbol{P}$ y luego $\boldsymbol{Q}$ a $\boldsymbol{\Phi}$, obtenemos $\boldsymbol{Q}(\boldsymbol{P}(\boldsymbol{\Phi}))$ o simplemente $\boldsymbol{QP\Phi}$, que es $\boldsymbol{QP}$ actuando sobre $\boldsymbol{\Phi}$.

Mi forma de resolver esta complicada cuestión de signos con los cuaterniones consistirá en tomar todo en el sentido dextrógiro estándar y adoptar este convenio matemático «habitual» de orden-inverso para la ordenación de los operadores. Ahora al lector le resultará fácil confirmar que todas las «ecuaciones del Puente Brougham» de Hamilton $\mathbf{i}^2 = \mathbf{j}^2 = \mathbf{k}^2 = \mathbf{ijk} = -1$ son satisfechas por nuestro «libro espinorial». Por supuesto, tenemos en cuenta que $\mathbf{ijk}$ representa ahora a «$\mathbf{k}$ seguido de $\mathbf{j}$ seguido de $\mathbf{i}$».[9]

11.4. ¿Cómo componer rotaciones?

Esta curiosa propiedad de que los ángulos de rotación tengan que ser el doble de lo que hubiera parecido geométricamente apropiado, puede demostrarse de otra manera. Una característica especial de las rotaciones (propias, i.e., no reflexivas) en tres dimensiones es que si combinamos cualquier número de ellas, entonces obtenemos siempre una rotación alrededor de *algún* eje. ¿Cómo podemos encontrar dicho eje de una manera geométrica sencilla, así como el valor de dicha rotación? Hamilton encontró una respuesta elegante.[10] Veamos cómo funciona. Mi presentación será un poco diferente de la ofrecida inicialmente por Hamilton.

Recordemos que cuando componemos dos desplazamientos diferentes que son simples *traslaciones*, podemos utilizar la ley del triángulo estándar (equivalente a la ley del paralelogramo ilustrada en la Fig. 5.1a) para obtener la respuesta. Así pues, podemos representar la primera traslación por un *vector* (por lo que entiendo aquí un segmento de recta orientado, cuya orientación está indicada por una punta de flecha en un extremo del segmento) y la segunda traslación por otro vector semejante, cuya cola coincide con la punta del primero. El vector que va desde la cola del primer vector hasta la punta del segundo representa la composición de los dos movimientos traslacionales (véase la Fig 11.4a).

¿Podemos hacer algo similar para las rotaciones? Resulta, notablemente, que sí podemos. Pensemos ahora en los «vectores» como si fueran arcos orientados de círculos máximos dibujados en una esfera —dibujados una vez más con una punta de flecha en un extremo para representar la orientación—. (Un círculo máximo en una esfera es la intersección de la esfera con un plano que pasa por su centro.) Podemos imaginar que puede utilizarse un «arco vector» semejante para representar una rotación en la dirección de la flecha. Esta rotación va a ser alrededor de un eje que pasa por el centro de la esfera y es perpendicular al plano del círculo máximo en el que reside la flecha.

¿Podemos considerar que la composición de dos rotaciones representadas de este modo está dada por una «ley del triángulo» similar a la situación que teníamos para las traslaciones ordinarias? En efecto, podemos hacerlo, pero hay una trampa. La rotación que va a representarse por

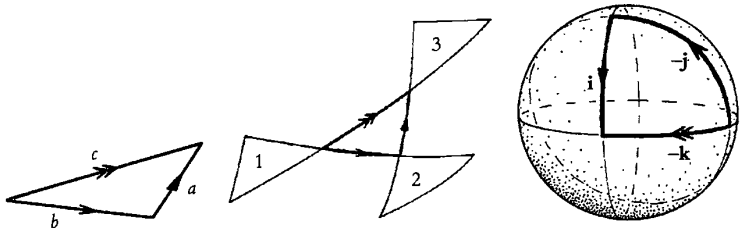


Fig. 11.4. (a) Traslaciones en el plano euclídeo representadas por segmentos de línea orientados. El segmento con flecha doble representa la composición de los otros dos, por la ley del triángulo. (b) Para rotaciones en el 3-espacio euclídeo, los segmentos son ahora arcos de círculo máximo trazados sobre la esfera unidad, cada uno de los cuales presenta una rotación del doble del ángulo medido por el arco (alrededor de un eje perpendicular a su plano). Para ver por qué funciona esto, reflejemos en cada vértice, por turno, el triángulo formado por los arcos. La primera rotación lleva el triángulo 1 al triángulo 2, la segunda lleva el triángulo 2 al triángulo 3, y la composición lleva el triángulo 1 al triángulo 3. (c) La relación cuaterniónica $\mathbf{ij} = \mathbf{k}$ (en la forma $\mathbf{i}(-\mathbf{j}) = -\mathbf{k}$), como caso especial. Cada rotación es de π, pero se representa por el semiángulo $\pi/2$.

nuestro «arco vector» debe ser de un ángulo que es exactamente el *doble* del ángulo representado por la longitud del arco. (Por conveniencia, podemos considerar que la esfera es de radio unidad. Entonces el ángulo representado por el arco es simplemente la distancia medida a lo largo del arco. Para que sea válida la «ley del triángulo», el ángulo de la rotación debe ser el doble que esta longitud de arco.) La razón de que esto funcione se ilustra en la Fig. 11.4b. El triángulo curvilíneo (esférico) en el centro ilustra la «ley del triángulo» y los tres triángulos externos son las respectivas reflexiones en sus tres vértices. Las dos rotaciones iniciales llevan uno de estos triángulos externos a un segundo triángulo y luego el segundo a un tercero; la rotación que es la composición de los dos lleva el primer triángulo al tercero. Notamos que cada una de estas rotaciones es de un ángulo que es precisamente el doble de la longitud del arco correspondiente en el triángulo curvilíneo original.[11.5] Veremos una variante de esta construcción en la física relativista, en §18.4 (Fig. 18.13).

[11.5] En la versión original de esta construcción, debida a Hamilton, se utiliza el triángulo esférico «dual» de este, cuyos vértices están en los puntos de corte de la esfera con los tres ejes de rotación involucrados en el problema. Dé una demostración directa de cómo funciona esto (quizá «dualizando» el argumento dado en el texto), representando las cantidades de las rotaciones como el *doble* de los ángulos de este triángulo dual.

Podemos examinar esto en la situación concreta que hemos considerado antes, y tratar de ilustrar la relación cuaterniónica $\mathbf{ij} = \mathbf{k}$. Cada una de las rotaciones descritas por $\mathbf{i}, \mathbf{j}$ y $\mathbf{k}$ es de un ángulo π. Así pues, utilizamos arcos de longitudes que son precisamente la mitad de este ángulo, a saber, $\frac{1}{2}\pi$, para mostrar la «ley del triángulo». Esta se ilustra por completo en la Fig. 11.4c (en la forma $\mathbf{i}(-\mathbf{j}) = -\mathbf{k}$, por claridad). También podemos ver la relación $\mathbf{i}^2 = -1$ ilustrada por el hecho de que un arco de círculo máximo, de longitud π, que se extiende desde un punto π de la esfera hasta su punto *antipodal* (que representa «–1»), es esencialmente diferente de un arco de longitud cero o de longitud 2π, pese al hecho de que cada uno de ellos representa una rotación de la esfera que la devuelve a su posición original. La descripción de «arco vector» representa correctamente las rotaciones de un «objeto espinorial».

11.5. Álgebras de Clifford

Para pasar a dimensiones más altas y a la idea de un álgebra de Clifford, debemos considerar cuál debe ser el análogo de una «rotación alrededor de un eje». En n dimensiones, tal rotación básica tiene un «eje» que es un espacio $(n-2)$-dimensional, en lugar del simple eje 1-dimensional que tenemos para las rotaciones 3-dimensionales ordinarias. Pero aparte de esto, una rotación alrededor de un eje $(n-2)$-dimensional se parece al caso familiar de una rotación 3-dimensional ordinaria alrededor de un eje 1-dimensional en el hecho de que la rotación está completamente determinada por la dirección del eje y por el valor del ángulo de rotación. Una vez más, tenemos objetos espinoriales con la propiedad de que, si un objeto semejante se rota de forma continua en un ángulo 2π, entonces no es devuelto a su estado original sino a lo que consideramos el «negativo» de dicho estado. Una rotación de 4π siempre devuelve al objeto a su estado original.

No obstante, hay un «nuevo ingrediente clave», que se ha mencionado antes, y es que en dimensiones mayores que 3 no es cierto que la composición de rotaciones básicas alrededor de ejes $(n-2)$-dimensionales será siempre otra rotación alrededor de un eje $(n-2)$-dimensional. En dimensiones más altas, las (composiciones de) rotaciones generales no pueden describirse de forma tan simple. Semejante rotación (generalizada) puede tener un «eje» (i.e., un espacio que queda invariante por el movimiento rotacional) cuya dimensión puede tomar varios valores diferentes. Así pues, para un álgebra de Clifford en n dimensiones, necesitamos una jerarquía de diferentes tipos de entidades que representen dichos tipos diferentes de rotaciones. De hecho, resulta que es mejor empezar con algo que es incluso más elemental que una rotación de π, a saber, una *reflexión* en un (hiper-)plano $(n-1)$-dimensional. Una composición de dos reflexiones semejantes (con respecto a dos de estos planos que son perpendiculares) proporciona una rotación de π, lo que da como entidades «secundarias» estas π-rotaciones antes básicas, mientras que ahora son las reflexiones las entidades primarias.[11.6]

Etiquetamos estas reflexiones básicas como $\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2, \boldsymbol{\gamma}_3, \ldots, \boldsymbol{\gamma}_n$, donde $\boldsymbol{\gamma}_r$ invierte el r-ésimo eje de coordenadas mientras conserva todos los demás. Para el tipo apropiado de «objeto espinorial», reflejarlo dos veces en la misma dirección da el negativo del objeto, de modo que tenemos n relaciones de tipo cuaternión

$$\boldsymbol{\gamma}_1^2 = -1, \quad \boldsymbol{\gamma}_2^2 = -1, \quad \boldsymbol{\gamma}_3^2 = -1, \quad \ldots, \boldsymbol{\gamma}_n^2 = -1,$$

satisfechas por estas reflexiones primarias. Las entidades secundarias, que representan nuestras π-rotaciones originales, son productos de pares de $\boldsymbol{\gamma}$'s distintas, y estos productos tienen propiedades de anticonmutación (de modo parecido a los cuaterniones)

$$\boldsymbol{\gamma}_p\boldsymbol{\gamma}_q = -\boldsymbol{\gamma}_q\boldsymbol{\gamma}_p \quad (p \neq q).$$

En el caso particular de tres dimensiones ($n = 3$), podemos definir las tres diferentes cantidades de «segundo orden»

$$\mathbf{i} = \boldsymbol{\gamma}_2\boldsymbol{\gamma}_3, \quad \mathbf{j} = \boldsymbol{\gamma}_3\boldsymbol{\gamma}_1, \quad \mathbf{k} = \boldsymbol{\gamma}_1\boldsymbol{\gamma}_2,$$

[11.6] Encuentre la naturaleza geométrica de la transformación, en el 3-espacio euclídeo, que es la composición de dos reflexiones en planos que no son perpendiculares.

y se comprueba fácilmente que estas tres cantidades $\mathbf{i}, \mathbf{j}$ y $\mathbf{k}$ satisfacen las leyes del álgebra cuaterniónica (las «ecuaciones del Puente Brougham» de Hamilton).[11.7]

El elemento general del álgebra de Clifford para un espacio n-dimensional es una suma de múltiplos reales (i.e., una combinación lineal) de productos de conjuntos de $\boldsymbol{\gamma}$'s distintas. Las entidades de primer orden («primarias») son las n diferentes cantidades individuales $\boldsymbol{\gamma}_p$. Las entidades de segundo orden («secundarias») son los $\frac{1}{2}n(n-1)$ productos independientes $\boldsymbol{\gamma}_p\boldsymbol{\gamma}_q$ (con $p < q$); hay $\frac{1}{6}n(n-1)(n-2)$ entidades independientes de tercer orden $\boldsymbol{\gamma}_p\boldsymbol{\gamma}_q\boldsymbol{\gamma}_r$ (con $p < q < r$), $\frac{1}{24}n(n-1)(n-2)(n-3)$ entidades independientes de cuarto orden, etc., y finalmente la única entidad de n-ésimo orden $\boldsymbol{\gamma}_1\boldsymbol{\gamma}_2\boldsymbol{\gamma}_3 \ldots \boldsymbol{\gamma}_n$. Tomando todas estas junto con la única entidad de orden cero 1, obtenemos

$$1 + n + \frac{1}{2}n(n-1) + \frac{1}{6}n(n-1)(n-2) + \ldots + 1 = 2^n$$

entidades en total,[11.8] y el elemento general del álgebra de Clifford es una combinación lineal de estas. Así pues, los elementos de un álgebra de Clifford constituyen un álgebra 2^n-dimensional sobre los reales, en el sentido descrito en §11.1. Forman un anillo con identidad, pero, a diferencia de los cuaterniones, no forman un anillo de división.

Una razón por la que las álgebras de Clifford son importantes es el papel que desempeñan en la definición de espinores. En física, los espinores hicieron su aparición en la famosa ecuación de Dirac para el electrón (Dirac 1928), pues el estado del electrón es una cantidad espinorial (véase el capítulo 24). Un espinor puede considerarse como un objeto sobre el que actúan como operadores los elementos del álgebra de Clifford, tal como sucede con las reflexiones o las rotaciones básicas del «objeto espinorial» que acabamos de considerar. La propia noción de un «objeto espinorial» es algo confusa y poco intuitiva, y algunos prefieren recurrir a una aproximación puramente (Clifford-)alge-

[11.7] Demuéstrelo.
[11.8] Explique este recuento. *Sugerencia*: Considere $(1 + 1)^n$.

braica[11] para su estudio. Esto tiene ciertamente sus ventajas, especialmente para una discusión general y rigurosa; pero tengo la sensación de que también es importante no perder de vista la geometría, y aquí he intentado resaltar este aspecto de las cosas.

En n dimensiones,[12] el espacio completo de espinores (a veces llamado *espacio de espín*) es $2^{n/2}$-dimensional, si n es par, y $2^{(n-1)/2}$-dimensional, si n es impar. Cuando n es par, el espacio de espinores se separa en dos espacios independientes (a veces llamados espacios de «espinores reducidos» o «semiespinores»), cada uno de los cuales es $2^{(n-1)/2}$-dimensional; esto es, cada elemento del espacio completo es la suma de dos elementos, uno de cada uno de los dos espacios reducidos. Una reflexión en el espacio n-dimensional (par) convierte a uno de estos espacios de espín reducidos en el otro. Los elementos de un espacio de espín reducido tienen cierta «quiralidad» o «mano»; los del otro espacio tienen la quiralidad opuesta. Esto parece tener una profunda importancia en física, donde aquí me refiero a los espinores para el espaciotiempo 4-dimensional ordinario. Cada uno de los dos espacios de espín reducidos es 2-dimensional: uno de ellos se refiere a entidades de mano derecha, y el otro a las de mano izquierda. Parece que la naturaleza asigna un papel diferente a cada uno de estos dos espacios de espín reducidos, y por medio de esto pueden emerger procesos físicos que no son invariantes por reflexión. De hecho, uno de los descubrimientos más sorprendentes (y algunos dirán «escandalosos») de la física del siglo XX (predicho teóricamente por Chen Ning Yang y Tsung Dao Lee y confirmado experimentalmente por Chien-Shiung Wu y su grupo, en 1957) fue el hecho de que existen realmente procesos fundamentales en la naturaleza que no ocurren en su forma especular. Volveré más adelante a estas cuestiones de fundamentos (§§25.3,4, §32.2, §§33.4,7,11,14).

Los espinores también tienen un valor matemático importante en varios contextos diferentes[13] (véanse §§22.8-11, §§23.3,4,5, §§24.6,7, §§32.3,4 y §§33.4,6,8,11), y pueden tener un uso práctico en ciertos tipos de computación. Debido a la relación «exponencial» entre la dimensión del espacio de espín ($2^{n/2}$, etc.) y la dimensión n del espacio original, no es sorprendente que los espinores sean mejores herramientas prácticas cuando n es razonablemente pequeño. En el caso del

espaciotiempo 4-dimensional ordinario, por ejemplo, cada uno de los espacios de espín reducidos tiene solamente dimensión 2, mientras que en la moderna «teoría-M» 11-dimensional (véase §31.14) el espacio de espín tiene 32 dimensiones.

11.6. Álgebras de Grassmann

Finalmente, permítanme volver al álgebra de Grassmann. Desde el punto de vista de la discusión anterior, podemos considerar el álgebra de Grassmann como una especie de caso degenerado de álgebra de Clifford, en el que tenemos elementos generadores básicos anticonmutativos $\boldsymbol{\eta}_1, \boldsymbol{\eta}_2, \boldsymbol{\eta}_3, \ldots, \boldsymbol{\eta}_n$, similares a los $\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2, \boldsymbol{\gamma}_3, \ldots, \boldsymbol{\gamma}_n$ del álgebra de Clifford, pero donde el cuadrado de cada $\boldsymbol{\eta}$ es *cero*, en lugar de ser el -1 que tenemos en el caso de Clifford:

$$\boldsymbol{\eta}_1^2 = 0, \quad \boldsymbol{\eta}_2^2 = 0, \ \ldots, \quad \boldsymbol{\eta}_n^2 = 0.$$

La ley anticonmutativa

$$\boldsymbol{\eta}_p\boldsymbol{\eta}_q = -\boldsymbol{\eta}_q\boldsymbol{\eta}_p$$

es válida como antes, excepto que el álgebra de Grassmann es ahora más «sistemática» que el álgebra de Clifford, porque no tenemos que especificar «$p \neq q$» en esta ecuación. El caso $\boldsymbol{\eta}_p\boldsymbol{\eta}_p = -\boldsymbol{\eta}_p\boldsymbol{\eta}_p$ simplemente reexpresa $\boldsymbol{\eta}_p^2 = 0$.

En realidad, las álgebras de Grassmann son más primitivas y universales que las álgebras de Clifford, en cuanto que dependen solo de una cantidad mínima de estructura local. Básicamente, la clave consiste en que el álgebra de Clifford necesita «saber» qué significa «perpendicular», de modo que puedan construirse rotaciones ordinarias a partir de reflexiones, mientras que la noción de «rotación» no es parte de lo que se describe según las álgebras de Grassmann. Para decirlo de otro modo, las nociones ordinarias de «álgebra de Clifford» y «espinor» requieren que haya una *métrica* en el espacio, mientras que esto no es necesario para el álgebra de Grassmann. (Las métricas se discutirán en §13.8 y §14.7.)

El álgebra de Grassmann está relacionada con la idea básica de un «elemento plano» para números diferentes de dimensiones. Considere-

mos que cada una de las cantidades básicas $\boldsymbol{\eta}_1, \boldsymbol{\eta}_2, \boldsymbol{\eta}_3, \ldots, \boldsymbol{\eta}_n$, define un *elemento de línea* o «vector» (más que un hiperplano de reflexión) en el origen de coordenadas en cierto espacio n-dimensional, estando cada $\boldsymbol{\eta}$ asociada a uno de los n diferentes ejes de coordenadas. (Estos pueden ser ejes «oblicuos», puesto que el álgebra de Grassmann no está interesada en la ortogonalidad; véase la Fig. 11.5.) El vector general en el origen será cierta combinación

$$\boldsymbol{a} = a_1\boldsymbol{\eta}_1 + a_2\boldsymbol{\eta}_2 + \ldots + a_n\boldsymbol{\eta}_n,$$

donde $a_1, a_2, \ldots, a_n$ son números reales. (Alternativamente, los a_i podrían ser números complejos, en el caso de un espacio complejo; pero los casos real e imaginario tienen un tratamiento algebraico similar.) Para describir el elemento plano 2-dimensional generado por dos de estos vectores $\boldsymbol{a}$ y $\boldsymbol{b}$, donde

$$\boldsymbol{b} = b_1\boldsymbol{\eta}_1 + b_2\boldsymbol{\eta}_2 + \ldots + b_n\boldsymbol{\eta}_n,$$

formamos el *producto de Grassmann* de $\boldsymbol{a}$ por $\boldsymbol{b}$. Para evitar confusiones con otras formas de producto, a partir de ahora adoptaré la notación (estándar) $\boldsymbol{a} \wedge \boldsymbol{b}$ para este producto (llamado «producto cuña») en lugar de utilizar la simple yuxtaposición de símbolos. En consecuencia, lo que he escrito previamente como $\boldsymbol{\eta}_p\boldsymbol{\eta}_q$, lo denotaré ahora por $\boldsymbol{\eta}_p \wedge \boldsymbol{\eta}_q$. La ley de anticonmutación de estas $\boldsymbol{\eta}$'s se escribe ahora

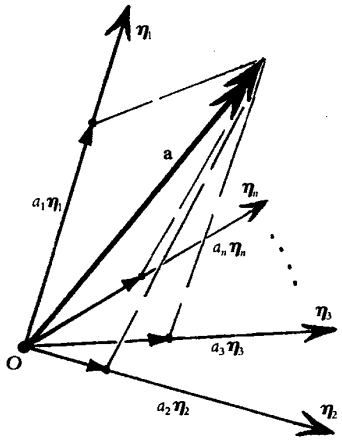


Fig. 11.5. Cada elemento básico $\boldsymbol{\eta}_1, \boldsymbol{\eta}_2, \boldsymbol{\eta}_3, \ldots, \boldsymbol{\eta}_n$, de un álgebra de Grassmann define un vector en un espacio n-dimensional, en un punto-origen **O**. Estos vectores pueden estar a lo largo de los diferentes ejes de coordenadas (que pueden ser «ejes» oblicuos; el álgebra de Grassmann no se interesa en la ortogonalidad). Un vector general en **O** es una combinación lineal $\mathbf{a} = a_1\boldsymbol{\eta}_1 + a_2\boldsymbol{\eta}_2 + \ldots + a_n\boldsymbol{\eta}_n$.

$$\boldsymbol{\eta}_p \wedge \boldsymbol{\eta}_q = -\boldsymbol{\eta}_q \wedge \boldsymbol{\eta}_p.$$

Adoptando la ley distributiva (véase §11.1) en la definición del producto $\boldsymbol{a} \wedge \boldsymbol{b}$, obtenemos consiguientemente la propiedad de anticonmutación más general[11.9]

$$\boldsymbol{a} \wedge \boldsymbol{b} = -\boldsymbol{b} \wedge \boldsymbol{a},$$

para vectores arbitrarios $\boldsymbol{a}, \boldsymbol{b}$. La cantidad $\boldsymbol{a} \wedge \boldsymbol{b}$ proporciona una representación algebraica del elemento plano generado por los vectores $\boldsymbol{a}$ y $\boldsymbol{b}$ (Fig. 11.6a). Nótese que esto contiene la información no solo de una orientación para el elemento plano (puesto que el signo de $\boldsymbol{a} \wedge \boldsymbol{b}$ tiene que ver con cuál de $\boldsymbol{a}$ o $\boldsymbol{b}$ viene antes), sino también de una «magnitud» asignada al elemento plano.

Podemos preguntarnos cómo va a representarse una cantidad tal como $\boldsymbol{a} \wedge \boldsymbol{b}$ como un conjunto de *componentes*, en correspondencia con el modo en que $\boldsymbol{a}$ puede representarse como $(a_1, a_2, a_3, \ldots, a_n)$ y $\boldsymbol{b}$ como $(b_1, b_2, b_3, \ldots, b_n)$, siendo estos los coeficientes que aparecen cuando $\boldsymbol{a}$ y $\boldsymbol{b}$ se presentan respectivamente como combinaciones lineales de $\boldsymbol{\eta}_1, \boldsymbol{\eta}_2, \boldsymbol{\eta}_3, \ldots, \boldsymbol{\eta}_n$. En correspondencia, la cantidad $\boldsymbol{a} \wedge \boldsymbol{b}$ puede presentarse como una combinación lineal de $\boldsymbol{\eta}_1 \wedge \boldsymbol{\eta}_2$, $\boldsymbol{\eta}_1 \wedge \boldsymbol{\eta}_3$, etc., y necesitamos saber la forma de los coeficientes que aparecen. Aquí hay implícita cierta elección de convenio porque, por ejemplo, $\boldsymbol{\eta}_1 \wedge \boldsymbol{\eta}_2$ y $\boldsymbol{\eta}_2 \wedge \boldsymbol{\eta}_1$ no son independientes (pues uno es el negativo del otro), de modo que podríamos querer distinguir uno de otro. Lo más sistemático es incluir ambos términos y dividir el coeficiente relevante entre ellos por igual. Entonces encontramos[11.10] que los coeficientes —i.e., las *componentes*— de $\boldsymbol{a} \wedge \boldsymbol{b}$ son las diversas cantidades $a_{[p}b_{q]}$, donde los paréntesis cuadrados alrededor de los índices denotan «antisimetrización», definida por

$$A_{[pq]} = \frac{1}{2}(A_{pq} - A_{qp}),$$

de donde

$$a_{[p}b_{q]} = \frac{1}{2}(a_pb_q - a_qb_p).$$

[11.9] Demuéstrelo.
[11.10] Escriba $\boldsymbol{a} \wedge \boldsymbol{b}$ completamente en el caso $n = 2$ para ver cómo resulta.

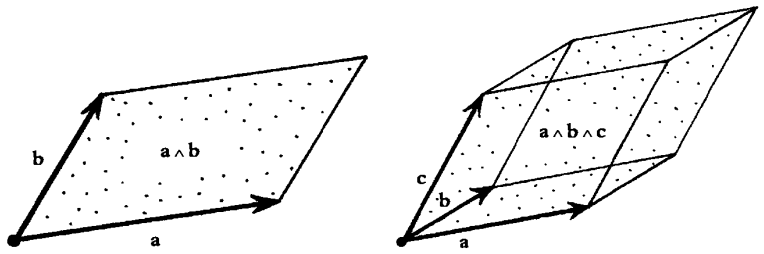


Fig. 11.6. (a) La cantidad $\mathbf{a} \wedge \mathbf{b}$ representa el elemento plano (orientado y escalado) generado por vectores independientes **a** y **b**. (b) El triple producto de Grassmann $\mathbf{a} \wedge \mathbf{b} \wedge \mathbf{c}$ representa el 3-elemento generado por vectores independientes **a**, **b** y **c**.

¿Qué pasa con un «elemento plano» 3-dimensional? Tomando tres vectores independientes $\boldsymbol{a}$, $\boldsymbol{b}$ y $\boldsymbol{c}$ que generan este 3-elemento, podemos formar el *triple* producto de Grassmann $\boldsymbol{a} \wedge \boldsymbol{b} \wedge \boldsymbol{c}$ como su representación (de nuevo con orientación y magnitud), y encontramos las propiedades de anticonmutación

$$\boldsymbol{a} \wedge \boldsymbol{b} \wedge \boldsymbol{c} = \boldsymbol{b} \wedge \boldsymbol{c} \wedge \boldsymbol{a} = \boldsymbol{c} \wedge \boldsymbol{a} \wedge \boldsymbol{b} = -\boldsymbol{b} \wedge \boldsymbol{a} \wedge \boldsymbol{c} = -\boldsymbol{a} \wedge \boldsymbol{c} \wedge \boldsymbol{b} = -\boldsymbol{c} \wedge \boldsymbol{b} \wedge \boldsymbol{a}$$

(véase la Fig. 11.6b). Las *componentes* de $\boldsymbol{a} \wedge \boldsymbol{b} \wedge \boldsymbol{c}$ se toman, de acuerdo con lo anterior,

$$a_{[p}b_qc_{r]} = \frac{1}{6}\,(a_pb_qc_r + a_qb_rc_p + a_rb_pc_q - a_qb_pc_r - a_pb_rc_q - a_rb_qc_p),$$

donde los paréntesis cuadrados denotan de nuevo antisimetrización, como ilustra la expresión del miembro derecho.

Expresiones similares definen r-elementos generales, donde r llega hasta la dimensión n del espacio entero. Las componentes del producto cuña de r-ésimo orden se obtienen tomando el producto antisimetrizado de las componentes de los vectores individuales.[11.11],[11.12] En realidad, un álgebra de Grassmann proporciona un poderoso medio de describir los elementos geométricos básicos de dimensión (finita) arbitraria.

El álgebra de Grassmann es un álgebra *graduada* en el sentido de que contiene elementos de r-ésimo orden (donde r es el número de $\boldsymbol{\eta}$'s que intervienen en el «producto cuña» dentro de la expresión). El número r (donde $r = 0, 1, 2, 3, \ldots, n$) se denomina *grado* del elemento del álgebra de Grassmann. Sin embargo, habría que advertir que el elemento general del álgebra de grado r no tiene por qué ser un único producto cuña (tal como $\boldsymbol{a} \wedge \boldsymbol{b} \wedge \boldsymbol{c}$ en el caso $r = 3$), sino que puede ser una suma de tales expresiones. En consecuencia, hay muchos elementos del álgebra de Grassmann que no describen directamente r-elementos geométricos. La importancia de tales elementos de Grassmann «no geométricos» aparecerá más adelante (§12.7).

En general, si $\boldsymbol{P}$ es un elemento de grado p y $\mathbf{Q}$ es un elemento de grado q, definimos su producto cuña $\boldsymbol{P} \wedge \mathbf{Q}$ de grado $p + q$ con componentes $P_{[a\ldots c}Q_{d\ldots f]}$, donde $P_{a\ldots c}$ y $Q_{d\ldots f}$ son las componentes de $\boldsymbol{P}$ y $\mathbf{Q}$ respectivamente. Entonces encontramos[11.13],[11.14]

$$\boldsymbol{P} \wedge \mathbf{Q} = \begin{cases} +\mathbf{Q} \wedge \boldsymbol{P} & \text{si } p \text{ o } q, \text{ o ambos, son pares,} \\ -\mathbf{Q} \wedge \boldsymbol{P} & \text{si } p \text{ y } q \text{ son ambos impares.} \end{cases}$$

La suma de elementos de un grado fijo r es una vez más un elemento de grado r; también podemos sumar elementos de grados diferentes para obtener una cantidad «mixta» que no tiene un grado concreto. Tales elementos del álgebra de Grassmann no tienen, sin embargo, estas interpretaciones directas.

[11.11] Escriba esta expresión explícitamente en el caso de un producto cuña de cuatro vectores.

[11.12] Demuestre que el producto cuña queda inalterado si $\boldsymbol{a}$ se reemplaza por $\boldsymbol{a}$ sumado a cualquier múltiplo de cualquiera de los otros vectores implicados en el producto cuña.

Notas

Sección 11.1

11.1. Según Eduard y Klein (1954), parece que ya Carl Friedrich Gauss había señalado la ley de multiplicación para cuaterniones alrededor de

[11.13] Demuéstrelo.

[11.14] Deduzca que $\boldsymbol{P} \wedge \boldsymbol{P} = 0$, si p es impar.

1820, pero no lo había publicado (Gauss, 1900). No obstante, esto es discutido por Tait (1900) y Knott (1900). Para más información, véase Crowe (1967).

11.2. El término «vector» tiene un amplio espectro de significados. Aquí no exigimos ninguna asociación con la noción de *diferenciación* de un «campo vectorial», descrita en §10.3.

Sección 11.2

11.3. No tengo claro hasta qué punto el propio Hamilton pudo ceder seriamente a esta tentación. Antes de su descubrimiento de los cuaterniones, se había interesado en el tratamiento algebraico del «paso del tiempo», y esto pudo tener alguna influencia sobre su predisposición a aceptar una cuarta dimensión en el álgebra cuaterniónica. Véase Crowe (1967), pp. 23-27.

11.4. De todas formas, se ha dedicado una buena cantidad de trabajo a la cuestión de los análogos cuaterniónicos de las nociones holomorfas y su valor en la teoría física. Véanse Gürsey (1983) y Adler (1995). Se podrían considerar las expresiones twistoriales (§§33.8,9) para resolver las ecuaciones de campos libres sin masa como un análogo 4-dimensional apropiado del método de solución mediante funciones holomorfas de la ecuación de Laplace. Sin embargo, este utiliza un análisis complejo, no cuaterniónico. Para una referencia general sobre cuaterniones y octoniones, véase Conway y Smith (2003).

11.5. Véase Adams y Atiyah (1966).

11.6. Véase Clifford (1878). Para referencias modernas, véanse Hestenes y Sobczyk (2001) y Lounesto (1999).

11.7. Véanse Grassmann (1844, 1862); Van der Waerden (1985), pp. 191-192, y Crowe (1967), cap. 1.

Sección 11.3

11.8. Lo pronunciamos como «espinor». Véase la nota 24.12.

11.9. Aunque no sé quién sugirió por primera vez esta forma de demostrar la multiplicación cuaterniónica, J. H. Conway la utilizó en demostraciones privadas en el Congreso Internacional de Matemáticos en Helsinki en 1978; véanse también Newman (1942), y Penrose y Rindler (1984), pp. 41-46.

Sección 11.4

11.10. Véase Pars (1968).

Sección 11.5

11.11. Para una aproximación a muchos problemas físicos a través del álgebra de Clifford, véase Lasenby *et al.* (2000), y las referencias allí contenidas.

11.12. Véanse Cartan (1966), Brauer y Weyl (1935), Penrose y Rindler (1986), apéndice, Harvey (1990), Budinich y Trautman (1988).

11.13. Véanse Lounesto (1999), Cartan (1966), Crumeyrolle (1990), Chevalley (1954) y Kamberov (2002), para algunos ejemplos.

12

Variedades de n dimensiones

12.1. ¿Por qué estudiar variedades de dimensiones más altas?

Vayamos ahora al procedimiento general para construir variedades de dimensiones más altas, donde la dimensión n puede ser cualquier entero positivo (o incluso cero, si nos permitimos considerar que un punto constituye una 0-variedad). Esta es una noción esencial para prácticamente todas las teorías modernas de la física fundamental. El lector podría preguntarse qué interés físico puede tener considerar n-variedades para las que n es mayor que 4, puesto que el espaciotiempo ordinario solo tiene cuatro dimensiones. De hecho, muchas teorías modernas, tales como la teoría de cuerdas, operan dentro de un «espaciotiempo» cuya dimensión es mucho mayor que 4. Más adelante (§15.1, §§31.4,10-12, 14-17) llegaremos a cuestiones de este tipo, en las que examinaremos la plausibilidad física de esta idea general. Pero con total independencia de la cuestión de si el «espaciotiempo» real podría ser descrito adecuadamente como una n-variedad, existen otras razones poderosas y muy diferentes para considerar n-variedades generales en física.

Por ejemplo, el *espacio de configuración* de un cuerpo rígido ordinario en el 3-espacio euclídeo —por el que entiendo un espacio $\mathcal{C}$ cuyos diferentes puntos representan las diferentes localizaciones físicas del cuerpo— es una 6-variedad no euclídea (véase la Fig. 12.1). ¿Por qué de seis dimensiones? Hay tres dimensiones (grados de libertad) en la posición del centro de gravedad y tres más en la orientación rotacional del cuerpo.[12.1] ¿Por qué no euclídea? Hay muchas razones, pero una

[12.1] Explique más explícitamente esta cuenta de dimensiones.

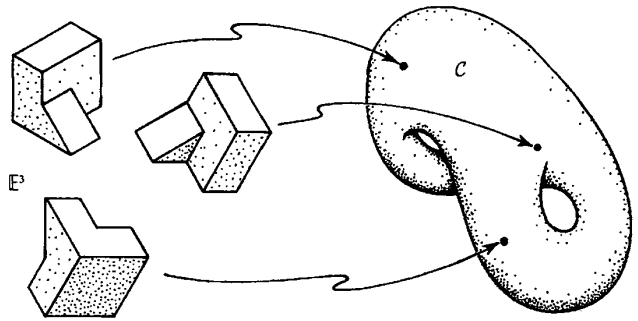


Fig. 12.1. Espacio de configuración $\mathcal{C}$, cada uno de cuyos puntos representa una localización posible de un cuerpo rígido dado en el 3-espacio euclídeo $\mathbb{E}^3$: $\mathcal{C}$ es una 6-variedad no euclídea.

particularmente llamativa es que incluso su *topología* es diferente de la del 6-espacio euclídeo. Esta «no trivialidad topológica» de $\mathcal{C}$ se manifiesta simplemente en el aspecto 3-dimensional del espacio que se refiere a la orientación *rotacional* del cuerpo. Llamemos $\mathcal{R}$ a este 3-espacio, de modo que cada punto de $\mathcal{R}$ representa una orientación rotacional particular del cuerpo. Recordemos nuestra exposición de las rotaciones de un libro en el capítulo anterior. Tomaremos este libro como nuestro «objeto» (que, por supuesto, debe permanecer cerrado, pues, de otra forma, el espacio de configuración tendría muchas más dimensiones correspondientes al movimiento de las páginas).

¿Cómo vamos a reconocer la «no trivialidad topológica»? Podemos suponer que esta no es una cuestión fácil para una variedad tridimensional. Sin embargo, hay varios procedimientos matemáticos para establecerlo. Recordemos que en nuestro examen de las superficies de Riemann, dado en §8.4 (véase la Fig. 8.9), hemos considerado varios tipos de 2-superficies topológicamente no triviales. Aparte de la esfera (de Riemann), la más simple de tales superficies es el toro (superficie de género 1). ¿Cómo podemos distinguir el toro de la esfera? Una forma de hacerlo es considerar lazos cerrados en la superficie. Es intuitivamente obvio que en el toro hay lazos que no pueden ser deformados de forma continua hasta que se contraigan (y reduzcan a un simple punto), mientras que en la esfera todo lazo cerrado puede contraerse de esta manera (véase la Fig. 12.2). También podrían contraerse lazos

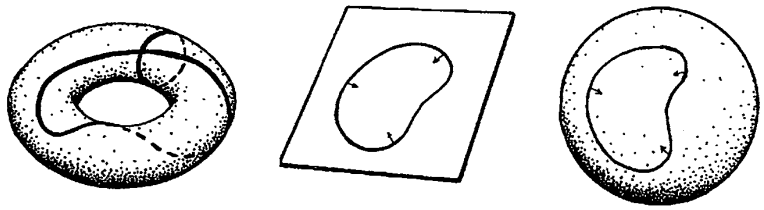


Fig. 12.2. Algunos lazos en el toro no pueden contraerse de forma continua (hasta reducirse a un único punto) mientras permanecen en la superficie; sin embargo, en el plano o la esfera cualquier lazo cerrado puede hacerlo. Por consiguiente, el plano y la esfera son simplemente conexos, pero el toro (y superficies de género más alto) son múltiplemente conexos.

en el plano euclídeo. Decimos que la esfera y el plano son *simplemente conexos* en virtud de esta propiedad de «contractibilidad». El toro, y las superficies de géneros más altos, son, por el contrario, *múltiplemente conexos* debido a la existencia de lazos no contraíbles[1] Esto nos proporciona una manera clara de distinguir, desde dentro de la propia superficie, el toro (y superficies de géneros más altos) de la esfera y del plano.

Podemos aplicar la misma idea para distinguir la topología de la 3-variedad $\mathcal{R}$ de la topología «trivial» del 3-espacio euclídeo, o la topología de la 6-variedad $\mathcal{C}$ de la del 6-espacio euclídeo «trivial». Volvamos a nuestro «libro» que, como en §11.3, imaginamos unido a una estructura fija mediante un cinturón imaginario. Cada orientación rotacional individual del libro se representará por un correspondiente punto en $\mathcal{R}$. Si rotamos el libro un ángulo 2π de forma continua, de modo que vuelva a su orientación original, encontramos que este movimiento está representado, en $\mathcal{R}$, por cierto lazo cerrado (véase la Fig. 12.3). ¿Podemos deformar este lazo cerrado de una manera continua hasta que se contraiga (y reduzca a un simple punto)? Semejante deformación del lazo correspondería a un cambio gradual de la orientación de nuestro libro hasta que no haya movimiento en absoluto. Pero recordemos nuestra atadura mediante el cinturón imaginario (que podemos implementar con un cinturón real). Nuestra 2π-rotación deja el cinturón retorcido; pero esto no puede deshacerse mediante un movimiento continuo del cinturón que deje el libro en reposo. Ahora esta 2π-torsión debe permanecer (o transformarse en un múltiplo impar

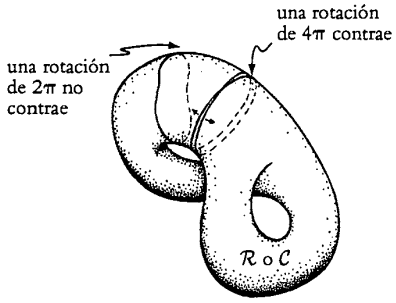


Fig. 12.3. La noción de conectividad múltiple, como se ilustra en la Fig. 12.2, distingue la topología de la 3-variedad $\mathcal{R}$ (espacio de rotación), o de la 6-variedad $\mathcal{C}$ (espacio de configuración), de las topologías «triviales» del 3-espacio y el 6-espacio euclídeos. Un lazo en $\mathcal{R}$ o $\mathcal{C}$ que representa una rotación continua de 2π no puede contraerse hasta un punto, de modo que $\mathcal{R}$ y $\mathcal{C}$ son múltiplemente conexos. Pese a todo, cuando se recorre dos veces (lo que representa una 4π-rotación) el lazo se contrae hasta un punto (torsión topológica). Véase la Fig. 11.3. (N.B. La 2-variedad mostrada, al ser solo esquemática, no tiene realmente esta última propiedad.)

de una 2π-torsión) durante toda la deformación gradual de la rotación del libro, así que concluimos que es imposible que una 2π-rotación pueda transformarse realmente de modo continuo hasta coincidir con una total ausencia de rotación. En correspondencia, no hay manera de que nuestro lazo cerrado escogido en $\mathcal{R}$ pueda ser deformado de forma continua hasta que se contraiga. En consecuencia, la 3-variedad $\mathcal{R}$ (y análogamente la 6-variedad $\mathcal{C}$) debe ser múltiplemente conexa y, por consiguiente, topológicamente diferente del 3-espacio (o el 6-espacio) euclídeo simplemente conexo.[2]

Puede advertirse que la múltiple conectividad de los espacios $\mathcal{R}$ y $\mathcal{C}$ es de una naturaleza más interesante que lo que ocurre en el caso del toro. En efecto, nuestro lazo que representa una 2π-rotación tiene la curiosa propiedad de que, si lo recorremos dos veces (una 4π-rotación), entonces obtenemos un lazo que ahora sí *puede* ser deformado continuamente hasta un punto.[12.2] (Esto no sucede, por supuesto, para el toro.) Esta curiosa característica de los lazos en $\mathcal{R}$ y $\mathcal{C}$ es un ejemplo de lo que se conoce como *torsión topológica*.

[12.2] Muestre cómo hacer esto, por ejemplo, apelando a la representación de $\mathcal{R}$ que se da en el ejercicio [12.8].

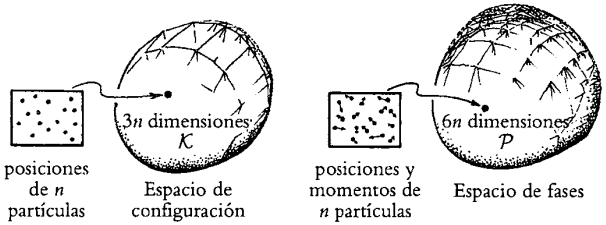


Fig. 12.4. (a) El espacio de configuración $\mathcal{K}$ para un sistema de n partículas puntuales en una región del 3-espacio tiene $3n$ dimensiones, y cada punto de $\mathcal{K}$ representa las posiciones de las n partículas. (b) El espacio de fases $\mathcal{P}$ tiene $6n$ dimensiones, y cada punto de $\mathcal{P}$ representa las posiciones y los momentos de las n partículas. (N.B. Momento = masa × velocidad.)

A partir de todo esto, vemos que sí tiene un interés físico estudiar los espacios, tales como la 6-variedad $\mathcal{C}$, que no solo son de dimensión mayor que la del espaciotiempo ordinario, sino que también pueden tener una topología no trivial. Además, tales espacios físicamente relevantes pueden tener una dimensión enormemente mayor que 6. Espacios de dimensiones muy grandes pueden aparecer como espacios de configuración, y también como lo que se denominan *espacios de fases*, para sistemas que incluyen grandes números de partículas individuales. El espacio de configuración $\mathcal{K}$ de un gas, cuyas partículas se describen como puntos individuales en el espacio 3-dimensional, tiene dimensión $3N$, donde N es el número de partículas en el gas. Cada punto de $\mathcal{K}$ representa una configuración del gas en la que está determinada la posición de cada partícula por separado (Fig. 12.4a). En el caso del espacio de fases $\mathcal{P}$ del gas, debemos identificar también el *momento* de cada partícula (que es la velocidad de la partícula multiplicada por su masa), que es una magnitud vectorial (3 componentes para cada partícula), de modo que el número total de dimensiones es $6N$. Así pues, cada punto en $\mathcal{P}$ representa no solo la posición de todas las partículas en el gas sino también el movimiento de cada partícula individual (Fig. 12.4b). En un dedal lleno de aire ordinario podría haber unas 10^{19} moléculas,[3] de modo que $\mathcal{P}$ tiene algo parecido a ¡60.000.000.000.000.000.000 dimensiones! Los espacios de fases son particularmente útiles en el estudio del com-

portamiento de sistemas físicos (clásicos) que incluyen muchas partículas, por lo que los espacios de dimensión tan alta pueden ser muy relevantes físicamente.

12.2. Variedades y cartas de coordenadas

Consideremos ahora cómo puede tratarse matemáticamente la estructura de una n-variedad. Una n-variedad $\mathcal{M}$ puede construirse de una forma completamente análoga a aquella que hemos utilizado en los capítulos 8 y 10 (véase §10.2) para construir la superficie $\mathcal{S}$ a partir de varias cartas de coordenadas. Sin embargo, ahora necesitamos en cada carta más coordenadas que solo un par de números (x, y) o (X, Y). De hecho, necesitamos n coordenadas por carta, donde n es un número fijo —la *dimensión* de $\mathcal{M}$— que puede ser cualquier entero positivo. Por esta razón, es conveniente no utilizar una letra independiente para cada coordenada, sino distinguir nuestras diferentes coordenadas

$$x^1,\ x^2,\ x^3,\ \ldots,\ x^n$$

mediante la utilización de un (super) índice numérico. Esto no debe llamar a error. No hay que interpretar que son diferentes potencias de una única cantidad x, sino números reales independientes. El lector podría encontrar extraño que haya buscado la complicación, de forma deliberada, utilizando un superíndice en lugar de un subíndice (por ejemplo, $x_1, x_2, x_3, \ldots, x_n$), lo que conduce a la inevitable confusión entre, por ejemplo, la coordenada x^3 y el cubo de cierta cantidad x. La confusión del lector está justificada. Yo mismo no solo lo encuentro confuso, sino también, en ocasiones, bastante irritante. Por alguna razón histórica, los convenios estándar para el *análisis tensorial* clásico (al que llegaremos de una forma más seria más adelante en este capítulo) han dado este resultado. Tales convenios incluyen reglas firmemente trenzadas que gobiernan la colocación arriba/abajo de los índices, y la colocación consistente de los índices en las propias coordenadas ha resultado ser en la posición superior. (Estas reglas funcionan bien en la práctica, pero es una pena que los convenios no se escogieran a la inversa. Me temo que esto es algo con lo que tenemos que vivir.)

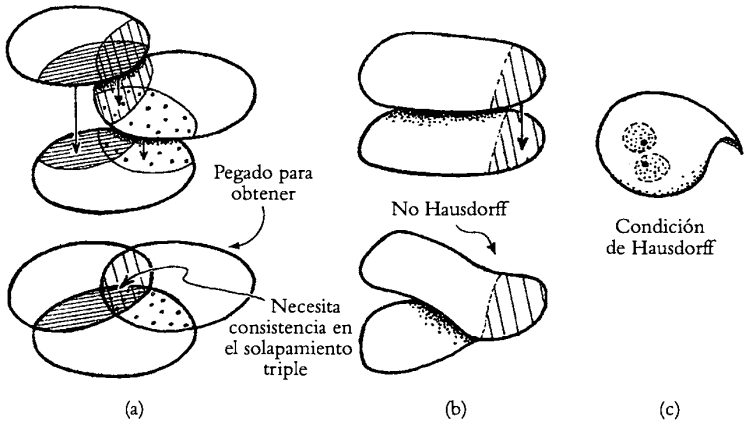


Fig. 12.5. (a) Las funciones de transición que traducen entre coordenadas en cartas que se solapan deben satisfacer una relación de consistencia en cada solapamiento triple. (b) Las regiones de solapamiento (conjunto abierto) entre pares de cartas deben ser adecuadas; de lo contrario, puede darse la «ramificación» que caracteriza a un espacio no Hausdorff. (c) Un espacio de Hausdorff es un espacio con la propiedad de que dos puntos distintos cualesquiera poseen entornos que no se solapan. (En (b), para que la parte «pegada» sea un conjunto abierto, su «borde», donde ocurre la ramificación, debe quedar separado, y es a lo largo de este donde falla la condición de Hausdorff.)

¿Cómo tenemos que imaginarnos nuestra variedad $\mathcal{M}$? La consideramos formada por varias cartas de coordenadas «pegadas», siendo cada carta una región abierta de $\mathbb{R}^n$. Aquí $\mathbb{R}^n$ representa el «espacio de coordenadas» cuyos puntos son simplemente las n-tuplas de números reales $(x^1, x^2, \ldots, x^n)$, donde podemos recordar de §6.1 que $\mathbb{R}$ representa el sistema de los números reales. En nuestro procedimiento de pegado habrá *funciones de transición* que expresan las coordenadas en una carta en términos de las coordenadas en otra, dondequiera que, en la variedad $\mathcal{M}$, encontremos una carta de coordenadas que se solapa con otra. Estas funciones de transición deben satisfacer ciertas condiciones mutuas para asegurar la consistencia del procedimiento global. El procedimiento se ilustra en la Fig. 12.5a. Pero debemos tener cuidado en que, para producir el tipo estándar de variedad,[4] sea un *espacio de Hausdorff*. (Las variedades no Hausdorff pueden «ramificarse» en formas tales como las indicadas en la Fig. 12.5b; véase también la Fig. 8.2c.) Un

espacio de Hausdorff tiene la propiedad definitoria de que para cualesquiera dos puntos distintos del espacio existen conjuntos abiertos que contienen a cada uno de ellos y que no se intersectan (Fig. 12.5c).

No obstante, es importante advertir que una variedad $\mathcal{M}$ no tiene por qué «saber» dónde están estas cartas individuales o cuáles podrían ser los valores concretos de las coordenadas en un punto. Una forma razonable de considerar $\mathcal{M}$ es suponer que se ha construido empalmando varias cartas de coordenadas, pero luego hemos «olvidado» la forma concreta en que se han introducido dichas cartas de coordenadas. La variedad es una estructura matemática por sí misma, y las coordenadas son tan solo elementos auxiliares que pueden ser reintroducidos a conveniencia cuando se desee. La definición matemática exacta de una variedad (de la que existen varias alternativas) nos distraería aquí.[5]

12.3. Escalares, vectores y covectores

Como en §10.2, tenemos la noción de una *función uniforme* Φ, definida en $\mathcal{M}$ (a veces llamada un *campo escalar* en $\mathcal{M}$) donde Φ está definida, en cualquier carta de coordenadas local, como una función suave de las n coordenadas en dicha carta. Aquí «suave» se tomará siempre en el sentido «C^∞-suave» (véase §6.3). En cada solapamiento entre dos cartas, las coordenadas en cada carta son funciones suaves de las coordenadas en la otra, de modo que la suavidad de Φ en términos de un conjunto de coordenadas en el solapamiento implica su suavidad en términos del otro. De esta manera, la definición («basada en cartas») local de suavidad de una función escalar Φ se extiende a la totalidad de $\mathcal{M}$, y podemos hablar simplemente de la *suavidad* de Φ en $\mathcal{M}$.

A continuación podemos definir la noción de un *campo vectorial* $\boldsymbol{\xi}$ en $\mathcal{M}$, que debería ser algo interpretable geométricamente como una familia de «flechas» en $\mathcal{M}$ (Fig. 10.5), donde $\boldsymbol{\xi}$ es algo que actúa sobre un campo escalar (suave) Φ para producir otro campo escalar $\boldsymbol{\xi}(\Phi)$ a la manera de un operador de diferenciación. La interpretación de $\boldsymbol{\xi}(\Phi)$ va a ser la «tasa de incremento» de Φ en la dirección indicada por las flechas que representan a $\boldsymbol{\xi}$, igual que para las 2-superficies de §10.3. Al ser un «operador de diferenciación», $\boldsymbol{\xi}$ satisface ciertas relaciones algebrai-

cas características (básicamente objetos que hemos visto en §6.5, es decir, $d(f+g) = df + dg$, $d(fg) = f dg + g df$, $da = 0$ si a es constante):

$$\boldsymbol{\xi}(\Phi + \Psi) = \boldsymbol{\xi}(\Phi) + \boldsymbol{\xi}(\Psi),$$
$$\boldsymbol{\xi}(\Phi\Psi) = \Phi\boldsymbol{\xi}(\Psi) + \Psi\boldsymbol{\xi}(\Phi),$$
$$\boldsymbol{\xi}(k) = 0 \text{ si } k \text{ es una constante.}$$

De hecho, hay un teorema que nos dice que estas propiedades algebraicas son *suficientes* para caracterizar a $\boldsymbol{\xi}$ como un campo vectorial.[6]

También podemos utilizar tales medios puramente algebraicos para definir una *1-forma*, o, dicho de otra forma, un *campo de covectores*. (Pronto llegaremos al significado geométrico de un covector.) Un campo de covectores $\boldsymbol{\alpha}$ puede considerarse como una aplicación de campos vectoriales en campos escalares, y la acción de $\boldsymbol{\alpha}$ sobre $\boldsymbol{\xi}$ se escribe en la forma $\boldsymbol{\alpha} \bullet \boldsymbol{\xi}$ (el *producto escalar* de $\boldsymbol{\alpha}$ por $\boldsymbol{\xi}$; véase §10.4), donde para cualesquiera campos vectoriales $\boldsymbol{\xi}$ y $\boldsymbol{\eta}$, y el campo escalar $\boldsymbol{\Phi}$, tenemos *linealidad*

$$\boldsymbol{\alpha} \bullet (\boldsymbol{\xi} + \boldsymbol{\eta}) = \boldsymbol{\alpha} \bullet \boldsymbol{\xi} + \boldsymbol{\alpha} \bullet \boldsymbol{\eta},$$
$$\boldsymbol{\alpha} \bullet (\Phi\boldsymbol{\xi}) = \Phi(\boldsymbol{\alpha} \bullet \boldsymbol{\xi}).$$

Estas relaciones definen a los covectores como objetos *duales* de los vectores (y a esto se refiere el prefijo «co»). La relación entre vectores y covectores es simétrica, de modo que tenemos las correspondientes expresiones

$$(\boldsymbol{\alpha} + \boldsymbol{\beta}) \bullet \boldsymbol{\xi} = \boldsymbol{\alpha} \bullet \boldsymbol{\xi} + \boldsymbol{\beta} \bullet \boldsymbol{\xi},$$
$$(\Phi\boldsymbol{\alpha}) \bullet \boldsymbol{\xi} = \Phi(\boldsymbol{\alpha} \bullet \boldsymbol{\xi}),$$

que llevan a la definición de la suma de dos covectores y el producto de un covector por un escalar. Cuando tomamos el dual del espacio de covectores, obtenemos de nuevo el espacio de vectores original. (En otras palabras, un «*co*-covector» sería un vector.)

Podemos tomar estas relaciones referidas a campos enteros o meramente a entidades definidas en un único punto de $\mathcal{M}$. Los vectores tomados en un punto fijo o concreto constituyen un *espacio vectorial*. (Tal como se ha descrito en §11.1, en un espacio vectorial podemos sumar elementos $\boldsymbol{\xi}$ y $\boldsymbol{\eta}$ para formar su suma $\boldsymbol{\xi} + \boldsymbol{\eta}$, con $\boldsymbol{\xi} + \boldsymbol{\eta} = \boldsymbol{\eta} + \boldsymbol{\xi}$ y $(\boldsymbol{\xi} + \boldsymbol{\eta}) + \boldsymbol{\zeta} = \boldsymbol{\xi} + (\boldsymbol{\eta} + \boldsymbol{\zeta})$, y podemos multiplicarlos por escalares

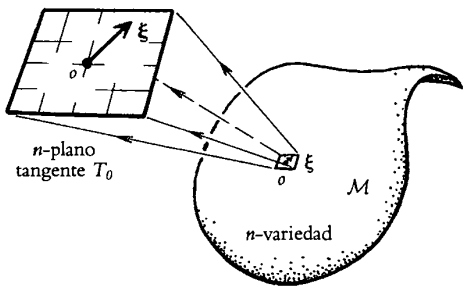


Fig. 12.6. El espacio tangente T_o a una n-variedad $\mathcal{M}$ en un punto o puede entenderse de manera intuitiva como el espacio limitante, cuando se examinan entornos de o en $\mathcal{M}$ cada vez más pequeños con aumentos correspondientemente cada vez mayores. (Compárese con la Fig. 10.6.) El T_o resultante es plano: un espacio vectorial n-dimensional.

—aquí, números reales f, g— donde $(f+g)\boldsymbol{\xi} = f\boldsymbol{\xi} + g\boldsymbol{\xi}$, $f(\boldsymbol{\xi} + \boldsymbol{\eta}) = f\boldsymbol{\xi} + f\boldsymbol{\eta}$, $f(g\boldsymbol{\xi}) = fg(\boldsymbol{\xi})$, $1\boldsymbol{\xi} = \boldsymbol{\xi}$.) Podemos considerar que este espacio vectorial (plano) proporciona la estructura de la variedad en la inmediata vecindad de o (véase la Fig. 12.6). Llamamos a este espacio vectorial el *espacio tangente* T_o a $\mathcal{M}$ en o. T_o puede entenderse intuitivamente como el espacio límite al que se llega cuando se examinan entornos cada vez más pequeños de o en $\mathcal{M}$ con una amplificación, en correspondencia, cada vez mayor. La inmediata vecindad de o, en $\mathcal{M}$, parece estar así infinitamente «dilatada» bajo este examen. En el límite, cualquier «curvatura» de $\mathcal{M}$ quedaría «alisada» para dar la estructura plana de T_o. El espacio vectorial T_o tiene la dimensión (finita) n, porque podemos encontrar un conjunto de n *elementos de base*, a saber, las cantidades $\partial/\partial x^1, \ldots, \partial/\partial x^n$, en el punto o, que apuntan a lo largo de los ejes de coordenadas, en cuyos términos puede expresarse linealmente cualquier elemento de T_o de forma unívoca (véase también §13.5).

Podemos formar el espacio vectorial *dual* de T_o (el espacio de covectores en o) de la forma antes descrita, y este se denomina el *espacio cotangente* T_o^* a $\mathcal{M}$ en o. Un caso particular de un *campo* covectorial es el *gradiente* (o derivada exterior) $d\Phi$ de un campo escalar Φ. (Ya hemos encontrado esta notación en el caso 2-dimensional; véase §10.3.) El covector $d\Phi$ (con componentes $\partial\Phi/\partial x^1, \ldots, \partial\Phi/\partial x^n$) tiene la propiedad definitoria

$$d\Phi \bullet \boldsymbol{\xi} = \boldsymbol{\xi}(\Phi).$$

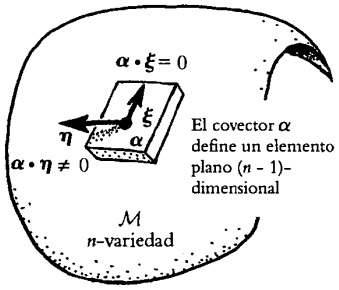


Fig. 12.7. Un covector $\boldsymbol{\alpha}$ (no nulo) en un punto de $\mathcal{M}$ determina un elemento plano $(n-1)$-dimensional. Los vectores $\boldsymbol{\xi}$ que satisfacen $\boldsymbol{\alpha} \cdot \boldsymbol{\xi} = 0$ definen las direcciones dentro del mismo.

(Véase también §10.4.)[12.3] Aunque no todos los covectores tienen la forma $\mathrm{d}\Phi$ para cierto Φ, todos pueden expresarse de tal manera en cualquier punto. Veremos a continuación por qué esto no se extiende a *campos* covectoriales.

¿Cuál es la diferencia geométrica entre un covector y un vector? En cada punto de $\mathcal{M}$, un covector (no nulo) determina un *elemento plano* $(n-1)$*-dimensional*. Las direcciones que yacen dentro de este elemento $(n-1)$-plano son las determinadas por vectores $\boldsymbol{\xi}$ para los que $\boldsymbol{\alpha} \cdot \boldsymbol{\xi} = 0$; véase la Fig. 12.7. En el caso particular en que $\boldsymbol{\alpha} = \mathrm{d}\Phi$, estos elementos $(n-1)$-planos son tangentes a la familia de superficies $(n-1)$-dimensionales[12.4] de Φ constante (lo que generaliza la noción de «curvas de nivel», tal como se ha ilustrado en la Fig. 10-8a). No obstante, en general los elementos $(n-1)$-planos definidos por un covector $\boldsymbol{\alpha}$ se retorcerán de un modo que les impedirá tocar consistentemente a cualquier familia semejante de $(n-1)$-superficies (Fig. 12.8).[7]

En cualquier carta de coordenadas concreta, con coordenadas $x^1, \ldots, x^n$, podemos representar el vector (del campo) $\boldsymbol{\xi}$ por su conjunto de *componentes* $(\zeta^1, \zeta^2, \ldots, \zeta^n)$, siendo este el conjunto de coeficientes en la representación explícita de $\boldsymbol{\xi}$ en términos de operadores de diferenciación parcial

$$\boldsymbol{\xi} = \zeta^1 \frac{\partial}{\partial x^1} + \zeta^2 \frac{\partial}{\partial x^2} + \ldots + \zeta^n \frac{\partial}{\partial x^n},$$

[12.3] Demuestre que «$\mathrm{d}\Phi$», definido de esta forma, satisface los requisitos de «linealidad» de un covector, como se ha especificado antes.
[12.4] ¿Por qué?

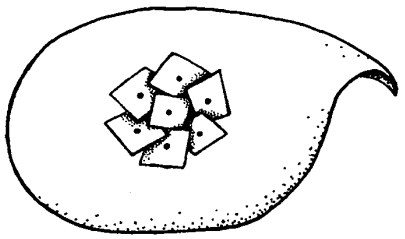


Fig. 12.8. Los elementos $(n-1)$-planos definidos por un campo covectorial $\boldsymbol{\alpha}$ se retorcerían en general de una forma que les impediría tocar consistentemente a una única familia de $(n-1)$-superficies, aunque en el caso particular $\boldsymbol{\alpha} = \mathrm{d}\Phi$ (para un campo escalar Φ), tocarían las superficies Φ = constante (lo que generaliza las «líneas de nivel» de la Fig. 10.8).

en la carta (véase §10.4). En el caso de un vector en un punto concreto, $\zeta^1, \ldots, \zeta^n$, serán simplemente n números reales; en el caso de un campo vectorial dentro de cierta carta de coordenadas, serán n funciones (suaves) de las coordenadas $x^1, \ldots, x^n$ (y se recuerda al lector que «ζ^n» *no* representa ahora «la n-ésima potencia de ξ», etc.). Recordemos que cada uno de los operadores «$\partial/\partial x$» representa «tomar la tasa de cambio en la dirección del r-ésimo eje de coordenadas (local)». La expresión anterior para $\boldsymbol{\xi}$ expresa simplemente dicho vector (que, como opera-

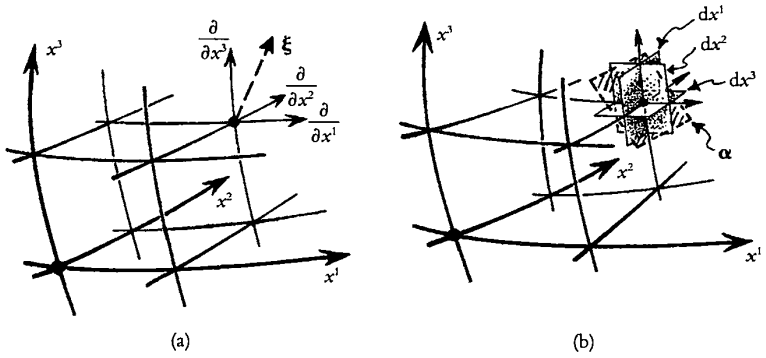


Fig. 12.9. Componentes en una carta de coordenadas $(x^1, \ldots, x^n)$ (con $n = 3$ aquí). (a) Para un (campo) vector $\boldsymbol{\xi}$, estos son los coeficientes $(\xi^1, \xi^2, \ldots, \xi^n)$ en $\boldsymbol{\xi} = \xi^1 \partial/\partial x^1 + \xi^2 \partial/\partial x^2 + \ldots + \xi^n \partial/\partial x^n$, donde «$\partial/\partial x$» representa la «tasa de cambio a lo largo del r-ésimo eje de coordenadas» (véase también la Fig. 10.9). (b) Para un (campo) covector $\boldsymbol{\alpha}$, estos son los coeficientes $(\alpha_1, \alpha_2, \ldots, \alpha_n)$ en $\boldsymbol{\alpha} = \alpha_1 \mathrm{d}x^1 + \ldots + \alpha_n \mathrm{d}x^n$, donde $\mathrm{d}x^r$ representa «el gradiente de x^r», y se refiere al elemento $(n-1)$-plano generado por los ejes de coordenadas (locales) excepto el eje x^r (local).

dor, recordemos que afirma «tomar la tasa de cambio en la $\boldsymbol{\xi}$-dirección») como combinación lineal de los vectores que apuntan a lo largo de cada uno de los ejes de coordenadas (véase la Fig. 12.9a).

De modo similar, un covector (de un campo) $\boldsymbol{\alpha}$ está representado, en la carta de coordenadas, por un conjunto de componentes $(\alpha^1, \alpha^2, \ldots, \alpha^n)$, donde ahora escribimos

$$\boldsymbol{\alpha} = \alpha_1 \mathrm{d}x^1 + \alpha_2 \mathrm{d}x^2 + \ldots + \alpha_n \mathrm{d}x^n,$$

que expresa $\boldsymbol{\alpha}$ como una combinación lineal de las 1-formas (covectores)[8] básicas $\mathrm{d}x^1, \mathrm{d}x^2, \ldots, \mathrm{d}x^n$. Desde un punto de vista geométrico, cada uno de los $\mathrm{d}x^r$ se refiere al elemento $(n-1)$-plano que abarca todos los ejes de coordenadas con excepción del eje x^r (véase la Fig. 12.9b)[12.5]. El producto escalar $\boldsymbol{\alpha} \bullet \boldsymbol{\xi}$ viene dado por la expresión[12.6]

$$\boldsymbol{\alpha} \bullet \boldsymbol{\xi} = \alpha_1 \xi^1 + \alpha_2 \xi^2 + \ldots + \alpha_n \xi^n.$$

12.4. Productos de Grassmann

Consideremos ahora la representación de elementos planos de otras dimensiones, utilizando la idea del *producto de Grassmann* definida en §11.6. Un elemento 2-plano en un punto de $\mathcal{M}$ (o un campo de elementos 2-planos sobre $\mathcal{M}$) estará representado por una cantidad

$$\boldsymbol{\xi} \wedge \boldsymbol{\eta},$$

donde $\boldsymbol{\xi}$ y $\boldsymbol{\eta}$ son dos vectores (o campos vectoriales) independientes que llenan el o los 2-planos (Figs. 11.6a o 12.10a). Una cantidad $\boldsymbol{\xi} \wedge \boldsymbol{\eta}$ se suele conocer como un *bivector*. Sus componentes, en términos de los de $\boldsymbol{\xi}$ y $\boldsymbol{\eta}$, son las expresiones

$$\xi^{[r}\eta^{s]} = \frac{1}{2}(\xi^r\eta^s - \xi^s\eta^r),$$

[12.5] Por ejemplo, demuestre que $\mathrm{d}x^2$ tiene componentes $(0, 1, 0, \ldots, 0)$ y representa los elementos hiperplano tangentes a $x^2 = \text{constante}$.

[12.6] Demuestre, mediante el uso de la regla de la cadena (véase §10.3), que esta expresión para $\boldsymbol{\alpha} \bullet \boldsymbol{\xi}$ es consistente con $\mathrm{d}\Phi \bullet \boldsymbol{\xi} = \boldsymbol{\xi}(\Phi)$, en el caso particular $\boldsymbol{\alpha} = \mathrm{d}\Phi$.

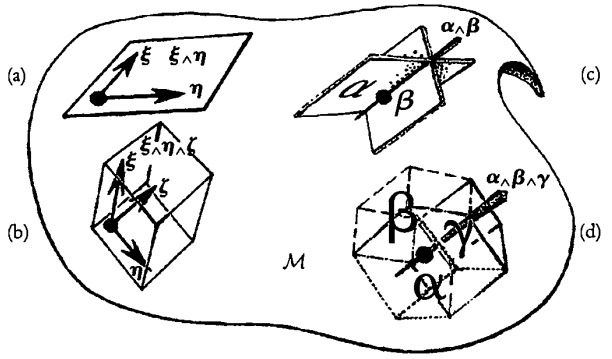


Fig. 12.10. (a) Un elemento 2-plano en un punto de $\mathcal{M}$, que está generado por vectores independientes $\boldsymbol{\xi}, \boldsymbol{\eta}$, se describe por el bivector $\boldsymbol{\xi} \wedge \boldsymbol{\eta}$. (b) Análogamente, un elemento 3-plano generado por $\boldsymbol{\xi}, \boldsymbol{\eta}, \boldsymbol{\zeta}$ se describe por $\boldsymbol{\xi} \wedge \boldsymbol{\eta} \wedge \boldsymbol{\zeta}$. (c) Dualmente, un elemento $(n-2)$-plano, la intersección de dos elementos $(n-1)$-planos especificados por 1-formas $\boldsymbol{\alpha}, \boldsymbol{\beta}$, se describe por $\boldsymbol{\alpha} \wedge \boldsymbol{\beta}$. (d) El elemento $(n-3)$-plano de intersección de los tres elementos $(n-1)$-planos especificados por $\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}$ se describe por $\boldsymbol{\alpha} \wedge \boldsymbol{\beta} \wedge \boldsymbol{\gamma}$.

descritas hacia el final del capítulo anterior. Una *suma* $\boldsymbol{\psi}$ de bivectores simples de la forma $\boldsymbol{\xi} \wedge \boldsymbol{\eta}$ se denomina también un *bivector*; sus componentes ψ^{rs} tienen la propiedad característica de que son *antisimétricas* en r y s, i.e., $\psi^{sr} = -\psi^{rs}$.

Análogamente, un elemento 3-plano (o un campo de ellos) estaría representado por un *trivector*

$$\boldsymbol{\xi} \wedge \boldsymbol{\eta} \wedge \boldsymbol{\zeta},$$

donde los vectores $\boldsymbol{\xi}, \boldsymbol{\eta}, \boldsymbol{\zeta}$ llenan el 3-plano (Fig. 12.10b), siendo sus componentes

$$\xi^{[r}\eta^s\zeta^{t]} = \frac{1}{6}(\xi^r\eta^s\zeta^t + \xi^s\eta^t\zeta^r + \xi^t\eta^r\zeta^s - \xi^r\eta^t\zeta^s - \xi^t\eta^s\zeta^r - \xi^s\eta^r\zeta^t).$$

El trivector general $\boldsymbol{\tau}$ tiene componentes totalmente antisimétricas τ^{rst}, y siempre sería una suma de tales trivectores simples. Podemos continuar de manera similar para definir elementos 4-planos, representados por simples 4-vectores, y así sucesivamente. El n-vector *general* tiene conjuntos de componentes que son completamente antisimétricas. Siempre serían expresables como una *suma* de simples n-vectores.

Aquí surge una cuestión que puede parecer enigmática. Parece que ahora tenemos dos formas *diferentes* de representar un elemento $(n-1)$-plano: o bien como una 1-forma (covector), o bien como una cantidad $(n-1)$-vectorial, obtenida mediante un «producto cuña» de $n-1$ vectores independientes que abarcan el $(n-1)$-plano. De hecho, hay una diferencia geométrica entre las cantidades descritas de estas dos maneras distintas, pero es algo sutil. La diferencia reside en que la 1-forma debería imaginarse como un «tipo» de densidad, mientras que el $(n-1)$-vector no. Para aclarar esto, será útil introducir antes la noción de una p-forma general.

En esencia, procederemos igual que para los multivectores anteriores, pero empezando con 1-formas en lugar de con vectores. Dado un número p de 1-formas (independientes) $\boldsymbol{\alpha}, \boldsymbol{\beta}, \ldots, \boldsymbol{\delta}$, podemos formar su producto cuña

$$\boldsymbol{\alpha} \wedge \boldsymbol{\beta} \wedge \ldots \wedge \boldsymbol{\delta},$$

cuyas componentes están dadas por

$$\alpha_{[r}\beta_{s} \ldots \delta_{u]}$$

en una carta de coordenadas (utilizando la notación general de paréntesis cuadrados que rodean índices de §11.6). Una cantidad semejante determina un elemento $(n-p)$-plano (o un campo de ellos), que es la *intersección* de los diversos elementos $(n-1)$-planos determinados por $\boldsymbol{\alpha}, \boldsymbol{\beta}, \ldots, \boldsymbol{\delta}$ individualmente (Fig. 12.10c,d). Esta cantidad se denomina una *p-forma simple*. Como sucedía con los p-vectores, la p-forma más general no es expresable como un producto cuña directo de covectores (excepto en los casos particulares $p = 0, 1, n-1, n$), sino que es una suma de términos que sí son expresables de ese modo. En términos de componentes, una p-forma general $\boldsymbol{\varphi}$ está representada (en cualquier carta de coordenadas) por un conjunto de cantidades

$$\varphi_{rs\ldots u}$$

(donde cada una de las $r, s, \ldots, u$ recorre los valores $1, \ldots, n$) que es *antisimétrico* en sus índices $r, s, \ldots, u$, cuyo número es p. Como antes, la antisimetría significa que si intercambiamos cualquier par de índices, obtenemos una cantidad que es exactamente la negativa de la que teníamos antes. Con nuestra notación de corchetes (§11.6), podemos expresar esta propiedad de antisimetría en la ecuación[12.7]

$$\varphi_{[rs\ldots u]} = \varphi_{rs\ldots u}.$$

Aquí también podría comentarse que la $(p+q)$-forma $\boldsymbol{\varphi} \wedge \boldsymbol{\chi}$, que es el producto cuña de la p-forma $\boldsymbol{\varphi}$ por una q-forma $\boldsymbol{\chi}$, tiene componentes

$$\varphi_{[rs\ldots u}\chi_{jk\ldots m]},$$

donde la antisimetrización se toma sobre todos los índices (y donde $\chi_{jk\ldots m}$ son las componentes de $\boldsymbol{\chi}$).[12.8] Una notación similar se aplica al producto cuña de un p-vector por un q-vector.

12.5. Integrales de formas

Volvamos ahora al aspecto «densidad» de una p-forma. Recordemos que en la física ordinaria la *densidad* de un objeto es su masa por la unidad de volumen. Dicha densidad es una propiedad del material del que está compuesto el objeto. Utilizamos esta noción de «densidad» cuando queremos calcular la masa total del objeto si conocemos su volumen total y la naturaleza de su material. Matemáticamente hablando, lo que haríamos sería integrar su densidad sobre el volumen que ocupa. Básicamente, lo importante de la densidad es que es el tipo de magnitud adecuado que podemos integrar sobre una región; es el tipo de magnitud que colocamos a continuación de un símbolo de integral. Sin embargo, aquí deberíamos tener cuidado en distinguir integrales sobre espacios de diferente dimensión. («Masa por unidad de área» es un tipo de magnitud diferente de «masa por unidad de volumen», por ejemplo.) Encontraremos que una p-forma es la magnitud apropiada para integrar sobre un espacio p-dimensional.

Empecemos con una 1-forma. Este es el caso más sencillo. Estamos

[12.7] Explique por qué funciona esto.

[12.8] Justifique el hecho de que $\varphi\wedge\chi = \alpha\wedge \ldots \wedge\gamma\wedge\lambda\wedge \ldots \wedge\nu$ donde $\varphi = \alpha\wedge \ldots \wedge\gamma$, $\chi = \lambda\wedge \ldots \wedge\nu$.

interesados en la integral de una magnitud sobre una variedad 1-dimensional, es decir, a lo largo de cierta curva γ. Recordemos de §6.6 que las integrales ordinarias (1-dimensionales) son cosas que se escriben

$$\int f(x)\ \mathrm{d}x,$$

donde x es una magnitud con valores reales que podemos tomar como un parámetro a lo largo de la curva γ. Debemos pensar que la cantidad «$f(x)\mathrm{d}x$» denota una 1-forma. La notación para las 1-formas se ha hecho a la medida con mucho cuidado para que sea consistente con la notación para las integrales ordinarias. Esta es una característica del cálculo infinitesimal del siglo XX conocida como *cálculo exterior*, introducida por el destacado matemático francés Élie Cartan (1869-1951), a quien volveremos a encontrarnos en los capítulos 13, 14 y 17, que encaja de forma muy bella con la notación introducida en el siglo XVII por Gottfried Wilhelm Leibniz (1646-1716). Sin embargo, en el esquema de Cartan no consideramos que «dx» denote una «cantidad infinitesimal», sino que nos proporciona el tipo apropiado de densidad (1-forma) que se puede integrar sobre una curva.

Una de las bellezas de esta notación es que trata automáticamente cualquier cambio de variable que decidamos invocar. Si cambiamos el parámetro x por otro X, por ejemplo, se considera que la 1-forma $\boldsymbol{\alpha} = f(x)\mathrm{d}x$ sigue siendo la misma —en el sentido de que $\int \boldsymbol{\alpha}$ sigue siendo la misma—, incluso si cambia su expresión funcional explícita en términos de la variable dada (x o X).[12.9] También podemos considerar que la 1-forma $\boldsymbol{\alpha}$ está definida en cierto espacio ambiente de dimensión mayor dentro del cual está nuestra curva. El parámetro x o el X podría tomarse como una de las coordenadas en una carta de coordenadas en este espacio ambiente, donde cambiamos sin problemas a una coordenada diferente cuando pasamos a otra carta de coordenadas. No hay que preocuparse de nada. Simplemente escribimos esta integral como

$$\int \boldsymbol{\alpha} \quad \text{o} \quad \int_{\mathcal{R}} \boldsymbol{\alpha},$$

[12.9] Demuéstrelo explícitamente explicando cómo tratar los límites para una integral definida $\int_a^b \boldsymbol{\alpha}$.

donde $\mathcal{R}$ representa cierta porción de la curva γ sobre la que va a tomarse la integral.

¿Qué sucede con integrales sobre regiones de dimensión más alta? Para una región 2-dimensional, necesitamos una 2-forma después del símbolo integral.[9] Esta podría ser alguna magnitud $f(x, y)\mathrm{d}x \wedge \mathrm{d}y$ (o una suma de objetos como este) y podemos escribir

$$\int_{\mathcal{R}} f(x, y)\ \mathrm{d}x \wedge \mathrm{d}y = \int_{\mathcal{R}} \boldsymbol{\alpha}$$

(o una suma de tales cantidades) donde $\mathcal{R}$ es ahora una región 2-dimensional sobre la que va a realizarse la integral, que yace dentro de una 2-superficie dada. Una vez más, los parámetros x e y, que localmente sirven de coordenadas para la superficie, pueden ser reemplazados por cualquier otro par semejante, y la notación hace todo lo demás. Esto se aplica perfectamente si la 2-forma habita en cierto espacio ambiente de dimensión más alta dentro del que se encuentra la 2-región $\mathcal{R}$. Todo esto funciona también para 3-formas integradas sobre regiones 3-dimensionales o 4-formas integradas sobre regiones 4-dimensionales, etc. El producto cuña en la notación de forma diferencial de Cartan (junto con la derivada exterior de §12.6) se ocupa de todo si decidimos cambiar nuestras coordenadas. (Esto elimina la mención explícita de cantidades complicadas conocidas como «jacobianos», que, de lo contrario, tendrían que ser incorporadas.)[12.10]

Recordemos, de §6.6, el *teorema fundamental del cálculo infinitesimal* que afirma, para integrales 1-dimensionales, que la integración es la inversa de la diferenciación, o, dicho de otra forma, que

$$\int_a^b \frac{\mathrm{d}f(x)}{\mathrm{d}x}\mathrm{d}x = f(b) - f(a).$$

¿Hay algún teorema análogo a este en dimensión más alta? En realidad, hay teoremas análogos para dimensiones diferentes que llevan nombres diversos (Ostrogradski, Gauss, Green, Kelvin, Stokes, etc.), pero el resultado general, que es parte del cálculo exterior de formas diferenciales de Cartan, será denominado aquí «el teorema fundamental del

[12.10] Sea $G = \int_{-\infty}^{\infty} e^{-x^2}\,\mathrm{d}x$. Explique por qué $G^2 = \int_{\mathbb{R}^2} e^{-(x^2+y^2)}\mathrm{d}x\wedge\mathrm{d}y$ y calcúlelo pasando a coordenadas polares (r, θ). (Véase §5.1.) De ahí, demuestre $G = \sqrt{\pi}$.

cálculo exterior».[10] Este depende de la noción general de *derivada exterior* que se debe a Cartan, que explicaremos a continuación.

12.6. Derivada exterior

Una ruta «libre de coordenadas» para definir esta importante noción consiste en construir la derivada exterior axiomáticamente como el operador único «d», que convierte p-formas en $(p + 1)$-formas, para cada $p = 0, 1, \ldots, n - 1$, que tiene las propiedades siguientes

$$\begin{aligned} \mathrm{d}(\boldsymbol{\alpha} + \boldsymbol{\beta}) &= \mathrm{d}\boldsymbol{\alpha} + \mathrm{d}\boldsymbol{\beta}, \\ \mathrm{d}(\boldsymbol{\alpha} \wedge \boldsymbol{\gamma}) &= \mathrm{d}\boldsymbol{\alpha} \wedge \boldsymbol{\gamma} + (-1)^p \boldsymbol{\alpha} \wedge \mathrm{d}\boldsymbol{\gamma}, \\ \mathrm{d}(\mathrm{d}\boldsymbol{\alpha}) &= 0, \end{aligned}$$

siendo $\boldsymbol{\alpha}$ una p-forma, y donde $\mathrm{d}\Phi$ tiene el mismo significado («gradiente de Φ») para una 0-forma (i.e., para un escalar) que el que tenía en nuestra discusión anterior (definido a partir de $\mathrm{d}\Phi \bullet \boldsymbol{\xi} = \boldsymbol{\xi}(\Phi)$; también la «d» en $\mathrm{d}x$ es esta misma operación). La última ecuación en la lista anterior se suele expresar simplemente como

$$\mathrm{d}^2 = 0,$$

que es una propiedad clave del operador derivada exterior d. (Podemos advertir que la «razón» para el término de apariencia complicada $(-1)^p$ en la segunda ecuación mostrada es que la «d» siguiente está realmente «situada en el lugar equivocado», al tener que «acompañar» a $\boldsymbol{\alpha}$, con sus p índices antisimétricos. Esto queda más manifiesto en las expresiones con índices que se dan más abajo.[12.11])

Una 1-forma $\boldsymbol{\alpha}$ que es un *gradiente* $\boldsymbol{\alpha} = \mathrm{d}\Phi$ debe satisfacer $\mathrm{d}\boldsymbol{\alpha} = 0$, por lo que ya se ha dicho.[12.12] Pero no todas las 1-formas satisfacen esta relación. De hecho, si una 1-forma $\boldsymbol{\alpha}$ satisface $\mathrm{d}\boldsymbol{\alpha} = 0$, entonces se sigue que *localmente* (i.e., en un conjunto abierto suficientemente pequeño que contiene a cualquier punto dado) tiene la forma $\boldsymbol{\alpha} = \mathrm{d}\Phi$

[12.11] Demuestre, utilizando las relaciones anteriores, que $\mathrm{d}(A\mathrm{d}x + B\mathrm{d}y) = (\partial B/\partial x - \partial A/\partial y)\mathrm{d}x \wedge \mathrm{d}y$.

[12.12] ¿Por qué?

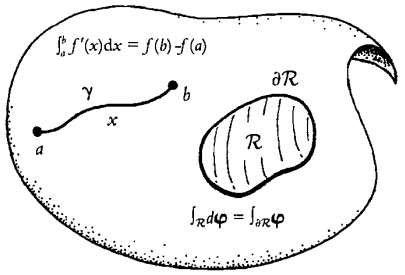


Fig. 12.11. El teorema fundamental del cálculo exterior $\int_{\mathcal{R}} \mathrm{d}\boldsymbol{\varphi} = \int_{\partial\mathcal{R}} \boldsymbol{\varphi}$. (a) El caso clásico (siglo XVII) $\int_a^b f'(x)\mathrm{d}x = f(b) - f(a)$, donde $\boldsymbol{\varphi} = f(x)$ y $\mathcal{R}$ es el segmento de una curva γ de a a b, parametrizada por x, de modo que $\partial\gamma$ consiste en los puntos extremos de γ, $x = a$ (que cuenta negativamente) y $x = b$ (positivamente). (b) El caso general, para una p-forma $\boldsymbol{\varphi}$, donde $\mathcal{R}$ es una región $(p + 1)$-dimensional compacta y orientada con frontera p-dimensional $\partial\mathcal{R}$.

para algún Φ. Este es un caso del importante *lema de Poincaré*,[11,12.13] que afirma que si una p-forma $\boldsymbol{\beta}$ satisface $\mathrm{d}\boldsymbol{\beta} = 0$, entonces *localmente* $\boldsymbol{\beta}$ tiene la forma $\boldsymbol{\beta} = \mathrm{d}\boldsymbol{\gamma}$, para alguna $(p - 1)$-forma $\boldsymbol{\gamma}$.

La derivada exterior se clarifica, y se hace explícita, mediante el uso de componentes. Consideremos una p-forma $\boldsymbol{\alpha}$. En una carta de coordenadas, con coordenadas $x^1, \ldots, x^n$, tenemos un conjunto antisimétrico de componentes $\alpha_{r\ldots t}$ $(= \alpha_{[r\ldots t]}$, donde $r, \ldots, t$ son p en número; véase §11.6) para representar $\boldsymbol{\alpha}$. Podemos escribir dicha representación

$$\boldsymbol{\alpha} = \sum \alpha_{r\ldots t}\, \mathrm{d}x^r \wedge \ldots \wedge \mathrm{d}x^t,$$

donde la suma (indicada por el símbolo Σ) se toma sobre todos los conjuntos de p números $r, \ldots, t$, cada uno de los cuales recorre el intervalo de valores $1, \ldots, n$. (Algunos prefieren evitar la redundancia en esta expresión que surge debido a que la antisimetría del producto cuña hace que cada término distinto de cero se repita $p!$ veces. Sin embargo, la notación funciona mucho mejor si vivimos sencillamente con esta redundancia, que es mi elección preferida.) La *derivada exterior* de la p-forma $\boldsymbol{\alpha}$ es una $(p + 1)$-forma que se escribe $\mathrm{d}\boldsymbol{\alpha}$, y que tiene componentes

[12.13] Suponiendo el resultado del ejercicio [12.11], demuestre el lema de Poincaré para $p = 1$.

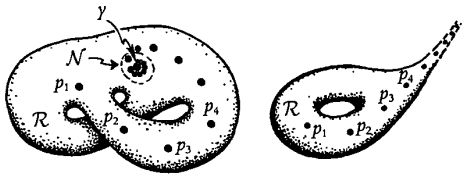


Fig 12.12. Compacticidad. (a) Un espacio compacto $\mathcal{R}$ tiene la propiedad de que cualquier secuencia infinita de puntos $p_1, p_2, p_3, \ldots$ en $\mathcal{R}$ debe acumularse finalmente en un punto y en $\mathcal{R}$, de modo que todo conjunto abierto $\mathcal{N}$ en $\mathcal{R}$ que contiene a y debe contener también (infinitos) miembros de la secuencia. (b) En un espacio no compacto esta propiedad falla.

$$(\mathrm{d}\boldsymbol{\alpha})_{qr\ldots t} = \frac{\partial}{\partial x^{[q}}\alpha_{r\ldots t]},$$

(La notación parece aquí algo complicada. La antisimetrización —que es la característica clave de la expresión— se extiende a todos los $p + 1$ índices, incluyendo el que aparece en el símbolo de derivada.)[12.14],[12.15]

Ahora estamos en condiciones de escribir el *teorema fundamental del cálculo exterior*. Este se expresa en la muy elegante (y potente) fórmula siguiente para una p-forma $\boldsymbol{\varphi}$ (véase la Fig. 12.11):

$$\int_{\mathcal{R}} \mathrm{d}\boldsymbol{\varphi} = \int_{\partial\mathcal{R}} \boldsymbol{\varphi},$$

donde $\mathcal{R}$ es una región (orientada) compacta $(p + 1)$-dimensional cuya frontera p-dimensional (en consecuencia, también compacta) está denotada por $\partial\mathcal{R}$.

Aquí he utilizado varios términos que todavía no he explicado. Para nuestros propósitos, «compacto» significa intuitivamente que la región $\mathcal{R}$ no «llega al infinito» y no tiene «agujeros recortados en ella» ni «trozos de su frontera eliminados». Con más precisión, una región *compacta* $\mathcal{R}$ es, para nuestros fines actuales,[12] una región con la propiedad de que cualquier secuencia infinita de puntos en $\mathcal{R}$ debe *acumularse* en algún punto dentro de $\mathcal{R}$ (Fig. 12.12a). Aquí, un *punto de acumulación* y tiene la propiedad de que todo conjunto abierto en $\mathcal{R}$ (véase §7.4) que contiene a y debe contener también miembros de la secuencia infinita (de modo que los puntos de la secuencia se aproximan cada vez más a y, sin límite). El plano euclídeo infinito no es compacto, pero la superficie de una esfera sí lo es, y también lo es el toro. También lo es el conjunto de puntos dentro o sobre el círculo unidad en el plano complejo (disco unidad cerrado); pero si eliminamos el propio círculo del conjunto, o incluso solo el centro del círculo, entonces el conjunto resultante no es compacto. Véase la Fig. 12.13.

El término «orientado» se refiere a la asignación de una «mano» consistente a cada punto de $\mathcal{R}$ (Fig. 12.14). En el caso de una 0-variedad, o conjunto de puntos *discretos*, la orientación asigna simplemente un «valor positivo» (+) o «negativo» (–) a cualquier punto (Fig. 12.14a). En el caso de una 1-variedad, o *curva*, esta orientación proporciona una «dirección» a lo largo de la curva. Esta puede representarse en un diagrama mediante la colocación de una «flecha» sobre la curva para indicar esta dirección (Fig. 12.14b). En el caso de una 2-variedad, la orientación puede representarse diagramáticamente por un minúsculo círculo o arco de círculo con una flecha (Fig. 12.14c); esta indica qué rotación de un vector tangente en un punto de la superficie se considera en la dirección positiva. En el caso de una 3-variedad, la orientación especifica qué tríada de vectores independientes en un punto va a considerarse «dextrógira» y cuál «levógira» (recuérdense §11.3 y la Fig. 11.1). Véase la Fig. 12.14d. Solo para espacios bastante inusuales no es posible asignar una orientación de forma consistente. Un ejemplo («no orientable») donde no puede hacerse es la *cinta de Moebius*, que se ilustra en la Fig. 12.15.

La *frontera* $\partial\mathcal{R}$ de una región $(p + 1)$-dimensional $\mathcal{R}$ (compacta y orientada) consiste en aquellos puntos de $\mathcal{R}$ que no están en su interior. Si $\mathcal{R}$ es apropiadamente no patológica, entonces $\partial\mathrm{R}$ es una región p-dimensional (compacta y orientada), aunque posiblemente vacía. *Su*

[12.14] Demuestre directamente que todos los «axiomas» para la derivada exterior son satisfechos por esta definición.

[12.15] Demuestre que esta definición de coordenadas da la misma cantidad $\mathrm{d}\boldsymbol{\alpha}$, cualquiera que sea la elección de coordenadas que se haya hecho, donde la transformación de las componentes $\alpha_{r\ldots t}$ de una forma está definida por el requisito de que la propia forma $\boldsymbol{\alpha}$ sea inalterada por el cambio de coordenadas. *Sugerencia*: Demuestre que esta transformación es idéntica a la transformación pasiva de componentes tensoriales $\binom{0}{p}$-valentes, como las que se dan en §13.8.

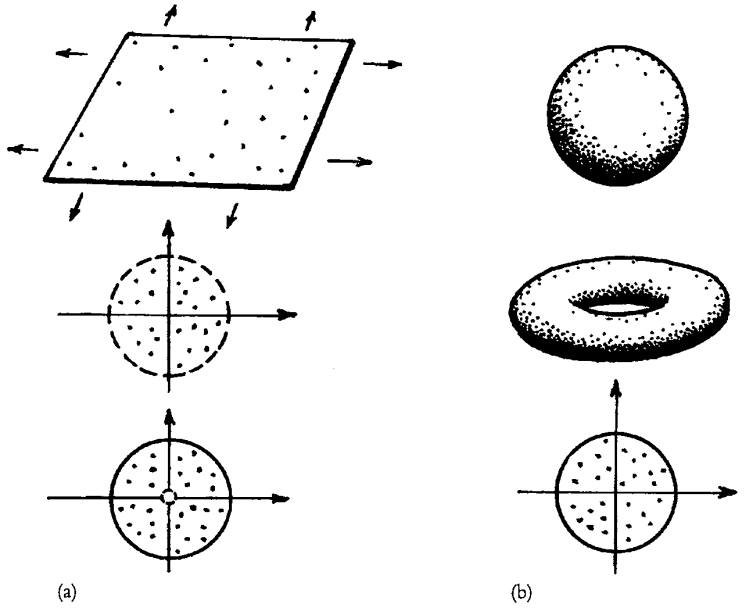


Fig. 12.13. (a) Algunos espacios no compactos: el plano euclídeo infinito, el disco unidad abierto y el disco unidad cerrado del que se ha eliminado el centro. (b) Algunos espacios compactos: la esfera, el toro y el disco unidad cerrado. (Las líneas frontera sólidas son parte del conjunto; las líneas frontera de trazos no lo son.)

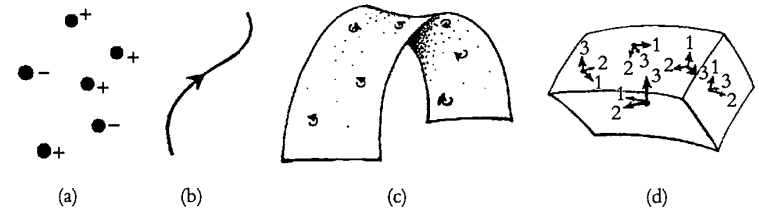


Fig. 12.14. Orientación. (a) Una 0-variedad (multicomponente) es un conjunto de puntos discretos; la orientación asigna simplemente un valor «positivo» (+) o «negativo» (–) a cada uno. (b) Para una 1-variedad, o curva, la orientación proporciona una «dirección» a lo largo de la curva, representada en un diagrama por la colocación de una flecha sobre ella. (c) Para una 2-variedad, la orientación puede indicarse por un minúsculo arco circular con una flecha, que indica la dirección de rotación «positiva» de un vector tangente. (d) Para una 3-variedad, la orientación especifica qué tríadas de vectores independientes en un punto deben considerarse «dextrógiras» (cf. Fig. 11.1).

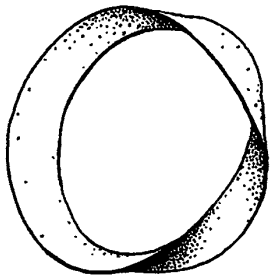


Fig. 12.15. La cinta de Moebius: un ejemplo de espacio no orientable.

frontera $\partial\partial\,\mathcal{R}$ es vacía. Así $\partial^2 = 0$, lo que complementa nuestra relación anterior $d^2 = 0$.

La frontera del disco unidad cerrado en el plano complejo es el círculo unidad; la frontera de la esfera unidad es vacía. La frontera de un cilindro finito (2-superficie cilíndrica) consiste en los dos círculos en sus extremos, pero la orientación de cada uno de ellos es opuesta a la del otro. La frontera de un segmento de línea finito consiste en sus dos puntos extremos, uno que cuenta positivamente y el otro negativamente. Véase la Fig. 12.16.[13] La versión original 1-dimensional del

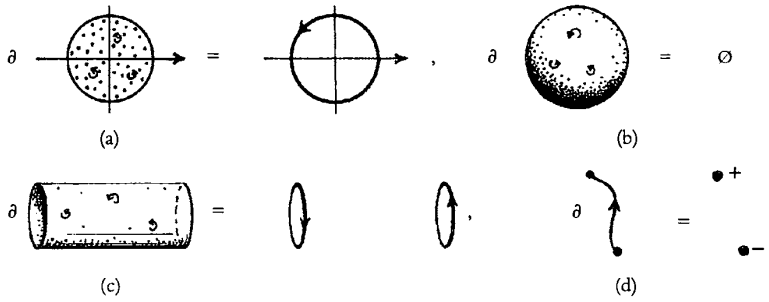


Fig. 12.16. La frontera $\partial\mathcal{R}$ de una región $\mathcal{R}$ $(p+1)$-dimensional orientada, compacta, y de buen comportamiento es una región p-dimensional (orientada y compacta, y posiblemente vacía), que consta de aquellos puntos de $\mathcal{R}$ que no yacen en el interior $(p+1)$-dimensional. (a) La frontera del disco unidad cerrado (dado por $|z| \leq 1$ en el plano complejo $\mathbb{C}$) es el círculo unidad. (b) La frontera de la esfera unidad es vacía ($\varnothing$ denota el conjunto vacío; véase §3.4). (c) La frontera de una longitud finita de superficie cilíndrica consiste en los dos círculos en ambos extremos, cuyas orientaciones son opuestas. (d) La frontera de un segmento de curva finito consiste en los dos puntos extremos, uno positivo y el otro negativo.

teorema fundamental del cálculo mostrada antes resulta un caso especial del teorema fundamental del cálculo exterior, cuando se toma como $\mathcal{R}$ tal segmento de línea.

12.7. El elemento de volumen; convenio de suma

Volvamos ahora a la distinción —y la relación— entre una p-forma y un $(n-p)$-vector, en una n-variedad $\mathcal{M}$. Para entender esta relación, es mejor ir primero al caso extremo en que $p = n$, de modo que estamos examinando la relación entre una n-forma y un campo escalar en $\mathcal{M}$. En el caso de una n-forma $\boldsymbol{\varepsilon}$, el elemento de n-superficie asociado en un punto o de $\mathcal{M}$ es precisamente el n-plano tangente entero en o. La *medida* que $\boldsymbol{\varepsilon}$ proporciona es simplemente una n-densidad, sin ninguna propiedad direccional. Una n-densidad semejante (que se supone que no es nula en ninguna parte) se conoce a veces como un *elemento de volumen* para la n-variedad $\mathcal{M}$. Un elemento de volumen puede utilizarse para convertir $(n-p)$-vectores en p-formas, y viceversa. (A veces hay un elemento de volumen asignado a una variedad, como parte de su «estructura» asignada; en tal caso, la distinción esencial entre una p-forma y una $(n-p)$-forma desaparece.)

¿Cómo podemos utilizar un elemento de volumen para convertir un $(n-p)$-vector en una p-forma? En términos de componentes, la n-forma $\boldsymbol{\varepsilon}$ estaría representada en cada carta por una cantidad con n subíndices antisimétricos:

$$\varepsilon_{r\ldots w}.$$

(Algunos preferirían incorporar a esto un factor $(n!)^{-1}$; véase §5.3 para la notación «$n!$». Sin embargo, no me voy a ocupar en los diversos factoriales complicados que aparecen aquí, pues nos distraerían de las ideas principales.) Podemos utilizar la cantidad $\varepsilon_{r\ldots w}$ para transformar la familia de componentes $\psi^{\mu\ldots w}$ de un $(n-p)$-vector $\boldsymbol{\psi}$ en la familia de componentes $\alpha_{r\ldots t}$ de una p-forma $\boldsymbol{\alpha}$. Lo hacemos aprovechando las operaciones del *álgebra tensorial*, a la que llegaremos con más detalle en la próxima sección. Esta álgebra nos permite «empalmar» los $n-p$ superíndices de $\psi^{\mu\ldots w}$ a $n-p$ de los n subíndices de $\varepsilon_{r\ldots w}$, lo que nos deja

con los p subíndices no ligados que necesitamos para $\alpha_{r\ldots t}$. La operación de «empalme» que interviene aquí es lo que se conoce como una «contracción» tensorial (o «transvección»), y permite emparejar un superíndice con un subíndice y «sumar» sobre los dos, de modo que desaparecen ambos índices de la expresión final.

El ejemplo arquetípico de esto es el *producto escalar* (§12.3), que combina las componentes β_r de un covector $\boldsymbol{\beta}$ con las componentes ξ^r de un vector $\boldsymbol{\xi}$, multiplicando los dos conjuntos de componentes y «sumando sobre» índices repetidos para obtener:

$$\boldsymbol{\beta} \bullet \boldsymbol{\xi} = \sum \beta_r \xi^r,$$

donde la suma se refiere al índice repetido r (uno arriba y otro abajo). Este procedimiento de suma se aplica también a cantidades pluriindexadas, y los físicos encuentran muy conveniente adoptar un convenio introducido por Einstein, conocido como *convenio de suma*. Este convenio consiste en omitir los símbolos de suma, dando por supuesto que se está haciendo una suma sobre un subíndice y un superíndice cuando quiera que la misma letra índice aparece en ambas posiciones en un término, haciendo que la suma recorra siempre los valores 1, ..., n del índice. En consecuencia, el producto escalar se escribirá ahora simplemente como:

$$\boldsymbol{\beta} \bullet \boldsymbol{\xi} = \beta_r \xi^r.$$

Utilizando este convenio, podemos escribir el procedimiento esbozado antes para expresar una p-forma en términos de un correspondiente $(n-p)$-vector y una forma volumen como

$$\alpha_{r\ldots t} \propto \varepsilon_{r\ldots tu\ldots w} \psi^{u\ldots w}$$

con contracción sobre los $n-p$ índices $u, \ldots, w$. Aquí estoy introduciendo el símbolo «$\propto$», que quiere decir «es proporcional a», lo que significa que cada lado es un múltiplo no nulo del otro. Esto se hace para que nuestras expresiones no se llenen confusamente con factoriales de apariencia complicada. Decimos a veces que el $(n-p)$-vector $\boldsymbol{\psi}$ y la p-forma $\boldsymbol{\alpha}$ son *duales*[14] uno de otro si se cumple esta relación (salvo un factor de proporcionalidad), en cuyo caso habrá también una correspondiente fórmula inversa

$$\psi^{u\ldots w} \propto \alpha_{r\ldots t}\, \epsilon^{r\ldots tu\ldots w}$$

para alguna apropiada forma de volumen (n-vector) $\boldsymbol{\epsilon}$ recíproca, a menudo «normalizada» frente a $\boldsymbol{\varepsilon}$ de acuerdo con

$$\boldsymbol{\varepsilon} \bullet \boldsymbol{\epsilon} = \varepsilon_{r\ldots w}\, \epsilon^{r\ldots w} = n!$$

(aunque aquí no estamos interesados en las cuestiones de normalización).

Estas fórmulas son parte del álgebra tensorial clásica (véase §12.8). Esta proporciona un potente procedimiento de manipulación (también extendido al cálculo tensorial, del que veremos más en el capítulo 14), que gana mucho del uso de una notación de índices combinada con el convenio de suma de Einstein. La notación de paréntesis cuadrados para la antisimetrización (véase §11.6) también desempeña un valioso papel en esta álgebra, como lo hace una notación adicional de paréntesis curvos para la *simetrización*

$$\psi^{(ab)} = \frac{1}{2}\,(\psi^{ab} + \psi^{ba}),$$

$$\psi^{(abc)} = \frac{1}{6}\,(\psi^{abc} + \psi^{acb} + \psi^{bca} + \psi^{bac} + \psi^{cab} + \psi^{cba})$$

etc.,

donde todos los signos menos que definen el paréntesis cuadrado están reemplazados por signos más.

Como un ejemplo más del valor de la notación de paréntesis, veamos cómo se escribe la condición de que una p-forma $\boldsymbol{\alpha}$ o un q-vector $\boldsymbol{\psi}$ sean *simples*, lo que significa que son el producto cuña de p 1-formas o de q vectores ordinarios individuales. En términos de componentes, esta condición resulta ser

$$\alpha_{[r\ldots t}\alpha_{u]v\ldots w} = 0 \quad \text{o} \quad \psi^{[r\ldots t}\psi^{u]v\ldots w} = 0,$$

donde todos los índices del primer factor están «recogidos» con solo un índice del segundo.[15] Si $\boldsymbol{\alpha}$ y $\boldsymbol{\psi}$ resultan ser duales uno del otro, entonces podríamos escribir una u otra condición alternativamente como

$$\psi^{r\ldots tu}\alpha_{uv\ldots w} = 0,$$

donde un único índice de $\boldsymbol{\psi}$ se contrae con un único índice de $\boldsymbol{\alpha}$. La simetría de esta expresión muestra que el dual de una p-forma simple es un $(n-p)$-vector simple, y recíprocamente.[12.16]

12.8. Tensores; notación de índices abstractos y notación diagramática

Aquí surge algo que a veces se ve como un conflicto entre «la notación del matemático» y «la notación del físico». Las dos notaciones están ejemplificadas por los dos miembros de la ecuación anterior $\boldsymbol{\beta} \bullet \boldsymbol{\xi} = \beta_r\xi^r$. La notación del matemático es manifiestamente independiente de las coordenadas, y vemos que la expresión $\boldsymbol{\beta} \bullet \boldsymbol{\xi}$ (para la que una notación tal como $(\boldsymbol{\beta}, \boldsymbol{\xi})$ o $\langle \boldsymbol{\beta}, \boldsymbol{\xi} \rangle$ podría ser más habitual en la literatura matemática) no hace referencia a ningún sistema de coordenadas, pues el producto escalar está definido en términos del todo geométrico-algebraicos. Por el contrario, la expresión $\beta_r\xi^r$ del físico hace referencia explícitamente a componentes en algún sistema de coordenadas. Estas componentes cambiarán cuando pasemos de una carta de coordenadas a otra; además, la notación depende del convenio de suma «cuestionable» (que está en conflicto con gran parte de la notación matemática estándar). Pese a todo, hay una gran flexibilidad en la notación del físico, en especial en la facilidad con que puede utilizarse para construir nuevas operaciones que no entran fácilmente dentro del alcance de las operaciones especificadas del matemático. Cálculos algo complicados (tales como el último par de fórmulas mostradas más arriba) son con frecuencia casi intratables si uno insiste en atenerse a expresiones libres de índices. Con frecuencia —cuando se necesita algún ingrediente esencial de cálculo en un argumento— los matemáticos puros se encuentran recurriendo (con cierto embarazo) a cálculos en «cartas de coordenadas», y en raras ocasiones utilizan el convenio de suma.

En mi opinión, este conflicto es básicamente artificial, y puede evi-

[12.16] Confirme la equivalencia de todas estas condiciones para la simplicidad; demuestre la suficiencia de $\alpha_{[r}\alpha_{u]v} = 0$ en el caso $p = 2$. (*Sugerencia*: Contraiga esta expresión con dos vectores.)

tarse efectivamente con un cambio de actitud. Cuando un físico utiliza una cantidad «ξ^a», normalmente tiene en mente la magnitud vectorial que he estado denotando por ξ, antes que su conjunto de componentes en algún sistema de coordenadas elegido arbitrariamente. Lo mismo se aplica a la magnitud «α_a», que se consideraría una 1-forma. De hecho, esta noción puede hacerse completamente rigurosa dentro del marco de lo que se ha dado en llamar la *notación de índices abstractos*.[16] En este esquema, los índices *no* representan uno de los $1, 2, \ldots, n$ referidos a algún sistema de coordenadas; en su lugar, son simplemente *marcadores abstractos* en cuyos términos se formula el álgebra. Esto nos permite retener las ventajas prácticas de la notación de índices sin la desventaja conceptual de tener que hacer referencia, de manera explícita o no, a un sistema de coordenadas. Además, la notación de índices abstractos resulta tener numerosas ventajas prácticas adicionales, particularmente en relación con los formalismos basados en espinores.[17]

Pese a todo, la notación de índices abstractos sigue adoleciendo del problema de que puede ser difícil distinguir detalles de suma importancia en una fórmula porque los índices tienden a ser pequeños y sus configuraciones precisas muy difíciles de discernir. Estas dificultades se reducen con la introducción de otra notación más para el álgebra tensorial que describiré brevemente a continuación. Se trata de la notación *diagramática*.

En primer lugar, deberíamos saber qué es realmente un *tensor*. En la notación de índices, un tensor se denota mediante una cantidad tal como

$$Q^{f\ldots h}_{a\ldots c},$$

que puede tener q subíndices y p superíndices, para cualesquiera p, $q \geq 0$, y no tiene por qué tener simetrías especiales. Llamamos a esto un tensor de *valencia*[18] $\left[{}^{p}_{q}\right]$ (o un tensor $\left[{}^{p}_{q}\right]$-valente o simplemente un $\left[{}^{p}_{q}\right]$-tensor). Algebraicamente, esto representaría una cantidad $\boldsymbol{Q}$ que puede considerarse como una función (de un tipo particular conocido como *multilineal*)[19] de q vectores $\boldsymbol{A}, \ldots, \boldsymbol{C}$ y p covectores $\boldsymbol{F}, \ldots, \boldsymbol{H}$, donde

$$\boldsymbol{Q}(\boldsymbol{A}, \ldots, \boldsymbol{C}; \boldsymbol{F}, \ldots, \boldsymbol{H}) = A^a \ldots C Q^{f\ldots h}_{a\ldots c} F_f \ldots H_h.$$

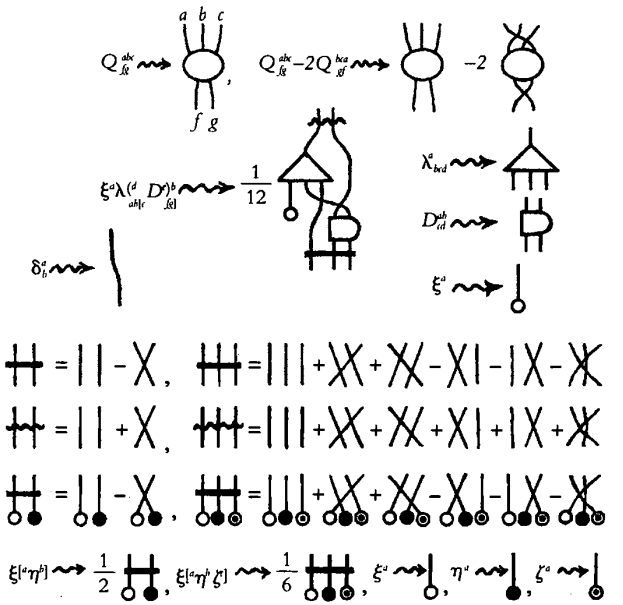


Fig. 12.17. Notación tensorial diagramática. El tensor $\left[{}^{3}_{2}\right]$-valente se representa por un óvalo con tres brazos y dos piernas, donde la imagen tensorial general $\left[{}^{p}_{q}\right]$-valente tendría p brazos y q piernas. En una expresión tal como $Q^{abc}_{fg} - 2Q^{bca}_{gf}$, la notación diagramática utiliza la posición en la página de los extremos de los brazos y piernas para seguir la pista de cuál es cada índice, en lugar de utilizar letras índice individuales. Las contracciones de índices tensoriales se representan por la unión de un brazo y una pierna, como se ilustra en el diagrama para $\xi^n \lambda^{(d}_{ab[c} D^{b)}_{fg]}$. Este diagrama ilustra también el uso de una barra gruesa que cruza líneas de índices para denotar antisimetrización y una barra ondulada para representar simitrización. El factor $\frac{1}{12}$ en el diagrama resulta del hecho de que (para facilitar los cálculos) el denominador factorial normal para simetrizadores y antisimetrizadores está omitido en la notación diagramática (de modo que aquí necesitamos $\frac{1}{2!} \times \frac{1}{3!} = \frac{1}{12}$). En la mitad inferior del diagrama, antisimetrizadores y simetrizadores se escriben como expresiones «incorpóreas» (mediante el uso de la representación diagramática de la delta de Kronecker δ^a_b que se introducirá en §13.3 y la Fig. 13.6c). Esta se utiliza entonces para expresar productos cuña (multivectoriales) $\boldsymbol{\xi} \wedge \boldsymbol{\eta}$ y $\boldsymbol{\xi} \wedge \boldsymbol{\eta} \wedge \boldsymbol{\zeta}$.

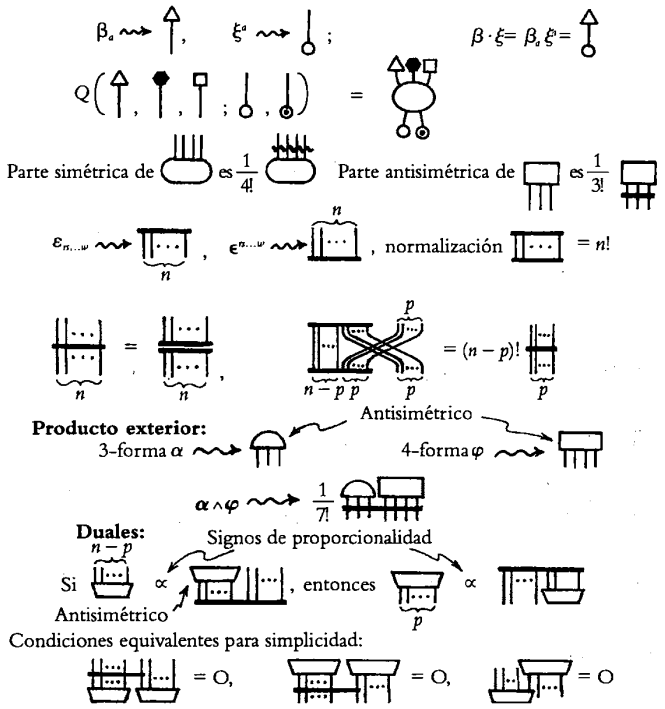


Fig. 12.18. Más notación tensorial diagramática. El diagrama para un covector $\boldsymbol{\beta}$ (1-forma) tiene una única pierna que cuando se une al único brazo de un vector $\boldsymbol{\xi}$ da su producto escalar. Con más generalidad, la forma multilineal definida por un tensor $Q\,[^{p}_{q}]$-valente se representa uniendo los p brazos a las piernas de p covectores variables y las q piernas a los brazos de q vectores variables (aquí $p = 3$ y $q = 2$). Las partes simétrica y antisimétrica de tensores generales pueden expresarse utilizando las líneas onduladas y barras gruesas de las operaciones de la Fig. 12.17. Además, la notación de barras se combina con una notación diagramática relacionada para la n-forma volumen $\boldsymbol{\varepsilon}_{n\ldots w}$ (para un espacio n-dimensional) y su n-vector dual $\boldsymbol{\epsilon}^{n\ldots w}$, normalizado de acuerdo con $\boldsymbol{\varepsilon}_{n\ldots w}\boldsymbol{\epsilon}^{n\ldots w} = n!$ También se expresan relaciones equivalentes a $n!\delta^{a}_{[r}\delta^{b}_{s}\ldots\delta^{f}_{w]} = \boldsymbol{\epsilon}^{ab\ldots f}\boldsymbol{\varepsilon}_{rs\ldots w}$ (n índices antisimetrizados) y $\boldsymbol{\varepsilon}_{a\ldots cu\ldots w}\boldsymbol{\epsilon}^{a\ldots ce\ldots f} = p!(n-p)!\delta^{e}_{[u}\ldots\delta^{f}_{w]}$ (véanse §13.3 y la Fig. 13.6c). Los productos exteriores de formas, la «dualidad» entre p-formas y $(n-p)$-vectores, y las condiciones para «simplicidad» se representan diagramáticamente de forma sucinta. (Para diagramas de derivadas exteriores, véase la Fig. 14.18.)

En la notación diagramática, el tensor $\boldsymbol{Q}$ estaría representado por un símbolo distintivo (digamos un rectángulo o un triángulo o un óvalo, según la conveniencia) al que están unidas p líneas que se extienden hacia abajo (las «piernas») y q líneas que se extienden hacia arriba (los «brazos»). En cualquier término de una expresión tensorial, los diversos elementos que están multiplicados se dibujan en algún tipo de yuxtaposición, pero no necesariamente ordenados linealmente en la página. Para cualesquiera dos índices que se contraen juntos, las líneas deben estar conectadas, superior con inferior. Algunos ejemplos se ilustran en las Figs. 12.17 y 12.18, incluyendo ejemplos de varias de las fórmulas que acabamos de encontrar. Como parte de esta notación se dibuja una barra que cruza líneas de índices para denotar la antisimetrización, lo que refleja el paréntesis cuadrado de la notación de índices (aunque resulta conveniente adoptar un convenio diferente con respecto a multiplicadores factoriales). En correspondencia, una barra «ondulada» refleja la simetrización. Aunque la notación diagramática es difícil de imprimir en un dispositivo de impresión ordinario, puede ser muy conveniente en muchos cálculos privados. ¡Yo mismo la he estado utilizando durante más de cincuenta años![20]

12.9. Variedades complejas

Finalmente, volvamos a la cuestión de las variedades complejas que hemos abordado en el capítulo 10. Cuando consideramos que una superficie de Riemann es 1-dimensional, estamos pensando solo en términos de operaciones holomorfas que se realizan sobre números complejos. Podemos adoptar exactamente la misma postura con variedades de dimensiones más altas, considerando ahora que nuestras coordenadas $x^1, \ldots, x^n$ son números complejos $z^1, \ldots, z^n$ y nuestras funciones de las mismas son funciones holomorfas. Consideramos de nuevo que nuestra variedad consiste en varias cartas de coordenadas «pegadas», y cada carta es ahora una región abierta del espacio de coordenadas $\mathbb{C}^n$, el espacio cuyos puntos son las n-tuplas $(z^1, z^2, \ldots, z^n)$ de números complejos (y recordemos de §10.2 que «$\mathbb{C}$» por sí mismo representa al sistema de los números complejos). Las *funciones de transición*, que expresan las trans-

formaciones de coordenadas cuando pasamos de una carta de coordenadas a otra, tienen que estar dadas ahora enteramente por *funciones holomorfas*. Podemos definir campos vectoriales holomorfos, covectores holomorfos, p-formas holomorfas, tensores holomorfos, etc., exactamente de la misma forma que hemos hecho antes en el caso de una n-variedad real.

Pero está también el punto de vista filosófico alternativo según el cual podríamos expresar todas nuestras coordenadas complejas en términos de sus partes real e imaginaria $z^j = x^j + \mathrm{i}y^j$ (o, equivalentemente, incluir la noción de conjugación compleja en nuestra categoría de función aceptable, de modo que las operaciones ya no tengan que ser exclusivamente holomorfas; véase §10.1). Entonces, nuestra «n-variedad compleja» ya no debe verse como un espacio n-dimensional, sino que debe verse, en su lugar, como una $2n$-variedad *real*. Por supuesto, es una $2n$-variedad con un tipo muy particular de estructura local, conocido como una *estructura compleja*.

Hay varias maneras de formular esta idea. En esencia, lo que se necesita es una versión en dimensión más alta de las ecuaciones de Cauchy-Riemann (§10.5), pero las cosas se expresan normalmente de una forma algo diferente. Consideremos la relación entre campos vectoriales complejos y campos vectoriales reales en la variedad. Podemos considerar que un campo vectorial complejo $\boldsymbol{\zeta}$ está representado de la forma

$$\boldsymbol{\zeta} = \boldsymbol{\xi} + \mathrm{i}\boldsymbol{\eta},$$

donde $\boldsymbol{\xi}$ y $\boldsymbol{\eta}$ son campos vectoriales reales ordinarios en la $2n$-variedad. Lo que la «estructura compleja» hace por nosotros es decirnos cómo tienen que estar relacionados entre sí estos campos vectoriales reales y qué ecuaciones diferenciales deben satisfacer para que $\boldsymbol{\zeta}$ pueda calificarse de «holomorfa». Consideremos ahora el nuevo campo vectorial complejo que aparece cuando el campo complejo $\boldsymbol{\zeta}$ se multiplica por i. Vemos que, por consistencia, debemos tener $\mathrm{i}\boldsymbol{\zeta} = -\boldsymbol{\eta} + \mathrm{i}\boldsymbol{\xi}$, de modo que el campo vectorial real $\boldsymbol{\xi}$ queda ahora reemplazado por $-\boldsymbol{\eta}$ y análogamente $\boldsymbol{\eta}$ debe ser reemplazado por $\boldsymbol{\xi}$. La operación $\boldsymbol{J}$ que efectúa estos reemplazamientos (i.e., $\boldsymbol{J}(\boldsymbol{\xi}) = -\boldsymbol{\eta}$ y $\boldsymbol{J}(\boldsymbol{\eta}) = \boldsymbol{\xi}$) es la que normalmente se conoce como la «estructura compleja».

Advertimos que si $\boldsymbol{J}$ se aplica dos veces, simplemente invierte el signo de aquello sobre lo que actúa (puesto que $\mathrm{i}^2 = -1$), de modo que podemos escribir:

$$\boldsymbol{J}^2 = -1.$$

Esta sola condición define lo que se conoce como una *estructura cuasicompleja*. Para particularizar esto a una estructura compleja real, de modo que pueda aparecer una noción consistente de «holomorfa» para la variedad, debe satisfacerse cierta ecuación diferencial[21] en la cantidad $\boldsymbol{J}$. Existe un teorema llamado de Newlander-Nirenberg[22] que nos dice que esto es *suficiente* (además de necesario) para que una variedad real $2n$-dimensional con esta $\boldsymbol{J}$-estructura pueda reinterpretarse como una n-variedad compleja. Este teorema nos permite movernos libremente entre los dos puntos de vista con respecto a las variedades complejas.

Notas

Sección 12.1

12.1. Esta «contractibilidad» se toma en el sentido de *homotopía* (véanse §7.2 y la Fig. 7.2), de modo que no se permite la «cancelación» de segmentos del lazo con orientación opuesta; esta múltiple conectividad es parte de la teoría de homotopía. Véanse Huggett y Jordan (2001) y Sutherland (1975).

12.2. Estrictamente hablando, este argumento es incompleto, puesto que no he presentado ninguna razón convincente por la que la 2π-torsión del cinturón no puede deshacerse de forma continua si los extremos se mantienen fijos.[12.17] Véase Penrose y Rindler (1984), pp. 41-44.

[12.17] Representando una rotación en el 3-espacio euclídeo ordinario como un vector que apunta a lo largo del eje de rotación y de longitud igual al ángulo de rotación, demuestre que la topología de $\mathcal{R}$ puede describirse como una bola maciza (de radio π) acotada por una esfera ordinaria, donde cada punto de la esfera se identifica con su punto antipodal. Dé un argumento directo para demostrar por qué un lazo cerrado que representa una 2π-rotación no puede ser deformado de forma continua hasta un punto.

12.3. Aquí tratamos las moléculas como partículas puntuales. La dimensión de $\mathcal{P}$ sería considerablemente mayor para moléculas con grados de libertad internos o rotacionales.

Sección 12.2

12.4. La noción usual de «variedad» presupone que nuestro espacio $\mathcal{M}$ es, en primera instancia, un *espacio topológico*. Asignar una *topología* a un espacio $\mathcal{M}$ es especificar exactamente cuáles de sus conjuntos de puntos van a llamarse «abiertos» (cf. §7.4). Los conjuntos abiertos deben tener la propiedad de que la intersección de dos cualesquiera de ellos es un conjunto abierto y la unión de cualquier número de ellos (finito o infinito) es de nuevo un conjunto abierto. Además de la condición de Hausdorff mencionada en el texto, es habitual exigir que la topología de $\mathcal{M}$ tenga algunas otras restricciones, y muy particularmente que satisfaga un requisito denominado «paracompacticidad». Para el significado de este y otros términos relacionados, el lector interesado puede consultar Kelley (1965), Engelking (1968) y cualquier otro texto estándar sobre topología general. Para nuestros fines actuales, basta con suponer que $\mathcal{M}$ está construida a partir de un atlas localmente finito de regiones abiertas de $\mathbb{R}^n$, donde «localmente finito» significa que cada carta es intersectada por solo un número finito de otras cartas.

Un requisito final que se introduce a veces en la definición de una variedad es que sea *conexa*, lo que significa que consta solo de «una pieza» (que aquí puede tomarse con el significado de que no es una unión disjunta de dos conjuntos abiertos no vacíos). No insistiré aquí en esto; si se requiere conectividad, entonces será establecido explícitamente (pero en cualquier caso la desconectividad solo se admitirá para un número finito de partes separadas).

12.5. Véanse, por ejemplo, Kobayashi y Nomizu (1963), Hicks (1965), Lang (1972) y Hawking y Ellis (1973). Un procedimiento interesante para definir una variedad $\mathcal{M}$ es reconstruir la propia $\mathcal{M}$ simplemente a partir del álgebra conmutativa de campos escalares definidos sobre $\mathcal{M}$; véanse Chevalley (1946), Nomizu (1956) y Penrose y Rindler (1984). Este tipo de idea generaliza las álgebras no conmutativas y conduce a la noción de «geometría no conmutativa» de Alain Connes (1994), que proporciona una de las aproximaciones modernas a la «geometría espaciotemporal cuántica» (véase §33.1).

Sección 12.3

12.6. Véanse Helgason (2001) y Frankel (2001).

12.7. La condición general para que la familia de elementos $(n-1)$-planos definidos por una 1-forma $\boldsymbol{\alpha}$ toque a una familia 1-paramétrica de $(n-1)$-superficies (de modo que $\boldsymbol{\alpha} = \lambda \mathrm{d}\Phi$ para ciertos campos escalares λ, Φ) es la *condición de Frobenius* $\boldsymbol{\alpha} \wedge \mathrm{d}\boldsymbol{\alpha} = 0$; véase Flanders (1963).

12.8. Fácilmente surge la confusión entre la idea «clásica» de que un objeto como «dx» debería representar un desplazamiento infinitesimal (vector), mientras que aquí parece que estamos viéndolo como un covector. De hecho, la notación es consistente, ¡pero hace falta tener la cabeza clara para verlo! La cantidad $\mathrm{d}x^r$ parece tener un carácter vectorial debido a su superíndice r, y este sería realmente el caso si se trata a r como un índice abstracto, de acuerdo con §12.8. Por el contrario, si se toma r como un índice numérico, digamos $r = 2$, entonces obtenemos un covector, a saber $\mathrm{d}x^2$, el gradiente de la cantidad escalar $y = x^2$ («x dos», y no «x al cuadrado»). Pero esto depende de interpretar «d» como representación del gradiente en lugar de denotar un infinitesimal, como se hubiera hecho en la tradición clásica. De hecho, si tratamos *a la vez* la r como abstracta y la d como gradiente, ¡entonces «$\mathrm{d}x^r$» representa simplemente la delta de Kronecker (abstracta)!

Sección 12.5

12.9. Esto representa un cambio de actitud con respecto al punto de vista «infinitesimal» acerca de las cantidades como «dx». Aquí las propiedades de anticonmutación de «$\mathrm{d}x\wedge\mathrm{d}y$» nos dicen que estamos operando con densidades con respecto a medidas de área orientadas.

12.10. Un nombre que me sugirió N.M.J. Woodhouse. A veces este teorema se denomina simplemente *teorema de Stokes*. Sin embargo, ello parece particularmente inapropiado, puesto que la única contribución de Stokes fue ponerlo como una pregunta de examen (en Cambridge) que aparentemente obtuvo de William Thompson (lord Kelvin).

Sección 12.6

12.11. Véase Flanders (1963). (En dicho libro, lo que yo he llamado «lema de Poincaré» es mencionado como el *inverso*.)

12.12. Existe una definición más general de compacidad de un espacio topológico, que, sin embargo, no es tan intuitiva como la que se da en el texto. Un espacio $\mathcal{R}$ es compacto si para cada manera en que puede

expresarse como una unión de conjuntos abiertos, existe una colección finita de dichos conjuntos cuya unión sigue siendo $\mathcal{R}$.

12.13. Para más información sobre estas cuestiones, véase Willmore (1959).

Sección 12.7

12.14. Esta noción de «dual» es bastante diferente de la que hace que un covector sea «dual» de un vector, como se ha descrito en §12.3. Sin embargo, está estrechamente relacionada con otro concepto de «dualidad»: el *dual de Hodge*. Este desempeña un papel en electromagnetismo (véase §19.2), y versiones del mismo tienen importancia en varias aproximaciones a la gravedad cuántica (véanse §§31.5,14, §32.2, §§33.11,12) y física de partículas (véase §25.8). Por desgracia, este es solo un lugar entre muchos, donde las limitaciones de la terminología matemática pueden crear confusión.

12.15. Véase Penrose y Rindler (1984), pp. 165 y 166.

Sección 12.8

12.16. Véanse Penrose (1968a), pp. 135-141; Penrose y Rindler (1984), pp. 68-103, y Penrose (1971).

12.17. Véanse Penrose (1968a), Penrose y Rindler (1984, 1986), Penrose (1971) y O'Donnell (2003).

12.18. A veces se utiliza el término *rango* para el valor de $p + q$, pero esto es confuso debido a que «rango» tiene otro significado independiente en conexión con las matrices; véanse la nota 13.10 y §13.8.

12.19. Esto significa lineal en cada uno de los $\boldsymbol{A}, \ldots, \boldsymbol{C}; \boldsymbol{F}, \ldots, \boldsymbol{H}$ por separado; véanse también §§13.7-10.

12.20. Véanse Penrose y Rindler (1984), apéndice, Penrose (1971b) y Cvitanović y Kennedy (1982).

Sección 12.9

12.21. Esto es la anulación de una expresión llamada «el tensor de Nijenhuis construido a partir de $\boldsymbol{J}$», que podemos expresar como $J^d_{[a}\partial J^c_{b]}/\partial x^d + J^c_{[a}\partial J^d_{b]}/\partial x^{b]} = 0$.

12.22. Newlander y Nirenberg (1957).

13

Grupos de simetría

13.1. Grupos de transformaciones

Los espacios que son simétricos tienen una importancia fundamental en la física moderna. ¿Por qué razón? Cabría pensar que la simetría completamente exacta es algo que solo aparecería de forma excepcional, o quizá tan solo como una aproximación conveniente. Aunque un objeto simétrico, tal como un cuadrado o una esfera, tiene una existencia precisa como una estructura matemática idealizada («platónica»; véase §1.3), cualquier realización *física* de un objeto semejante sería considerada únicamente como un tipo de representación aproximada de este ideal platónico, que, por consiguiente, no posee ninguna simetría real que pueda considerarse exacta. Pese a todo, y curiosamente, según las teorías físicas de gran éxito del siglo XX, todas las interacciones físicas, incluida la gravedad, actúan de acuerdo con una idea que, estrictamente hablando, depende de forma crucial de ciertas estructuras físicas que poseen una simetría que, en un nivel de descripción fundamental, ¡es necesariamente exacta!

¿Cuál es esta idea? Es un concepto que ha llegado a conocerse como una «conexión gauge». Dicho nombre dice muy poco, pero la idea es importante, y nos permite encontrar una noción sutil («retorcida») de diferenciación que se aplica a entidades generales en una variedad (entidades que son en realidad más generales que aquellas —las p-formas— que están sujetas a una diferenciación exterior, como se ha descrito en el capítulo 12). Estas cuestiones constituirán el tema de nuestros dos próximos capítulos; pero como prerrequisito debemos explorar antes la noción básica de un *grupo de simetría*. Esta noción tam-

bién tiene muchas otras áreas de aplicación importantes en física, química y cristalografía, y asimismo dentro de muchas áreas diferentes de las propias matemáticas.

Tomemos un ejemplo sencillo. ¿Cuáles son las simetrías de un *cuadrado*? La pregunta tiene dos respuestas diferentes dependiendo de si admitimos o no simetrías que invierten la orientación del cuadrado (i.e., para las que el cuadrado se vuelve del revés). Consideremos primero el caso en el que no se permiten estas simetrías que invierten la orientación. Entonces, las simetrías del cuadrado se generan a partir de una simple rotación de un ángulo recto en el plano del cuadrado, repetida varias veces. Por conveniencia, podemos representar estos movimientos en términos de números complejos, como hemos hecho en el capítulo 5. Si queremos, podemos considerar que los vértices del cuadrado ocupan los puntos 1, i, −1, −i en el plano complejo (Fig. 13.1a), y nuestra notación básica está representada por la multiplicación por i (i.e., por «i×»). Las diversas *potencias* de i representan todas nuestras rotaciones, de las que hay cuatro diferentes en total:

$$i^0 = 1, \quad i^1 = i, \quad i^2 = -1, \quad i^3 = -i$$

(Fig. 13.1b). La cuarta potencia $i^4 = 1$ nos lleva de nuevo al punto de partida, de modo que ya no tenemos más elementos. El producto de dos cualesquiera de estos cuatro elementos es de nuevo uno de ellos.

Estos cuatro elementos nos proporcionan un ejemplo sencillo de un *grupo*. Este consiste en un conjunto de elementos y una «ley de multiplicación» definida entre los pares de ellos (denotada por la yuxtaposición de símbolos) para la que se cumple la ley *asociativa* de la multiplicación

$$a(bc) = (ab)c,$$

donde hay un elemento *identidad*, 1, que satisface

$$1a = a1 = a,$$

y donde cada elemento a tiene un *inverso* a^{-1}, tal que[13.1]

[13.1] Demuestre que si suponemos simplemente que $1a = a$ y $a^{-1}a = 1$ para todo a, junto con asociatividad $a(bc) = (ab)c$, entonces puede *deducirse* $a1 = a$ y $aa^{-1} = 1$. (*Sugerencia*: Por supuesto, a no es el único elemento del que se afirma que tiene un inverso.) Demuestre por qué, por el contrario, $a1 = a$, $a^{-1}a = 1$, y $a(bc) = (ab)c$ son insuficientes.

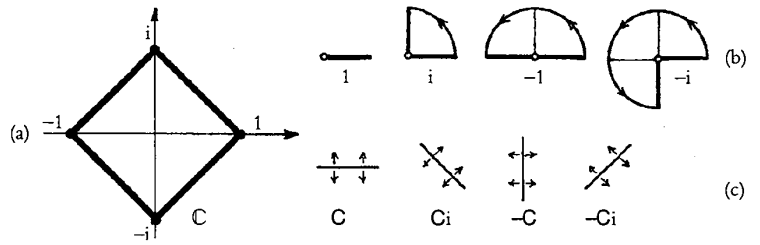


Fig. 13.1. Simetría de un cuadrado. (a) Podemos representar los vértices de un cuadrado por los puntos 1, i, −1, −i en el plano complejo ℂ. (b) El grupo de simetrías no reflexivas se representa, en ℂ, como multiplicación por $1 = i^0$, $i = i^1$, $-1 = i^2$, $-i = i^3$, respectivamente. (c) Las simetrías reflexivas están dadas, en C, por C (conjugación compleja), Ci, −C y −Ci.

$$a^{-1}a = aa^{-1} = 1.$$

Las operaciones de simetría que transforman un objeto en sí mismo satisfacen siempre dichas leyes, llamadas *axiomas del grupo*.

Recordemos los convenios recomendados en el capítulo 11, donde considerábamos que, en el producto ab, b actúa primero y a después. Podemos considerarlas como operaciones que se realizan sobre cierto objeto que aparece a su derecha. Así pues, podríamos considerar el movimiento, b, que expresa una simetría de un objeto $\boldsymbol{\Phi}$, como $\boldsymbol{\Phi} \mapsto b(\boldsymbol{\Phi})$, que iría seguido por otro movimiento semejante a, dando $b(\boldsymbol{\Phi}) \mapsto a(b(\boldsymbol{\Phi}))$. Esto da como resultado la acción combinada $\boldsymbol{\Phi} \mapsto ab(\boldsymbol{\Phi})$, que podemos escribir simplemente $\boldsymbol{\Phi} \mapsto ab(\boldsymbol{\Phi})$, correspondiente al movimiento ab. La operación identidad simplemente deja el objeto inalterado (lo que, evidentemente, es siempre una simetría) y el inverso es sencillamente la operación inversa de una simetría dada, que devuelve el objeto a su posición de partida.

En nuestro ejemplo particular de las rotaciones no reflexivas del cuadrado, tenemos la propiedad *conmutativa* adicional

$$ab = ba.$$

Los grupos que son conmutativos en este sentido se denominan *abelianos*, por el matemático noruego Niels Henrik Abel, que murió muy joven en trágicas circunstancias.[1] Evidentemente, cualquier gru-

po que pueda representarse simplemente por la multiplicación de números complejos debe ser abeliano (puesto que la multiplicación de números complejos individuales siempre conmuta). Hemos visto otros ejemplos de esto al final del capítulo 5, cuando hemos considerado el caso general de un grupo finito *cíclico* $\mathbb{Z}_n$, generado por una única raíz n-ésima de la unidad.[13.2]

Admitamos ahora las *reflexiones* que invierten la orientación de nuestro cuadrado. Podemos seguir usando la representación anterior del cuadrado en términos de números complejos, pero necesitaremos una nueva operación, que denoto por C, a saber, la *conjugación compleja*. (Esta hace que el cuadrado dé la vuelta sobre una línea horizontal; véanse §10.1 y la Fig. 10.1.) Ahora encontramos (véase la Fig. 13.1c) las leyes de «multiplicación»[13.3]

$$\mathrm{Ci} = (-\mathrm{i})\mathrm{C}, \quad \mathrm{C}(-1) = (-1)\mathrm{C}, \quad \mathrm{C}(-\mathrm{i}) = \mathrm{iC}, \quad \mathrm{CC} = 1$$

(donde[2] en lo sucesivo escribiré (–i)C como –iC, etc.). De hecho, podemos obtener las leyes de multiplicación para todo el grupo simplemente a partir de las relaciones básicas[13.4]

$$\mathrm{i}^4 = 1, \quad \mathrm{C}^2 = 1, \quad \mathrm{Ci} = \mathrm{i}^3\mathrm{C},$$

y ahora el grupo no es abeliano, como queda de manifiesto en la última ecuación. El número total de elementos diferentes en un grupo se denomina *orden* del grupo. El orden de este grupo concreto es 8.

Consideremos ahora otro ejemplo sencillo, a saber, el grupo de las simetrías rotacionales de una esfera ordinaria. Como antes, podemos considerar primero el caso en el que están excluidas las reflexiones. Esta vez, nuestro grupo de simetría tendrá un número infinito de elementos, porque podemos rotar alrededor de cualquier eje en el 3-espacio. El grupo de simetría constituye en realidad un espacio 3-dimensional, a saber, la 3-variedad denotada por $\mathcal{R}$ en el capítulo 12.

[13.2] Explique por qué cualquier espacio vectorial es un grupo abeliano —llamado un grupo abeliano *aditivo*—, donde la operación de «multiplicación» en el grupo es la operación «suma» en el espacio vectorial.
[13.3] Verifique estas relaciones teniendo en cuenta que Ci representa «la operación i×, seguida de la operación C, etc. (*Sugerencia*: Puede comprobar las relaciones confirmando simplemente sus efectos sobre 1 e i. ¿Por qué?)
[13.4] Demuéstrelo.

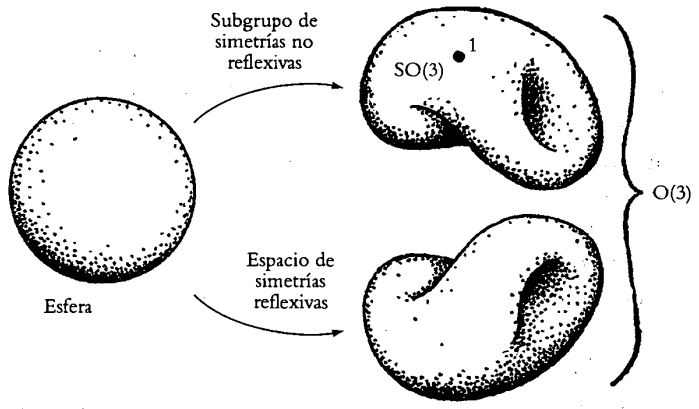


Fig. 13.2. Simetría rotacional de una esfera. El grupo de simetría entero, O(3), es una 3-variedad inconexa, que consiste en dos piezas. La componente que contiene el elemento identidad 1 es el subgrupo (normal) SO(3) de simetrías no reflexivas de la esfera. La componente restante es la 3-variedad de simetrías reflexivas.

Permítanme dar ahora a este grupo (3-variedad) su nombre oficial. Se denomina SO(3),[3] el grupo ortogonal no reflexivo en 3 dimensiones. Si ahora incluimos las reflexiones, entonces obtenemos todo un nuevo conjunto de simetrías —que vale por otra 3-variedad— que están desconectadas de las del primer grupo, a saber, aquellas que implican una inversión de la orientación de la esfera. La familia entera de los elementos del grupo constituye de nuevo una 3-variedad, pero ahora es una 3-variedad no conexa, consistente en dos piezas conexas separadas (véase la Fig. 13.2). Este grupo espacial entero se denomina O(3).

Estos dos ejemplos ilustran dos de las categorías más importantes de grupos: los grupos finitos y los grupos continuos (o grupos de *Lie*; véase §13.6).[4] Aunque existe una gran diferencia entre estos dos tipos de grupos, hay muchas propiedades importantes que son comunes a ambos.

13.2. Subgrupos y grupos simples

De particular importancia es la noción de *subgrupo* de un grupo. Para mostrar un subgrupo, seleccionamos alguna colección de elementos

dentro del grupo que por sí mismos forman un grupo, utilizando las mismas operaciones de multiplicación e inversión que en el grupo entero. Los subgrupos son importantes en muchas teorías modernas de física de partículas. Se supone que existe alguna simetría fundamental que relaciona diferentes tipos de partículas entre sí y también relaciona diferentes interacciones entre partículas. Pese a todo, no se puede ver que este grupo completo actúe como una simetría de ninguna forma manifiesta; en su lugar, uno encuentra que esta simetría está «rota» en algún subgrupo del grupo original y es el *subgrupo* el que desempeña un papel manifiesto como una simetría. Así pues, es importante conocer cuáles son realmente los posibles subgrupos de un presunto grupo de simetría «fundamental», para que aquellas simetrías que son manifiestas en la naturaleza puedan considerarse como subgrupos de este presunto grupo. Abordaré cuestiones de este tipo en §§25.5-8, §26.11 y §28.1.

Examinemos algunos casos particulares de subgrupos, para los ejemplos que hemos estado considerando. Las simetrías *no reflexivas* del cuadrado constituyen un subgrupo de 4 elementos $\{1, i, -1, -i\}$ del grupo completo de simetrías del cuadrado que tiene 8 elementos. Análogamente, el grupo de rotaciones no reflexivas SO(3) constituye un subgrupo del grupo completo O(3). Otro subgrupo de las simetrías del cuadrado consta de los 4 elementos $\{1, -1, C, -C\}$; pero otro tiene solo los 2 elementos $\{1, -1\}$.[13.5] Además, está siempre el subgrupo «trivial», que consiste solo en la identidad $\{1\}$ (y el propio grupo completo es, de forma igualmente trivial, siempre un subgrupo).

Todos los diferentes subgrupos que acabo de describir tienen una propiedad especial de particular importancia. Son ejemplos de lo que se denominan subgrupos *normales*. La importancia de un subgrupo normal reside en que, en un sentido apropiado, la acción de cualquier elemento del grupo completo conserva cualquier subgrupo normal, o, dicho de forma más técnica, cada elemento del grupo completo *conmuta* con el subgrupo normal. Voy a ser más explícito. Llamaremos $\mathcal{G}$ al grupo completo y $\mathcal{S}$ al subgrupo. Si selecciono cualquier elemento

[13.5] Verifique que todos los que aparecen en este párrafo son subgrupos (y tenga en cuenta la nota 13.4).

particular g del grupo $\mathcal{G}$, entonces puedo denotar por $\mathcal{S}g$ el conjunto consistente en todos los elementos de $\mathcal{S}$ multiplicados cada uno de ellos individualmente por g a la derecha (lo que se denomina *posmultiplicado* por g). Así pues, en el caso del subgrupo particular $\mathcal{S} = \{1, -1, C, -C\}$, del grupo de simetrías del cuadrado, si escogemos $g = i$, obtenemos $\mathcal{S}i = \{i, -i, Ci, -Ci\}$. Análogamente, la notación $g\mathcal{S}$ denotará el conjunto consistente en todos los elementos de $\mathcal{S}$ multiplicados cada uno de ellos individualmente por g por la izquierda (*premultiplicados* por g). Así pues, en nuestro ejemplo, tenemos ahora $i\mathcal{S} = \{i, -i, iC, -iC\}$. La condición para que $\mathcal{S}$ sea un subgrupo normal de $\mathcal{G}$ es que estos dos conjuntos sean iguales, i.e.,

$$\mathcal{S}g = g\mathcal{S}, \quad \text{para todo } g \text{ perteneciente a } \mathcal{S}.$$

En nuestro ejemplo particular, vemos que este es realmente el caso (puesto que $Ci = -iC$ y $-Ci = iC$), donde debemos tener en cuenta que la colección de objetos dentro de los corchetes debe tomarse como un conjunto *no ordenado* (de modo que no importa que cuando se escriben $\mathcal{S}i$ e $i\mathcal{S}$ explícitamente los elementos $-iC$ e iC aparezcan en orden inverso en la colección de elementos).

Podemos mostrar un subgrupo *no* normal del grupo de simetrías del cuadrado, que es el subgrupo de dos elementos $\{1, C\}$. Es no normal porque $\{1, C\}i = \{i, Ci\}$, mientras que $i\{1, C\} = \{i, -Ci\}$. Nótese que este subgrupo aparece como el nuevo grupo de simetría (reducido) si marcamos nuestro cuadrado con una flecha horizontal que apunte hacia la derecha (véase la Fig. 13.3a). Podemos obtener otro subgrupo no normal, a saber, $\{1, Ci\}$, si en su lugar marcamos el cuadrado con una flecha que apunte en diagonal hacia abajo y a la derecha (Fig. 13.3b).[13.6] En el caso de O(3), resulta que solamente hay un grupo normal no trivial, a saber, el SO(3),[13.7] pero hay muchos subgrupos no normales. Ejemplos no normales se obtienen si seleccionamos algún conjunto finito de puntos apropiado en la esfera, y buscamos las simetrías de la esfera con estos puntos marcados. Si marcamos

[13.6] Compruebe estas afirmaciones, y encuentre otros dos grupos no normales, demostrando que no hay ninguno más.

[13.7] Demuéstrelo. (*Sugerencia*: ¿Qué *conjuntos* de rotaciones pueden ser invariantes por rotación?)

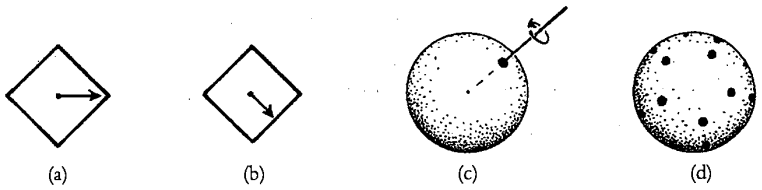


Fig. 13.3. (a) Marcar el cuadrado de la Fig. 13.1 con una flecha que apunta a la derecha reduce su grupo de simetría a un subgrupo no normal $\{1, C\}$. (b) Marcarlo con una flecha que apunta en diagonal hacia abajo y la derecha da un subrupo no normal diferente $\{1, Ci\}$. (c) Marcar la esfera de la Fig. 13.2 con un único punto reduce su simetría a un subgrupo (no normal) O(2) de O(3): rotaciones en torno al eje que une el origen a dicho punto. (d) Si la esfera se marca con los vértices de un poliedro regular (aquí un dodecaedro), su grupo de simetrías es un subgrupo (no normal) de O(3).

solo un punto, entonces el subgrupo consiste en las rotaciones de la esfera alrededor del eje que une dicho punto con el origen (Fig. 13.3c). Alternativamente podríamos marcar, por ejemplo, puntos que sean los vértices de un poliedro regular. Entonces, el subgrupo es finito, y consiste en el grupo de simetría de ese poliedro particular (Fig. 13.3d).

Una razón por la que los subgrupos normales son importantes es que si un grupo $\mathcal{G}$ posee un subgrupo normal no trivial, entonces podemos descomponer $\mathcal{G}$ en cierto sentido en grupos más pequeños. Supongamos que $\mathcal{S}$ es un subgrupo normal de $\mathcal{G}$. Entonces, los diferentes conjuntos $\mathcal{S}g$, donde g recorre todos los elementos de $\mathcal{G}$, formarán por sí mismos un grupo. Nótese que para un conjunto dado $\mathcal{S}g$ la elección de g no es en general única; podemos tener $\mathcal{S}g_1 = \mathcal{S}g_2$, para diferentes elementos g_1, g_2 de $\mathcal{G}$. Los conjuntos de la forma $\mathcal{S}g$, para *cualquier* subgrupo $\mathcal{S}$, se denominan *co-conjuntos* de $\mathcal{G}$; pero cuando $\mathcal{G}$ es normal, los co-conjuntos forman un grupo. La razón para esto es que, si tenemos dos de tales co-conjuntos $\mathcal{S}g$ y $\mathcal{S}h$ (siendo g y h elementos de $\mathcal{G}$), entonces podemos definir el «producto» de $\mathcal{S}g$ por $\mathcal{S}h$ como

$$(\mathcal{S}g)(\mathcal{S}h) = \mathcal{S}(gh),$$

y encontramos que todos los axiomas de grupo se satisfacen con tal de que $\mathcal{S}$ sea normal, esencialmente porque el segundo miembro está bien definido, con independencia de qué g y h se escojan en la representación de los co-conjuntos en el primer miembro de esta ecuación.[13.8] El grupo resultante definido de esta forma se denomina el *grupo cociente* de $\mathcal{G}$ por su subgrupo normal $\mathcal{S}$. El grupo cociente de $\mathcal{G}$ por $\mathcal{S}$ se escribe $\mathcal{G}/\mathcal{S}$. Podemos seguir escribiendo $\mathcal{G}/\mathcal{S}$ para el *espacio* cociente (no un grupo) de co-conjuntos diferentes $\mathcal{S}g$ incluso cuando S no es normal.[13.9]

Los grupos que no poseen ningún subgrupo normal no trivial se denominan *grupos simples*. El grupo SO(3) es un ejemplo de grupo simple. Los grupos simples son, en un sentido evidente, los bloques constituyentes básicos de la teoría de grupos. Por ello es un logro importante de las matemáticas de los siglos XIX y XX el que se conozcan ahora todos los grupos simples finitos y todos los grupos simples continuos. En el caso de los continuos (es decir, los grupos de Lie), esta fue una soberbia hazaña iniciada por el influyente matemático alemán Wilhelm Killing (1847-1923), cuyos artículos básicos aparecieron en 1888-1890, y que fue esencialmente completada, en 1894, en uno de los más importantes artículos matemáticos jamás escritos,[5] por el soberbio geómetra y algebrista Élie Cartan (de quien ya he hablado en el capítulo 12, y volveremos a hacerlo en el capítulo 17). Esta clasificación sigue desempeñando hoy día un papel fundamental en muchas áreas de las matemáticas y la física. Resulta que existen cuatro familias, conocidas como A_m, B_m, C_m, D_m (para $m = 1, 2, 3, \ldots$), de dimensiones respectivas $m(m+2)$, $m(2m+1)$, $m(2m+1)$, $m(2m-1)$, llamados *grupos clásicos* (véase el final de §13.10) y cinco *grupos excepcionales* conocidos como E_6, E_7, E_8, F_4, G_2, de dimensiones respectivas 78, 133, 248, 52, 14.

La clasificación de los grupos simples finitos es una hazaña más reciente (y aún más difícil), llevada a cabo durante muchos años del siglo XX por un número considerable de matemáticos (con ayuda de computadores en los casos más recientes), que no fue completada hasta 1982.[6] De nuevo hay algunas familias sistemáticas y una colección finita de grupos simples finitos *excepcionales*. El mayor de estos grupos excepcionales se conoce como el *monstruo*, cuyo orden es

[13.8] Verifique esto y demuestre que los axiomas fallan si $\mathcal{S}$ no es normal.
[13.9] Explique por qué el número de elementos en $\mathcal{G}/\mathcal{S}$, para cualquier subgrupo finito $\mathcal{S}$ de un grupo finito $\mathcal{G}$, es el orden de $\mathcal{G}$ dividido por el orden de $\mathcal{S}$.

$= 8080174247945128758864599049617107570057543680000000000.$
$= 2^{46} \times 3^{20} \times 5^9 \times 7^6 \times 11^2 \times 13^3 \times 17 \times 19 \times 23 \times 29 \times 31 \times 41 \times 47 \times$
$\times 59 \times 71.$

Los grupos excepcionales parecen tener un particular atractivo para muchos físicos teóricos modernos. El grupo E_8 desempeña un papel importante en la teoría de cuerdas (§31.14), mientras que varias personas han expresado su confianza en que el enorme pero finito monstruo pueda tener un papel en alguna teoría futura.[7]

La clasificación de los grupos simples puede considerarse un paso fundamental hacia la clasificación de los grupos en general, puesto que, como se ha indicado antes, los grupos generales pueden considerarse construidos a partir de grupos simples (junto con los abelianos). De hecho, esta no es realmente toda la historia porque hay información adicional de cómo se construye un grupo simple a partir de otro. No pretendo entrar aquí en detalles sobre esta cuestión, pero vale la pena mencionar la forma más sencilla en que esto puede suceder. Si $\mathcal{G}$ y $\mathcal{H}$ son dos grupos cualesquiera, entonces pueden combinarse para formar lo que se denomina el *grupo producto* $\mathcal{G} \times \mathcal{H}$, cuyos elementos son simplemente *pares* (g, h), donde g pertenece a $\mathcal{G}$ y h pertenece a $\mathcal{H}$, estando definida la regla de multiplicación del grupo entre elementos (g_1, h_1) y (g_2, h_2) de $\mathcal{G} \times \mathcal{H}$ como

$$(g_1, h_1)(g_2, h_2) = (g_1 g_2, h_1 h_2),$$

y es muy fácil verificar que se satisfacen los axiomas de grupo. Muchos de los grupos que intervienen en la física de partículas son, de hecho, grupos producto de grupos simples (o modificaciones elementales de los mismos).[13.10]

13.3. Transformaciones lineales y matrices

Dentro del estudio general de los grupos se ha encontrado que existe una clase especial de grupos de simetría que desempeñan un papel cen-

[13.10] Verifique que, para dos grupos $\mathcal{G}$ y $\mathcal{H}$ cualesquiera, $\mathcal{G} \times \mathcal{H}$ es un grupo, y que podemos identificar el grupo cociente $(\mathcal{G} \times \mathcal{H})/\mathcal{G}$ con $\mathcal{H}$.

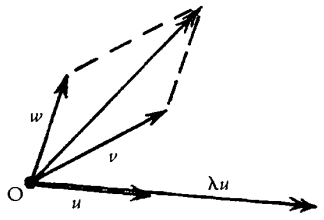


Fig. 13.4. Una transformación lineal conserva la estructura de espacio vectorial del espacio sobre el que actúa. Dicha estructura está definida por las operaciones de suma (ilustrada por la ley del paralelogramo) y multiplicación por un escalar λ (que podría ser un número real o, en el caso de un espacio vectorial complejo, un número complejo). Una transformación semejante conserva la «rectitud» de las líneas y la noción de «paralela», manteniendo el origen O fijo.

tral. Son los grupos de simetrías de espacios vectoriales. Las simetrías de un espacio vectorial se expresan mediante las *transformaciones lineales* que conservan la estructura del espacio de fases.

Recordemos de §11.1 y §12.3 que en un espacio vectorial V tenemos, definiendo su estructura, una noción de suma de vectores y de multiplicación de vectores por números. Podemos tomar nota del hecho de que la imagen geométrica de la suma se obtiene mediante el uso de la ley del paralelogramo, mientras que la multiplicación por un número se visualiza como un cambio en la escala del vector, ampliándolo (o reduciéndolo) en un factor dado por dicho número (Fig. 13.4). Aquí estamos representándolo como un número real, pero también se admiten los espacios vectoriales complejos (¡y son particularmente importantes en muchos contextos, a causa de la magia compleja!), aunque difíciles de mostrar en un diagrama. Una transformación lineal de V es una transformación que lleva V sobre sí mismo, conservando su estructura, tal como está definida por estas nociones básicas de espacio vectorial. Con más generalidad, podemos considerar también transformaciones lineales que llevan un espacio vectorial a otro.

Una transformación lineal puede describirse de forma explícita utilizando un conjunto de números llamado *matriz*. Las matrices son importantes en muchos contextos matemáticos. Examinaremos estas entidades extraordinariamente útiles con sus elegantes reglas algebraicas en esta sección (y en §§13.4,5). De hecho, §§13.3-7 pueden considerarse una sucinta introducción a la teoría de matrices y su aplicación a la teoría de los grupos continuos. Las nociones descritas aquí son vitales para una comprensión adecuada de la teoría cuántica, pero los lec-

tores que ya estén familiarizados con tales temas —o quienes prefieran un conocimiento menos detallado de la teoría cuántica cuando lleguemos a ella— quizá prefieran saltarse estas secciones, al menos por el momento.

Para ver qué aspecto tiene una transformación lineal, consideremos primero el caso de un espacio vectorial 3-dimensional y veamos su relevancia para el grupo de rotación O(3) (o SO(3)), que se ha planteado en §13.1, que da las simetrías de la esfera. Podemos considerar esta esfera inmersa en un 3-espacio euclídeo $\mathbb{E}^3$ (que consideramos un espacio vectorial con respecto al origen O en el centro de la esfera)[8] como el lugar geométrico

$$x^2 + y^2 + z^2 = 1$$

en términos de coordenadas cartesianas ordinarias (x, y, z).[13.11] Las rotaciones de la esfera se expresan ahora en términos de transformaciones lineales de $\mathbb{E}^3$, pero de un tipo muy especial conocido como *ortogonales* a las que llegaremos en §13.8 (véase también §13.1).

Sin embargo, las transformaciones lineales *generales* aplastarían o estirarían la esfera convirtiéndola en un *elipsoide*, como se ilustra en la Fig. 13.5. En términos geométricos, una transformación lineal es una transformación que conserva la «rectitud» de las líneas y la noción de líneas «paralelas», manteniendo fijo el origen O. Pero no tiene por qué conservar los ángulos rectos u otros ángulos, de modo que las formas pueden resultar aplastadas o estiradas de una forma uniforme pero anisótropa.

¿Cómo expresar las transformaciones lineales en términos de las coordenadas x, y, z? La respuesta es que cada nueva coordenada se expresa como una combinación lineal (homogénea) de las originales, i.e., por una expresión independiente como $\alpha x + \beta y + \gamma z$, donde α, β y γ son números constantes.[13.12] Tenemos 3 de tales expresiones, una para cada una de las nuevas coordenadas. Para escribir todo esto de forma compacta, será útil tomar contacto con la *notación de índices* del

[13.11] Demuestre cómo esta ecuación, que da los puntos a distancia unidad de O, se sigue del teorema de Pitágoras de §2.1.

[13.12] ¿Puede explicar por qué? Por simplicidad, hágalo solo en el caso 2-dimensional.

capítulo 12. Para esto, reetiquetamos las coordenadas como (x^1, x^2, x^3), donde

$$x^1 = x, \quad x^2 = y, \quad x^3 = z$$

(teniendo en mente, una vez más, que estos superíndices *no* denotan potencias; véase §12.2). Un punto general en nuestro 3-espacio euclídeo tiene coordenadas x^a, donde $a = 1, 2, 3$. Una ventaja de utilizar la notación de índices es que la discusión se aplicará también en cualquier otro número de dimensiones, de modo que podemos considerar que a (y todas nuestras otras letras índices) recorre los valores $1, 2, \ldots, n$, donde n es cierto entero positivo fijo. En el caso recién considerado, $n = 3$.

En la notación de índices, con el convenio de suma de Einstein (§12.7), la transformación lineal general toma ahora la forma[9,13.13]

$$x^a \mapsto T^a{}_b\, x^b.$$

Llamando a esta transformación lineal $\boldsymbol{T}$, vemos que $\boldsymbol{T}$ está determinada por este conjunto de *componentes* $T^a{}_b$. Dicho conjunto se conoce como una *matriz* $n \times n$, normalmente dispuestos como una formación cuadrada —o, en otros contextos (véase *infra*) $m \times n$-rectangular— de números. La ecuación que se ha mostrado antes se escribe entonces, en el caso 3-dimensional,

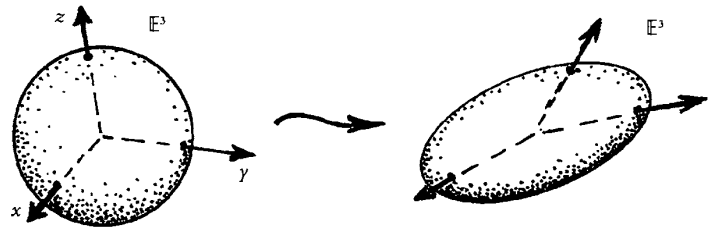


Fig. 13.5. Una transformación lineal que actúa sobre $\mathbb{E}^3$ (expresado en términos de coordenadas cartesians x, y, z) en general comprimiría o estiraría la esfera unidad $x^2 + y^2 + z^2 = 1$ para dar un elipsoide. El grupo ortogonal O(3) consiste en las transformaciones lineales de $\mathbb{E}^3$ que conservan la esfera unidad.

[13.13] Demuéstrelo explícitamente en el caso 3-dimensional.

$$\begin{pmatrix} x^1 \\ x^2 \\ x^3 \end{pmatrix} \mapsto \begin{pmatrix} T^1{}_1 & T^1{}_2 & T^1{}_3 \\ T^2{}_1 & T^2{}_2 & T^2{}_3 \\ T^3{}_1 & T^3{}_2 & T^3{}_3 \end{pmatrix} \begin{pmatrix} x^1 \\ x^2 \\ x^3 \end{pmatrix},$$

que representa tres ecuaciones independientes, empezando por $x^1 \mapsto$ $\mapsto T^1{}_1 x^1 + T^1{}_2 x^2 + T^1{}_3 x^3$.[13.14]

También es posible escribir esto sin índices ni coordenadas explícitas, como $\boldsymbol{x} \mapsto \boldsymbol{Tx}$. Si lo preferimos, podemos adoptar la notación de *índices abstractos* (§12.8) en la que «$x^a \mapsto T^a{}_b x^b$» no es una expresión en componentes, sino que en realidad representa esta transformación abstracta $\boldsymbol{x} \mapsto \boldsymbol{Tx}$. (Cuando sea importante que una expresión indexada se lea de forma abstracta o como componentes, esto se aclarará por el enunciado.) Alternativamente, podemos utilizar la notación *diagramática*, como se muestra en la Fig. 13.6a. En mis descripciones, la matriz de números $(T^a{}_b)$ o la transformación lineal abstracta $\boldsymbol{T}$ serán utilizadas de forma intercambiable cuando no esté interesado en las distinciones técnicas entre estos dos conceptos (el primero dependiente de una descripción en coordenadas específicas de nuestro espacio vectorial V, y el segundo no).

Consideremos una segunda transformación lineal $\boldsymbol{S}$, aplicada a continuación de la aplicación de $\boldsymbol{T}$. El producto $\boldsymbol{R}$ de las dos, escrito $\boldsymbol{R} = \boldsymbol{ST}$, tendrá una descripción en componentes

$$R^a{}_c = S^a{}_b T^b{}_c$$

(¡convenio de suma para las componentes!).[13.15] La forma diagramática del producto $\boldsymbol{ST}$ está dada en la Fig. 13.6b. Nótese que en la notación diagramática, para formar un producto sucesivo de transformaciones lineales, las encadenamos en una línea descendente. Esto funciona de forma conveniente en la notación, pero perfectamente podríamos adoptar un convenio diferente en el que las «líneas índices» conectoras están dibujadas horizontalmente. (Entonces habría una correspondencia más estrecha entre la notación algebraica y la diagramática.)

[13.14] Escriba todo esto en detalle, explicando cómo se expresa $x^a \mapsto T^a{}_b x^b$.

[13.15] ¿Cuál es esta operación entre $\boldsymbol{R}$, $\boldsymbol{S}$ y $\boldsymbol{T}$, escrita explícitamente en términos de los elementos 3×3 de las matrices cuadradas de componentes? Puede reconocer aquí la ley normal para la «multiplicación de matrices», si esta le es familiar.

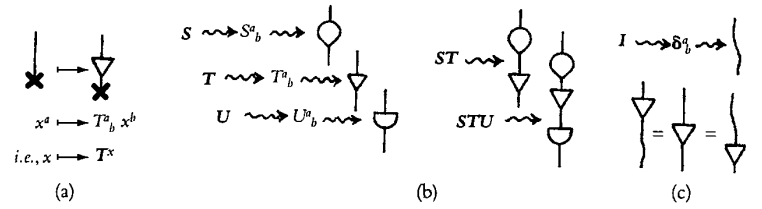


Fig. 13.6. (a) La transformación lineal $x^a \mapsto T^a{}_b x^b$, o escrita sin índices como $\boldsymbol{x} \mapsto \boldsymbol{Tx}$ (o leída con los índices abstractos, como en §12.8), de forma diagramática. (b) Diagramas para transformaciones lineales $\boldsymbol{S}$, $\boldsymbol{T}$, $\boldsymbol{U}$, y sus productos $\boldsymbol{ST}$ y $\boldsymbol{STU}$. En un producto sucesivo, los encadenamos en una línea hacia abajo. (c) La delta de Kronecker δ^a_b, o transformación identidad $\boldsymbol{I}$, se representa como una línea «incorpórea», de modo que las relaciones $T^a{}_b \delta^b_c = T^a{}_c = \delta^a_b T^b{}_c$ se hacen automáticas en la notación (véase también la Fig. 12.17).

La transformación lineal *identidad* $\boldsymbol{I}$ tiene componentes que normalmente se escriben δ^a_b (la *delta de Kronecker*; ahora el convenio estándar es que estos índices no estén escalonados), para la que

$$\delta^a_b = \begin{cases} 1 & \text{si } a = b, \\ 0 & \text{si } a \neq b, \end{cases}$$

y tenemos[13.16]

$$T^a{}_b \delta^b_c = T^a{}_c = \delta^a_b T^b{}_c$$

para dar las relaciones algebraicas $\boldsymbol{TI} = \boldsymbol{T} = \boldsymbol{IT}$. La matriz cuadrada de componentes δ^a_b tiene 1s en la que se denomina la *diagonal principal*, que se extiende desde la esquina superior izquierda a la inferior derecha. En el caso $n = 3$, esta es

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

En la notación diagramática, representamos simplemente la delta de Kronecker mediante una línea «incorpórea», y las relaciones algebraicas anteriores se hacen automáticas en la notación; véase la Fig. 13.6c.

[13.16] Verifíquelo.

Aquellas transformaciones lineales que aplican el espacio vectorial entero en una región (subespacio) de dimensión menor dentro de dicho espacio se denominan *singulares*.[10] Una condición equivalente para que $\boldsymbol{T}$ sea singular es la existencia de un vector $\boldsymbol{v}$ no nulo tal que[13.17]

$$\boldsymbol{T}\boldsymbol{v} = 0.$$

Con tal de que la transformación sea no singular, tendrá una inversa,[13.18] donde la inversa de $\boldsymbol{T}$ se escribe $\boldsymbol{T}^{-1}$, de modo que

$$\boldsymbol{T}\boldsymbol{T}^{-1} = \boldsymbol{I} = \boldsymbol{T}^{-1}\,\boldsymbol{T},$$

como se exige de una inversa. Es posible dar la expresión explícita para esta inversa de forma conveniente en la notación diagramática; véase la Fig. 13.7, en la que he introducido una notación útil para los símbolos antisimétricos (de *Levi-Civita*) $\varepsilon_{a\ldots c}$ y $\in^{a\ldots c}$ (con normalización $\varepsilon_{a\ldots c}\in^{a\ldots c} = n!$) que han sido abordados en §12.7 y la Fig. 12.18.[13.19]

El álgebra de matrices (iniciada por el prolífico matemático y abogado inglés Arthur Cayley en 1858)[11] encuentra un amplio abanico de aplicaciones (por ejemplo, estadística, ingeniería, cristalografía, psicología, computación, por no mencionar la mecánica cuántica). Esto generaliza el álgebra de cuaterniones y las álgebras de Clifford y Grassmann §§11.3,5,6. Utilizo letras negritas *verticales* (**A**, **B**, **C**, ...) para los conjuntos de componentes que constituyen matrices reales (en lugar de transformaciones lineales abstractas, para las que estoy utilizando letras negritas *itálicas*).

Restringiendo la atención a matrices $n \times n$ para n fijo, tenemos un sistema en el que están definidas las nociones de suma y multiplicación, donde son válidas las leyes matemáticas estándar

[13.17] ¿Por qué? Demuestre que esto sucedería en particular si el conjunto de componentes tiene toda una columna de 0s o dos columnas idénticas. ¿Por qué sucede también esto si hay dos filas idénticas? *Sugerencia*: Para la última parte, considere la condición del determinante que se da más abajo.

[13.18] Demuestre por qué, sin utilizar expresiones explícitas.

[13.19] Demuestre directamente, utilizando las relaciones diagramáticas que se dan en la Fig. 12.18, que esta definición da $\boldsymbol{T}\boldsymbol{T}^{-1} = \boldsymbol{I} = \boldsymbol{T}^{-1}\,\boldsymbol{T}$. *Sugerencia*:Véase la Fig. 13.8.

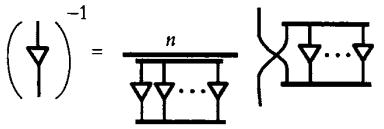


Fig. 13.7. La inversa $\boldsymbol{T}^{-1}$ de una matriz $(n \times n)$ no singular $\boldsymbol{T}$ que se da aquí explícitamente de forma diagramática, utilizando la forma diagramática de las cantidades antisimétricas $\varepsilon_{a\ldots c}$ y $\in^{a\ldots c}$ (normalizadas por $\varepsilon_{a\ldots c}\in^{a\ldots c} = n!$), que se han introducido en §12.7 y mostrado en la Fig. 12.18.

$$\mathbf{A} + \mathbf{B} = \mathbf{B} + \mathbf{A}, \quad \mathbf{A} + (\mathbf{B} + \mathbf{C}) = (\mathbf{A} + \mathbf{B}) + \mathbf{C}, \quad \mathbf{A}(\mathbf{BC}) = (\mathbf{AB})\mathbf{C},$$
$$\mathbf{A}(\mathbf{B} + \mathbf{C}) = \mathbf{AB} + \mathbf{AC}, \quad (\mathbf{A} + \mathbf{B})\mathbf{C} = \mathbf{AC} + \mathbf{BC}$$

(Cada elemento de $\mathbf{A} + \mathbf{B}$ es simplemente la suma de los correspondientes elementos de **A** y **B**.) Sin embargo, normalmente no tenemos la ley conmutativa de la multiplicación, de modo que en general $\mathbf{AB} \neq \mathbf{BA}$. Además, como hemos visto antes, las matrices $n \times n$ no nulas no siempre tienen inversas.

Habría que señalar que el álgebra también se extiende a los casos rectangulares de matrices $m \times n$, donde m no tiene por qué ser igual a n. Sin embargo, la suma está definida entre una matriz $m \times n$ y una matriz $p \times q$ solo si $m = p$ y $n = q$; la multiplicación está definida solo cuando $n = p$, y el resultado es una matriz $m \times q$. Esta álgebra ampliada engloba productos como el $\mathbf{Tx}$ que se ha visto antes, donde el «vector columna» $\mathbf{x}$ se considera como una matriz $n \times 1$.[13.20]

El *grupo lineal general* GL(n) es el grupo de simetrías de un espacio vectorial n-dimensional, y es realizado explícitamente como el grupo multiplicativo de matrices $n \times n$ no singulares. Si queremos resaltar que nuestro espacio vectorial es *real*, y que los números que aparecen en nuestras matrices son en correspondencia números reales, entonces llamamos a estre grupo lineal completo GL($n, \mathbb{R}$). También podemos considerar el caso complejo, y obtener el grupo lineal completo *complejo* GL($n, \mathbb{C}$). Cada uno de estos grupos tiene un subgrupo normal, escrito respectivamente SL($n, \mathbb{R}$) y SL($n, \mathbb{C}$) —o más brevemente, cuando el campo subyacente (véase §16.1) $\mathbb{R}$ o $\mathbb{C}$ está sobrentendido, SL(n)—, llamado grupo lineal *especial*. Estos se obtienen restringiéndonos a las

[13.20] Explique esto y dé las reglas algebraicas completas para matrices rectangulares.

matrices que tengan sus *determinantes* iguales a 1. La noción de determinante se expone a continuación.

13.4. Determinantes y trazas

¿Qué es el determinante de una matriz $n \times n$? Es un *simple número* calculado a partir de los elementos de la matriz, que se anula si y solo si la matriz es singular. La notación diagramática describe convenientemente el determinante de forma explícita; véase la Fig. 13.8a. La forma para esto en notación de índices es

$$\frac{1}{n!}\epsilon^{ab\ldots d}\, T^{e}{}_{a} T^{f}{}_{b} \ldots T^{h}{}_{d}\varepsilon_{ef\ldots h}$$

donde las cantidades $\epsilon^{ab\ldots d}$ y $\varepsilon_{ef\ldots h}$ son tensores antisimétricos (de Levi-Civita) normalizados de acuerdo con

$$\epsilon^{ab\ldots d}\varepsilon_{ef\ldots h} = n!$$

para un espacio n-dimensional (y recordemos que $n! = 1 \times 2 \times 3 \times \times \ldots \times n$), donde cada uno de los conjuntos $a, \ldots, d$ y $e, \ldots, h$ contiene n índices.

Podemos denotar a este determinante como det $(T^{a}{}_{b})$ o det $\boldsymbol{T}$ (o a veces $|\boldsymbol{T}|$) o como el conjunto de elementos que constituyen la matriz, pero con barras verticales reemplazando a los paréntesis). En los casos particulares de una matriz 2×2 y una matriz 3×3, el determinante está dado por[13.21]

$$\det \begin{pmatrix} a & b \\ c & d \end{pmatrix} = ad - bc,$$

$$\det \begin{pmatrix} a & b & c \\ d & e & f \\ g & h & j \end{pmatrix} = aej - afh + bfg - bdj + cdh - ceg.$$

El determinante satisface la importante y muy notable relación

$$\det \mathbf{AB} = \det \mathbf{A} \det \mathbf{B},$$

cuya validez puede verse de forma muy clara en la notación diagramática (Fig. 13.8b). Los ingredientes clave son las fórmulas ilustradas en la Fig. 12.18,[13.22] que cuando se escriben en la notación de índices toman la forma

$$\epsilon^{a\ldots c}\varepsilon_{f\ldots h} = n!\, \delta^{[a}_{f} \ldots \delta^{c]}_{h}$$

(véase §11.6 para la notación paréntesis/índice) y

$$\epsilon^{ab\ldots c}\varepsilon_{fb\ldots c} = (n-1)!\delta^{a}_{f}.$$

Tenemos también la noción de la *traza* de una matriz (o de una transformación lineal)

$$\text{traza } \mathbf{T} = T^{a}{}_{a} = T^{1}{}_{1} + T^{2}{}_{2} + \ldots + T^{n}{}_{n}$$

(i.e., la suma de los elementos de la diagonal principal; véase §13.3), lo que se ilustra diagramáticamente en la Fig. 13.9. A diferencia de lo que sucede con un determinante, no hay una relación especial entre la traza del producto **AB** de las dos matrices y las trazas de **A** y **B** por separado. En su lugar, tenemos la relación[13.23]

$$\text{traza } (\mathbf{A} + \mathbf{B}) = \text{traza } \mathbf{A} + \text{traza } \mathbf{B}.$$

Hay una conexión importante entre el determinante y la traza que tiene que ver con el determinante de una transformación lineal «infi-

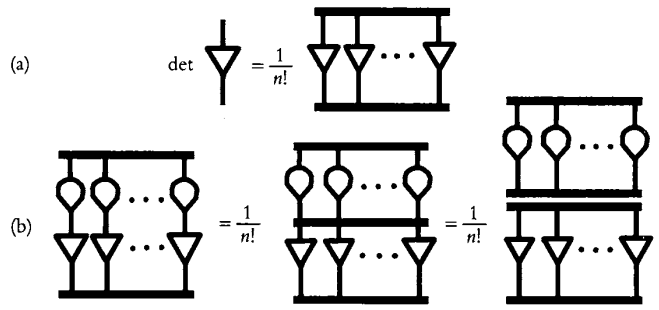


Fig. 13.8. (a) Notación diagramática para det $(T^{a}{}_{b})$ = det $\boldsymbol{T}$ = $|\boldsymbol{T}|$. (b) Demostración diagramática de que det $(\boldsymbol{ST})$ = det $\boldsymbol{S}$ det $\boldsymbol{T}$. La barra antisimetrizadora puede insertarse en el término medio porque ya hay antisimetría en las líneas de índices que cruza. Véanse las Figs. 12.17 y 12.18.

[13.21] Obténgalos a partir de la expresión de la Fig. 13.8a.
[13.22] Demuestre por qué se cumplen.
[13.23] Demuéstrelo.

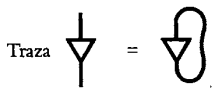


Fig. 13.9. Notación diagramática para traza $T(= T^{a}_{a})$.

nitesimal», dada por una matriz $n \times n$ de la forma $\mathbf{I} + \varepsilon\mathbf{A}$, donde el número ε se considera «infinitesimalmente pequeño», de modo que podemos ignorar su cuadrado ε^2 (y también las potencias superiores ε^3, ε^4, etc.). Entonces encontramos[13.24]

$$\det(\mathbf{I} + \varepsilon\mathbf{A}) = 1 + \varepsilon \operatorname{traza} \mathbf{A}$$

(ignorando ε^2, etc.). En particular, los elementos infinitesimales de SL(n), i.e., elementos de SL(n) que representan rotaciones infinitesimales, y son de determinante *unidad* (contrariamente a los de GL(n)), están caracterizados porque la $\mathbf{A}$ en $\mathbf{I} + \varepsilon\mathbf{A}$ tiene traza nula. Veremos la importancia de esto en §13.10. De hecho, la fórmula anterior puede extenderse a transformaciones lineales *finitas* (esto es, no infinitesimales) a través de la expresión[13.25]

$$\det e^{\mathbf{A}} = e^{\text{traza } \mathbf{A}},$$

donde «$e^{\mathbf{A}}$» para matrices tiene exactamente la misma definición que para números ordinarios (véase §5.3), i.e.,

$$e^{\mathbf{A}} = \mathbf{I} + \mathbf{A} + 1/2\mathbf{A}^2 + 1/6\mathbf{A}^3 + 1/24\mathbf{A}^4 + \ldots$$

Volveremos a estas cuestiones en §13.6 y §14.6.

13.5. Autovalores y autovectores

Entre las nociones más importantes asociadas a las transformaciones lineales están las que se denominan «autovalores» y «autovectores» (o también «valores propios» y «vectores propios»). Estos son vitales para

[13.24] Demuéstrelo.

[13.25] Establezca esta expresión. *Sugerencia*: Utilice la «forma canónica» para una matriz en términos de sus autovalores —descritos en §13.5— suponiendo primero que estos autovalores no son iguales (y véase el ejercicio [13.27]). Utilice luego un argumento general para demostrar que la igualdad de algunos autovalores no puede invalidar identidades de este tipo.

la mecánica cuántica, como veremos en §21.5 y §§22.1,5, y para muchas otras áreas de las matemáticas y sus aplicaciones. Un *autovector* de una transformación lineal $\boldsymbol{T}$ es un vector complejo no nulo $\boldsymbol{v}$ que $\boldsymbol{T}$ transforma en un múltiplo de sí mismo. Es decir, existe un número complejo λ, el correspondiente *autovector*, para el que

$$\boldsymbol{T}\boldsymbol{v} = \lambda\boldsymbol{v}, \text{ i.e., } \boldsymbol{T}^{a}_{\ b}v^{b} = \lambda v^{a}.$$

También podemos escribir esta ecuación como $(\boldsymbol{T}-\lambda\boldsymbol{I})\boldsymbol{v} = 0$, de modo que, si λ va a ser un autovalor de $\boldsymbol{T}$, la cantidad $\boldsymbol{T}-\lambda\boldsymbol{I}$ debe ser *singular*. Recíprocamente, si $\boldsymbol{T}-\lambda\boldsymbol{I}$ es singular, entonces λ es un autovalor de $\boldsymbol{T}$. Nótese que si $\boldsymbol{v}$ es un autovector, entonces también lo es cualquier múltiplo no nulo de $\boldsymbol{v}$. El espacio 1-dimensional complejo de estos múltiplos es invariable por la transformación $\boldsymbol{T}$, una propiedad que caracteriza a $\boldsymbol{v}$ como autovector (Fig. 13.10).

A partir de lo anterior, vemos que esta condición para que λ sea un autovalor de $\boldsymbol{T}$ es

$$\det(\boldsymbol{T}-\lambda\boldsymbol{I}) = 0.$$

Al calcular este determinante, obtenemos una ecuación polinómica[13.26] de grado n en λ. Por el «teorema fundamental del álgebra», §4.2, podemos factorizar el polinomio en λ $\det(\boldsymbol{T}-\lambda\boldsymbol{I})$ en factores lineales. Esto reduce la ecuación anterior a

$$(\lambda_1 - \lambda)(\lambda_2 - \lambda)(\lambda_3 - \lambda)\ldots(\lambda_n - \lambda) = 0,$$

donde los números complejos $\lambda_1, \lambda_2, \lambda_3, \ldots, \lambda_n$ son los diversos autovalores de $\boldsymbol{T}$. En casos particulares, algunos de estos factores pueden coincidir, en cuyo caso tenemos un autovalor *múltiple*. La multiplicidad m de un autovalor λ_r es el número de veces que aparece el factor $(\lambda_r - \lambda)$ en el producto anterior. El número total de autovalores de $\boldsymbol{T}$, contado adecuadamente con multiplicidades, es siempre igual a n, para una matriz $n \times n$.[13.27]

Para un autovalor particular λ de multiplicidad r, el espacio de autovectores correspondientes constituye un espacio lineal, de dimensionali-

[13.26] Trate de expresar los coeficientes de este polinomio de forma diagramática. Calcúlelos para $n = 2$ y $n = 3$.

[13.27] Demuestre que $\det \boldsymbol{T} = \lambda_1\lambda_2 \ldots \lambda_n$ y traza $\boldsymbol{T} = \lambda_1 + \lambda_2 + \ldots + \lambda_n$.

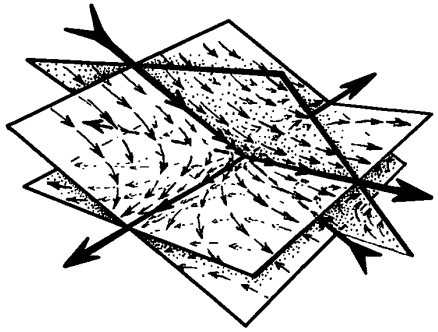


Fig. 13.10. La acción de una transformación T. Sus autovectores constituyen siempre espacios lineales que pasan por el origen (aquí tres líneas). Estos espacios son invariables por T. (En este ejemplo, hay dos autovalores (desiguales) positivos —flechas que apuntan hacia fuera— y uno negativo —flechas hacia dentro.

dad d, donde $1 \leq d \leq r$. Para ciertos tipos de matrices, incluyendo matrices unitarias, hermíticas y normales de gran interés en mecánica cuántica (véanse §13.9 y §§22.4,6), tenemos siempre la máxima dimensionalidad $d = r$ (pese al hecho de que $d = 1$ es el caso más «general», para r dado). Esto es una suerte, porque los casos (más generales) para los que $d < r$ son más difíciles de tratar. En mecánica cuántica, las multiplicidades en los autovalores se conocen como *degeneraciones* (cf. §§22.6,7).

Una *base* de un espacio vectorial n-dimensional $\mathbf{V}$ es un conjunto ordenado $\boldsymbol{e} = (\boldsymbol{e}_1, \ldots, \boldsymbol{e}_n)$ de n vectores $\boldsymbol{e}_1, \ldots, \boldsymbol{e}_n$ que son linealmente independientes, lo que significa que no existe una relación de la forma $\alpha_1\boldsymbol{e}_1 + \ldots + \alpha_n\boldsymbol{e}_n = 0$ en la que al menos uno de los $\alpha_1, \ldots, \alpha_n$ es diferente de cero. Cada elemento de $\mathbf{V}$ es entonces unívocamente una combinación lineal de estos elementos de la base.[13.28] De hecho, esta propiedad es la que caracteriza a una base en el caso más general en que $\mathbf{V}$ puede ser de dimensión infinita, cuando la independencia lineal por sí sola no es suficiente.

Así pues, dada una base $\boldsymbol{e} = (\boldsymbol{e}_1, \ldots, \boldsymbol{e}_n)$, cualquier elemento $\boldsymbol{x}$ de $\mathbf{V}$ puede escribirse de forma unívoca

$$\begin{aligned}\boldsymbol{x} &= x^1\boldsymbol{e}_1 + x^2\boldsymbol{e}_2 + \ldots + x^n\boldsymbol{e}_n \\ &= x^j\boldsymbol{e}_j,\end{aligned}$$

(los índices j *no* son abstractos aquí), donde $(x^1, x^2, \ldots, x^n)$ es el conjunto ordenado de *componentes* de $\boldsymbol{x}$ con respecto a $\boldsymbol{e}$ (compárelo con

[13.28] Demuéstrelo.

§12.3). Una transformación lineal no singular T siempre transforma una base en otra base; además, si $\boldsymbol{e}$ y $\boldsymbol{f}$ son dos bases dadas cualesquiera, entonces existe una única T que transforma cada $\boldsymbol{e}_j$ en su correspondiente $\boldsymbol{f}_j$:

$$T\boldsymbol{e}_j = \boldsymbol{f}_j.$$

En términos de componentes tomadas con respecto a $\boldsymbol{e}$, las componentes de los propios elementos de la base $\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_n$ son, respectivamente, $(1, 0, 0, \ldots, 0)$, $(0, 1, 0, \ldots, 0)$, $\ldots$, $(0, 0, \ldots, 0, 1)$. En otras palabras, las componentes de $\boldsymbol{e}_j$ son $(\delta^1_j, \delta^2_j, \delta^3_j, \ldots, \delta^n_j)$.[13.29] Cuando todas las componentes se toman con respecto a la base $\boldsymbol{e}$, encontramos que T está representada como la matriz $(T^i{}_j)$, donde las componentes de $\boldsymbol{f}_j$ en la base $\boldsymbol{e}$ serían[13.30]

$$(T^1{}_j, T^2{}_j, T^3{}_j, \ldots, T^n{}_j).$$

Habría que recordar que la diferencia conceptual entre una transformación lineal y una matriz es que la última se refiere a una representación dependiente de una base, mientras que la primera es abstracta, sin dependencia de una base.

Ahora, con tal de que cada autovalor múltiple de T (si los hay) satisfaga $d = r$, i.e., la dimensionalidad de su autoespacio iguala a su multiplicidad, es posible encontrar una base $(\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n)$ para $\mathbf{V}$, cada uno de cuyos vectores es un autovector de T.[13.31] Sean $\lambda_1, \lambda_2, \ldots, \lambda_n$ los correspondientes autovalores:

$$T\boldsymbol{e}_1 = \lambda_1\boldsymbol{e}_1, \quad T\boldsymbol{e}_2 = \lambda_2\boldsymbol{e}_2, \ldots, T\boldsymbol{e}_n = \lambda_n\boldsymbol{e}_n.$$

Si, como antes, T transforma la base $\boldsymbol{e}$ en la base $\boldsymbol{f}$, entonces los elementos de la base $\boldsymbol{f}$ son como antes, de modo que tenemos $\boldsymbol{f}_1 = \lambda_1\boldsymbol{e}_1, \boldsymbol{f}_2 = \lambda_2\boldsymbol{e}_2, \ldots, \boldsymbol{f}_n = \lambda_n\boldsymbol{e}_n$. Se sigue que T, referida a la base $\boldsymbol{e}$, toma la forma de matriz *diagonal*

[13.29] Explique esta notación.
[13.30] ¿Por qué? ¿Cuáles son las componentes de $\boldsymbol{e}_j$ en la base $\boldsymbol{f}$?
[13.31] Trate de demostrarlo. *Sugerencia*: Para cada autovalor de multiplicidad r, escoja r autovectores linealmente independientes. Demuestre que una relación lineal entre vectores de toda esta colección conduce a una contradicción cuando esta relación es multiplicada por T sucesivamente.

$$\begin{pmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & 0 \\ \vdots & \vdots & \vdots\vdots & \vdots \\ 0 & 0 & \dots & \lambda_n \end{pmatrix},$$

esto es, $T^1_1 = \lambda_1$, $T^2_2 = \lambda_2$, ..., $T^n_n = \lambda_n$, y las demás componentes son nulas. Esta *forma canónica* para una transformación lineal es muy útil tanto desde el punto de vista conceptual como de cálculo.[12]

13.6. Teoría de la representación y álgebras de Lie

Hay un cuerpo importante de ideas (particularmente importante para la teoría cuántica) conocido como la *teoría de representación* de grupos. Hemos visto un ejemplo muy sencillo de una representación de un grupo en §13.1, cuando hemos observado que las simetrías no reflexivas de un cuadrado pueden representarse por números complejos, y el grupo de multiplicación es representado fielmente como una multiplicación real de los números complejos. Pero, nada tan sencillo es aplicable a los grupos no abelianos, puesto que la multiplicación de números complejos es conmutativa. Por el contrario, las transformaciones lineales (o matrices) no conmutan normalmente, de modo que podemos considerar una perspectiva razonable representar grupos no abelianos en sus términos. En realidad, ya hemos encontrado algo de este tipo al comienzo de §13.1, donde representábamos el grupo de rotación O(3) en términos de transformaciones lineales en tres dimensiones.

Como veremos en el capítulo 22, la mecánica cuántica tiene mucho que ver con las transformaciones lineales. Además, varios grupos de simetría tienen importancia crucial en la física de partículas moderna, tal como el grupo de rotación O(3), los grupos de simetría de la teoría de la relatividad (capítulo 18) y las simetrías subyacentes a las interacciones entre partículas (capítulo 25). Por consiguiente, no es sorprendente que representaciones de estos grupos en particular, en términos de transformaciones lineales, desempeñen papeles fundamentales en teoría cuántica.

Sucede que la teoría cuántica (resulta especial la teoría cuántica de campos del capítulo 26) se interesa con frecuencia en transformaciones

lineales de espacios de dimensión *infinita*. Sin embargo, por simplicidad aquí formularé las cosas solo para representaciones mediante transformaciones lineales en el caso de dimensión finita. La mayor parte de las ideas que encontraremos se aplican también en el caso de representaciones en dimensión infinita, aunque hay diferencias que pueden ser importantes en ciertas circunstancias.

¿Qué es una representación de un grupo? Consideremos un grupo $\mathcal{G}$. La teoría de la representación se interesa en encontrar un subgrupo de GL(n) (i.e., un grupo multiplicativo de matrices $n \times n$) con la propiedad de que, para cualquier elemento g en $\mathcal{G}$, existe una correspondiente transformación lineal $\boldsymbol{T}(g)$ (perteneciente a GL(n)) tal que la ley de multiplicación en $\mathcal{G}$ es conservada por las operaciones de GL(n), i.e., para cualesquiera dos elementos g, h de $\mathcal{G}$, tenemos

$$\boldsymbol{T}(g)\boldsymbol{T}(h) = \boldsymbol{T}(gh).$$

La representación se denomina *fiel* si $\boldsymbol{T}(g)$ es diferente de $\boldsymbol{T}(h)$ cuando quiera que g sea diferente de h. En este caso tenemos una *copia idéntica* del grupo $\mathcal{G}$, como un *subgrupo* de GL(n).

De hecho, todo grupo *finito* tiene una representación fiel en GL($n, \mathbb{R}$), donde n es el orden de $\mathcal{G}$,[13.32] y existen con frecuencia muchas representaciones no fieles. Por el contrario, no es totalmente cierto que todo grupo continuo (de dimensión finita) tenga una representación fiel en algún GL(n). Sin embargo, si no nos preocupamos por los aspectos globales del grupo, entonces una representación *es* siempre posible (localmente).[13]

Existe una bella teoría, debida al profundamente original matemático noruego Sophus Lie (1842-1899), que conduce a un tratamiento completo de la teoría local de los grupos continuos. (De hecho, a los grupos continuos se les llama normalmente «grupos de Lie»; véase §13.1.) Esta teoría depende de un estudio de elementos de grupo *infinitesimales*.[14] Estos elementos infinitesimales definen un tipo de álge-

[13.32] Demuéstrelo. *Sugerencia*: Etiquete cada columna de la matriz de representación por un elemento independiente del grupo finito $\mathcal{G}$, y etiquete también cada fila por el elemento del grupo correspondiente. Coloque un 1 en cualquier posición en la matriz para la que se cumple cierta relación (¡encuéntrela!) entre el elemento de $\mathcal{G}$ que etiqueta la fila, el que etiqueta la columna y el elemento de $\mathcal{G}$ que está representando esta matriz concreta. Coloque un 0 cuando no se cumple dicha relación.

bra —que se conoce como *álgebra de Lie*— que nos proporciona información completa de la estructura local del grupo. Aunque el álgebra de Lie quizá no nos proporcione toda la estructura *global* del grupo, esto se considera normalmente una cuestión de menor importancia.

¿Qué es el álgebra de Lie? Supongamos que tenemos una matriz (o transformación lineal) $\boldsymbol{I} + \varepsilon \boldsymbol{A}$ que representa un elemento «infinitesimal» a de un grupo continuo $\mathcal{G}$, donde ε se toma «pequeño» (compárese con la parte final de §13.4). Cuando formamos la matriz producto de $\boldsymbol{I} + \varepsilon \boldsymbol{A}$ e $\boldsymbol{I} + \varepsilon \boldsymbol{B}$ para representar el producto ab de estos dos elementos a y b, obtenemos

$$(\boldsymbol{I} + \varepsilon \boldsymbol{A})\,(\boldsymbol{I} + \varepsilon \boldsymbol{B}) = \boldsymbol{I} + \varepsilon(\boldsymbol{A} + \boldsymbol{B}) + \varepsilon^2 \boldsymbol{A}\boldsymbol{B} = \\ = \boldsymbol{I} + \varepsilon(\boldsymbol{A} + \boldsymbol{B})$$

si se nos permite ignorar la cantidad ε^2, que es «demasiado pequeña para contar». De acuerdo con esto, la matriz *suma* $\boldsymbol{A} + \boldsymbol{B}$ representa el *producto de grupo* ab de dos elementos infinitesimales a y b.

En realidad, la operación suma es parte del álgebra de Lie de las cantidades $\boldsymbol{A}, \boldsymbol{B}, \ldots$ Pero la suma es conmutativa, mientras que el grupo $\mathcal{G}$ muy bien podría ser no abeliano, de modo que no captamos mucho de la estructura del grupo si consideramos solo sumas (de hecho, solo la dimensión de $\mathcal{G}$). La naturaleza no abeliana de $\mathcal{G}$ se expresa en los *conmutadores del grupo* que son las expresiones[13.33]

$$a\,b\,a^{-1}\,b^{-1}.$$

Escribamos esto en términos $\boldsymbol{I} + \varepsilon \boldsymbol{A}$, etc., tomando nota de la expresión en serie de potencias $(\boldsymbol{I} + \varepsilon \boldsymbol{A})^{-1} = \boldsymbol{I} - \varepsilon \boldsymbol{A} + \varepsilon^2 \boldsymbol{A}^2 - \varepsilon^3 \boldsymbol{A}^3 + \ldots$ (se puede comprobar fácilmente esta serie multiplicando ambos miembros por $\boldsymbol{I} + \varepsilon \boldsymbol{A}$). Ahora es ε^3 lo que ignoramos por ser «demasiado pequeño para contar», pero conservamos ε^2, de donde[13.34]

$$(\boldsymbol{I} + \varepsilon \boldsymbol{A})\,(\boldsymbol{I} + \varepsilon \boldsymbol{B})\,(\boldsymbol{I} + \varepsilon \boldsymbol{A})^{-1}\,(\boldsymbol{I} + \varepsilon \boldsymbol{B})^{-1} = \\ = (\boldsymbol{I} + \varepsilon \boldsymbol{A})\,(\boldsymbol{I} + \varepsilon \boldsymbol{B})\,(\boldsymbol{I} - \varepsilon \boldsymbol{A} + \varepsilon^2 \boldsymbol{A}^2)\,(\boldsymbol{I} - \varepsilon \boldsymbol{B} + \varepsilon^2 \boldsymbol{B}^2) = \\ = \boldsymbol{I} + \varepsilon^2(\boldsymbol{A}\boldsymbol{B} - \boldsymbol{B}\boldsymbol{A}).$$

[13.33] ¿Por qué esta expresión es precisamente el elemento identidad del grupo cuando a y b conmutan?

[13.34] Explique en detalle este cálculo a «orden ε^2».

Esto nos dice que si queremos seguir la pista de la forma exacta en que el grupo $\mathcal{G}$ es no abeliano, debemos tomar nota de los conmutadores, o *paréntesis de Lie*

$$[\boldsymbol{A}, \boldsymbol{B}] = \boldsymbol{A}\boldsymbol{B} - \boldsymbol{B}\boldsymbol{A}.$$

El álgebra de Lie se construye ahora por medio de la aplicación repetida de las operaciones $+$, su inversa $-$ y la operación paréntesis $[\ ,\ ]$, donde también es costumbre permitir la multiplicación por números ordinarios (que podrían ser reales o complejos). El aspecto «aditivo» del álgebra tiene la estructura usual de espacio vectorial (como sucede con los cuaterniones, en §11.1). Además, el paréntesis de Lie satisface la distributividad, etc., a saber

$$[\boldsymbol{A} + \boldsymbol{B}, \boldsymbol{C}] = [\boldsymbol{A}, \boldsymbol{C}] + [\boldsymbol{B}, \boldsymbol{C}], [\lambda \boldsymbol{A}, \boldsymbol{B}] = \lambda[\boldsymbol{A}, \boldsymbol{B}],$$

la propiedad de antisimetría

$$[\boldsymbol{A}, \boldsymbol{B}] = -[\boldsymbol{B}, \boldsymbol{A}],$$

(de donde también $[\boldsymbol{A}, \boldsymbol{C} + \boldsymbol{D}] = [\boldsymbol{A}, \boldsymbol{C}] + [\boldsymbol{A}, \boldsymbol{D}], [\boldsymbol{A}, \lambda \boldsymbol{B}] = \lambda[\boldsymbol{A}, \boldsymbol{B}]$), y una elegante relación conocida como la *identidad de Jacobi*[13.35]

$$[\boldsymbol{A}, [\boldsymbol{B}, \boldsymbol{C}]] + [\boldsymbol{B}, [\boldsymbol{C}, \boldsymbol{A}]] + [\boldsymbol{C}, [\boldsymbol{A}, \boldsymbol{B}]] = 0$$

(una forma más general de la cual se encontrará en §14.6).

Podemos escoger una base $(\boldsymbol{E}_1, \boldsymbol{E}_2, \ldots, \boldsymbol{E}_N)$ para el espacio vectorial de nuestras matrices $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}, \ldots$ (donde N es la dimensión del grupo $\mathcal{G}$, si la representación es fiel). Formando sus diversos conmutadores $[\boldsymbol{E}_\alpha, \boldsymbol{E}_\beta]$, expresamos estos en términos de los elementos de la base, para obtener relaciones (utilizando el convenio de suma)

$$[\boldsymbol{E}_\alpha, \boldsymbol{E}_\beta] = \gamma_{\alpha\beta}{}^{\chi}\, \boldsymbol{E}_\chi.$$

Las N^3 cantidades componentes $\gamma_{\alpha\beta}{}^{\chi}$ se denominan *constantes de estructura* para $\mathcal{G}$. No todas son independientes, pues satisfacen (véase §11.6 para la notación de paréntesis)

$$\gamma_{\alpha\beta}{}^{\chi} = -\gamma_{\beta\alpha}{}^{\chi}, \quad \gamma_{[\alpha\beta}{}^{\xi}\gamma_{\chi]\xi}{}^{\zeta} = 0,$$

[13.35] Demuestre todo esto.

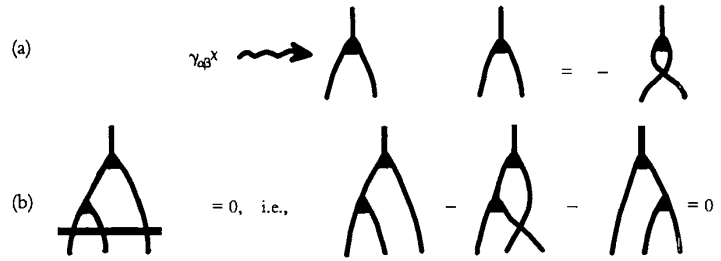


Fig. 13.11. (a) Constantes de estructura $\gamma_{\alpha\beta}{}^{\chi}$ de forma diagramática, que muestran antisimetría en α, β, y (b) la identidad de Jacobi.

en virtud de la antisimetría anterior y la identidad de Jacobi.[13.36] Estas relaciones están dadas de forma diagramática en la Fig. 13.11.

Es un hecho notable que la estructura del álgebra de Lie para una representación fiel (básicamente, el conocimiento de las constantes de estructura $\gamma_{\alpha\beta}{}^{\chi}$) es suficiente para determinar la naturaleza local precisa del grupo $\mathcal{G}$. Aquí «local» significa en una región $\mathcal{N}$ abierta N-dimesional (suficientemente pequeña) que rodea al elemento identidad $\boldsymbol{I}$ en la «variedad del grupo» $\widetilde{\mathcal{G}}$ cuyos puntos representan los diferentes elementos de $\mathcal{G}$ (véase la Fig. 13.12). De hecho, partiendo de un elemento $\boldsymbol{A}$ del grupo de Lie podemos construir un correspondiente grupo *finito* (i.e., no infinitesimal) por medio de la operación de exponenciación $e^{\boldsymbol{A}}$ definida al final de §13.4. (Esto será considerado con más detalle en §14.6.) Así pues, la teoría de representaciones de grupos continuos mediante transformaciones lineales (o mediante matrices) puede ser transferida en general al estudio de representaciones de álgebras de Lie mediante tales transformaciones, que, de hecho, es la práctica habitual en física.

Esto es particularmente importante en mecánica cuántica, en la que los propios elementos del álgebra de Lie tienen con frecuencia, y de una manera notable, interpretaciones directas como magnitudes físicas (tales como el momento angular, cuando el grupo $\mathcal{G}$ es el grupo de rotaciones, como veremos más adelante en §22.8).

Las matrices del álgebra de Lie tienden a tener una estructura con-

[13.36] Demuéstrelo.

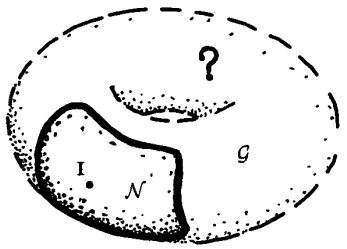


Fig. 13.12. El álgebra de Lie para una representación (fiel) de un grupo de Lie $\mathcal{G}$ (básicamente, conocimiento de las constantes de estructura $\gamma_{\alpha\beta}{}^{\chi}$) determina la estructura local de $\mathcal{G}$, es decir, fija la estructura de $\mathcal{G}$ dentro de una región abierta $\mathcal{N}$ (suficientemente pequeña) que rodea al elemento identidad $\mathbf{I}$, pero no nos dice nada de la naturaleza global de $\mathcal{G}$.

siderablemente más simple que las correspondientes matrices del grupo de Lie, al estar sujetas a restricciones lineales en lugar de restricciones no lineales (véase §13.10 para el caso de los grupos clásicos). ¡Este procedimiento es muy apreciado por los físicos cuánticos!

13.7. Espacios de representación tensoriales; reducibilidad

Existen formas de construir representaciones más elaboradas de un grupo $\mathcal{G}$, partiendo de una concreta. ¿Cómo lo hacemos? Supongamos que $\mathcal{G}$ está representado por alguna familia $\mathcal{T}$ de transformaciones lineales que actúan sobre un espacio vectorial n-dimensional $\mathbf{V}$. Este $\mathbf{V}$ se denomina *espacio de representación* para $\mathcal{G}$. Cualquier elemento t de $\mathcal{G}$ está ahora representado por una correspondiente transformación lineal $\boldsymbol{T}$ en $\mathcal{T}$, donde $\boldsymbol{T}$ efectúa $\boldsymbol{x} \mapsto \boldsymbol{T}\boldsymbol{x}$ para cada $\boldsymbol{x}$ perteneciente a $\mathbf{V}$. En la notación de índices (abstractos) (§12.7) escribimos esto $x^a \mapsto T^a{}_b x^b$, como en §13.3, o de forma diagramática, como en la Fig. 13.6a. Veamos cómo podemos encontrar otros espacios de representación para $\mathcal{G}$, partiendo del V dado.

Como primer ejemplo, recordemos de §12.3 la definición del *espacio dual* $\mathbf{V}^*$ de $\mathbf{V}$. Los elementos de $\mathbf{V}^*$ están definidos como aplicaciones lineales de $\mathbf{V}$ sobre los escalares. Podemos escribir la acción de $\boldsymbol{y}$ (en $\mathbf{V}^*$) sobre un elemento x de $\mathbf{V}$ como $y_a x^a$, en la notación de índices (§12.7). La notación $\boldsymbol{y} \bullet \boldsymbol{x}$ hubiera sido utilizada antes (§12.3) para esto ($\boldsymbol{y} \bullet \boldsymbol{x} = y_a x^a$), pero ahora también podemos utilizar la notación matricial

$$\mathbf{y}\mathbf{x} = y_a x^a,$$

donde tomamos **y** como un vector *fila* (i.e., una matriz $1 \times n$) y **x** como un vector *columna* (una matriz $n \times 1$). De acuerdo con nuestra transformación $\mathbf{x} \mapsto \boldsymbol{T}\mathbf{x}$, ahora considerada como una transformación *matricial*, el espacio dual $\mathbf{V}^*$ experimenta la transformación lineal

$$\mathbf{y} \mapsto \mathbf{yS}, \quad \text{i.e.,} \quad \gamma_a \mapsto \gamma_b S^b{}_a,$$

donde **S** es la *inversa* de **T**:

$$S = T^{-1}, \quad \text{de modo que} \quad S^a{}_b T^b{}_c = \delta^a_c,$$

puesto que, si $\mathbf{x} \mapsto \mathbf{Tx}$, necesitamos $\mathbf{y} \mapsto \mathbf{yT}^{-1}$ para asegurar que **yx** es conservado por $\mapsto$.

El uso de un vector fila, en lo que precede, nos da un orden de multiplicación no estándar. Es más habitual escribir las cosas a la inversa, empleando la notación de la *traspuesta* $\mathbf{A}^\mathsf{T}$ de una matriz **A**. Los elementos de la matriz $\mathbf{A}^\mathsf{T}$ son los mismos que los de **A**, pero con las filas y las columnas intercambiadas. Si **A** es cuadrada ($n \times n$), entonces también lo es $\mathbf{A}^\mathsf{T}$, y sus elementos son los de **A** reflejados en su diagonal principal (véase §13.3). Si **A** es rectangular ($m \times n$), entonces $\mathbf{A}^\mathsf{T}$ es $n \times m$, correspondientemente reflejada. Así pues $\mathbf{y}^\mathsf{T}$ es un vector columna estándar, y podemos escribir el anterior $\mathbf{y} \mapsto \mathbf{yS}$ como

$$\mathbf{y}^\mathsf{T} \mapsto \mathbf{S}^\mathsf{T}\mathbf{y}^\mathsf{T},$$

puesto que la operación traspuesta $^\mathsf{T}$ invierte el orden de multiplicación: $(\mathbf{AB})^\mathsf{T} = \mathbf{B}^\mathsf{T}\mathbf{A}^\mathsf{T}$. Vemos entonces que el espacio dual $\mathbf{V}^*$ de cualquier espacio de representación V es en sí mismo un espacio de representación de $\mathcal{G}$. Nótese que la operación inversa $^{-1}$ también invierte el orden de multiplicación, $(\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}$,[13.37] de modo que se recupera el orden de multiplicación necesario para una representación.

Consideraciones del mismo tipo son también válidas para los diversos espacios vectoriales de tensores construidos a partir de **V**; véase §12.8. Recordemos que un tensor **Q** de valencia $\left[{}^p_q\right]$ (sobre el espacio vectorial **V**) tiene una descripción de índices como una cantidad

$$Q^{f\ldots h}_{a\ldots c},$$

[13.37] ¿Por qué?

con q subíndices y p superíndices. Podemos sumar tensores a otros tensores de la misma valencia y podemos multiplicar tensores por escalares; los tensores de valencia $\left[{}^p_q\right]$ dada forman un espacio vectorial de dimensión n^{p+q} (el número total de componentes).[13.38] En abstracto, consideramos que **Q** pertenece a un espacio vectorial al que llamamos *producto tensorial*

$$\mathbf{V}^* \otimes \mathbf{V}^* \otimes \ldots \otimes \mathbf{V}^* \otimes \mathbf{V} \otimes \mathbf{V} \otimes \ldots \otimes \mathbf{V}$$

de q copias del espacio dual $\mathbf{V}^*$ y p copias de **V** ($p, q \geq 0$). (Llegaremos a esta noción de «producto tensorial» con un poco más de detalle en §23.3.) Recordemos la definición abstracta de un tensor, dada en §12.8, como una función multilineal. Esto bastará para nuestro propósito (aunque hay algunas sutilezas en el caso de un **V** de dimensión infinita, relevantes para las aplicaciones a estados cuánticos de muchas-partículas, que se necesitan en §23.8).[15]

Cuando quiera que se aplica una transformación lineal $x^a \mapsto T^a{}_b x^b$ a **V**, esto induce una correspondiente transformación lineal en el anterior espacio producto tensorial, dada explícitamente por[13.39]

$$Q^{f\ldots h}_{a\ldots c} \mapsto S^{a'}{}_a \ldots S^{c'}{}_c T^f{}_{f'} \ldots T^h{}_{h'} Q^{f'\ldots h'}_{a'\ldots c'}.$$

Todos estos índices requieren buena vista y un examen cuidadoso, para asegurarse de qué se suma con qué; así que yo recomiendo la notación diagramática, que es más clara, como se ilustra en la Fig. 13.13. Vemos que cada subíndice de $Q^{\cdots}_{\cdots}$ se transforma por la matriz inversa $\mathbf{S} = \mathbf{T}^{-1}$ (o, mejor, por $\mathbf{S}^\mathsf{T}$), como sucede con γ_a y cada superíndice por **T**, como sucede con x^a. En consecuencia, el espacio de tensores $\left[{}^p_q\right]$-valentes sobre V es también un espacio de representación para $\mathcal{G}$, de dimensión n^{p+q}.

No obstante, es probable que estos espacios de representación sean lo que se denomina *reducibles*. Para ilustrar esta situación, consideremos el caso de un tensor $\left[{}^2_0\right]$-valente Q^{ab}. Cualquier tensor semejante puede dividirse en su parte *simétrica* $Q^{(ab)}$ y su parte *antisimétrica* $Q^{[ab]}$ (§11.6 y §12.7):

$$Q^{ab} = Q^{(ab)} + Q^{[ab]},$$

[13.38] ¿Por qué este número?
[13.39] Demuéstrelo.

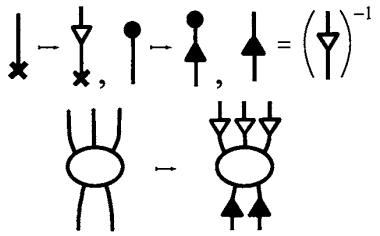


Fig. 13.13. La transformación lineal $x^a \mapsto T^a{}_b x^b$, aplicada a $\mathbf{x}$ en el espacio vectorial $\mathbf{V}$ (con T representada como un triángulo blanco), se extiende al espacio dual $\mathbf{V}^*$ por el uso de la inversa $S = T^{-1}$ (mostrada como un triángulo negro) y de ahí a los espacios $\mathbf{V}^* \otimes \ldots \otimes \mathbf{V}^* \otimes \mathbf{V} \otimes \ldots \otimes \mathbf{V}$ de tensores $Q[{}^p_q]$-valentes. Se ilustra el caso $p = 3$, $q = 2$ con Q representado como un óvalo con tres brazos y dos piernas $Q_{ab}{}^{cde} \propto S^{a'}{}_a S^{b'}{}_b T^c{}_{c'} T^d{}_{d'} T^e{}_{e'} Q_{a'b'}{}^{c'd'e'}$.

donde

$$Q^{(ab)} = \tfrac{1}{2}(Q^{ab} + Q^{ba}), \qquad Q^{[ab]} = \tfrac{1}{2}(Q^{ab} - Q^{ba}).$$

La dimensión del espacio *simétrico* $\mathbf{V}_+$ es $\frac{1}{2}n(n+1)$, y la del espacio *antisimétrico* $\mathbf{V}_-$ es $\frac{1}{2}n(n-1)$.[13.40] No es difícil ver que, bajo la transformación $x^a \mapsto T^a{}_b x^b$, de modo que $Q^{ab} \mapsto T^a{}_c T^b{}_d Q^{cd}$, las partes simétrica y antisimétrica se transforman en tensores que son de nuevo, respectivamente, simétrico y antisimétrico.[13.41] En consecuencia, los espacios $\mathbf{V}_+$ y $\mathbf{V}_-$ son, independientemente, espacios de representación para $\mathcal{G}$. Escogiendo una base para $\mathbf{V}$ donde los $\frac{1}{2}n(n+1)$ primeros elementos de la base están en $\mathbf{V}_+$ y los $\frac{1}{2}n(n-1)$ elementos restantes están en $\mathbf{V}_-$, obtenemos nuestra representación en la que todas las matrices son de la forma «diagonal por bloques» $n^2 \times n^2$

$$\begin{pmatrix} \mathbf{A} & \mathbf{O} \\ \mathbf{O} & \mathbf{B} \end{pmatrix},$$

[13.40] Demuéstrelo.
[13.41] Explíquelo.

donde $\mathbf{A}$ representa una matriz $\frac{1}{2}n(n+1) \times \frac{1}{2}n(n+1)$, $\mathbf{B}$ representa una matriz $\frac{1}{2}n(n-1) \times \frac{1}{2}n(n-1)$, y los dos $\mathbf{O}$ representan los apropiados bloques rectangulares de ceros.

Una representación de esta forma se conoce como la *suma directa* de la representación dada por las matrices $\mathbf{A}$ y la dada por las matrices $\mathbf{B}$. La representación en términos de tensores $[{}^2_0]$-valentes es por lo tanto *reducible* en este sentido.[13.42] La noción de «suma directa» se extiende también a cualquier número (tal vez infinito) de representaciones más pequeñas.

De hecho, hay un significado más general para el término «representación reducible», a saber, uno para el que hay una elección de base para la que todas las matrices de la representación pueden escribirse de la forma algo más complicada

$$\begin{pmatrix} \mathbf{A} & \mathbf{C} \\ \mathbf{O} & \mathbf{B} \end{pmatrix},$$

donde $\mathbf{A}$ es $p \times p$, $\mathbf{B}$ es $q \times q$, y $\mathbf{C}$ es $p \times q$, con $p, q \geq 1$ (para p y q dados). Nótese que si todas las matrices tienen esta forma, entonces las matrices $\mathbf{A}$ y las matrices $\mathbf{B}$, cada una por separado, constituyen una representación (más pequeña) de $\mathcal{G}$.[13.43] Si todas las matrices $\mathbf{C}$ son cero, tenemos el caso anterior en el que la representación es la suma directa de estas dos representaciones más pequeñas. Una representación se llama *irreducible* si no es reducible (esté o no $\mathbf{C}$ presente). Una representación se llama *completamente reducible* si nunca se da la situación anterior (con $\mathbf{C}$ no nula), de modo que es una suma directa de representaciones irreducibles.

Hay una clase importante de grupos continuos conocidos como grupos *semisimples*. Esta clase estudiada en extenso incluye los grupos simples mencionados en §13.2. Los grupos semisimples compactos tienen la propiedad de que todas sus representaciones son completamente reducibles. (Véanse §12.6 y la Fig. 12.12 para la definición de «com-

[13.42] Demuestre que el espacio de representación de los tensores $[{}^1_1]$-valentes es también reducible. *Sugerencia*: Descomponga un tensor semejante en una parte «libre de traza» y una parte «traza».
[13.43] Confírmelo.

pacto».) Es suficiente estudiar las representaciones irreducibles de un grupo semejante, pues cada representación es simplemente una suma directa de representaciones irreducibles. De hecho, toda representación irreducible de un grupo semejante es de dimensión finita (lo que no ocurre si permitimos que un grupo semisimple sea no compacto, caso en que también pueden darse representaciones que no son completamente reducibles).

¿Qué es un grupo semisimple? Recordemos las «constantes de estructura» $\gamma_{\alpha\beta}^{\chi}$ de §13.6, que especifican los paréntesis de Lie y definen la estructura local del grupo $\mathcal{G}$. Hay una cantidad de gran importancia conocida[16] como la «forma de Killing» κ que puede construirse a partir de las $\gamma_{\alpha\beta}{}^{\chi}$:[13.44]

$$\kappa_{\alpha\beta} = \gamma_{\alpha\zeta}{}^{\xi}\,\gamma_{\beta\xi}{}^{\zeta} = \kappa_{\beta\alpha}.$$

La forma diagramática de esta expresión se da en la Fig. 13.14. La condición para que $\mathcal{G}$ sea semisimple es que la matriz $\kappa_{\alpha\beta}$ sea no singular.

Aquí es necesario hacer algunos comentarios concernientes a la condición de compacticidad de un grupo semisimple. Para un conjunto dado de constantes de estructura $\gamma_{\alpha\beta}{}^{\chi}$, suponiendo que podemos tomarlas como números reales, podríamos considerar el álgebra de Lie real o el álgebra compleja obtenida a partir de ellas. En el caso complejo, no obtenemos un grupo compacto $\mathcal{G}$, pero sí podríamos hacerlo en el caso real. De hecho, la compacticidad ocurre en el caso real cuando $-\kappa_{\alpha\beta}$ es lo que se denomina *definido positivo* (el significado de este término se dará en §13.8). Para $\gamma_{\alpha\beta}{}^{\chi}$ dadas, en el caso de un grupo real $\mathcal{G}$, siempre podemos construir la *complexificación* $\mathbb{C}\mathcal{G}$ (al menos localmente) de $\mathcal{G}$ que resulta simplemente de utilizar las mismas $\gamma_{\alpha\beta}{}^{\chi}$, pero con coeficientes complejos en el álgebra de Lie. Sin embargo, diferentes grupos reales $\mathcal{G}$ a veces pueden dar lugar al mismo[17] $\mathbb{C}\mathcal{G}$. Estos distintos grupos reales se denominan diferentes *formas reales* del grupo complejo. Veremos ejemplos importantes de ello más adelante en §18.2, donde se comparan los movimientos euclídeos en 4 dimensiones y las simetrías Lorentz/Poincaré de la relatividad especial. Una propiedad notable de cualquier grupo de Lie semisimple complejo es que tiene exactamente una forma real $\mathcal{G}$ que es compacta.

[13.44] ¿Por qué hace $\kappa_{\alpha\beta} = \kappa_{\beta\alpha}$?

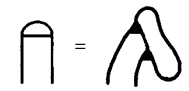


Fig. 13.14. La «forma de Killing» $\kappa_{\alpha\beta}$ definida a partir de las constantes de estructura $\gamma_{\alpha\zeta}{}^{\xi}$ por $\kappa_{\alpha\beta} = \gamma_{\alpha\zeta}{}^{\xi}\,\gamma_{\beta\xi}{}^{\zeta}$.

13.8. Grupos ortogonales

Volvamos ahora al grupo ortogonal. Ya hemos visto al principio de §13.3 cómo representar O(3) o SO(3) fielmente como transformaciones lineales de un espacio vectorial real 3-dimensional, con coordenadas cartesianas ordinarias (x, y, z), en las que la esfera

$$x^2 + y^2 + z^2 = 1$$

queda invariante (aquí el superíndice 2 significa el «cuadrado» normal). Escribamos esta ecuación en la notación de índices (§12.7), de modo que podamos generalizar a n dimensiones. La ecuación de nuestra esfera puede escribirse ahora

$$g_{ab}x^a x^b = 1,$$

que representa $(x^1)^2 + (x^2)^2 + \ldots + (x^n)^2 = 1$, estando dadas las componentes g_{ab} por

$$g_{ab} = \begin{cases} 1 & \text{si } a = b, \\ 0 & \text{si } a \neq b. \end{cases}$$

En la notación diagramática, yo recomendaba utilizar simplemente un «arco» para g_{ab}, como se indica en la Fig. 13.15a. También utilizaré la notación g_{ab} (con las *mismas* componentes explícitas que g_{ab}) para la cantidad *inversa* («arco invertido» en la Fig. 13.15a):

$$g_{ab}\,g^{bc} = \delta_a^c = g^{cb}\,g_{ba}.$$

El lector intrigado podría preguntarse muy razonablemente por qué he introducido dos notaciones nuevas, a saber, g_{ab} y g^{ab} para, exactamente, las mismas componentes matriciales que he denotado mediante δ_b^a en §13.3. La razón tiene que ver con la consistencia de la no-

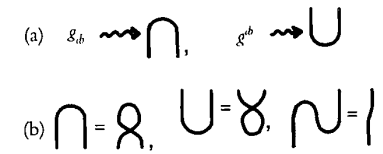


Fig. 13.15. (a) La métrica g_{ab} y su inversa g^{ab} en la notación diagramática de «arco». (b) Las relaciones $g_{ab} = g_{ba}$ (i.e., $\mathbf{g}^{\mathrm{T}} = \mathbf{g}$), $g_{ab} = g^{ba}$ y $g_{ab}g^{bc} = \delta^c_a$ en notación diagramática.

tación y con lo que sucede cuando se aplica una *transformación lineal* a las coordenadas, de acuerdo con cierto reemplazamiento

$$x^a \mapsto t^a{}_b x^b,$$

siendo $t^a{}_b$ no singular, de modo que tiene una inversa $s^a{}_b$:

$$t^a{}_b s^b{}_c = \delta^a_c = s^a{}_b t^b{}_c.$$

Este es formalmente el mismo tipo de transformación lineal que se ha considerado en §§13.3,7, pero ahora lo estamos haciendo de una forma muy diferente. En esas secciones nuestra transformación lineal se consideraba *activa*, de modo que el espacio vectorial **V** era realmente *movido* (sobre sí mismo). Aquí estamos considerando la transformación como *pasiva* en el sentido de que los objetos en consideración —y, de hecho, el propio espacio **V**— permanecen fijos, pero cambian las representaciones en términos de coordenadas. Otra manera de decirlo es que la base $(\boldsymbol{e}_1, \ldots, \boldsymbol{e}_n)$ que habíamos estado utilizando previamente (para la representación de magnitudes vectoriales/tensoriales en términos de componentes)[18] debe reemplazarse por alguna otra base. Véase la Fig. 13.16.

En correspondencia directa con lo que hemos visto en §13.7 para la transformación activa de un tensor, encontramos que el correspondiente cambio pasivo en las componentes $Q^{a\ldots c}_{p\ldots r}$ de un tensor **Q** está dado por[13.45]

$$Q^{a\ldots c}_{p\ldots r} \mapsto t^a{}_d \ldots t^c{}_f \; Q^{d\ldots f}_{j\ldots l} \; s^j{}_p \ldots s^l{}_r.$$

[13.45] Utilice la nota 13.18 para establecer esto.

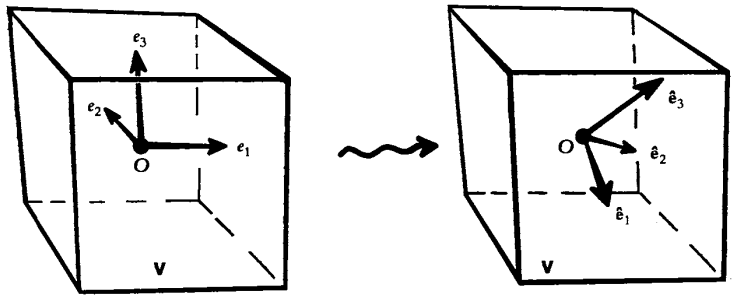


Fig. 13.16. Una transformación pasiva en un espacio vectorial **V** deja **V** fijo, pero cambia su descripción en coordenadas, i.e., la base $\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_n$ es reemplazada por otra base (caso $n = 3$ ilustrado).

Cuando aplicamos esto a δ^b_a encontramos que sus componentes quedan completamente inalteradas,[13.46] mientras que *no* sucede así con g_{ab} Además, tras un cambio de coordenadas general, las componentes g^{ab} serán por completo diferentes de g_{ab} (son matrices inversas). Así pues, la razón para los símbolos adicionales g^{ab} y g_{ab} es simplemente que estos solo pueden representar la misma matriz de coeficientes que δ^b_a en tipos especiales de sistemas de coordenadas (los «cartesianos») y que, en general, las componentes son *diferentes*. Ello es de particular importancia para la relatividad general, en la que el sistema de coordenadas no puede disponerse en general de modo que tenga esta forma (cartesiana) especial.

Un cambio de coordenadas general puede hacer que la matriz de componentes g_{ab} sea una matriz más complicada aunque no completamente general. Retiene la propiedad de simetría entre a y b dando una matriz *simétrica*. El término simétrica nos dice que la disposición cuadrada de componentes es simétrica respecto a su diagonal principal, i.e., $\mathbf{g}^{\mathrm{T}} = \mathbf{g}$ (utilizando la notación «traspuesta» de §13.7). En términos de notación de índices, esta simetría se expresa como una u otra de las dos formas equivalentes[13.47]

$$g_{ab} = g_{ba}, \; g^{ab} = g^{ba},$$

y véase la Fig. 13.15b para la forma diagramática de estas relaciones.

[13.46] ¿Por qué?
[13.47] ¿Por qué equivalentes?

¿Qué ocurre si vamos en la dirección opuesta? ¿Puede reducirse cualquier *matriz* $n \times n$ simétrica no singular a la forma de componentes de una delta de Kronecker? No del todo, no mediante una transformación lineal real de coordenadas. A lo que puede reducirse por tales medios es a esa misma forma excepto que puede haber algunos términos + 1 y algunos términos −1 en la diagonal principal. El número, p, de estos términos + 1, y el número, q, de términos −1 es un *invariante*, lo que quiere decir que no podemos obtener un número diferente ensayando alguna otra transformación lineal real. Este invariante (p, q) se denomina *signatura* de $\mathbf{g}$. (A veces es la diferencia $p - q$ la que se denomina signatura; a veces simplemente se escribe + ... + − ... − el número apropiado de veces para cada uno.) De hecho, esto funciona también para una $\mathbf{g}$ *singular*, pero entonces necesitamos algunos 0s en la diagonal principal y el número de 0s se convierte en parte de la signatura tanto como el número de + 1s y el número de −1s. Si solo tenemos + 1s, de modo que $\mathbf{g}$ es no singular y $q = 0$, decimos que $\mathbf{g}$ es *definida positiva*. Una $\mathbf{g}$ no singular para la que $p = 1$ y $q \neq 0$ (o $q = 1$ y $p \neq 0$) se denomina *lorentziana*, en honor del físico holándes H.A. Lorentz (1853-1928), cuya importante obra en relación con esto proporcionó una de las piedras fundacionales de la teoría de la relatividad; véanse §§17.6-9 y §§18.1-3.

Una caracterización alternativa de una matriz definida positiva $\mathbf{A}$, de considerable importancia en algunos otros contextos (véanse §20.3, §24.3 y §29.3), es que la matriz simétrica real $\mathbf{A}$ satisface

$$\mathbf{x}^{\mathsf{T}}\mathbf{A}\mathbf{x} > 0$$

para todo $\mathbf{x} \neq 0$. En notación de índices, esto es: «$A_{ab}x^a x^b > 0$, a menos que el vector x^a se anule».[13.48] Decimos que $\mathbf{A}$ es *definida no negativa* (o *semidefinida positiva*) si se da esto pero con $\geq$ en lugar de $>$ (de modo que ahora admitimos $\mathbf{x}^{\mathsf{T}}\mathbf{A}\mathbf{x} = 0$ para algún $\mathbf{x}$ no nulo).

En circunstancias adecuadas, un $\left[{}^{0}_{2}\right]$-tensor no singular simétrico g_{ab}, se denomina una *métrica* —o a veces una *pseudométrica* cuando $\boldsymbol{g}$ no es definida positiva. Esta terminología se aplica si vamos a utilizar la cantidad ds, definida por su cuadrado $\mathrm{d}s^2 = g_{ab}\mathrm{d}x^a\mathrm{d}x^b$, que nos proporciona

[13.48] ¿Puede confirmar esta caracterización?

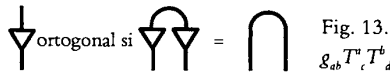


Fig. 13.17. $\boldsymbol{T}$ es una transformación ortogonal si $g_{ab}T^a{}_cT^b{}_d = g_{cd}$.

una noción de «distancia» a lo largo de curvas. En §14.7 veremos cómo se aplica esta noción a variedades curvas (véanse §10.2 y §§12.1,2), y en §17.8 cómo, en el caso lorentziano, nos proporciona una medida de «distancia» que es en realidad el *tiempo* de la teoría de la relatividad. A veces nos referimos a la cantidad

$$|\boldsymbol{v}| = (g_{ab}v^a v^b)^{1/2}$$

como la longitud del vector $\boldsymbol{v}$, con la forma de índices v^a.

Volvamos a la definición del grupo ortogonal $\mathrm{O}(n)$. Este es simplemente el grupo de transformaciones lineales en n dimensiones —llamadas transformaciones *ortogonales*— que conservan una métrica dada $\boldsymbol{g}$ definida positiva. «Conservar» $\boldsymbol{g}$ significa que una transformación ortogonal $\boldsymbol{T}$ tiene que satisfacer

$$g_{ab}T^a{}_cT^b{}_d = g_{cd}.$$

Este es un ejemplo de la regla de «cambio de coordenadas» para un tensor descrita en §13.7, tal como se aplica a g_{ab} (y véase la Fig. 13.17 para la forma diagramática de esta ecuación). Otra forma de decirlo es que la forma métrica $\mathrm{d}s^2$ del párrafo anterior no cambia bajo transformaciones ortogonales. Si queremos, podemos insistir en que el sistema de componentes g_{ab} sea en realidad el mismo que la delta de Kronecker —que, en efecto, proporciona la definición de O(3) que se da en §§13.1,3—, pero el grupo que se obtiene es el mismo[19] cualquiera que sea la g_{ab} ($n \times n$) definida positiva que escojamos.[13.49]

Con la realización en componentes concreta de g_{ab} como la delta de Kronecker, las matrices que describen nuestras transformaciones ortogonales son las que satisfacen[13.50]

$$\boldsymbol{T}^{-1} = \boldsymbol{T}^{\mathsf{T}},$$

[13.49] Explique por qué.
[13.50] Explique esto. ¿Qué es $\boldsymbol{T}^{-1}$ en los casos pseudoortogonales (definidos en el siguiente párrafo)?

llamadas *matrices ortogonales*. Las matrices $n \times n$ ortogonales reales proporcionan una realización concreta para el grupo $O(n)$. Para particularizar al grupo no reflexivo $SO(n)$, exigimos que el determinante sea igual a la unidad:[13.51]

$$\det \boldsymbol{T} = 1.$$

También podemos considerar los correspondientes grupos *pseudo-ortogonales* $O(p, q)$ y $SO(p, q)$ que se obtienen cuando $\boldsymbol{g}$, aunque no singular, no es necesariamente definida positiva y tiene una signatura (p, q) más general. El caso $p = 1$, $q = 3$ (o, de manera equivalente, $p = 3$, $q = 1$), llamado el grupo de *Lorentz*, desempeña un papel fundamental en la teoría de la relatividad, como se ha indicado antes. Asimismo, veremos (si ignoramos las reflexiones) que el grupo de Lorentz es el mismo que el grupo de simetrías del 3-espacio hiperbólico que se ha descrito en §2.7, y también (si ignoramos las reflexiones espaciales) el mismo que el grupo de simetrías de la esfera de Riemann, conseguido mediante las transformaciones bilineales (de Moebius) estudiadas en §8.2. Será mejor dejar para más adelante las explicaciones de estos hechos notables hasta que investiguemos la geometría espaciotemporal de Minkowski de la teoría de la relatividad especial (§§18.4,5). En §33.2 también veremos que estos hechos tienen una importancia seminal para la teoría de twistores.

¿Hasta qué punto son «diferentes» los diversos grupos $O(p, q)$, para $p + q = n$, con n fijo? (Los casos definido positivo y lorentziano se contrastan, para $n = 2$ y $n = 3$, en la Fig. 13.18.) Todos están íntimamente relacionados, pues todos tienen la misma dimensión, $\frac{1}{2}n(n-1)$; son lo que se denominan *formas reales* de un mismo grupo complejo $O(n, \mathbb{C})$, la *complexificación* de $O(n)$. Este grupo complejo se define del mismo modo que $O(n)$ $(= O(n, \mathbb{R}))$, pero ahora se permite que las transformaciones lineales sean *complejas*. En realidad, aunque he expresado mis consideraciones anteriores en términos de transformaciones lineales reales, existe una discusión paralela donde «complejo» reemplaza a «real» en

[13.51] Explique por qué esto es equivalente a conservar la forma de volumen $\boldsymbol{\varepsilon}_{a\ldots c}$, i.e., $\boldsymbol{\varepsilon}_{a\ldots c} T^a_p \ldots T^c_r = \boldsymbol{\varepsilon}_{p\ldots r}$. Además, ¿por qué basta con la conservación del signo?

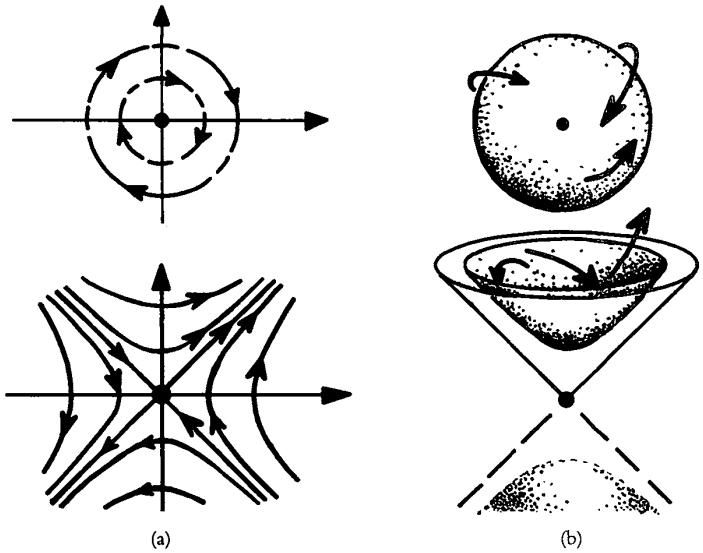


Fig. 13.18. (a) Se contrastan O(2,0) y O(1,1). (b) Análogamente, se contrastan O(3,0) y O(1,2), ilustrándose la «esfera unidad» en cada caso. Para O(1,2) esta «esfera» es un plano hiperbólico (o dos copias de tal plano) (véanse §§2.4,5 y §18.4).

todas partes. (Por ejemplo, las coordenadas x^a se vuelven complejas y así lo hacen también las componentes de nuestras matrices.) La única diferencia esencial con lo que se ha dicho antes reside con el concepto de *signatura*. Existen transformaciones de coordenadas lineales complejas que pueden convertir un -1 de una realización diagonal de g_{ab} en un $+1$ y *viceversa*,[13.52] de modo que ahora no tenemos una noción significativa de signatura. El único invariante[20] de $\boldsymbol{g}$, en el caso complejo, es lo que se denomina su *rango*, que es el número de términos no nulos en su realización diagonal. Para una $\boldsymbol{g}$ no singular, el rango tiene que ser máximo, i.e., n.

¿Cuándo es importante la diferencia entre estas diversas formas reales y cuándo no lo es? Esta puede ser una cuestión delicada, pero los físicos suelen ser bastante displicentes acerca de las distinciones, inclu-

[13.52] ¿Por qué?

so si pueden ser importantes. El caso definido positivo tiene la virtud de que el grupo es compacto, y buena parte de las matemáticas son más fáciles en tales situaciones (véase §13.7). A veces, la gente trasvasa alegremente resultados del caso compacto a los casos no compactos ($p \neq 0 \neq q$), pero con frecuencia esto no está justificado. (Por ejemplo, en el caso compacto solo hay que estar interesado en las representaciones que son de dimensión finita, pero en el caso no compacto aparecen representaciones adicionales de dimensión infinita.) Por otra parte, se dan otras situaciones en las que pueden obtenerse intuiciones considerables ignorando las distinciones. (Podemos comparar esto con el descubrimiento de Lambert de la fórmula para el área del triángulo hiperbólico en términos de ángulos dada en §2.4. Él obtuvo su fórmula permitiendo que su esfera tuviera un radio imaginario. Esto es similar a un cambio de signatura, lo que equivale a permitir que algunas coordenadas tengan valores imaginarios. En §18.4, Fig. 18.9, trataré de argumentar que la aproximación de Lambert a la geometría no euclídea es perfectamente justificable.)

Las diferentes formas reales posibles de $O(n, \mathbb{C})$ se distinguen por cierto conjunto de *desigualdades* sobre los elementos de matriz (tales como $\det \boldsymbol{T} > 0$). Una de las características de la *teoría cuántica* es que tales desigualdades son a menudo *violadas* en procesos físicos. Por ejemplo, las magnitudes *imaginarias* pueden, en cierto sentido, tener una importancia físicamente *real* en mecánica cuántica, de modo que la distinción entre signaturas diferentes puede quedar difuminada. Por otra parte, tengo la impresión de que los físicos suelen ser algo menos cuidadosos sobre estas cuestiones de lo que, en mi opinión, deberían ser. En realidad, esta cuestión tendrá una considerable relevancia para nosotros cuando analicemos varias teorías modernas (§28.9, §31.13 y §32.3). Pero lo veremos con detalle más adelante. ¡Este es el «cajón de problemas» que he insinuado en §11.2!

13.9. Grupos unitarios

El grupo $O(n, \mathbb{C})$ nos proporciona *una* manera de generalizar la noción de un «grupo de rotación» desde los números reales a los complejos.

Pero hay otra manera que, en ciertos contextos, tiene una importancia aún mayor. Esta es la noción de un grupo *unitario*.

¿Qué significa «unitario»? El grupo ortogonal concierne a la conservación de una *forma cuadrática*, que podemos escribir de forma equivalente como $g_{ab}x^a x^b$ o $\mathbf{x}^{\mathsf{T}}\mathbf{g}\mathbf{x}$. En el caso de un grupo unitario, utilizamos transformaciones lineales *complejas* que, en su lugar, conservan lo que se denomina una forma *hermítica* (llamada así por el importante matemático francés del siglo XIX Charles Hermite 1822–1901).

¿Qué es una forma hermítica? Volvamos primero al caso ortogonal. En lugar de una forma cuadrática (en $\boldsymbol{x}$), igualmente podríamos haber utilizado la forma bilineal simétrica (en $\boldsymbol{x}$ e $\boldsymbol{y}$)

$$\boldsymbol{g}(\boldsymbol{x}, \boldsymbol{y}) = g_{ab}x^a y^b = \mathbf{x}^{\mathsf{T}}\mathbf{g}\mathbf{y}.$$

Esto aparece como un ejemplo concreto de la definición de un tensor como «forma multilineal» dada en §12.8, aplicada al $\left[{}^0_2\right]$-tensor $\boldsymbol{g}$ (y poniendo $\boldsymbol{y} = \boldsymbol{x}$, recuperamos la forma cuadrática anterior). La simetría de $\boldsymbol{g}$ se expresaría entonces como

$$\boldsymbol{g}(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{g}(\boldsymbol{y}, \boldsymbol{x}),$$

y la linealidad en la segunda variable $\boldsymbol{y}$ como

$$\boldsymbol{g}(\boldsymbol{x}, \boldsymbol{y} + \boldsymbol{w}) = \boldsymbol{g}(\boldsymbol{x}, \boldsymbol{y}) + \boldsymbol{g}(\boldsymbol{x}, \boldsymbol{w}), \quad \boldsymbol{g}(\boldsymbol{x}, \lambda\boldsymbol{y}) = \lambda\boldsymbol{g}(\boldsymbol{x}, \boldsymbol{y}).$$

Para la *bilinealidad* exigimos también linealidad en la *primera* variable $\boldsymbol{x}$, pero esto se sigue ahora de la simetría.

Una *forma hermítica* $\boldsymbol{h}(\boldsymbol{x}, \boldsymbol{y})$ satisface en su lugar la simetría hermítica

$$\boldsymbol{h}(\boldsymbol{x}, \boldsymbol{y}) = \overline{\boldsymbol{h}(\boldsymbol{y}, \boldsymbol{x})},$$

junto con linealidad en la segunda variable $\boldsymbol{y}$:

$$\boldsymbol{h}(\boldsymbol{x}, \boldsymbol{y} + \boldsymbol{w}) = \boldsymbol{h}(\boldsymbol{x}, \boldsymbol{y}) + \boldsymbol{h}(\boldsymbol{x}, \boldsymbol{w}), \quad \boldsymbol{h}(\boldsymbol{x}, \lambda\boldsymbol{y}) = \lambda\boldsymbol{h}(\boldsymbol{x}, \boldsymbol{y}).$$

La simetría hermítica implica ahora lo que se denomina *antilinealidad* en la primera variable:

$$\boldsymbol{h}(\boldsymbol{x} + \boldsymbol{w}, \boldsymbol{y}) = \boldsymbol{h}(\boldsymbol{x}, \boldsymbol{y}) + \boldsymbol{h}(\boldsymbol{w}, \boldsymbol{y}), \quad \boldsymbol{h}(\lambda\boldsymbol{x}, \boldsymbol{y}) = \bar{\lambda}\boldsymbol{h}(\boldsymbol{x}, \boldsymbol{y}).$$

Mientras que un grupo ortogonal conserva una forma bilineal simétrica (no singular), las transformaciones lineales complejas que

conservan una forma hermítica no singular nos dan un grupo unitario.

¿Para qué nos sirven tales formas? Una forma bilineal no singular (no necesariamente hermítica) $\boldsymbol{g}$ nos proporciona un medio de identificar el espacio vectorial $\mathbf{V}$, al que pertenecen $\boldsymbol{x}$ e $\boldsymbol{y}$, con el espacio dual $\mathbf{V}^*$. Así, si $\boldsymbol{v}$ pertenece a $\mathbf{V}$, entonces $\boldsymbol{g}(\boldsymbol{v},)$ nos proporciona una aplicación lineal en $\mathbf{V}$, que aplica el elemento $\boldsymbol{x}$ de $\mathbf{V}$ en el número $\boldsymbol{g}(\boldsymbol{v}, \boldsymbol{x})$. En otras palabras, $\boldsymbol{g}(\boldsymbol{v},)$ es un elemento de $\mathbf{V}^*$ (véase §12.3). De forma indexada, este elemento de $\mathbf{V}^*$ es el covector $v^a g_{ab}$, que es costumbre escribir con la misma letra v, pero con el índice descendido (véase también §14.7) mediante g_{ab}, de acuerdo con

$$v_b = v^a g_{ab}.$$

La inversa de esta operación se consigue elevando el índice de v_a mediante el uso del $\left[{}^{2}_{0}\right]$-tensor métrico inverso g^{ab}:

$$v^a = g^{ab} v_b.$$

Necesitaremos el análogo de esto en el caso hermítico. Como antes, cada elección de elemento v del espacio vectorial $\mathbf{V}$ nos proporciona un elemento $\boldsymbol{h}(\boldsymbol{v},)$ del espacio dual $\mathbf{V}^*$. Sin embargo, la diferencia reside en que ahora $\boldsymbol{h}(\boldsymbol{v},)$ depende antilinealmente de $\boldsymbol{v}$ y no linealmente: así $\boldsymbol{h}(\lambda \boldsymbol{v},) = \bar{\lambda} \boldsymbol{h}(\boldsymbol{v},)$.

Una forma equivalente de decirlo es que $\boldsymbol{h}(\boldsymbol{v},)$ es *lineal* en $\bar{v}$, siendo esta cantidad vectorial $\bar{v}$ el complejo conjugado de v. Consideramos que estos vectores complejo conjugados constituyen un espacio vectorial $\bar{v}$ independiente. Este punto de vista es particularmente útil para la notación de índices (abstractos), donde se utiliza un «alfabeto» de índices independiente, digamos a', b', c', ..., para estos elementos complejo conjugados, donde no se permiten sumas (contracciones) entre índices primados y no primados. La operación de conjugación compleja intercambia los índices primados con los no primados. En la notación de índices, nuestra forma hermítica se representa por una matriz de cantidades $h_{a'b}$ con un subíndice primado de cada tipo, de modo que

$$\boldsymbol{h}(\boldsymbol{x}, \boldsymbol{y}) = h_{a'b} \bar{x}^{a'} y^b$$

(siendo $\bar{x}^{a'}$ el complejo conjugado del elemento x^a), donde la hermiticidad se expresa por

$$h_{a'b} = \overline{h_{b'a}}.$$

La colección de cantidades $h_{a'b}$ nos permite bajar o subir un índice, pero no cambia índices primados por no primados, y viceversa, de modo que nos remite al espacio dual del espacio complejo conjugado:

$$\bar{v}_a = \bar{v}^{a'} h_{a'b}, \quad v_{a'} = h_{a'b} v^b.$$

Para las inversas de estas operaciones —donde se supone que la forma hermítica es no singular (i.e., la matriz de componentes $h^{ab'}$ es no singular)—, necesitamos la inversa $h^{ab'}$ de $h_{a'b}$

$$h^{ab'} h_{b'c} = \delta^a_c, \quad h_{a'b} h^{bc'} = \delta^{c'}_{a'},$$

de donde[13.53]

$$\bar{v}^{a'} = \bar{v}_b h^{ba'}, \quad v^a = h^{ab'} v_{b'}.$$

Nótese que todos los índices primados pueden eliminarse utilizando $h_{a'b}$ (y la inversa correspondiente $h^{ab'}$) en virtud de las relaciones anteriores, que puede aplicarse índice a índice a cualquier cantidad tensorial. El espacio complejo conjugado es por ello «identificado» con el espacio dual, en lugar de tener que ser un espacio totalmente separado.

La operación de «conjugación compleja» —normalmente llamada *conjugación hermítica*— que incorpora esta identificación con el dual dentro de la noción de conjugación compleja (aunque no escrita habitualmente en la notación de índices) tiene una importancia fundamental en la mecánica cuántica, así como en muchas otras áreas de las matemáticas y de la física (tales como la teoría de twistores; véase §33.5). En la literatura mecanocuántica, esta se denota a menudo con una *daga* «†», pero a veces mediante un *asterisco* «*».

Yo prefiero el asterisco, que es más habitual en la literatura matemática, de modo que aquí lo utilizaré —en negrita—. El asterisco es apropiado porque intercambia los papeles del espacio vectorial $\mathbf{V}$ y su dual $\mathbf{V}^*$. Un tensor complejo de valencia $\left[{}^{p}_{q}\right]$ (del que se han eliminado

[13.53] Verifique estas relaciones explicando la consistencia notacional de $h^{ab'}$.

todos los índices primados, como antes) se aplica mediante $*$ en un tensor de valencia $\left[{}^{q}_{p}\right]$. Así pues, los superíndices se convierten en subíndices y los subíndices en superíndices bajo la acción de $*$. Cuando se aplica a escalares, $*$ es simplemente la operación ordinaria de conjugación compleja. La operación $*$ es una noción equivalente para la propia forma hermítica $\boldsymbol{h}$.

La operación de conjugación hermítica más familiar (que ocurre cuando se toma la delta de Kronecker como componentes $h_{a'b}$) toma simplemente el complejo conjugado de cada componente, reorganizando las componentes de modo que se lean superíndices como subíndices y subíndices como superíndices. En consecuencia, la matriz de componentes de una transformación lineal se toma como la traspuesta de su conjugada compleja (a veces llamada la *traspuesta conjugada* de la matriz), de modo que en el caso 2×2 tenemos

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}^* = \begin{pmatrix} \bar{a} & \bar{c} \\ \bar{b} & \bar{d} \end{pmatrix}.$$

Una *matriz hermítica* es una matriz que es igual a su conjugada hermítica en este sentido. Este concepto, y el más general y abstracto de *operador hermítico*, son de gran importancia en teoría cuántica.

Notemos que $*$ es *antilineal* en el sentido

$$(\boldsymbol{T} + \boldsymbol{U})^* = \boldsymbol{T}^* + \boldsymbol{U}^*,$$
$$(z\boldsymbol{T})^* = \bar{z}\boldsymbol{T}^*,$$

aplicada a tensores $\boldsymbol{T}$ y $\boldsymbol{U}$, ambos de la misma valencia, y para cualquier número complejo z. La acción de $*$ debe conservar también los productos de tensores, pero debido a la inversión de la posición de los índices, invierte el orden de las contracciones; en particular, cuando se aplica $*$ a transformaciones lineales (consideradas como tensores con un subíndice y un superíndice), el orden de multiplicación se invierte:

$$(\boldsymbol{LM})^* = \boldsymbol{M}^* \boldsymbol{L}^*.$$

En la notación diagramática es muy cómodo representar esta operación de conjugación como una reflexión en un plano horizontal. Esto intercambia subíndices y superíndices, como se requiere; véase la Fig. 13.19.

La operación $*$ nos permite definir un *producto escalar hermítico* en-

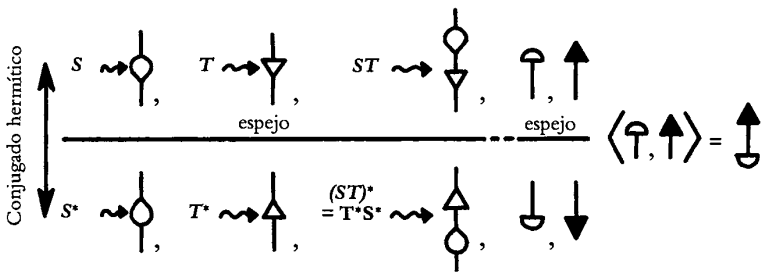


Fig. 13.19. Operación de conjugación hermítica ($*$) convenientemente representada como reflexión en un plano horizontal. Esta intercambia «brazos» y «piernas» e invierte el orden de multiplicación: $(\boldsymbol{ST})^* = \boldsymbol{T}^*\boldsymbol{S}^*$. Se da la expresión diagramática para el producto escalar hermítico $\langle v|w\rangle = v^*w$ (de modo que tomar su conjugado complejo reflejaría el diagrama en el extremo derecho al revés).

tre dos elementos v y w, de $\mathbf{V}$, a saber, el producto escalar del covector v^* con el vector w (las diferentes notaciones son útiles en contextos diferentes):

$$\langle v \mid w\rangle = v^* \bullet w = \boldsymbol{h}(v, w)$$

(y véase la Fig. 13.19), y tenemos

$$\langle v \mid w\rangle = \overline{\langle w \mid v\rangle}.$$

En el caso particular $w = v$, obtenemos la *norma* de v con respecto a $*$:

$$\|v\| = \langle v \mid v\rangle.$$

Podemos escoger una base $(\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_n)$ para $\mathbf{V}$, y entonces las componentes $h_{a'b}$ en esta base son simplemente los n^2 números complejos

$$h_{a'b} = \boldsymbol{h}(\boldsymbol{e}_a, \boldsymbol{e}_b) = \langle \boldsymbol{e}_a \mid \boldsymbol{e}_b\rangle,$$

que constituyen los elementos de una matriz hermítica. La base $(\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_n)$ se denomina *pseudo-ortonormal* con respecto a $*$, si

$$\langle \boldsymbol{e}_i \mid \boldsymbol{e}_j\rangle = \begin{cases} \pm 1 & \text{si } i = j; \\ 0 & \text{si } i \neq j; \end{cases}$$

en el caso en que todos los signos $\pm$ son $+$, i.e., cuando cada ± 1 es exactamente 1, la base es *ortonormal*.

Siempre puede encontrarse una base pseudo-ortonormal, pero hay muchas elecciones posibles. Con respecto a cualquiera de estas bases, la matriz $h_{a'b}$ es diagonal, con solo 1s y −1s debajo de la diagonal. El número total de 1s, p, siempre es el mismo, para una $*$ dada, con independencia de la elección particular de base, y también lo es el número total de −1s, q. Esto nos permite definir la noción invariante de *signatura* (p, q) para la operación $*$.

Si $q = 0$, decimos que $*$ es *definida positiva*. En este caso,[21] la norma de cualquier vector no nulo es siempre positiva:[13.54]

$$v \neq 0 \quad implica \quad \|v\| > 0.$$

Nótese que esta noción de «definida positiva» generaliza la de §13.8 al caso complejo.

Una transformación lineal $\boldsymbol{T}$ cuya inversa es $\boldsymbol{T}^*$, de modo que

$$\boldsymbol{T}^{-1} = \boldsymbol{T}^*, \text{ i.e., } \boldsymbol{T}\,\boldsymbol{T}^* = \boldsymbol{I} = \boldsymbol{T}^*\,\boldsymbol{T},$$

se llama *unitaria* en el caso definido positivo y *pseudounitaria* en caso contrario.[13.55] El término «matriz unitaria» se refiere a una matriz $\boldsymbol{T}$ que satisface la relación anterior cuando $*$ representa la operación traspuesta conjugada habitual, de modo que $\boldsymbol{T}^{-1} = \bar{\boldsymbol{T}}$.

El grupo de transformaciones unitarias en n dimensiones, o de matrices unitarias $(n \times n)$, se denomina *grupo unitario* $\mathrm{U}(n)$. De forma más general, obtenemos el grupo pseudounitario $\mathrm{U}(p, q)$ cuando $*$ tiene signatura (p, q).[22] Si las transformaciones tienen determinante unidad, entonces obtenemos correspondientemente $\mathrm{SU}(p)$ y $\mathrm{SU}(p, q)$. Las transformaciones unitarias desempeñan un papel esencial en mecánica cuántica (y también tienen gran valor en diversos contextos puramente matemáticos).

13.10. Grupos simplécticos

En las dos secciones precedentes hemos encontrado los grupos ortogonales y unitarios. Estos son ejemplos de los denominados *grupos clásicos*,

[13.54] Demuéstrelo.

[13.55] Demuestre que estas transformaciones son precisamente las que conservan la correspondencia entre vectores v y covectores v^*, y que son las que conservan $h_{ab'}$.

a saber, los grupos simples de Lie distintos de los excepcionales; véase §13.2. La lista de grupos clásicos se completa con la familia de los grupos *simplécticos*. Estos tienen gran importancia en la física *clásica*, como veremos sobre todo en §20.4, y también en física cuántica, especialmente en el caso de la dimensión infinita (§26.3).

¿Qué es un grupo simpléctico? Volvamos de nuevo a la noción de una forma bilineal, pero donde en lugar de la simetría $(\boldsymbol{g}(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{g}(\boldsymbol{y}, \boldsymbol{x}))$ requerida para definir el grupo ortogonal, imponemos *antisimetría*

$$\boldsymbol{s}(\boldsymbol{x}, \boldsymbol{y}) = -\boldsymbol{s}(\boldsymbol{y}, \boldsymbol{x}),$$

junto con linealidad

$$\boldsymbol{s}(\boldsymbol{x}, \boldsymbol{y} + \boldsymbol{w}) = \boldsymbol{s}(\boldsymbol{x}, \boldsymbol{y}) + \boldsymbol{s}(\boldsymbol{x}, \boldsymbol{w}), \quad \boldsymbol{s}(\boldsymbol{x}, \lambda\boldsymbol{y}) = \lambda\boldsymbol{s}(\boldsymbol{x}, \boldsymbol{y}),$$

donde la linealidad en la primera variable $\boldsymbol{x}$ se sigue ahora de la *anti*simetría. Podemos escribir nuestra forma antisimétrica de diversas maneras como

$$\boldsymbol{s}(\boldsymbol{x}, \boldsymbol{y}) = x^a s_{ab} y^b = \mathbf{x}^{\mathsf{T}}\mathbf{S}\mathbf{y},$$

igual que en el caso simétrico, pero donde s_{ab} es *antisimétrica*:

$$s_{ba} = -s_{ab} \quad \text{i.e.,} \quad \mathbf{S}^{\mathsf{T}} = -\mathbf{S},$$

siendo $\mathbf{S}$ la matriz de componentes de s_{ab}. Exigimos que $\mathbf{S}$ sea no singular. Entonces s_{ab} tiene una inversa s^{ab}, que satisface[23]

$$s_{ab}s^{bc} = \delta_a^c = s^{cb}s_{ba},$$

donde $s^{ab} = -s^{ba}$.

Notemos que, por analogía con una matriz simétrica, una matriz antisimétrica $\mathbf{S}$ es igual a *menos* su traspuesta. Es importante observar que una matriz $n \times n$ antisimétrica $\mathbf{S}$ puede ser no singular solo si n es par.[13.56] Aquí n es la dimensión del espacio $\mathbf{V}$ al que pertenecen $\mathbf{x}$ e $\mathbf{y}$, y, de hecho, tomamos n par.

Los elementos $\mathbf{T}$ de $\mathrm{GL}(n)$ que conservan una s_{ab} antisimétrica no singular (o, lo que es equivalente, la forma bilineal $\mathbf{s}$), en el sentido

$$s_{ab}T^a{}_c\,T^b{}_d = s_{cd}, \quad \text{i.e.,} \quad \mathbf{T}^{\mathsf{T}}\mathbf{S}\,\mathbf{T} = \mathbf{S},$$

[13.56] Demuéstrelo.

se denominan *simplécticos*, y el grupo de dichos elementos se denomina *grupo simpléctico* (un grupo de importancia muy considerable en mecánica clásica, como veremos en §20.4). Sin embargo, existe alguna confusión en la literatura con respecto a esta terminología. Es más preciso desde el punto de vista matemático definir un grupo simpléctico (real) como una forma real del grupo simpléctico *complejo* $\mathrm{Sp}(\frac{1}{2}n, \mathbb{C})$, definido como el grupo de las $T^a{}_b$ *complejas* (o **T**) que satisfacen la relación anterior. La forma real que se acaba de definir es no compacta; pero, de acuerdo con los comentarios que se han hecho al final de §13.7 —y siendo $\mathrm{Sp}(\frac{1}{2}n, \mathbb{C})$ semisimple—, existe otra forma real de este grupo complejo que *es* compacta, y es *esta* la que normalmente se conoce como el grupo simpléctico (real) $\mathrm{Sp}(\frac{1}{2}n)$.

¿Cómo encontramos estas diferentes formas reales? De hecho, como sucede en el caso de los grupos ortogonales, existe una noción de *signatura* que no es tan conocida como en los casos de los grupos ortogonales y los unitarios. El grupo simpléctico de transformaciones reales que conservan s_{ab} sería el caso de «signatura dividida» de signatura $(\frac{1}{2}n, \frac{1}{2}n)$. En el caso compacto, el grupo simpléctico tiene signatura $(n, 0)$ o $(0, n)$.

¿Cómo se define esta signatura? Para cada par de números naturales p y q tales que $p + q = n$, podemos definir una correspondiente «forma real» del grupo complejo $\mathrm{Sp}(\frac{1}{2}n, \mathbb{C})$ tomando solo aquellos elementos que son también pseudounitarios para la signatura (p, q), i.e., que pertenecen a $\mathrm{U}(p, q)$ (véase §13.9). Esto nos da[24] el grupo pseudosimpléctico $\mathrm{Sp}(p, q)$. (Otra manera de decirlo es decir que $\mathrm{Sp}(p, q)$ es la intersección de $\mathrm{Sp}(\frac{1}{2}n, \mathbb{C})$ con $\mathrm{U}(p, q)$.) En notación de índices, podemos definir $\mathrm{Sp}(p, q)$ como el grupo de transformaciones lineales complejas $T^a{}_b$ que conservan la s_{ab} antisimétrica, como antes, y también una matriz hermítica **H** de componentes $h_{a'b}$, en el sentido de que

$$\bar{T}^{a'}{}_{b'}\, T^a{}_b h_{a'a} = h_{b'b},$$

donde **H** tiene signatura (p, q) (de modo que podemos encontrar una base pseudo-ortonormal para la que **H** es diagonal con p entradas 1 y q entradas -1; véase §13.9).[25] El grupo simpléctico *clásico* compacto $\mathrm{Sp}(\frac{1}{2}n)$ es mi $\mathrm{Sp}(n, 0)$ (o $\mathrm{Sp}(0, n)$), pero la forma de mayor importancia en física clásica es $\mathrm{Sp}(\frac{1}{2}n, \frac{1}{2}n)$.[13.57]

Igual que con los grupos ortogonales y unitarios, podemos encontrar elecciones de base para las que las componentes s_{ab} tienen una forma particularmente simple. Sin embargo, ahora no podemos tomar una forma diagonal ¡porque la única matriz diagonal antisimétrica es nula! En su lugar, podemos tomar la matriz s_{ab} consistente en bloques 2×2 bajo la diagonal principal, de la forma

$$\begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}.$$

En el caso familiar con signatura dividida $\mathrm{Sp}(\frac{1}{2}n, \frac{1}{2}n)$, podemos tomar las transformaciones lineales *reales* que conservan dicha forma. El caso general $\mathrm{Sp}(p, q)$ se muestra tomando, en lugar de transformaciones reales, transformaciones pseudounitarias de signatura (p, q).[13.58]

Para varios valores (pequeños) de p y q, algunos de los grupos ortogonales, unitarios y simplécticos son iguales («isomorfos»), o al menos localmente iguales («localmente isomorfos»), en el sentido de tener las mismas álgebras de Lie (cf. §13.6).[26] El ejemplo más elemental es el grupo SO(2), que describe el grupo de simetrías no reflexivas de un círculo, que es el mismo que el grupo unitario U(1), el grupo multiplicativo de números complejos $e^{i\theta}$ (θ real) de módulo unidad.[13.59] De particular importancia para la física es el hecho de que SU(2) y Sp(1) son iguales, y son localmente iguales a SO(3) (pues son el recubrimien-

[13.57] Encuentre descripciones explícitas de Sp(1) y Sp(1,1) utilizando esta fórmula. ¿Puede ver por qué los grupos $\mathrm{Sp}(0, n)$ son compactos?
[13.58] Muestre por qué estas dos descripciones diferentes para el caso $p = q = \frac{1}{2}$ son equivalentes.
[13.59] ¿Por qué son iguales?

to doble de este último grupo, de acuerdo con la naturaleza doble de la representación cuaterniónica de las rotaciones en el 3-espacio, como se ha descrito en §11.3). Esto tiene gran importancia para la física cuántica del *espín* (§22.8). De importancia en la teoría de la relatividad es el hecho de que SL(2, ℂ), siendo el mismo que Sp(1, ℂ), es localmente igual a la parte no reflexiva del grupo de Lorentz O(1, 3) (de nuevo un recubrimiento doble del mismo). También encontramos que SU(1, 1), Sp(1, 1) y SO(2, 1) son iguales, y hay varios ejemplos más. Especialmente valiosa para la teoría de twistores es la identidad local entre SU(2, 2) y la parte no reflexiva del grupo O(2, 4) (véase §33.3).

El álgebra de Lie de un grupo simpléctico se obtiene buscando soluciones $\mathbf{X}$ de la ecuación matricial

$$\mathbf{X}^{\mathsf{T}}\mathbf{S} + \mathbf{S}\,\mathbf{X} = 0, \quad \text{i.e.,} \quad \mathbf{S}\,\mathbf{X} = (\mathbf{S}\,\mathbf{X})^{\mathsf{T}},$$

de modo que la transformación infinitesimal (elemento del álgebra de Lie) $\mathbf{X}$ es simplemente $\mathbf{S}^{-1}$ multiplicado por una matriz $n \times n$ simétrica. Esto permite ver directamente la dimensionalidad $\frac{1}{2}n(n+1)$ del grupo simpléctico. Nótese que $\mathbf{X}$ tiene que ser libre de traza (i.e., traza $\mathbf{X} = 0$; véase §13.4).[13.60] Las álgebras de Lie para grupos ortogonales y unitarios se obtienen también fácilmente, en términos, respectivamente, de matrices antisimétricas y múltiplos imaginarios puros de matrices hermíticas, cuyas dimensiones respectivas son $n(n-1)/2$ y n^2.[13.61]

Notemos de §13.4 que, para que las transformaciones tengan determinante unidad, la traza del elemento infinitesimal $\boldsymbol{X}$ debe anularse. Esto es automático en el caso simpléctico (señalado antes), y en el caso ortogonal todos los elementos infinitesimales tienen determinante unidad.[13.62] En el caso unitario, la restricción a SU(n) es una condición adicional (traza $\boldsymbol{X} = 0$), de modo que la dimensión del grupo se reduce a $n^2 - 1$.

[13.60] Explique de dónde procede la ecuación $\boldsymbol{X}^{\mathsf{T}}\boldsymbol{S} + \boldsymbol{S}\boldsymbol{X} = 0$ y por qué $\boldsymbol{S}\boldsymbol{X} = (\boldsymbol{S}\boldsymbol{X})^{\mathsf{T}}$. ¿Por qué se anula la traza $\boldsymbol{X}$? Dé el álgebra de Lie explícitamente. ¿Por qué es de esta dimensión?

[13.61] Describa dichas álgebras de Lie y obtenga dichas dimensiones.

[13.62] ¿Por qué, y qué significa esto geométricamente?

Los *grupos clásicos* mencionados en §13.2, a veces etiquetados A_m, B_m, C_m, D_m (para $m = 1, 2, 3, \ldots$), son simplemente los grupos respectivos SU($m+1$), SO($2m+1$), Sp(m) y SO($2m$) que hemos examinado en §§13.8-10, y vemos de lo anterior que tienen dimensiones respectivas $m(m+2)$, $m(2m+1)$, $m(2m+1)$, y $m(2m-1)$, como se ha afirmado en §13.2. Así, el lector ha tenido la oportunidad de hacerse una idea de todos los grupos simples clásicos. Como hemos visto, tales grupos, y algunas de las otras «formas reales» (o sus complexificaciones), desempeñan un papel importante en física. Nos familiarizaremos un poco más con todo esto en el próximo capítulo. Como se ha mencionado al principio de este capítulo, según la física moderna, todas las interacciones están gobernadas por «conexiones gauge» que, técnicamente, dependen crucialmente de espacios que tienen simetrías exactas. Sin embargo, aún tenemos que saber qué es realmente una «teoría gauge». Esto se abordará en el capítulo 15.

Notas

Sección 13.1

13.1. Abel nació en 1802 y murió de tuberculosis en 1829, a los veintiséis años. La teoría de grupos no abelianos ($ab \neq ba$) más general fue introducida por el matemático francés Évariste Galois (1811-1832), que murió de manera aún más trágica en un duelo antes de cumplir los veintiún años, tras haber pasado la noche anterior a su muerte desarrollando febrilmente sus revolucionarias ideas sobre el uso de grupos para investigar la solubilidad de las ecuaciones algebraicas, hoy día denominada *teoría de Galois*.

13.2. Deberíamos también tomar nota de que «–C» significa «tomar el complejo conjugado, y multiplicarlo por –1», i.e., –C = (–1)C.

13.3. La S viene de «especial» (que significa «de determinante unidad»), lo que, en el contexto presente, nos dice simplemente que están excluidos los movimientos que invierten la orientación. La O viene de «ortogonal», lo que tiene que ver con el hecho de que los movimientos que representa conservan la «ortogonalidad» (es decir, la naturaleza de ángu-

lo recto) de los ejes de coordenadas. El 3 se refiere a que estamos considerando rotaciones en tres dimensiones.

13.4. Existe un notable teorema que nos dice que no solo todo grupo continuo es también *suave* (i.e., C^0 implica C^1, en la notación de §§6.3,6, e incluso C^0 implica C^∞), sino que es también *analítico* (i.e., C^0 implica C^ω). Este famoso resultado, que representó la solución de lo que se llegó a conocer como «el quinto problema de Hilbert», fue obtenido por Andrew Mattei Gleason, Deane Montgomery, Leo Zippin y Hidehiko Yamabe en 1953; véase Montgomery y Zippin (1955). Esto justifica el uso de series de potencias en §13.6.

Sección 13.2

13.5. Véase Van der Waerden (1985), pp. 166-174.

13.6. Véase Devlin (1988).

13.7. Véanse Conway y Norton (1972) y Dolan (1996).

Sección 13.3

13.8. Veremos en §14.1. que un espacio euclídeo es un ejemplo de espacio *afín*. Si seleccionamos un punto concreto (origen) O, se convierte en un espacio vectorial.

13.9. En muchos lugares de este libro será conveniente —y a veces esencial— escalonar los índices en un símbolo de tipo tensorial. En el caso de una transformación lineal, necesitamos hacerlo para expresar el orden de la multiplicación de matrices.

13.10. Esta región es un espacio vectorial de dimensión r (donde $r < n$). Llamamos a r el *rango* de la matriz o de la transformación lineal $\boldsymbol{T}$. Una matriz $n \times n$ no singular tiene rango n. (El concepto de «rango» se aplica también a matrices rectangulares.) Compárese con la nota 12.18.

13.11. Para una historia de la teoría de matrices, véase MacDuffe (1933).

Sección 13.5

13.12. En aquellas situaciones degeneradas en las que los autovectores no llenan todo el espacio (es decir, algún d es menor que el correspondiente r), aún podemos encontrar una forma canónica, pero ahora permitimos que aparezcan 1s inmediatamente por encima de la diagonal principal, que residen dentro de bloques cuadrados cuyos términos diagonales son autovalores *iguales* (*forma normal de Jordan*); véase Anton y Busby (2003). Al parecer, Weierstrauss había encontrado,

efectivamente, esta forma normal en 1868, dos años antes de Jordan; véase también Hawkins (1977).

Sección 13.6

13.13. Para ilustrar este punto, consideremos $SL(n, \mathbb{R})$ (i.e., los elementos de determinante unidad del propio $GL(n, \mathbb{R})$). Este grupo tiene un «recubrimiento doble» $\widetilde{SL}(n, \mathbb{R})$ (siempre que $n \geq 3$), que se obtiene a partir de $SL(n, \mathbb{R})$ básicamente de la misma forma con que, en efecto, encontramos el recubrimiento doble $\widetilde{SO}(3)$ de $SO(3)$ cuando considerábamos las rotaciones de un libro, con el cinturón unido, en §11.3. Así, $\widetilde{SO}(3)$ es el grupo de rotaciones (no reflexivas) de un *objeto espinorial* en un 3-espacio ordinario. De la misma manera, podemos considerar «objetos espinoriales» que están sujetos a las transformaciones lineales más generales que permiten «comprimir» o «estirar», como se ha discutido en §13.3. De esta manera, llegamos al grupo $\widetilde{SL}(n, \mathbb{R})$, que es *localmente* igual que $SL(n, \mathbb{R})$, pero que, de hecho, no puede ser representado con fidelidad en ningún $GL(m)$. Véase la nota 15.9.

13.14. Esta noción está bien definida; cf. nota 13.4.

Sección 13.7

13.15. Véase Thirring (1983).

13.16. Aquí, de nuevo, tenemos un ejemplo de la manera caprichosa de dar nombre a los conceptos matemáticos. Mientras que muchas nociones de gran importancia en esta materia, a las que se une convencionalmente el nombre de Cartan (por ejemplo, «subálgebra de Cartan, entero de Cartan»), se debían originalmente a Killing (véase §13.2), lo que conocemos como «forma de Killing» se debe realmente a Cartan (y Hermann Weyl); véase Hawkins (2000), sección 6.2. Sin embargo, el «vector de Killing» que encontraremos en §30.6 se debe en realidad a Killing (Hawkins, 2000, nota 20, p. 128).

13.17. Estoy siendo (deliberadamente) un poco resbaladizo desde el punto de vista matemático en mi uso de la expresión «igual» en este tipo de contexto. El término matemático estricto es «isomorfo».

Sección 13.8

13.18. No he sido muy explícito sobre este procedimiento hasta este momento. Una base $\boldsymbol{e} = (\boldsymbol{e}_1, \ldots, \boldsymbol{e}_n)$ para $\mathbf{V}$ está asociada con una base dual —que es una base $\boldsymbol{e}^* = (\boldsymbol{e}^1, \ldots, \boldsymbol{e}^n)$ para $\mathbf{V}^*$— con la propiedad de

que $\boldsymbol{e}^i \bullet \boldsymbol{e}^j = \delta^i_j$. Las componentes de un tensor $\mathbf{Q}[{}^p_q]$-valentes se obtienen aplicando la función multilineal de §12.8 a los diversos conjuntos de p elementos de la base dual y q elementos de la base: $\mathbf{Q}^{f\ldots h}_{a\ldots c} = \mathbf{Q}(\boldsymbol{e}^f, \ldots, \boldsymbol{e}^h; \boldsymbol{e}_a, \ldots, \boldsymbol{e}_c)$.

13.19. Véase la nota 13.3.

13.20. Véase la nota 13.10. El lector puede preguntarse por qué los $T^a{}_b$ de §13.5 pueden tener muchas invariantes, a saber, todos sus *autovalores* λ_1, λ_2, λ_3, ... λ_n, mientras que g_{ab} no los tiene. La respuesta reside simplemente en la diferencia en el comportamiento de la transformación implícito en la diferente posición de los índices.

Sección 13.9

13.21. Nótese que, en el caso definido positivo, $(\boldsymbol{e}^*_1, \boldsymbol{e}^*_2, \ldots, \boldsymbol{e}^*_n)$ es una *base dual* de $(\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots \boldsymbol{e}_n)$, en el sentido de la nota 13.18.

13.22. Los grupos $\mathrm{U}(p, q)$, para $p + q = n$ fijo, así como $\mathrm{GL}(n, \mathbb{R})$, tienen todos la misma complexificación, a saber, $\mathrm{GL}(n, \mathbb{C})$, y todos pueden ser considerados como diferentes formas reales de este grupo complejo.

Sección 13.10

13.23. Podemos entonces utilizar s_{ab} y s^{ab} para subir y bajar índices de tensores, igual que con g_{ab} y g^{ab}, de modo que $\nu_a = s_{ab}\nu^b$, $\nu^a = s^{ab}\nu_b$ (véase §13.8); pero, debido a la antisimetría, debemos tener mucho cuidado de hacer consistente el orden de los índices. Los lectores que estén familiarizados con el cálculo 2-espinorial (véase Penrose y Rindler, 1984, vol. 1) notarán una ligera discrepancia notacional entre nuestro s_{ab} y el ε_{AB} de dicho cálculo.

13.24. No conozco ninguna terminología o notación estándar para estas diversas formas reales, de modo que la notación $\mathrm{Sp}(p, q)$ se ha inventado para los fines presentes.

13.25. De hecho, todo elemento de $\mathrm{Sp}(\frac{1}{2}n, \mathbb{C})$ tiene determinante unidad, de modo que no necesitamos un «$\mathrm{SSp}(\frac{1}{2}n)$» por analogía con $\mathrm{SO}(n)$ y $\mathrm{SU}(n)$. La razón es que hay una expresión (el «pfaffiano») para el $\varepsilon\ldots$ de Levi-Civita en términos de los s_{ab}, que debe conservarse cualesquiera que sean los s_{ab}.

13.26. Véase la nota 13.17.

14

Cálculo infinitesimal en variedades

14.1. ¿Diferenciación en una variedad?

En el capítulo anterior (en §§13.3,6-10) hemos visto cómo pueden actuar los grupos de simetría sobre espacios vectoriales, representados por transformaciones lineales de dichos espacios. Para un grupo específico podemos considerar que el espacio vectorial posee cierta *estructura* concreta que es conservada por las transformaciones. Esta noción de estructura es importante. Por ejemplo, podría ser una estructura métrica, en el caso del grupo ortogonal (§13.8), o una estructura hermítica, que es conservada por un grupo unitario (§13.9). Como se ha señalado antes, la teoría de representación de grupos como acciones sobre espacios vectoriales tiene de un modo general gran importancia en muchas áreas de las matemáticas y la física, especialmente en teoría cuántica, donde, como veremos más adelante (especialmente, en §22.3), los espacios vectoriales con una estructura (producto escalar) hermítica forman la base esencial de dicha teoría.

Sin embargo, un espacio vectorial es en sí mismo un tipo muy especial de espacio, y se necesita algo mucho más general para las matemáticas de gran parte de la física moderna. Ni siquiera la antigua geometría de Euclides es un espacio vectorial, porque un espacio vectorial tiene que tener un punto privilegiado particular, a saber, el *origen* (dado por el vector cero), mientras que en la geometría euclídea todos los puntos están en pie de igualdad. De hecho, el espacio euclídeo es un ejemplo de lo que se denomina un espacio *afín*. Un espacio afín es similar a un espacio vectorial, salvo que «olvidamos» el origen; en efecto, es un espacio en el que existe una noción consistente de paralelo-

gramo.[14.1],[14.2] En cuanto especificamos un punto particular como origen, esto nos permite definir la suma de vectores mediante la «ley del paralelogramo» (véanse §13.3 y la Fig. 13.4).

El espaciotiempo curvo de la extraordinaria teoría de la relatividad general de Einstein es por supuesto más general que un espacio vectorial; es una 4-variedad. Pero su noción de geometría espaciotemporal requiere cierta estructura (local), y no tan solo la de una variedad suave (como se ha estudiado en el capítulo 12). Análogamente, los espacios de configuración o los espacios de fases de los sistemas físicos (brevemente considerados en §12.1) también poseen estructuras locales. ¿Cómo asignamos esta estructura necesaria? Una estructura local semejante podría proporcionar una medida de «distancia» entre puntos (en el caso de una estructura métrica), o de «área de una superficie» (como se especifica en el caso de una estructura simpléctica; cf. §13.10), o de «ángulo» entre curvas (como sucede con la estructura conforme de una superficie de Riemann; cf. §8.2), etc. En todos estos ejemplos, las nociones de espacio vectorial son las que se necesitan para decirnos cuál es esta geometría local, siendo el espacio vectorial en cuestión el *espacio tangente* n-dimensional $\mathcal{T}_p$ de un punto típico p de la variedad $\mathcal{M}$ (donde podemos considerar $\mathcal{T}_p$ como la vecindad inmediata de p en $\mathcal{M}$ «infinitamente estirada»; véase la Fig. 12.6).

En consecuencia, las diversas estructuras de grupo y entidades tensoriales que hemos encontrado en el capítulo 13 pueden tener una relevancia local en los puntos individuales de una variedad. Veremos que el espaciotiempo curvo de Einstein tiene una estructura local que está dada por una (pseudo)métrica lorentziana (§13.8) en cada espacio tangente, mientras que los espacios de fases (cf. §12.1) de la mecánica clásica tienen estructuras simplécticas locales (§13.10). Estos dos ejemplos

[14.1] Representemos por $[a, b; c, d]$ el enunciado «*abcd* forma un paralelogramo» (donde *abcd* se toman cíclicamente, como en §5.1). Tomemos como axiomas (i) para cualesquiera a, b y c, existe d tal que $[a, b; c, d]$; (ii) si $[a, b; c, d]$, entonces $[b, a; d, c]$ y $[a, c; b, d]$; (iii) si $[a, b; c, d]$ y $[a, b; e, f]$, entonces $[c, d; e, f]$. Demuestre que cuando se singulariza cualquier punto escogido y se etiqueta como el origen esta estructura algebraica se reduce a la de un «espacio vectorial», aunque sin la operación de «multiplicación por un escalar», dada en §11.1, es decir, obtenemos las reglas de un grupo abeliano aditivo; véase el ejercicio [13.2].

[14.2] ¿Puede ver cómo generalizar esto al caso no abeliano?

de variedades con estructura desempeñan papeles clave en la teoría física moderna. Pero ¿qué forma de cálculo infinitesimal puede aplicarse dentro de tales espacios?

Como se acaba de comentar, las variedades n-dimensionales que hemos estudiado en el capítulo 12 solo necesitaban ser suaves, sin ninguna otra estructura local especificada. En semejante variedad suave y sin estructura $\mathcal{M}$ hay relativamente pocas operaciones significativas basadas en el cálculo. Y, lo que es más importante, ni siquiera tenemos una noción general de diferenciación que pueda aplicarse en general dentro de $\mathcal{M}$.

Debería aclarar este punto. En cualquier carta de coordenadas particular podríamos, por supuesto, diferenciar simplemente las diversas cantidades de interés con respecto a cada una de las coordenadas $x^1, x^2, \ldots, x^n$ en dicha carta, mediante el uso de operadores de derivación (parcial) $\partial/\partial x^1, \partial/\partial x^2, \ldots, \partial/\partial x^n$. (véase §10.2). Pero en la mayoría de los casos las respuestas no tendrían significado geométrico porque dependen de la elección concreta (arbitraria) de coordenadas que se ha hecho, y las respuestas no empalmarían en general cuando pasamos de una carta a otra (cf. Fig. 10.7).

Sin embargo, en §12.6 hemos tomado nota de una importante noción de diferenciación que realmente sí se aplica en una n-variedad (sin estructura) suave general —que coincide de una carta a la siguiente—, a saber, la *derivada exterior* de una forma diferencial. Pero esta operación tiene un alcance algo limitado, en la medida en que se aplica solo a p-formas y, además, no da mucha información acerca de cómo varía una p-forma. ¿Podemos dar una noción más completa de «derivada» de una cantidad en una variedad suave general, por ejemplo, de un campo vectorial o tensorial? Una noción semejante tendría que estar definida independientemente de cualesquiera coordenadas particulares que pudieran haberse escogido para etiquetar puntos en una carta. De hecho, sería bueno tener algún tipo de cálculo infinitesimal independiente de las coordenadas que pueda aplicarse a estructuras en variedades, y que nos permita expresar cómo varía un campo vectorial o tensorial cuando nos movemos de un lugar a otro. Pero ¿cómo puede lograrse esto?

14.2. Transporte paralelo

Recordemos de §10.3 y §12.3 que en el caso de un campo *escalar* Φ en una n-variedad suave general $\mathcal{M}$, hemos sido capaces de proporcionar una medida apropiada de su «tasa de cambio», a saber, la 1-forma $d\Phi$, donde $d\Phi = 0$ es la condición de que Φ sea constante (a lo largo de regiones conexas de $\mathcal{M}$). Sin embargo, esta idea no funcionará para una cantidad tensorial general. Ni siquiera funcionará para un campo vectorial ξ. ¿Por qué es así? Una dificultad está en que en una variedad general no tenemos una noción apropiada de la constancia de ξ (como veremos enseguida), mientras que cualquier operación de diferenciación («gradiente») digna que se aplica a ξ debería tener la propiedad de que su anulación señala la constancia de ξ (como, de hecho, $d\Phi = 0$ señala la constancia de un campo escalar Φ). Con más generalidad, esperaríamos que para un ξ «no constante», dicha operación de derivación debería estar midiendo la *desviación* de ξ respecto de la constancia.

¿Por qué hay un problema con esta noción de «constancia» vectorial, en una n-variedad general $\mathcal{M}$? Un campo vectorial constante ξ, en un espacio euclídeo ordinario, debería tener la propiedad de que todas las «flechas» de su descripción geométrica son paralelas entre sí. Así pues, alguna noción de «paralelismo» tendría que formar parte de la estructura de $\mathcal{M}$. Uno podría estar preocupado por esto, teniendo en cuenta la cuestión del quinto postulado de Euclides —el postulado de las paralelas—, capital en la discusión del capítulo 2. La geometría hiperbólica, por ejemplo, no admite campos vectoriales que puedan considerarse inequívocamente «paralelos» en todas partes. En cualquier caso, una noción de «paralelismo» no es algo que $\mathcal{M}$ posea meramente en virtud de ser una variedad suave. En la Fig. 14.1 se ilustra esta dificultad en el caso de una 2-variedad compuesta a partir de dos cartas empalmadas del espacio euclídeo. La noción euclídea normal de «paralela» no es compatible entre una carta y la siguiente.

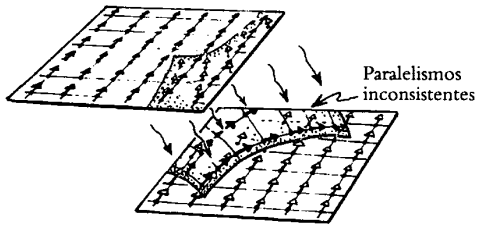


Fig. 14.1. Es probable que la noción euclídea de «paralela» sea inconsistente en el solapamiento entre cartas de coordenadas.

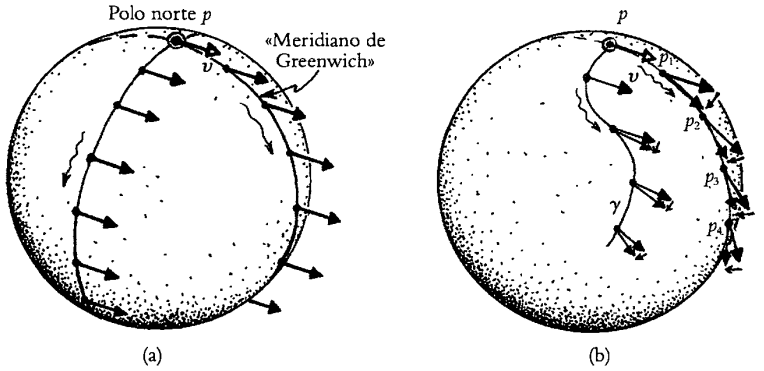


Fig. 14.2. Paralelismo sobre la esfera S^2. Escojamos p en el polo norte, con vector tangente $\boldsymbol{v}$ que apunta a lo largo del meridiano de Greenwich. ¿Qué vectores tangentes, en otros puntos de S^2, van a ser considerados «paralelos» a $\boldsymbol{v}$? (a) La noción euclídea directa de «paralela», a partir de la inmersión de S^2 en $\mathbb{E}^3$, no funciona porque (excepto a lo largo del meridiano perpendicular al meridiano de Greenwich) las paralelas $\boldsymbol{v}$ no permanecen tangentes a S^2. (b) Esto se remedia moviendo $\boldsymbol{v}$ paralelo a lo largo de una curva dada γ, proyectando continuamente a la tangencia con la esfera. (Considérese γ formado por un gran número de minúsculos segmentos $p_0p_1, p_1p_2, p_2p_3, \ldots$, que se proyectan en cada etapa. Luego tómese el límite a medida que los segmentos se hacen cada vez más pequeños.) Esta noción de *transporte paralelo* se indica para el meridiano de Greenwich, pero también para una curva general γ.

Para hacernos una idea de qué tipo de noción de «paralelismo» es adecuada, sería muy útil examinar primero la geometría intrínseca de una esfera 2-dimensional ordinaria S^2. Escojamos un punto particular p en S^2 (por ejemplo, el polo norte, por claridad) y un vector tangente concreto $\boldsymbol{v}$ en p (por ejemplo, apuntando en la dirección del meridiano de Greenwich; véase la Fig. 14.2a). ¿Qué otros vectores tangentes, en otros puntos de S^2, van a considerarse «paralelos» a $\boldsymbol{v}$? Si utilizamos simplemente la noción euclídea de «paralela» que se hereda de la inmersión estándar de S^2 en el 3-espacio euclídeo, entonces encontramos que en la mayoría de los puntos q de S^2 no hay ningún vector tangente a S^2 que sea «paralelo» a $\boldsymbol{v}$ en este sentido, puesto que el plano tan-

gente en q no contiene normalmente la dirección de $\boldsymbol{v}$. (Solo el círculo máximo que pasa por p y es perpendicular al meridiano de Greenwich en p contiene puntos en los que hay vectores tangentes a S^2 que serían «paralelos» a $\boldsymbol{v}$ en este sentido.) La noción apropiada de paralelismo, en S^2, debería referirse solo a vectores tangentes, de modo que debemos hacer todo lo posible para llevar de nuevo la dirección de $\boldsymbol{v}$ al plano tangente de q, cuando alejamos poco a poco q de p. De hecho, esta idea funciona, y funciona de forma muy bella, pero ahora hay un nuevo aspecto, pues la noción de paralelismo que obtenemos *depende del camino* a lo largo del cual alejamos q de p.[1] Esta dependencia del camino en el concepto de paralelismo es el nuevo ingrediente esencial, y versiones de la misma subyacen en todas las modernas teorías satisfactorias de las interacciones entre partículas, además de la relatividad general de Einstein.

Intentemos comprender esto un poco mejor. Consideremos un camino γ en S^2, que parte del punto p y termina en algún otro punto q en S^2. Supondremos que γ está formado por un gran número, N, de minúsculos segmentos $p_0p_1, p_1p_2, p_2p_3, \ldots, p_{N-1}p_N$, donde el punto de partida es $p_0 = p$ y el segmento final termina en $p_N = q$. Imaginemos que movemos $\boldsymbol{v}$ a lo largo de γ, y que a lo largo de cada uno de estos segmentos $p_{r-1}p_r$ movemos $\boldsymbol{v}$ paralelo a sí mismo —en nuestro sentido anterior de utilizar el 3-espacio euclídeo ambiente— y entonces proyectamos $\boldsymbol{v}$ sobre el espacio tangente en p_r. Véase la Fig. 14.2b. Por este procedimiento terminamos con un vector tangente en q que, de forma aproximada, podemos considerar que ha sido deslizado paralelo a sí mismo a lo largo de γ desde p a q, lo más próximo que sea posible a hacerlo totalmente dentro de la superficie. De hecho, este procedimiento dependerá un poco de cómo se aproxima γ por la sucesión de segmentos, pero puede demostrarse que en el límite, a medida que los segmentos se hacen cada vez más pequeños, obtenemos una respuesta bien definida que no depende de la forma detallada precisa en que dividimos γ en segmentos. Este procedimiento se conoce como *transporte paralelo* de $\boldsymbol{v}$ a lo largo de γ. En la Fig. 14.3 se indica qué aspecto tendría este transporte paralelo a lo largo de cinco caminos diferentes (todos ellos círculos máximos) que parten de p.

¿Cuál es entonces esta dependencia del camino antes mencionada?

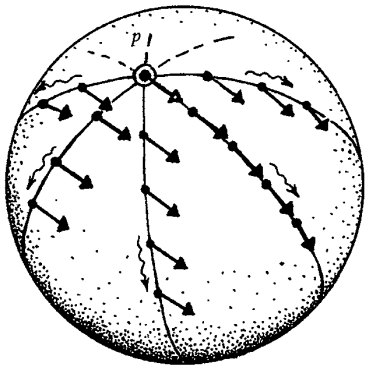


Fig. 14.3. Transporte paralelo de $\boldsymbol{v}$ a lo largo de cinco caminos diferentes (todos ellos círculos máximos).

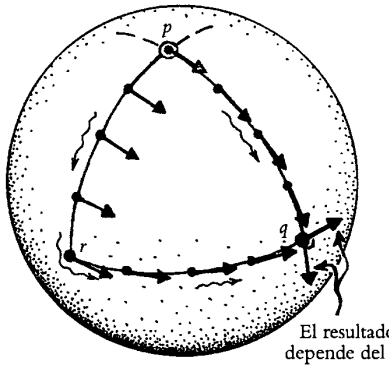


Fig. 14.4. Dependencia del camino del transporte paralelo. Esta se ilustra utilizando dos caminos distintos de p a q, uno de los cuales es una ruta directa de círculo máximo y el otro consiste en un par de arcos de círculo máximo unidos en el punto intermedio r. El transporte paralelo a lo largo de estos dos caminos da resultados en q que difieren en una rotación de un ángulo recto.

En la Fig. 14.4 he marcado puntos p, q en S^2 y dos caminos de p a q, uno de los cuales es la ruta directa por el círculo máximo y otro que consiste en un par de arcos de círculo máximo unidos en el punto intermedio r. De la geometría de la Fig. 14.3 vemos que el transporte paralelo a lo largo de estos dos caminos (uno de ellos con una esquina, pero eso no es importante) da dos resultados finales completamente diferentes que, en este caso, difieren en una rotación de un ángulo recto. Nótese que la discrepancia es solo una rotación de la dirección del vector. Existen razones generales por las que una noción de transporte paralelo definida de esta forma concreta conservará siempre la longitud del vector. (Sin embargo, existen otros tipos de «transporte paralelo» para los que no se da esto. Estas cuestiones serán importantes en sec-

ciones posteriores (§14.7, §§15.7,8 y §19.4). Podemos ver esta discrepancia angular en una forma extrema cuando nuestro camino γ es un lazo cerrado (de modo que $p = q$), en cuyo caso es probable que exista una discrepancia entre las direcciones inicial y final del vector tangente transportado paralelamente. De hecho, para una esfera geométrica de radio unidad, esta discrepancia es un ángulo de rotación que, cuando se mide en radianes, es exactamente igual al área total del lazo (donde las regiones rodeadas en sentido negativo cuentan negativamente).[14.3]

14.3. Derivada covariante

¿Cómo podemos utilizar un concepto de «transporte paralelo» tal como este para definir una noción apropiada de *diferenciación* de campos vectoriales (y así de tensores en general)? La idea esencial consiste en que podemos comparar el comportamiento real de un campo vectorial (o tensorial) en cierta dirección que se aleja de un punto p con el transporte paralelo del mismo vector llevado en la misma dirección desde p, y restar el segundo del primero. Podríamos aplicar esto a un desplazamiento finito a lo largo de cierta curva γ, pero para definir una derivada (primera) de un campo vectorial exigimos solo un desplazamiento infinitesimal a partir de p, y esto depende únicamente de la forma en que la curva «parte de» p, i.e., depende solo del vector tangente $\boldsymbol{w}$ de γ en p (Fig. 14.5). Para denotar la noción de diferenciación que surge de este modo, es habitual utilizar un símbolo ∇ conocido como *operador de derivada covariante* o simplemente una *conexión*.

Un requisito fundamental de un operador semejante (y que resulta ser cierto para la noción antes esbozada para S^2) es que dependa *linealmente* del vector $\boldsymbol{w}$. Así pues, escribiendo $\underset{w}{\nabla}$ para la derivada covariante definida por (la dirección de) el desplazamiento de $\boldsymbol{w}$, para dos de tales vectores desplazamiento $\boldsymbol{w}$ y $\boldsymbol{u}$, debe satisfacer

[14.3] Vea si puede confirmar esta afirmación en el caso de un triángulo esférico (un triángulo en S^2 formado por arcos de círculo máximo), donde puede suponer la fórmula de Hariot de 1603 para el área de un triángulo esférico dada en 2.6.

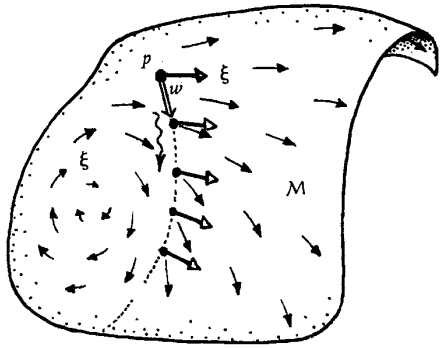


Fig. 14.5. La noción de derivada covariante puede entenderse en relación con el transporte paralelo. La forma en que un campo vectorial ξ en $\mathcal{M}$ varía de un punto a otro (flechas con punta negra) se mide por su diferencia respecto al patrón que proporciona el transporte paralelo (flechas con punta blanca). Esta comparación puede hacerse a lo largo de una curva γ (que parte de p), pero para la derivada primera covariante $\underset{w}{\nabla}$ en p solo necesitamos conocer el vector tangente $\boldsymbol{w}$ a γ en p, que determina la derivada covariante $\underset{w}{\nabla}\xi$ de ξ en p en la dirección $\boldsymbol{w}$.

$$\underset{w+u}{\nabla} = \underset{w}{\nabla} + \underset{u}{\nabla},$$

y para un multiplicador escalar λ

$$\underset{\lambda w}{\nabla} = \lambda \underset{w}{\nabla}.$$

Puede parecer que colocar el símbolo del vector por *debajo* de la ∇ parece una notación complicada —¡y, de hecho, lo es!—. Sin embargo, existe una genuina confusión entre la «notación del matemático» y la «notación del físico» en el uso de una expresión como «∇_w». Para nuestro «matemático», esto denotaría probablemente la operación que estoy utilizando aquí «$\underset{w}{\nabla}$», mientras que nuestro «físico» interpretaría probablemente la «$\boldsymbol{w}$» como un *índice* y no como un campo vectorial. En la notación del físico expresaríamos el operador $\underset{w}{\nabla}$ como

$$\underset{w}{\nabla} = w^a \nabla_a,$$

y la linealidad anterior simplemente refleja una consistencia en la notación

$$(w^a + u^a)\nabla_a = w^a\nabla_a + u^a\nabla_a \text{ y } (\lambda w^a)\nabla_a = \lambda(w^a\nabla_a).$$

La colocación de un subíndice en ∇ es consistente con que sea una entidad dual de un campo vectorial (como se refleja en la linealidad anterior; véase §12.3), i.e., ∇ es un operador *covectorial* (lo que significa un operador de valencia $\left[{}^{0}_{1}\right]$). Así, cuando ∇ actúa sobre un campo vectorial $\boldsymbol{\xi}$ (valencia $\left[{}^{1}_{0}\right]$), la cantidad resultante $\nabla\boldsymbol{\xi}$ es un tensor $\left[{}^{1}_{1}\right]$-valente. Esto se hace manifiesto en la notación de índices por el uso de la notación $\nabla_a\xi^b$ para la expresión en componentes (o en índices abstractos) del tensor $\nabla\boldsymbol{\xi}$. De hecho, hay una forma natural de extender el alcance del operador ∇ de vectores a tensores de valencia general, y la actuación de ∇ sobre un tensor $\boldsymbol{T}$ $\left[{}^{p}_{q}\right]$-valente da un tensor $\nabla\boldsymbol{T}$ $\left[{}^{p}_{q+1}\right]$-valente. Las reglas para conseguirlo pueden expresarse convenientemente en la notación de índices, pero hay una complicación en «la notación del matemático» que comentaremos enseguida.

En su actuación sobre campos vectoriales, ∇ satisface el tipo de reglas que satisface el operador diferencial «d» de §12.6:

$$\nabla(\boldsymbol{\xi} + \boldsymbol{\eta}) = \nabla\boldsymbol{\xi} + \nabla\boldsymbol{\eta}$$

y la «ley de Leibniz»

$$\nabla(\lambda\boldsymbol{\xi}) = \lambda\nabla\boldsymbol{\xi} + \boldsymbol{\xi}\nabla\lambda,$$

donde $\boldsymbol{\xi}$ y $\boldsymbol{\eta}$ son campos vectoriales y λ es un campo escalar. Como parte de los requisitos normales de una conexión, la actuación de ∇ sobre un escalar tiene que ser idéntica a la actuación del gradiente (derivada exterior) d sobre dicho escalar:

$$\nabla\Phi = \mathrm{d}\Phi.$$

La extensión de ∇ a un campo tensorial general está unívocamente determinada[14.4] por los dos requisitos naturales siguientes. El primero es la *aditividad* (para tensores $\boldsymbol{T}$, $\boldsymbol{U}$ de la misma valencia)

$$\nabla(\boldsymbol{T} + \boldsymbol{U}) = \nabla\boldsymbol{T} + \nabla\boldsymbol{U},$$

y el segundo es que sea válida la forma apropiada de la «ley de Leibniz». Esta ley es un poco difícil de enunciar, especialmente en «la notación del matemático», que evita los índices. La forma aproximada de esta ley (para tensores $\boldsymbol{T}$, $\boldsymbol{U}$ de valencia arbitraria) es

$$\nabla(\boldsymbol{T}\bullet\boldsymbol{U}) = (\nabla\boldsymbol{T})\bullet\boldsymbol{U} + \boldsymbol{T}\bullet\nabla\boldsymbol{U},$$

pero esto necesita alguna explicación. El punto «•» es para indicar alguna forma de producto contraído, en el que un conjunto de subíndices y superíndices de $\boldsymbol{T}$ se contrae con un conjunto de subíndices y superíndices de $\boldsymbol{U}$ (permitiendo que el conjunto pueda ser vacío, de modo que el producto se convierte en un producto *exterior*, sin ninguna contracción). En la fórmula anterior, la contracción en ambos términos del segundo miembro debe ser un reflejo exacto de los del primer miembro, y la letra índice en la ∇ debe ser la misma en toda la expresión.

Hay una dificultad especial con «la notación del matemático» —donde no se hace referencia a los «índices»— al escribir la fórmula que expresa precisamente lo que entendemos por la ley de Leibniz tensorial. Esta se mitiga ligeramente si utilizamos «$\overset{\boldsymbol{w}}{\nabla}$» en lugar de «$\nabla$», puesto que la «$\boldsymbol{w}$» sigue la pista del índice en la ∇, y podemos hacer algo similar con los otros índices si lo queremos, contrayendo cada uno de ellos con un campo de vectores o covectores (sobre el que no actúa ∇). En mi opinión, las cosas son más claras con índices, pero lo son mucho más en la notación diagramática en la que la diferenciación se denota dibujando un anillo alrededor de la cantidad que está siendo diferenciada. En la Fig. 14.6 se ilustra esto con un ejemplo representativo de la ley de Leibniz tensorial.

Todas estas propiedades también serían ciertas del operador «derivada en coordenadas» $\partial/\partial x^a$ en lugar de ∇_a. De hecho, en cualquier

[14.4] Explique por qué es única. *Sugerencia*: Considere la acción de ∇ sobre $\boldsymbol{\alpha}\bullet\boldsymbol{\xi}$, etc.

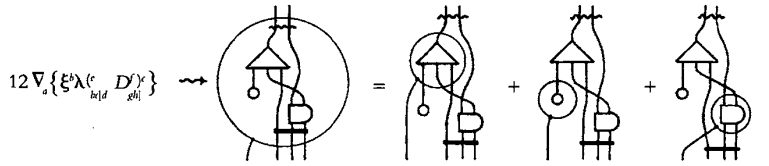


Fig. 14.6. En la notación diagramática, la derivada covariante se denota de forma conveniente dibujando un anillo alrededor de la cantidad a derivar. Esto se ilustra aquí con el ejemplo de la ley de Leibniz tensorial aplicada a $\nabla_a\{\xi^b\lambda^{(c}_{\ \ b]d}D^{)c}_{gh]}\}$ (véase la Fig. 12.17). (Los factores de antisimetría dan el «12».)

carta podemos utilizar $\partial/\partial x^a$ para definir una conexión particular en dicha región, que llamaré la *conexión en coordenadas*. No es una conexión muy interesante, puesto que las coordenadas son arbitrarias. (Proporciona una noción de «paralelismo» en la que todas las líneas coordenadas cuentan como «paralelas».) En el solapamiento entre las dos cartas, la conexión definida por las coordenadas en una carta no suele coincidir con la definida en la otra (véase la Fig. 14.1). Aunque la conexión en coordenadas no sea «interesante» (desde luego, no es físicamente interesante), suele ser bastante útil en expresiones explícitas. La razón tiene que ver con el hecho de que si tomamos la diferencia entre dos conexiones, la actuación de dicha diferencia sobre alguna cantidad tensorial $\boldsymbol{T}$ puede expresarse siempre de forma completamente algebraica (i.e., sin ninguna diferenciación) en términos de $\boldsymbol{T}$ y cierta cantidad $\boldsymbol{\Gamma}$ de valencia $\left[{}^{1}_{2}\right]$.[14.5] Esto nos permite expresar la actuación de $\boldsymbol{\nabla}$ sobre cualquier tensor $\boldsymbol{T}$ explícitamente en términos de las derivadas en coordenadas[2] de las componentes $T^{a\ldots c}_{d\ldots f}$ junto con algunos términos adicionales que incluyen las componentes Γ^{a}_{bc}.[14.6]

14.4. Curvatura y torsión

Una conexión en coordenadas es un tipo bastante especial de conexión en la medida que, a diferencia del caso general, define un paralelismo que es independiente del camino. Esto tiene que ver con el hecho (ya apuntado §10.2, en la forma $\partial^2 f/\partial x\partial y = \partial^2 f/\partial y\partial x$) de que los operadores de derivadas en coordenadas conmutan:

$$\frac{\partial^2}{\partial x^a \partial x^b} = \frac{\partial^2}{\partial x^b \partial x^a}.$$

Otra forma de decirlo es que la cantidad $\partial^2/\partial x^a \partial x^b$ es simétrica (en sus índices *ab*). Enseguida veremos que esto tiene que ver con la independencia de camino del paralelismo. Para una conexión general $\boldsymbol{\nabla}$, esta propiedad de simetría no se satisface para $\nabla_a \nabla_b$, y su parte antisimétrica $\nabla_{[a}\nabla_{b]}$ da lugar a dos tensores especiales, uno de valencia $\left[{}^{1}_{2}\right]$ llamado tensor de *torsión* $\boldsymbol{\tau}$ y otro de valencia $\left[{}^{1}_{3}\right]$ llamado tensor de *curvatura* $\boldsymbol{R}$. La torsión está presente cuando la acción de $\nabla_{[a}\nabla_{b]}$ sobre una cantidad escalar no se anula. En la mayoría de las teorías físicas, $\boldsymbol{\nabla}$ se toma *libre de torsión*, i.e., $\boldsymbol{\tau} = 0$, y esto ciertamente facilita las cosas. Pero existen algunas teorías, tales como la supergravedad y las teorías de espín/torsión de Einstein-Cartan-Sciama-Kibble que emplean una torsión no nula que desempeña un papel físico importante; véanse la nota 19.10 y §31.3. Cuando la torsión *está* presente, su expresión con índices $\tau_{ab}{}^{c}$, antisimétrica en *ab*, está definida por[14.7]

$$(\nabla_a \nabla_b - \nabla_b \nabla_a)\boldsymbol{\Phi} = \tau_{ab}{}^{c}\nabla_c \boldsymbol{\Phi}.$$

El tensor de *curvatura* $\boldsymbol{R}$, en el caso libre de torsión,[14.8] puede definirse[3] por[14.9]

$$(\nabla_a \nabla_b - \nabla_b \nabla_a)\xi^d = R_{abc}{}^{d}\,\xi^c.$$

Como suele suceder en este tema, tropezamos con expresiones amedrentadoras con muchos índices pequeños, de modo que recomiendo la versión diagramática de estas expresiones clave, por ejemplo, la Fig. 14.7a,b. En cualquier caso, también recomiendo que se lean las cantidades con índices, donde sea conveniente, como tensores con índices abstractos, como en §12.8 (Existen numerosos convenios dife-

[14.5] Intente demostrarlo encontrando la expresión explícitamente. *Sugerencia*: Considere, en primer lugar, la actuación de la diferencia entre dos conexiones sobre un campo vectorial $\boldsymbol{\xi}$, dando la respuesta en la forma de índices $\xi^c \Gamma^{a}_{bc}$; en segundo lugar, demuestre que esta diferencia de conexiones actuando sobre un covector $\boldsymbol{\alpha}$ tiene la forma de índices $-\alpha_c \Gamma^{c}_{ba}$; tercero, utilizando la definición de un tensor $\left[{}^{p}_{q}\right]$-valente $\boldsymbol{T}$ como una función multilineal de q vectores en p covectores (cf. §12.8), encuentre la expresión de índices general para la diferencia entre las conexiones que actúan sobre $\boldsymbol{T}$.

[14.6] Como aplicación de esto, considere que las dos conexiones son $\boldsymbol{\nabla}$ y la conexión coordenada. Encuentre una expresión explícita para la actuación de $\boldsymbol{\nabla}$ sobre cualquier tensor, demostrando cómo obtener las componentes Γ^{a}_{bc} explícitamente a partir de $\Gamma^{a}_{b1} = \nabla_b \delta^{a}_{1}, \ldots, \Gamma^{a}_{bn} = \nabla_b \delta^{a}_{n}$, i.e., en términos de la acción de $\boldsymbol{\nabla}$ sobre cada uno de los vectores coordenados $\delta^{a}_{1}, \ldots, \delta^{a}_{n}$. (Aquí a es un índice vectorial, que puede considerarse como un «índice abstracto» de acuerdo con §12.8, de modo que «δ^{a}_{1}», etc., denota en realidad vectores y no simplemente conjuntos de componentes, pero n denota solamente la dimensión del espacio. Nótese que la conexión coordenada aniquila cada uno de estos vectores coordenados.)

[14.7] Explique por qué el segundo miembro debe tener esta forma general; encuentre las componentes τ^{a}_{bc} en términos de Γ^{a}_{bc}. Véase el ejercicio [14.6].

[14.8] Muestre qué término extra es necesario para hacer consistente esta expresión, en presencia de torsión.

[14.9] ¿Cuál es la expresión correspondiente para $\nabla_a \nabla_b - \nabla_b \nabla_a$ actuando sobre un covector? Obtenga la expresión para un tensor general de valencia $\left[{}^{p}_{q}\right]$.

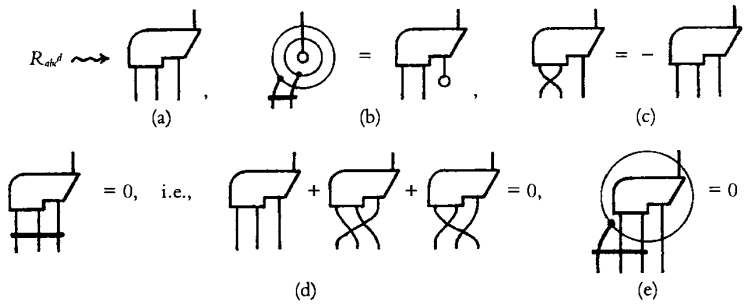


Fig. 14.7. (a) Una notación diagramática conveniente para el tensor de curvatura $R_{abc}{}^{d}$. (b) La identidad de Ricci $(\nabla_a\nabla_b - \nabla_b\nabla_a)\xi^d = R_{abc}{}^{d}\xi^c$. (c) La antisimetría $R_{bac}{}^{d} = -R_{abc}{}^{d}$. (d) La simetría de Bianchi $R_{[abc]}{}^{d} = 0$, que se reduce a $R_{abc}{}^{d} + R_{bca}{}^{d} + R_{cab}{}^{d} = 0$. (e) La identidad de Bianchi $\nabla_{[a}R_{bc]d}{}^{e} = 0$.

rentes en la literatura acerca de signos, ordenamiento de índices, etc. Estoy imponiendo al lector los que yo mismo tiendo a utilizar —¡al menos en artículos de los que soy el único autor!) El hecho de que $R_{abc}{}^{d}$ sea antisimétrico en su primer par de índices *ab*, a saber

$$R_{bac}{}^{d} = -R_{abc}{}^{d},$$

(véase la Fig. 14.7c) es evidente a partir de la correspondiente antisimetría de $(\nabla_a\nabla_b - \nabla_b\nabla_a) = 2\nabla_{[a}\nabla_{b]}$. Enseguida veremos el significado de esta antisimetría. En el caso libre de torsión tenemos una relación de simetría adicional.[14.10]

$$R_{[abc]}{}^{d} = 0, \quad \text{i.e.,}\ R_{abc}{}^{d} + R_{bca}{}^{d} + R_{cab}{}^{d} = 0.$$

Esta relación se denomina a veces «primera identidad de Bianchi». Yo la llamaré la *simetría de Bianchi*. El término *identidad de Bianchi* (Fig. 14.7e) se reserva normalmente para la «segunda» de tales identidades que, en ausencia de torsión, es[14.11]

$$\nabla_{[a}R_{bc]d}{}^{e} = 0, \quad \text{i.e.,}\ \nabla_a R_{bcd}{}^{e} + \nabla_b R_{cad}{}^{e} + \nabla_c R_{abd}{}^{e} = 0.$$

[14.10] En primer lugar, explique el «i.e.»; obtenga luego esto a partir de la ecuación que define el $R_{abc}{}^{d}$ anterior desarrollando $\nabla_{[a}\nabla_b(\xi^d\nabla_{d]}\Phi)$. (Los diagramas pueden ayudar.)

[14.11] Obtenga esto a partir de la ecuación que define la $R_{abc}{}^{d}$ anterior desarrollando $\nabla_{[a}\nabla_b\nabla_{d]}\xi^e$ de dos formas. (Los diagramas pueden ayudar una vez más.)

La identidad de Bianchi es el alma de las ecuaciones de campo de Einstein, como veremos en §19.6.

La curvatura es la cantidad esencial que expresa la dependencia de camino de la conexión (al menos en la escala local). Si imaginamos que transportamos un vector alrededor de un pequeño lazo en el espacio $\mathcal{M}$, utilizando la noción de transporte paralelo definido por ∇, entonces encontramos que es $\boldsymbol{R}$ lo que mide cuánto ha cambiado el vector cuando volvemos al punto de partida. Es más fácil pensar en el lazo como si fuera un «paralelogramo infinitesimal» dibujado en el espacio $\mathcal{M}$. (Como veremos, tales paralelogramos «existen» apropiadamente cuando ∇ es libre de torsión.) Sin embargo, aquí hay varias nociones que necesitan una clarificación.

14.5. Geodésicas, paralelogramos y curvatura

En primer lugar, para construir un paralelogramo consideremos el concepto de una *geodésica*, definida por la conexión ∇. Las geodésicas son importantes para nosotros por otras razones. Son las análogas a las líneas rectas de la geometría euclídea. En nuestro ejemplo de la esfera S^2 considerado antes (Figs. 14.2-14.4), las geodésicas son los círculos máximos en la esfera. Con más generalidad, para una superficie curva en el espacio euclídeo, las curvas de longitud mínima (como se obtendrían extendiendo una cuerda tirante a lo largo de la superficie) son geodésicas. Más adelante (§17.9) veremos que las geodésicas tienen una importancia fundamental para la relatividad general de Einstein, pues representan las trayectorias que describen en el espaciotiempo los cuerpos en caída libre. ¿Cómo nos proporciona nuestra conexión ∇ una noción de geodésica? Básicamente, una geodésica es una curva γ que continúa «paralela a sí misma», según el paralelismo definido por ∇. ¿Cómo vamos a expresar este requisito de forma precisa? Supongamos que el vector $\boldsymbol{t}$ (i.e., t^a) es tangente a γ, a lo largo de toda la γ. Podemos expresar el requisito de que su dirección permanece paralela a sí mismo a lo largo de γ como[4]

$$\nabla_{\boldsymbol{t}}\boldsymbol{t} \propto t, \quad \text{i.e.,}\ t^a\nabla_a t^b \propto t^b$$

(donde el símbolo «∝» significa «proporcional a»; véase §12.7). Cuando se satisface esta condición, t puede estirarse o contraerse a medida que lo seguimos a lo largo de γ, pero «sigue apuntando la misma dirección», de acuerdo con la noción de paralelismo definida por ∇. Si queremos afirmar que este «estiramiento o contracción» no tiene lugar, de modo que el propio vector t permanece constante a lo largo de γ, entonces exigimos la condición más fuerte de que el vector tangente $\boldsymbol{t}$ sea *transportado paralelamente* a lo largo de γ, i.e.,

$$\nabla_{\boldsymbol{t}} \boldsymbol{t} = 0, \quad \text{i.e.,} \quad t^a \nabla_a t^b = 0,$$

sea válido a lo largo de toda la γ, donde el vector $\boldsymbol{t}$ (con forma de índices t^a) es *tangente* a γ a lo largo de γ.

De acuerdo con esta ecuación más fuerte, no solo la dirección de $\boldsymbol{t}$, sino también la «escala» de $\boldsymbol{t}$ se mantiene constante a lo largo de γ. ¿Qué significa esto? Lo primero que hay que señalar es que cualquier curva (no necesariamente una geodésica), parametrizada por una coordenada u (adecuadamente suave), está asociada con una elección concreta de escalamiento para sus vectores tangentes t a lo largo de la curva. Esta es tal que t representa diferenciación (d/du) con respecto a u a lo largo de la curva. Podemos escribir esta condición de forma alternativa como

$$\boldsymbol{t}(u) = 1$$

o como

$$\nabla_{\boldsymbol{t}} u = 1, \quad \text{i.e.,} \quad t^a \nabla_a u = 1,$$

a lo largo de la curva.[14.12]

En el caso de una geodésica γ, la condición más fuerte de $\boldsymbol{t}$-escalamiento para $\nabla_{\boldsymbol{t}} \boldsymbol{t}$ está asociada con un tipo concreto de parámetro u, conocido como un parámetro *afín*[14.13] a lo largo de γ. Véase la Fig. 14.8. Cuando tenemos una noción apropiada de «distancia» a lo largo de curvas, podemos escoger normalmente nuestro parámetro afín para que

[14.12] Demuestre la equivalencia de todas estas condiciones.

[14.13] Demuestre que si u y v son dos parámetros afines sobre γ, con respecto a dos elecciones diferentes de $\boldsymbol{t}$, entonces $v = Au + B$, donde A y B son constantes a lo largo de γ.

sea esta medida de distancia. Pero los parámetros afines son más generales. Por ejemplo, en la teoría de la relatividad resulta que necesitamos tales parámetros para *rayos de luz*, siendo inútil aquí la «medida de distancia» apropiada, ¡porque es *cero*! (Véanse §17.8 y §18.1.)

Ahora tratemos de construir un paralelogramo a partir de geodésicas. Empecemos en algún punto p en $\mathcal{M}$, y tracemos dos geodésicas λ, μ, en $\mathcal{M}$ a partir de p, con vectores tangentes respectivos $\boldsymbol{L}$ y $\boldsymbol{M}$ en p, y parámetros afines respectivos ℓ y m. Escojamos un número positivo ε y midamos una distancia afín $\ell = \varepsilon$ a lo largo de λ desde p hasta llegar al punto q y también una distancia afín $m = \varepsilon$ a lo largo de μ desde p hasta llegar a r; véase la Fig. 14.9a. (Intuitivamente, podemos pensar que los segmentos de geodésica pq y pr tienen las «longitudes de las flechas» de $\varepsilon\boldsymbol{L}$ y $\varepsilon\boldsymbol{M}$, respectivamente, para un ε pequeño.) Para completar el paralelogramo, necesitamos movernos desde q a lo largo de una nueva geodésica μ' en una dirección que sea «paralela» a $\boldsymbol{M}$. Para conseguir esta condición «paralela», desplazamos $\boldsymbol{M}$ desde p a q a lo largo de λ por transporte paralelo (lo que significa que exigimos que $\boldsymbol{M}$ satisfaga ${}_{\boldsymbol{L}}\nabla \boldsymbol{M} = 0$ a lo largo de λ). Ahora tratemos de localizar el vértice final del paralelogramo en el punto s que es medido a partir de q por una distancia afín $m = \varepsilon$ a lo largo de μ'. No obstante, podríamos tratar alternativamente de situar este vértice final procediendo en sentido contrario: moviéndonos desde r una distancia afín $\ell = \varepsilon$ a lo largo de λ'

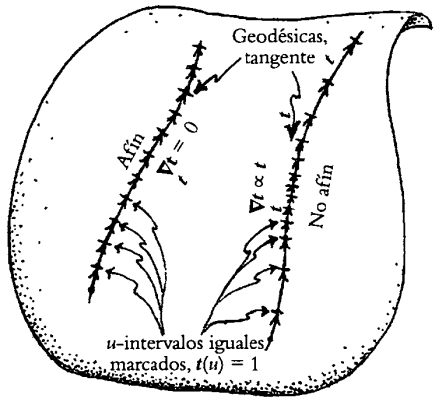


Fig. 14.8. Para cualquier parámetro u (adecuadamente suave) definido a lo largo de una curva γ, un campo de vectores tangentes $\boldsymbol{t}$ a γ está asociado naturalmente a u, de modo que, a lo largo de γ, t representa d/du (equivalentemente $\boldsymbol{t}(u) = 1$ o $t^a \nabla_a u = 1$). Si γ es una geodésica, u se denomina un parámetro afín si $\boldsymbol{t}$ es transportado paralelamente a lo largo de γ, de modo que $\nabla_{\boldsymbol{t}} \boldsymbol{t} = 0$ en lugar de simplemente $\nabla_{\boldsymbol{t}} \boldsymbol{t} \propto \boldsymbol{t}$. Un parámetro afín está «uniformemente espaciado» a lo largo de γ, según ∇.

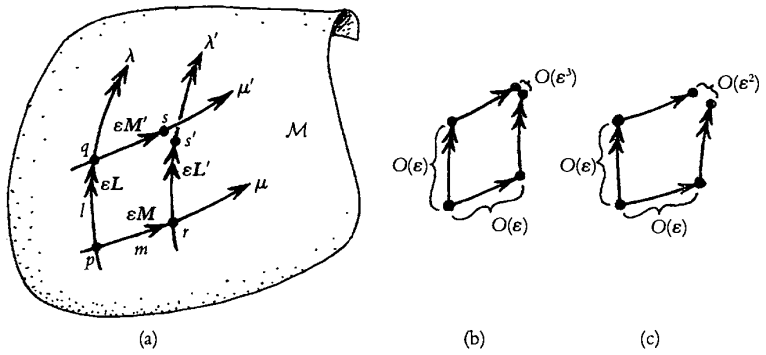


Fig. 14.9. (a) Tratemos de formar un paralelogramo a partir de geodésicas. Tómense dos geodésicas λ, μ, que pasan por p, en $\mathcal{M}$, con vectores tangentes respectivos $\boldsymbol{L}, \boldsymbol{M}$ en p y correspondientes parámetros afines l, m. Tómese q una distancia afín $l = \varepsilon$ a lo largo de λ a partir de p, y r una distancia afín $m = \varepsilon$ a lo largo de μ a partir de p (con $\varepsilon > 0$ un número pequeño dado). Los segmentos geodésicos pq y pr tienen «longitudes de flecha» respectivas $\varepsilon\boldsymbol{L}$, $\varepsilon\boldsymbol{M}$. Para formar el paralelogramo, movemos $\boldsymbol{M}$ de p a q a lo largo de λ por transporte paralelo ($\nabla_{\boldsymbol{L}}\boldsymbol{M} = 0$ a lo largo de λ), lo que nos da una geodésica μ' vecina a μ, que se extiende de q a s a lo largo de μ' por una distancia afín ε a lo largo de la nueva flecha «paralela» $\varepsilon\boldsymbol{M}'$. Del mismo modo, movemos $\boldsymbol{L}$ de p a r por transporte paralelo a lo largo de μ, y lo extendemos de r a s' por una flecha paralela $\varepsilon\,\boldsymbol{L}'$ medida a partir de q una distancia afín $m = \varepsilon$ a lo largo de λ'. (b) En general $s \neq s'$ y el paralelogramo no se cierran exactamente, pero su hueco es solo $O(\varepsilon^3)$ si la torsión τ se anula. (c) Si hay una torsión τ no nula, esto se manifestará como un término $O(\varepsilon^2)$.

hasta un punto final s', siendo λ' una geodésica que parte de r en la dirección de $\boldsymbol{M}$ que ha sido llevada desde p a r a lo largo de μ por transporte paralelo. ¡Para un paralelogramo completamente convincente exigiríamos que estos vértices finales alternativos s y s' sean el mismo punto ($s = s'$)!

Sin embargo, excepto en casos muy especiales (tales como la geometría euclídea), estos dos puntos serán diferentes. (¡Recordemos nuestros intentos de construir un cuadrado en §2.1!) Estos puntos no serán «muy» diferentes, en cierto sentido, si los vectores $\varepsilon\boldsymbol{L}$ y $\varepsilon\boldsymbol{M}$ se toman apropiadamente «pequeños». Pero *cuán* diferentes son exactamente, es algo que tiene que ver con la *torsión* τ. Para entender esto del modo adecuado necesitamos bastantes más nociones de cálculo que las que he dado hasta ahora. El punto esencial está en que podemos considerar

que las desviaciones relevantes respecto de la geometría euclídea se manifiestan en una escala que depende de la elección de nuestra pequeña cantidad ε. No estamos interesados en el tamaño real de estas medidas de desviaciones de la planitud, sino en el ritmo al que tienden a cero a medida que ε se hace cada vez más pequeño. Así, no estamos especialmente interesados en los valores precisos de estas cantidades, sino que queremos saber si una cantidad semejante $\mathbf{Q}$ se aproxima quizá a cero *tan rápidamente* como ε, o ε^2, o ε^3, o quizá alguna otra función concreta de ε. (Ya hemos visto algo parecido en §13.6.) Aquí «tan rápidamente como» significa que cuando se expresan en cierto sistema de coordenadas los valores absolutos de las componentes de $\mathbf{Q}$ son menores que una constante positiva multiplicada por ε, o por ε^2, o por ε^3, o por alguna otra función concreta de ε, según sea el caso. (Por lo tanto, «tan rápidamente como» ¡incluye «más rápidamente que»!). En estos casos, diremos, respectivamente, que $\mathbf{Q}$ es *del orden de* ε, o ε^2, o ε^3, etc., y lo escribiremos como $O(\varepsilon)$, u $O(\varepsilon^2)$, u $O(\varepsilon^3)$, etc. Esto es independiente de la elección de coordenadas concreta, razón por la que esta noción de «orden de pequeñez» es una noción práctica y potente. Mis descripciones han sido muy breves, y remito al lector interesado y no iniciado a la literatura relativa a este tema notable y ubicuo.[5] Intuitivamente, solo necesitamos tener en cuenta que «$O(\varepsilon^3)$» significa mucho más pequeño que «$O(\varepsilon^2)$», que a su vez es mucho más pequeño que «$O(\varepsilon)$», etc.

Volvamos a nuestro intento de paralelogramo. Supongamos que los vectores originales $\varepsilon\boldsymbol{L}$ y $\varepsilon\boldsymbol{M}$, en p, son ambos $O(\varepsilon)$, de modo que los lados pq y pr son ambos $O(\varepsilon)$, y también lo serán qs y rs'. ¿Qué tamaño esperamos que tenga el «hueco» ss'? La respuesta es que si la conexión está *libre de torsión*, entonces ss' es siempre $O(\varepsilon^3)$. Véase la Fig. 14.9b. De hecho, esta propiedad caracteriza por completo la condición libre de torsión. Si está presente una torsión τ no nula, entonces esto se manifestará en (algunos) paralelogramos, como un término $O(\varepsilon^2)$. Véase la Fig. 14.9c.[14.14] A veces decimos (de forma bastante vaga) que la anulación de la torsión es la condición de que los paralelogramos se cierren (por lo que entendemos «se cierran hasta orden ε^2»).

[14.14] Encuentre este término.

Supongamos ahora que la torsión se anula. ¿Podemos utilizar nuestro paralelogramo para interpretar la curvatura? De hecho, sí podemos. Supongamos que tenemos un tercer vector $\mathbf{N}$ en p, y lo llevamos mediante transporte paralelo alrededor de nuestro paralelogramo desde p a s, vía q, y lo comparamos con su transporte desde p a s', vía r. (Esta comparación tiene sentido a orden ε^2, cuando la torsión se anula, porque entonces el hueco entre s y s' es $O(\varepsilon^3)$ y puede ser ignorado. Cuando la torsión no se anula, tenemos que preocuparnos por el término de torsión adicional; véase el ejercicio [14.7].) Encontramos que la respuesta para la diferencia entre el resultado del transporte pqs y el transporte prs' es

$$\varepsilon^2 L^a M^b N^c R_{abc}{}^d.$$

Esto nos proporciona una interpretación geométrica muy directa del tensor de curvatura $\boldsymbol{R}$; véase la Fig. 14.10. (Se obtiene una versión equivalente de esta interpretación si pensamos en transportar N por todo el camino alrededor del paralelogramo, empezando y terminando en el mismo punto p, donde ignoramos discrepancias $O(\varepsilon^3)$ en los vértices del paralelogramo. La diferencia entre los valores de partida y de llegada de $\boldsymbol{N}$ es una vez más la cantidad anterior $\varepsilon^2 L^a M^b N^c R_{abc}{}^d$.)

Recordemos la antisimetría de $R_{abc}{}^d$ en ab. Esto significa que la expresión anterior es sensible solo a la parte antisimétrica, $L^{[a}M^{b]}$, de L^aM^b, i.e., del producto cuña $\boldsymbol{L} \wedge \boldsymbol{M}$; véase §11.6. Así pues, es el elemento 2-plano generado por $\boldsymbol{L}$ y $\boldsymbol{M}$ en p el que tiene relevancia. En el

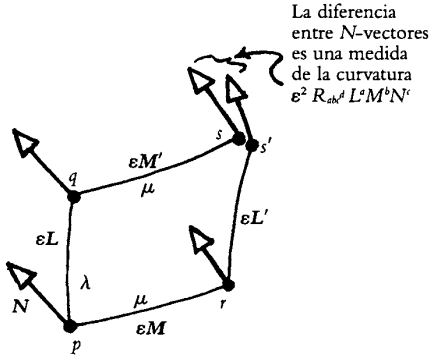


Fig. 14.10. Utilizamos el paralelogramo para interpretar la curvatura cuando $\tau = 0$. Llevamos un tercer vector $\boldsymbol{N}$, por transporte paralelo de p a s vía q, comparando esto con transportarlo de p a s' vía r. El término $O(\varepsilon^2)$ que mide la diferencia es $\varepsilon^2 L^a M^b N^c R_{abc}{}^d$, i.e., $\varepsilon^2 \boldsymbol{R}(\boldsymbol{L}, \boldsymbol{M}, \boldsymbol{N})$, que proporciona una interpretación geométrica directa del tensor de curvatura $\boldsymbol{R}$.

caso en que el propio $\mathcal{M}$ es una 2-superficie, hay solo una componente de la curvatura independiente (puesto que el elemento 2-plano tiene que ser tangente a $\mathcal{M}$ en p). Esta componente nos proporciona la *curvatura gaussiana* de una 2-superficie a la que he aludido en §2.6, y que sirve para distinguir las geometrías locales de la esfera, el plano euclídeo y el espacio hiperbólico. En dimensiones más altas, el asunto se complica, y hay más componentes de la curvatura que surgen de las diferentes elecciones posibles del elemento 2-plano $\boldsymbol{L} \wedge \boldsymbol{M}$.

Existe una versión particular de esta interpretación geométrica de la curvatura que tiene un significado especial. Esto ocurre si se escoge que el vector $\boldsymbol{N}$ sea el mismo que el $\boldsymbol{L}$. Entonces podemos considerar que los lados pq y rs' de nuestro paralelogramo son segmentos de dos geodésicas próximas γ y γ', respectivamente, y el vector $\boldsymbol{L}$ es tangente a dichas geodésicas. El vector $\varepsilon\boldsymbol{M}$ en p mide el desplazamiento de γ respecto a γ' en el punto p. A veces se llama a $\boldsymbol{M}$ un *vector conectante*. Las geodésicas γ y γ' parten paralelas una a otra (si se comparan en los dos «extremos» de este vector conectante, i.e., a lo largo de pr). Llevar el vector $\boldsymbol{L}$ (= $\boldsymbol{N}$) a s' por transporte paralelo a lo largo de la segunda ruta prs' lo deja tangente a la geodésica γ' en el punto s'. Pero si llevamos $\boldsymbol{L}$ a s por transporte paralelo a lo largo de la primera ruta pqs, entonces llegamos al vector de partida para otra geodésica γ'' próxima a γ, donde γ'' empieza paralela a γ en el punto q ligeramente «posterior». La diferencia $O(\varepsilon^2)$ entre estas dos versiones de $\boldsymbol{L}$ (una en s' y la

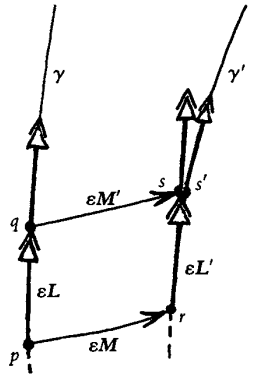


Fig. 14.11. Desviación geodésica: escojamos $\boldsymbol{N} = \boldsymbol{L}$ en el paralelogramo de la Fig. 14.10 Los lados pq y rs' son segmentos de dos geodésicas vecinas γ y γ' (γ siendo λ y g' siendo λ') que parten de p y r, respectivamente, con vectores tangentes $\boldsymbol{L}$ y $\boldsymbol{L}'$ propagados paralelamente, siendo $\boldsymbol{M}$ el vector que conecta en p. La desviación geodésica entre γ y γ' se mide por la diferencia entre los resultados del desplazamiento paralelo de $\boldsymbol{L}$ a lo largo de las rutas prs' y pqs, que es básicamente $\varepsilon^2 L^a M^b L^c R_{abc}{}^d$.

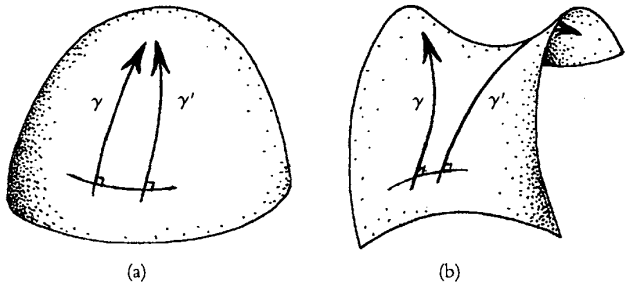


Fig. 14.12. Desviación geodésica cuando $\mathcal{M}$ es una 2-superficie (a) de curvatura (gaussiana) positiva, cuando las geodésicas γ, γ' se juntan, y (b) de curvatura negativa, cuando se separan.

otra en s), a saber $\varepsilon^2 L^a M^b L^c R_{abc}{}^d$, mide la «aceleración relativa» o «desviación geodésica» de γ' respecto de γ. Véase la Fig. 14.11. (Esta desviación geodésica viene descrita matemáticamente por lo que se conoce como *ecuación de Jacobi*.) En la Fig. 14.12 he ilustrado esta desviación geodésica cuando $\mathcal{M}$ es una 2-superficie de curvatura (gaussiana) positiva y negativa, respectivamente. Cuando la curvatura es positiva, las geodésicas vecinas, que empiezan paralelas, se curvan unas hacia las otras; cuando es negativa, se separan. Veremos la gran importancia que tiene esto para la relatividad general de Einstein en §17.5 y §19.6.

14.6. Derivada de Lie

En la discusión anterior sobre la dependencia del camino del paralelismo para una conexión ∇ he utilizado la notación de índices del físico. En la notación del matemático, los análogos directos de estas expresiones particulares no son tan fáciles de escribir. En su lugar, lo normal es seguir una ruta algo diferente. (Resulta curioso cómo las diferencias en notación pueden dirigir a veces un tema en direcciones conceptualmente diferentes.) Esta ruta implica otra operación de diferenciación, conocida como un *paréntesis de Lie*, que es una forma más general de la operación del mismo nombre que se ha introducido en §13.6. Este, a

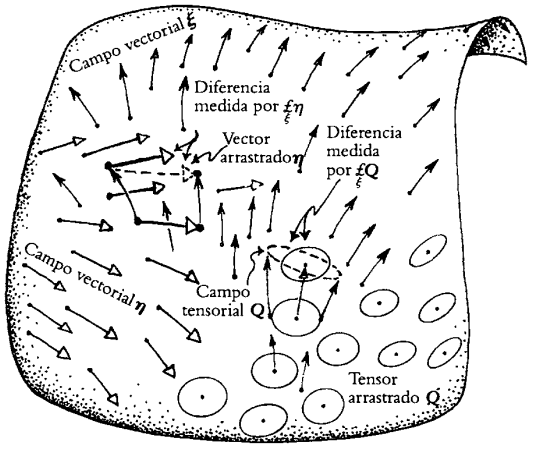


Fig. 14.13. Derivada de Lie definida en una variedad general $\mathcal{M}$ tomada con respecto a un campo vectorial suave $\boldsymbol{\xi}$ en $\mathcal{M}$ dado. Entonces, $\underset{\xi}{\pounds}\mathbf{Q}$ mide cómo una cantidad $\mathbf{Q}$ (por ejemplo, un campo vectorial $\boldsymbol{\eta}$ o un campo tensorial $\mathbf{Q}$) cambia realmente, comparado con la cantidad «arrastrada» por $\boldsymbol{\xi}$.

su vez, es un caso particular de un concepto importante conocido como *derivada de Lie*. Estas nociones son realmente independientes de cualquier elección particular de conexión (y por lo tanto se aplican en una variedad general, suave y sin estructura), y será pertinente discutir dichas operaciones en general antes de volver a su relevancia para la curvatura y la torsión al final de esta sección.

No obstante, para que una derivada de Lie esté definida en una variedad $\mathcal{M}$, exigimos que sea preasignado un campo vectorial $\boldsymbol{\xi}$ sobre $\mathcal{M}$. La derivada de Lie, escrita $\underset{\xi}{\pounds}$, es entonces una operación que se toma con respecto al campo vectorial ξ. La derivada $\underset{\xi}{\pounds}\mathbf{Q}$ mide cómo cambia cierta cantidad $\mathbf{Q}$ comparado con lo que sucedería si fuera simplemente «arrastrada» por el campo vectorial ξ. Véase la Fig. 14.13. Se aplica a tensores en general (e incluso a algunas entidades diferentes de los tensores, como las conexiones). Para empezar, consideremos simplemente la derivada de Lie de un campo *vectorial* $\boldsymbol{\eta}$ (= $\mathbf{Q}$) con respecto a otro campo vectorial ξ. Nos damos cuenta de que esta es la misma operación que la que he llamado «paréntesis de Lie» en §13.6, pero en un contexto más general. Veremos cómo generalizar esto a un campo tensorial $\mathbf{Q}$.

Recordemos de §12.3 que un campo vectorial puede interpretarse como un operador diferencial que actuando sobre campos *escalares* Φ, Ψ,... satisface las tres leyes (i) $\xi(\Phi + \Psi) = \xi(\Phi) + \xi(\Psi)$,

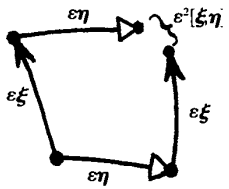


Fig. 14.14. El paréntesis de Lie $[\boldsymbol{\xi}, \boldsymbol{\eta}]$ $(= \pounds_{\xi}\boldsymbol{\eta})$ entre dos campos vectoriales $\boldsymbol{\xi}, \boldsymbol{\eta}$ mide el hueco $O(\varepsilon^2)$ en un cuadrilátero incompleto de «flechas» $O(\varepsilon)$ hecho alternativamente a partir de $\varepsilon\boldsymbol{\xi}$ y $\varepsilon\boldsymbol{\eta}$.

(ii) $\boldsymbol{\xi}(\Phi\Psi) = \Psi\boldsymbol{\xi}(\Phi) + \Phi\boldsymbol{\xi}(\Psi)$, y (iii) $\boldsymbol{\xi}(k) = 0$ si k = constante. Se puede demostrar directamente[14.15] que el operador $\boldsymbol{\omega}$, definido mediante

$$\boldsymbol{\omega}(\Phi) = \boldsymbol{\xi}(\boldsymbol{\eta}(\Phi)) - \boldsymbol{\eta}(\boldsymbol{\xi}(\Phi))$$

satisface estas mismas tres leyes, con tal de que lo hagan $\boldsymbol{\xi}$ y $\boldsymbol{\eta}$, de modo que $\boldsymbol{\omega}$ debe ser también un campo vectorial. El anterior *conmutador* de las dos operaciones $\boldsymbol{\xi}$ y $\boldsymbol{\eta}$ suele escribirse (como en §13.6) en la notación de *paréntesis de Lie*

$$\boldsymbol{\omega} = \boldsymbol{\xi}\boldsymbol{\eta} - \boldsymbol{\eta}\boldsymbol{\zeta} = [\boldsymbol{\xi}, \boldsymbol{\eta}].$$

El significado geométrico del conmutador entre dos campos vectoriales $\boldsymbol{\xi}$ y $\boldsymbol{\eta}$ se ilustra en la Fig. 14.14. Tratemos de formar un cuadrilátero de «flechas» hecho alternadamente de $\boldsymbol{\xi}$ y $\boldsymbol{\eta}$ (tomado cada uno $O(\varepsilon)$) y encontraremos que $\boldsymbol{\omega}$ mide el «hueco» (a orden $O(\varepsilon^2)$). Podemos verificar[14.16] que la conmutación satisface las relaciones siguientes

$$[\boldsymbol{\xi}, \boldsymbol{\eta}] = -[\boldsymbol{\eta}, \boldsymbol{\xi}], \quad [\boldsymbol{\xi} + \boldsymbol{\eta}, \boldsymbol{\zeta}] = [\boldsymbol{\xi}, \boldsymbol{\zeta}] + [\boldsymbol{\eta}, \boldsymbol{\zeta}],$$
$$[\boldsymbol{\xi}, [\boldsymbol{\eta}, \boldsymbol{\zeta}]] + [\boldsymbol{\eta}, [\boldsymbol{\zeta}, \boldsymbol{\xi}]] + [\boldsymbol{\zeta}, [\boldsymbol{\xi}, \boldsymbol{\eta}]] = 0,$$

como lo hacía el conmutador de dos elementos infinitesimales de un grupo de Lie, como hemos visto en §13.6.

¿Cómo se relaciona la operación de conmutación que acabo de definir con el álgebra (§13.6) de elementos infinitesimales de un grupo de Lie? Permítaseme una breve digresión para explicarlo. Consideremos el grupo como una variedad $\mathcal{G}$ (llamada una variedad de *grupo*), cuyos puntos son los elementos de nuestro grupo de Lie. Con más generalidad podemos considerar cualquier variedad $\mathcal{H}$ sobre la que ac-

[14.15] Demuéstrelo.
[14.16] Hágalo.

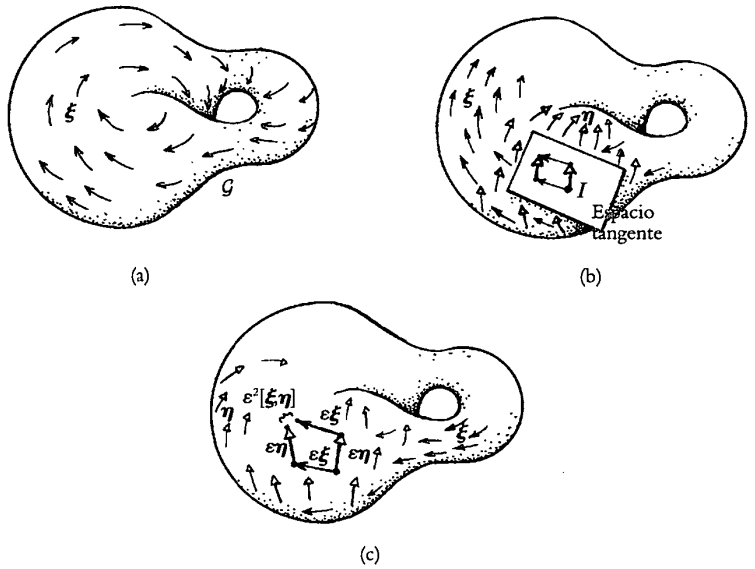


Fig. 14.15. Operaciones del álgebra de Lie, interpretadas geométricamente en la variedad del grupo continuo $\mathcal{G}$. (a) La premultiplicación de cada elemento de $\mathcal{G}$ por un elemento de grupo infinitesimal $\boldsymbol{\xi}$ (elemento del álgebra de Lie) da una desplazamiento infinitesimal de $\mathcal{G}$, i.e., un campo vectorial $\boldsymbol{\xi}$ en $\mathcal{G}$. (b) A primer orden, el producto de dos movimientos infinitesimales semejantes $\boldsymbol{\xi}$ y $\boldsymbol{\eta}$ da solo $\boldsymbol{\xi} + \boldsymbol{\eta}$, lo que refleja meramente la estructura del espacio tangente (en I). (c) La estructura de grupo local aparece a segundo orden, $\varepsilon^2[\boldsymbol{\xi}, \boldsymbol{\eta}]$, lo que proporciona el hueco $O(\varepsilon^2)$ en el «paralelogramo» con lados alternos $\varepsilon\boldsymbol{\xi}$ y $\varepsilon\boldsymbol{\eta}$ en I.

túan los elementos como transformaciones suaves (tales como la esfera S^2. En el caso del grupo de rotaciones $\mathcal{G}$ = SO(3), véase la Fig. 13.2). Pero de momento estamos interesados fundamentalmente en la variedad de grupo $\mathcal{G}$, antes que en la situación más general de $\mathcal{H}$, puesto que nos interesa cómo se relaciona el grupo entero $\mathcal{G}$ con la estructura de su álgebra de Lie. Los elementos de grupo infinitesimales deben imaginarse como campos vectoriales particulares en $\mathcal{G}$ (o, en realidad, $\mathcal{H}$). Es decir, consideremos «mover $\mathcal{G}$» infinitesimalmente a lo largo del campo vectorial relevante $\boldsymbol{\xi}$ en $\mathcal{G}$, para expresar la transformación que corresponde a premultiplicar cada elemento del grupo por el elemento infinitesimal representado por $\boldsymbol{\xi}$. Véase la Fig. 14.15a.

Escogiendo una pequeña cantidad escalar positiva ε podemos considerar que $\varepsilon\boldsymbol{\xi}$ es un movimiento de orden $O(\varepsilon)$ de $\mathcal{G}$ a lo largo del campo vectorial $\boldsymbol{\xi}$, correspondiendo el elemento identidad $\boldsymbol{I}$ del grupo al movimiento cero. El producto de dos de estas pequeñas acciones del grupo $\varepsilon\boldsymbol{\xi}$ y $\varepsilon\boldsymbol{\eta}$ viene dado, hasta $O(\varepsilon)$, por la *suma* $\varepsilon\boldsymbol{\xi} + \varepsilon\boldsymbol{\eta}$ de ambas, de modo que las «flechas» que representan a $\varepsilon\boldsymbol{\xi}$ y $\varepsilon\boldsymbol{\eta}$ simplemente se suman de acuerdo con la ley del paralelogramo (Fig. 14.15b). Pero esto nos da poca información sobre la estructura del grupo $\mathcal{G}$ (solo su *dimensión*, de hecho, puesto que únicamente estamos revelando la estructura aditiva del espacio tangente en el elemento identidad $\boldsymbol{I}$ del grupo). Para obtener la estructura del grupo necesitamos ir a orden $O(\varepsilon^2)$, y esto se hace, como en §13.6, considerando el conmutador $\boldsymbol{\xi\eta} - \boldsymbol{\eta\xi} = [\boldsymbol{\xi}, \boldsymbol{\eta}]$. Ahora $\varepsilon^2[\boldsymbol{\xi}, \boldsymbol{\eta}]$ corresponde a un hueco $O(\varepsilon^2)$ en el «paralelogramo» cuyos lados iniciales son $\varepsilon\boldsymbol{\xi}$ y $\varepsilon\boldsymbol{\eta}$ en el origen $\boldsymbol{I}$. La noción relevante de «paralelismo» procede de la acción de grupo, que suministra la noción necesaria de «transporte paralelo», que realmente da una conexión con torsión pero sin curvatura.[14.17] Véase la Fig. 14.15c.

Como se ha señalado en §13.6, el álgebra de Lie de estos campos vectoriales proporciona la estructura (local) entera del grupo. Aquí puede señalarse el procedimiento mediante el que uno obtiene un elemento de grupo finito (i.e., no infinitesimal) ordinario x a partir de un elemento $\boldsymbol{\xi}$ del álgebra de Lie. Se denomina exponenciación (cf. §5.3 y §13.4):

$$x = e^{\boldsymbol{\xi}} = \boldsymbol{I} + \boldsymbol{\xi} + \frac{1}{2}\boldsymbol{\xi}^2 + \frac{1}{6}\boldsymbol{\xi}^3 + \dots$$

Aquí $\boldsymbol{\xi}^2$ significa «el operador derivada segunda que consiste en aplicar $\boldsymbol{\xi}$ dos veces», etc. (e $\boldsymbol{I}$ es el operador identidad). Esta es básicamente una forma del teorema de Taylor, como se ha descrito en §6.4.[14.18] El producto de dos elementos de grupo finitos x e y se obtiene entonces de la expresión $e^{\boldsymbol{\xi}}e^{\boldsymbol{\eta}}$. Esta difiere de $e^{\boldsymbol{\xi}+\boldsymbol{\eta}}$ (compárese con §5.3) en una expresión que está construida enteramente a partir de la expresión del álgebra de Lie[6] en $\boldsymbol{\xi}$ y $\boldsymbol{\eta}$.

[14.17] Trate de explicar por qué hay torsión pero no curvatura.
[14.18] Explique (a un nivel formal) por qué $e^{a\mathrm{d}/\mathrm{d}y}f(y) = f(y + a)$ cuando a es una constante.

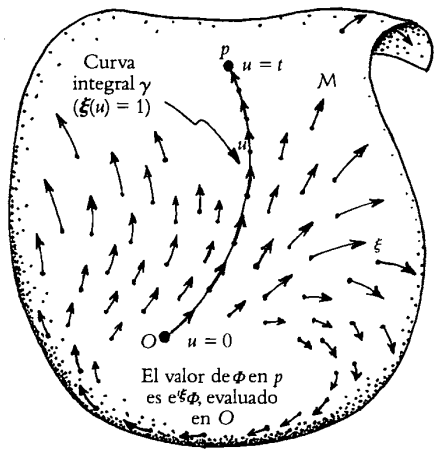
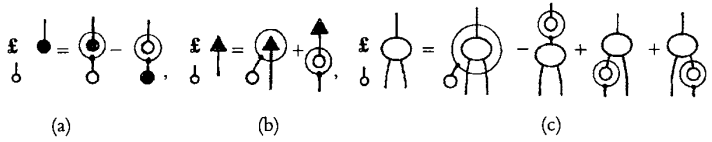


Fig. 14.16. Una curva integral de un campo vectorial $\boldsymbol{\xi}$ en $\mathcal{M}$ es una curva γ que «sigue las $\boldsymbol{\xi}$-flechas»; i.e., cuyos vectores tangentes son $\boldsymbol{\xi}$-vectores, con parámetro asociado u, en el sentido $\boldsymbol{\xi}(u) = 1$ (cf. §14.5 y la Fig. 14.8). Supongamos que $\mathcal{M}$ y $\boldsymbol{\xi}$ son analíticos (i.e., C^ω), como lo es el campo escalar Φ, y que γ se estira desde un punto base $O(u = 0)$ a otro punto $p(u = t)$. Entonces (suponiendo convergencia), el valor de Φ en p está dado por la cantidad $e^{t\boldsymbol{\xi}}(\Phi)$ evaluada en O, donde $e^{t\boldsymbol{\xi}} = 1 + t\boldsymbol{\xi} + \frac{1}{2}t^2\boldsymbol{\xi}^2 + \frac{1}{6}t^3\boldsymbol{\xi}^3 + \dots$ y donde $\boldsymbol{\xi}^r$ representa la derivada r-ésima $\mathrm{d}^r/\mathrm{d}u^r$ en O a lo largo de γ.

Puede señalarse que una versión de esta operación de exponenciación $e^{\boldsymbol{\xi}}$ se aplica también a un campo vectorial $\boldsymbol{\xi}$ en una variedad general $\mathcal{M}$ (donde $\mathcal{M}$ y $\boldsymbol{\xi}$ se suponen analíticas, i.e., C^ω-suaves; véase §6.4). Recordemos de §12.3 (y la Fig. 10.6) que, con ε escogida pequeña, $\varepsilon\boldsymbol{\xi}(\Phi)$ mide el incremento $O(\varepsilon)$ de una función escalar Φ desde la cola a la punta de la «flecha» que representa $\varepsilon\boldsymbol{\xi}$. Más exactamente, la cantidad $e^{t\boldsymbol{\xi}}(\Phi)$ mide el valor *total* Φ que se alcanza cuando seguimos las «$\boldsymbol{\xi}$-flechas» desde un punto de partida O a un punto final dado por el valor del parámetro $u = t$, donde el parámetro u está escalado de modo que $\boldsymbol{\xi}(u) = 1$ (cf. §14.5 y Fig. 14.8). Todas las derivadas (es decir, la r-ésima derivada, en el caso de $\boldsymbol{\xi}^r(\Phi)$) en el desarrollo en serie de potencias para $e^{t\boldsymbol{\xi}}(\Phi)$ tienen que ser evaluadas en O (supuesta la convergencia). «Seguir las flechas» significaría seguir lo que se llama



(a) (b) (c)

Fig. 14.17. Diagrama para la derivada de Lie (a) de un vector $\boldsymbol{\eta}$: $(\underset{\xi}{\pounds}\boldsymbol{\eta})^a = \xi^u\nabla_u\eta^b - \eta^u\nabla_u\xi^b$; (b) de un convector α: $(\underset{\xi}{\pounds}\boldsymbol{\alpha})_a = \xi^b\nabla_b\alpha^a + \alpha_b\nabla_a\xi^b$; y (c) de un tensor $\mathbf{Q}(\left[{}^2_1\right]$-valente): $\underset{\xi}{\pounds}Q^c_{ab} = \xi^u\nabla_u Q^c_{ab} + Q^c_{ub}\nabla_a\xi^u + Q^c_{au}\nabla_b\xi^u - Q^u_{ab}\nabla_u\xi^c$.

una «curva integral» de $\boldsymbol{\xi}$, es decir, una curva cuyos vectores tangentes son $\boldsymbol{\xi}$-vectores. Véase la Fig. 14.16.[7]

¿Cuál es entonces la definición de *derivada* de Lie? En primer lugar, reescribimos el paréntesis de Lie como una operación $\underset{\xi}{\pounds}$ (que depende de $\boldsymbol{\xi}$) que actúa sobre el campo vectorial $\boldsymbol{\eta}$:

$$\underset{\xi}{\pounds}\boldsymbol{\eta} = [\boldsymbol{\xi}, \boldsymbol{\eta}].$$

Esta va a ser la definición de la derivada de Lie $\underset{\xi}{\pounds}$ (con respecto a $\boldsymbol{\xi}$) de un $\left[{}^1_0\right]$-tensor $\boldsymbol{\eta}$ (i.e., un vector). Queremos escribir esto en términos de una conexión ∇ dada libre de torsión. La expresión requerida (véase la Fig. 14.17a, para la forma diagramática)

$$\underset{\xi}{\pounds}\boldsymbol{\eta} = \underset{\xi}{\nabla}\boldsymbol{\eta} - \underset{\eta}{\nabla}\boldsymbol{\xi} \quad \text{i.e., } (\underset{\xi}{\pounds}\boldsymbol{\eta})^a = \xi^u\nabla_u\eta^b - \eta^u\nabla_u\xi^b,$$

puede obtenerse directamente utilizando $\xi(\Phi) = \xi^u\nabla_u\Phi$, etc.[14.19],[14.20] Para obtener la derivada de Lie de un tensor general, empleamos la regla de que (excepto en ausencia de linealidad en ξ) $\underset{\xi}{\pounds}$ satisface reglas similares a las de una conexión $\underset{\xi}{\nabla}$. Estas son: $\underset{\xi}{\pounds}\Phi = \xi(\Phi)$ para un escalar Φ; $\underset{\xi}{\pounds}(\boldsymbol{T} + \boldsymbol{U}) = \underset{\xi}{\pounds}\boldsymbol{T} + \underset{\xi}{\pounds}\boldsymbol{U}$ para tensores $\boldsymbol{T}$ y $\boldsymbol{U}$ de la misma valencia; $\underset{\xi}{\pounds}(\boldsymbol{T}\bullet\boldsymbol{U}) = (\underset{\xi}{\pounds}\boldsymbol{T})\bullet\boldsymbol{U} + \boldsymbol{T}\bullet\underset{\xi}{\pounds}\boldsymbol{U}$ siendo la disposición de las contracciones la misma en cada término. A partir de estas, y $\underset{\xi}{\pounds}\boldsymbol{\eta} = [\boldsymbol{\xi}, \boldsymbol{\eta}]$, se sigue unívocamente la acción de $\underset{\xi}{\pounds}$ sobre cualquier tensor.[8] En particular, para un covector $\boldsymbol{\alpha}$ (valencia $\left[{}^0_1\right]$),

$$\underset{\xi}{\pounds}\boldsymbol{\alpha} = \underset{\xi}{\nabla}\boldsymbol{\alpha} + \boldsymbol{\alpha}\bullet(\nabla\boldsymbol{\xi}), \quad \text{i.e., } (\underset{\xi}{\pounds}\boldsymbol{\alpha})_a = \xi^b\nabla_b\alpha_a + \alpha_b\nabla_a\xi^b$$

[14.19] Deduzca esta fórmula para $\underset{\xi}{\pounds}\boldsymbol{\eta}$.
[14.20] ¿Cómo modifica la torsión $\underset{\xi}{\pounds}$ la fórmula del ejercicio [14.9]?

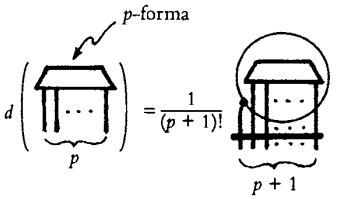


Fig. 14.18. Diagrama para la derivada exterior de una p-forma: $(\mathrm{d}\boldsymbol{\alpha})_{ab\ldots d} = \nabla_{[a}\alpha_{b\ldots d]}$.

(siendo ∇ libre de torsión); véase la Fig. 14.17b. Para un tensor $\boldsymbol{Q}$ de valencia $\left[{}^1_2\right]$, por ejemplo, tenemos entonces (Fig. 14.17c)[14.21]

$$\underset{\xi}{\pounds}Q^c_{ab} = \xi^u Q^c_{ab} + Q^c_{ub}\nabla_a\xi^u + Q^c_{au}\nabla_b\xi^u - Q^u_{ab}\nabla_u\xi^c.$$

Notamos que la derivada de Lie, considerada como una función tanto de ξ como de la cantidad $\mathbf{Q}$ (campo tensorial) sobre la que actúa, es *independiente* de la conexión, i.e., es la misma cualquiera que sea el operador ∇_a libre de torsión que escojamos. (Esto se sigue de que $\underset{\xi}{\pounds}$ está unívocamente definida a partir del operador gradiente «d».) En particular, podríamos utilizar el operador de derivada en coordenadas $\partial/\partial x^a$ (en cualquier sistema de coordenadas que escojamos) en lugar de ∇_a, y la respuesta sería la misma. Incluso si tenemos una conexión con torsión, aún podríamos utilizarla, expresándola en términos de una segunda conexión, unívocamente definida por la dada y que es libre de torsión, obtenida «restando» la torsión de la conexión dada.[14.22]

La derivada de Lie comparte con la derivada exterior (véase §12.6) esta propiedad de independencia de la conexión, por la que para cualquier p-forma $\boldsymbol{\alpha}$, con expresión de índices $\alpha_{b\ldots d}$,

$$(\mathrm{d}\boldsymbol{\alpha})_{ab\ldots d} = \nabla_{[a}\alpha_{b\ldots d]},$$

donde ∇ es cualquier conexión libre de torsión; véase la Fig. 14.18. Esta es la misma que la expresión dada en §12.6, excepto que allí se utilizaba explícitamente la *conexión en coordenadas* $\partial/\partial x^a$. Se ve inme-

[14.21] Establezca la unicidad, verificando la fórmula covectorial anterior, y dé explícitamente la derivada de Lie de un tensor general.
[14.22] Muestre cómo encontrar esta segunda conexión, tomando la «Γ» para que la diferencia entre las dos conexiones sea antisimétrica en sus dos subíndices. (Véase el ejercicio [14.5].)

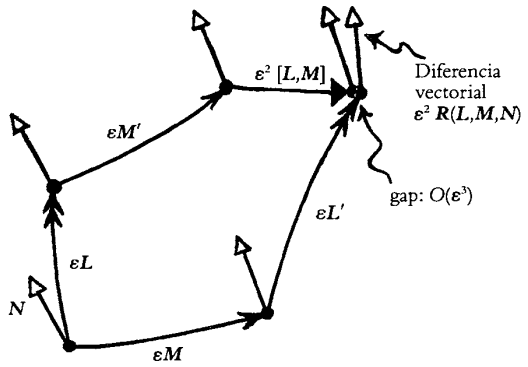


Fig. 14.19. Curvatura en la «notación del matemático» $(\underset{L}{\nabla}\underset{M}{\nabla} - \underset{M}{\nabla}\underset{L}{\nabla} - \underset{[M,L]}{\nabla})N = \boldsymbol{R}(\boldsymbol{L}, \boldsymbol{M}, \boldsymbol{N})$, a partir de la discrepancia $O(\varepsilon^2)$ en el transporte paralelo de un vector $\boldsymbol{N}$ alrededor del «cuadrilátero» (incompleto) con lados $\varepsilon\boldsymbol{L}$, $\varepsilon\boldsymbol{M}$, $\varepsilon\boldsymbol{L}'$, $\varepsilon\boldsymbol{M}'$. La contribución del paréntesis de Lie $\varepsilon^2[\boldsymbol{L}, \boldsymbol{M}]$ llena un hueco $O(\varepsilon^2)$. (La forma con índices del vector $\boldsymbol{R}(\boldsymbol{L}, \boldsymbol{M}, \boldsymbol{N})$ es $L^aM^bN^cR_{abc}{}^d$.)

diatamente que la expresión anterior es realmente independiente de la elección de conexión libre de torsión.[14.23] Además, la propiedad clave $d^2\alpha = 0$ se sigue inmediatamente de esta expresión.[14.24] Hay también algunas otras expresiones que son independientes de la conexión en este sentido.[9]

Por último, volviendo a la cuestión de la curvatura, en nuestra variedad $\mathcal{M}$, con conexión $\boldsymbol{\nabla}$, vemos que necesitamos el paréntesis de Lie para la definición del tensor de curvatura en la notación del matemático:

$$(\underset{L}{\boldsymbol{\nabla}}\underset{M}{\boldsymbol{\nabla}} - \underset{M}{\boldsymbol{\nabla}}\underset{L}{\boldsymbol{\nabla}} - \underset{[L,M]}{\boldsymbol{\nabla}})\, \boldsymbol{N} = \boldsymbol{R}(\boldsymbol{L}, \boldsymbol{M}, \boldsymbol{N})$$

(donde $\boldsymbol{R}(\boldsymbol{L}, \boldsymbol{M}, \boldsymbol{N})$ significa el vector $L^aM^bN^cR_{abc}{}^d$).[14.25] Aunque la inclusión de un término conmutador extra puede considerarse una desventaja de esta notación, tiene la ventaja compensatoria de que la torsión está automáticamente permitida (en contraste con la necesidad de un término adicional de torsión en la notación del físico). Recordemos el significado geométrico del término conmutador (Fig. 14.14). Permite un «hueco» $O(\varepsilon^2)$ en el cuadrilátero $O(\varepsilon)$ construido a partir de los campos vectoriales $\boldsymbol{L}$ y $\boldsymbol{M}$. De hecho, tiene la ventaja adicional de que no es necesario considerar que el lazo alrededor del cual llevamos nuestro vector $\boldsymbol{N}$ es un «paralelogramo» (hasta el orden previamente requerido), sino tan solo un cuadrilátero (curvilíneo). Véase la Fig. 14.19. Si $[\boldsymbol{L}, \boldsymbol{M}] = 0$, entonces este cuadrilátero se cierra (hasta orden $O(\varepsilon^2)$).

[14.23] Establezca esto y muestre cómo modifica la expresión la presencia de un tensor de torsión τ.

[14.24] Demuéstrelo.

[14.25] Demuestre la equivalencia (si se anula la torsión) con la expresión del físico anterior.

14.7. Lo que una métrica puede hacer por usted

Hasta ahora hemos considerado que la conexión $\boldsymbol{\nabla}$ ha sido sencillamente *asignada* a nuestra variedad $\mathcal{M}$. Esto proporciona a $\mathcal{M}$ cierto tipo de estructura. Sin embargo, es bastante habitual considerar una conexión más como una estructura secundaria que surge de una *métrica* definida en $\mathcal{M}$. Recordemos de §13.8 que una métrica (o pseudométrica) es un tensor $\boldsymbol{g}$ simétrico, no singular, $[^0_2]$-valente. Exigimos que $\boldsymbol{g}$ sea un *campo* tensorial suave, de modo que $\boldsymbol{g}$ se aplica a los espacios tangentes en los diversos puntos de $\mathcal{M}$. Una variedad con una métrica asignada de este modo se denomina *riemanniana*, o quizá *pseudorriemanniana*.[10] (Ya he hablado del gran matemático Bernhard Riemann en los capítulos 7 y 8. Él dio origen a este concepto de una variedad n-dimensional con una métrica, siguiendo el estudio anterior de Gauss de 2-variedades «riemannianas».) Normalmente, el término «riemanniana» se reserva para el caso en que $\boldsymbol{g}$ es definida positiva (véase §13.8). En este caso hay una medida (positiva) de *distancia* a lo largo de cualquier curva suave, definida por la integral de ds a lo largo de la misma, (Fig. 14.20) donde

$$ds^2 = g_{ab}\, dx^a\, dx^b.$$

Esto es un objeto adecuado para integrar a lo largo de una curva y definir así una *longitud* para dicha curva, que es una «longitud» en el sentido familiar de la palabra cuando $\boldsymbol{g}$ es definida positiva. Aunque ds

no es una 1-forma, comparte bastantes propiedades de una 1-forma para que sea una cantidad legítima para integrar a lo largo de una curva. La longitud ℓ de una curva que conecta un punto A y un punto B se expresa entonces como[11]

$$\ell = \int_A^B \mathrm{d}s, \quad \text{donde } \mathrm{d}s = (g_{ab}\mathrm{d}x^a\mathrm{d}x^b)^{\frac{1}{2}}.$$

Puede advertirse que en el caso del espacio euclídeo, esta es precisamente la definición ordinaria de longitud de una curva, vista más fácilmente en el sistema de coordenadas cartesianas, donde las componentes g_{ab} toman la forma estándar de «delta de Kronecker» de §13.3 (i.e., 1 si $a = b$, y 0 si $a \neq b$). La expresión para $\mathrm{d}s$ es básicamente un reflejo del teorema de Pitágoras (§2.1), como se ha señalado en §13.3 (véase el ejercicio [13.11]), pero operando en el nivel infinitesimal. Sin embargo, en una variedad riemanniana general la medida de longitud de una curva, de acuerdo con la fórmula anterior, nos proporciona una geometría que difiere de la de Euclides. Esto refleja el fallo del teorema de Pitágoras para intervalos finitos (en oposición a infinitesimales). De todas formas, llama la atención que este antiguo teorema siga desempeñando un papel fundamental, ahora en el nivel infinitesimal. (Recordemos el párrafo final de §2.7.)

En §17.7 veremos que el caso de signatura $+---$ tiene particular importancia en la relatividad, donde la (pseudo)métrica mide ahora directamente el tiempo que registraría un reloj ideal. Además, cualquier vector $\boldsymbol{v}$ tiene una *longitud* $|\boldsymbol{v}|$ definida por

$$|\boldsymbol{v}|^2 = g_{ab}v^a v^b,$$

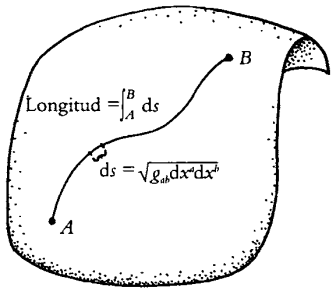


Fig. 14.20. La longitud de una curva suave es $\int \mathrm{d}s$, donde $\mathrm{d}s^2 = g_{ab}\mathrm{d}x^a\mathrm{d}x^b$.

que, para una $\boldsymbol{g}$ definida positiva, es positiva siempre que $\boldsymbol{v}$ no se anule. Sin embargo, en la teoría de la relatividad necesitamos en su lugar una métrica *lorentziana* (véase §13.8), y $|\boldsymbol{v}|^2$ puede ser de uno u otro signo. Veremos la importancia de esto más adelante (§17.9, §18.3).

¿Cómo una (pseudo)métrica $\boldsymbol{g}$ no singular determina unívocamente una conexión ∇ libre de torsión? Una manera de expresar el requisito sobre ∇ consiste en decir que el transporte paralelo de un vector debe conservar siempre su longitud (una propiedad que he establecido en §14.2 para el transporte paralelo en la esfera S^2). De modo equivalente, podemos expresar este requisito como

$$\nabla \boldsymbol{g} = 0.$$

Esta condición (junto con la anulación de la torsión) basta para fijar ∇ por completo.[14.26] La conexión ∇ recibe indistintamente los nombres de conexión *riemanniana*, de *Christoffel* o de *Levi-Civita* (por Bernhard Riemann, 1826-1866; Elwin Christoffel, 1829-1900, y Tulio Levi-Civita, 1873-1941, que aportaron ideas importantes en relación con esta noción).[14.27]

Hay otra forma de entender el hecho de que una métrica $\boldsymbol{g}$ (digamos definida positiva) determine una conexión. La noción de una *geodésica* puede obtenerse directamente a partir de la métrica. Una curva en $\mathcal{M}$ que *minimiza* su longitud $\int \mathrm{d}s$ (la cantidad que se ilustra en la Fig. 14.20) entre dos puntos fijos es en realidad una geodésica para la métrica $\boldsymbol{g}$. Conocer el lugar geométrico de las geodésicas es más de lo que se necesita para conocer la conexión ∇. El resto de la información necesaria para fijar ∇ por completo es un conocimiento de los *parámetros afines* a lo largo de las geodésicas. Estos son los parámetros que miden la longitud de arco a lo largo de las curvas, y los múltiplos constantes

[14.26] Obtenga explícitamente la expresión en componentes $\Gamma^n_{bc} = \frac{1}{2}g^{nd}(\partial g_{bd}/\partial x^c + \partial g_{cd}/\partial x^b - \partial g_{cb}/\partial x^d)$ para las cantidades Γ^n_{bc} de la conexión (símbolos de Christoffel). (Véase el ejercicio [14.6].)

[14.27] Obtenga la expresión clásica $R_{abc}{}^d = \partial\Gamma^d_{cb}/\partial x^a - \partial\Gamma^d_{ca}/\partial x^b + \Gamma^n_{cb}\Gamma^d_{na} - \Gamma^n_{ca}\Gamma^d_{nb}$ para el tensor de curvatura en términos de símbolos de Christoffel. *Sugerencia*: Utilice la definición de §14.4 del tensor de curvatura, donde ξ^d es cada uno de los vectores de coordenadas $\delta^a_1, \ldots, \delta^a_n$. (Como en el ejercicio [14.6], las cantidades δ^a_1, δ^a_2, etc., deben considerarse como vectores individuales, donde el superíndice a puede verse como un índice abstracto, de acuerdo con §12.8.)

de tales parámetros, y esto queda de nuevo fijado por $\boldsymbol{g}$.[14.28] Cuando $\boldsymbol{g}$ no es definida positiva, el argumento es básicamente el mismo, pero ahora las geodésicas no minimizan $\int ds$, y la integral es lo que se denomina «estacionaria» para una geodésica. (Esta cuestión será abordada de nuevo más adelante; véanse §17.9 y §20.1.)

En geometría (pseudo)riemanniana, la métrica g_{ab}, y su inversa g^{ab} (definida por $g^{ab}g_{bc} = \delta^a_c$) pueden utilizarse para subir o bajar los índices de un tensor. En particular, los vectores pueden convertirse en covectores y los covectores en vectores, como en §13.9:

$$v_a = g_{ab}v^b \quad \text{y} \quad \alpha^a = g^{ab}\alpha_b.$$

Es normal atenerse al mismo símbolo para el núcleo (aquí v y α) y utilizar la posición de los índices para distinguir el carácter geométrico de la cantidad. Aplicando este procedimiento para bajar el superíndice del tensor de curvatura, definimos el tensor de *Riemann* o de *Riemann-Christoffel*

$$R_{abcd} = R_{abc}{}^e g_{ed},$$

que tiene valencia $\left[{}^0_4\right]$. Posee algunas notables simetrías, además de las dos relaciones (antisimetría en *ab* y simetría de Bianchi, es decir, anulación de la parte antisimétrica en *abc*) que teníamos antes. También tenemos[14.29] antisimetría en *cd* y simetría bajo intercambio de *ab* por *cd*:

$$R_{abcd} = -R_{abdc} = R_{cdab}.$$

Véase la Fig. 14.21 para la representación diagramática de estos objetos. Un tensor $\left[{}^0_4\right]$-valente general en una n-variedad tiene n^4 componentes, pero en el caso de un tensor de Riemann, debido a estas simetrías, solo $\frac{1}{12}n^2(n^2-1)$ de estas componentes son independientes.[14.30]

Llegados a este punto, es oportuno llamar la atención del lector so-

[14.28] Dé detalles para este argumento.
[14.29] Establezca estas relaciones, obteniendo primero la antisimetría en *cd* a partir de $\nabla_{[a}\nabla_{b]}g_{cd} = 0$ y utilice luego las dos antisimetrías y la simetría de Bianchi para obtener la simetría de intercambio.
[14.30] Verifique que las simetrías permiten solo 20 componentes independientes cuando $n = 4$.

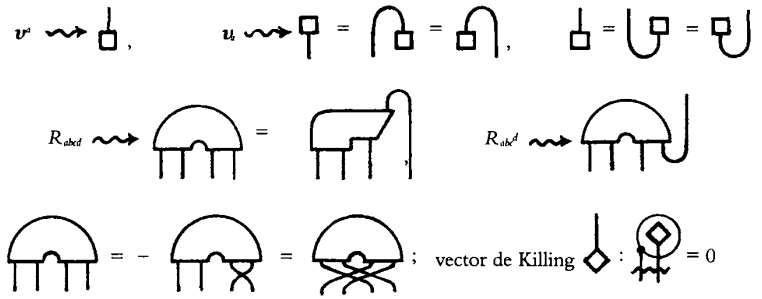


Fig. 14.21. Elevando y bajando índices en la notación de «aro»: $v_a = g_{ab}v^b = v^b g_{ba}$, $v^a = g^{ab}v_b = v_b g^{ba}$, $R_{abcd} = R_{abc}{}^e g_{ed}$, $R_{abc}{}^d = R_{abce}g^{ed}$, $R_{abcd} = -R_{abdc} = R_{cdab}$; $\boldsymbol{\kappa}^a$ es un vector de Killing si $\nabla_{(a}\kappa_{b)} = 0$.

bre la noción de un *vector de Killing* sobre una variedad (pseudo)riemanniana $\mathcal{M}$. Este es un campo vectorial $\boldsymbol{\kappa}$ que tiene la propiedad de que la derivada de Lie con respecto al mismo aniquila la métrica:

$$\pounds_{\boldsymbol{\kappa}} g = 0.$$

Esta ecuación puede reescribirse en la notación de índices (con paréntesis redondos que denotan simetrización, como en §12.7; véase también la Fig. 14.21) como

$$\nabla_a\kappa_b + \nabla_b\kappa_a = 0 \quad \text{i.e.,} \quad \nabla_{(a}\kappa_{b)} = 0,$$

donde $\boldsymbol{\nabla}$ es la conexión de Levi-Civita estándar.[14.31] Un vector de Killing sobre una variedad (pseudo)riemanniana $\mathcal{M}$ es el generador de una *simetría* continua de $\mathcal{M}$ (que solo puede ser una simetría local[12] si $\mathcal{M}$ es no compacta). Si $\mathcal{M}$ contiene más de un vector de Killing independiente, entonces el conmutador de los dos es otro vector de Killing.[14.32] Los vectores de Killing tienen particular importancia en la teoría de la relatividad, como veremos en §19.5 y §§30.4,6,7.

[14.31] Obtenga esta ecuación.
[14.32] Verifique este hecho «geométricamente obvio» por cálculo directo. (¿Y por qué es obvio?)

14.8. Variedades simplécticas

Habría que comentar que no existen muchas estructuras tensoriales locales que definan una conexión única, de modo que tenemos suerte de que las métricas (o las pseudométricas) sean con frecuencia cosas que nos son dadas físicamente. No obstante, se obtiene una importante familia de ejemplos para los que esto *no* es así cuando tenemos una estructura dada por un campo tensorial *antisimétrico* (no singular) $\boldsymbol{S}$, dado por sus componentes S_{ab}. Semejante estructura está presente en los *espacios de fases* de la mecánica clásica (§20.1). Más adelante, en §§20.2,4 y §27.3, haré más comentarios sobre estos notables espacios. Son ejemplos de lo que se conocen como *variedades simplécticas*. Además de ser antisimétrica y no singular, la *estructura simpléctica* $\boldsymbol{S}$ debe satisfacer[14.33]

$$\mathrm{d}\boldsymbol{S} = 0.$$

(Este sería el caso estándar de una forma simpléctica *real* en una variedad real $2m$-dimensional, donde la simetría local estaría dada por el habitual grupo simpléctico con «signatura dividida» $\mathrm{Sp}(m, m)$; véase §13.10. No me consta que se hayan estudiado extensamente «variedades simplécticas de otras signaturas.)

La inversa S^{ab}, de S_{ab} (definida por $S^{ab}S_{bc} = \delta^a_c$), define lo que se conoce como el «*paréntesis de Poisson*» (que debe su nombre al distinguido matemático francés Siméon Denis Poisson, que vivió de 1781 a 1840). Este combina dos campos escalares $\boldsymbol{\Phi}$, $\boldsymbol{\Psi}$ en un espacio de fases para dar un tercero:

$$\{\boldsymbol{\Phi}, \boldsymbol{\Psi}\} = -\frac{1}{2} S^{ab} \nabla_a \boldsymbol{\Phi} \nabla_b \boldsymbol{\Psi}$$

(donde el factor $-1/2$ se ha insertado simplemente por consistencia con las convencionales expresiones en coordenadas). Esta es una cantidad importante en mecánica clásica. Más adelante veremos (en §20.4) cómo codifica las *ecuaciones de Hamilton*, que ofrecen un método general fundamental que engloba las ecuaciones de la mecánica clásica y

[14.33] Explique por qué esto puede escribirse $\nabla_a S_{bc} + \nabla_b S_{ca} + \nabla_c S_{ab} = 0$, utilizando cualquier conexión ∇ libre de torsión.

proporciona el enlace con la mecánica cuántica. La antisimetría de $\boldsymbol{S}$ y la condición $\mathrm{d}\boldsymbol{S} = 0$ nos proporcionan las elegantes relaciones[14.34]

$$\{\boldsymbol{\Phi}, \boldsymbol{\Psi}\} = -\{\boldsymbol{\Psi}, \boldsymbol{\Phi}\}, \quad \{\boldsymbol{\Theta}, \{\boldsymbol{\Phi}, \boldsymbol{\Psi}\}\} + \{\boldsymbol{\Phi}, \{\boldsymbol{\Psi}, \boldsymbol{\Theta}\}\} + \{\boldsymbol{\Psi}, \{\boldsymbol{\Theta}, \boldsymbol{\Phi}\}\} = 0,$$

que pueden compararse con las correspondientes identidades con conmutadores (paréntesis de Lie) de §14.6. (Recordemos la identidad de Jacobi.) Volveremos a la extraordinariamente rica geometría de las variedades simplécticas cuando consideremos la descripción geométrica de la mecánica clásica en §20.4.

La estructura local de una variedad simpléctica es un ejemplo de lo que podría llamarse una estructura «blanda». No hay, por ejemplo, ninguna noción de curvatura para una variedad simpléctica que pudiera servir para distinguir localmente una variedad simpléctica de otra. Si tenemos dos variedades simplécticas reales de la misma dimensión (y la misma «signatura»; cf. §13.10), entonces son completamente idénticas localmente (en el sentido de que para cualquier punto p en una variedad y cualquier punto q en la otra existen conjuntos abiertos de p y q que son idénticos.[13] Esto está en abierto contraste con el caso de variedades (pseudo)riemannianas, o variedades en las que meramente se especifica una conexión. En tales casos, el tensor de curvatura (y, por ejemplo, sus diversas derivadas covariantes) define una estructura característicamente local que quizá es diferente para diferentes variedades.

Existen otros ejemplos de tales estructuras «blandas», entre los que se encuentra la estructura compleja definida en §12.9 que hace posible que una variedad real $2m$-dimensional sea reinterpretada como una variedad compleja m-dimensional. En este caso, la blandura es evidente, porque no hay ningún rasgo claro, aparte de la dimensión compleja m, que distinga localmente una variedad compleja de otra (o de $\mathbb{C}^m$). Seguiría siendo blanda si se le asignara una estructura simpléctica compleja (holomorfa)[14.35] (y ahora ni siquiera tenemos que preocuparnos de una noción de «signatura» para la S_{ab} compleja; véase §13.10).

Pueden especificarse muchos otros ejemplos de estructuras blan-

[14.34] Demuestre estas relaciones, estableciendo primero que $S^{a[b}\nabla_a S^{cd]} = 0$.
[14.35] Explique por qué.

das. Una de estas sería una variedad real con un campo vectorial en ella que no se anula en ninguna parte. Por el contrario, una variedad real con *dos* campos vectoriales generales no sería blanda.[14.36] La cuestión de la blandura tiene cierta importancia para la *teoría de twistores*, como veremos en §33.11.

Notas

Sección 14.2

14.1. De hecho, existe una razón topológica por la que no hay forma de asignar una «paralela» a $\boldsymbol{v}$ en todos los puntos de S^2 de una forma continua (¡es el problema de peinar el pelo de un perro esférico!). Sin embargo, el enunciado análogo para S^3 no es cierto, como muestra la construcción de paralelas de Clifford (dada en §15.4).

Sección 14.3

14.2. En buena parte de la literatura física y la más antigua literatura matemática, la derivada en coordenada $\partial/\partial x^a$ se indica añadiendo un subíndice a, precedido por una coma, al extremo derecho de la lista de índices unida a la cantidad que se está derivando. En el caso de ∇_a, se suele utilizar un punto y coma en lugar de la coma.

La notación «∇_a» funciona con la notación de índices abstractos (§12.8) y las ecuaciones siguientes en el texto principal de este libro pueden (y deberían) leerse de esta manera. Las expresiones en coordenadas también pueden tratarse en esta notación, pero se necesitan dos tipos de índices distinguibles, componente y abstracto (véanse Penrose, 1968a y Penrose y Rindler, 1984).

Sección 14.4

14.3 El escalonamiento de índices es necesario cuando se introduce una métrica (§14.7), puesto que se necesitan espacios para subir y bajar índices.

[14.36] Explique por qué en cada caso. *Sugerencia*: Construya un sistema de coordenadas con $\boldsymbol{\xi} = \partial/\partial x^1$; luego tome derivadas de Lie repetidamente para construir un sistema, etc.

Sección 14.5

14.4. Estrictamente, ∇ actúa sobre campos definidos en $\mathcal{M}$, y no solo a lo largo de curvas que yacen dentro de $\mathcal{M}$. Pero esta ecuación cobra sentido porque el operador deriva solo en la dirección a lo largo de la curva. Si queremos, podemos considerar que la región de definición de $\boldsymbol{t}$ se extiende suavemente fuera de γ en $\mathcal{M}$ de alguna forma arbitraria. La forma precisa en que se haga esto es irrelevante, puesto que solo estamos pidiendo que la ecuación para $\boldsymbol{t}$ sea válida a lo largo de γ.

14.5. Véanse, por ejemplo, Nayfeh (1993) y Simmonds y Mann (1998).

Sección 14.6

14.6. Vemos el papel explícito del álgebra de Lie de los conmutadores en la fórmula de *Baker-Campbell-Hausdorff*, cuyos primeros términos están dados explícitamente en $e^{\boldsymbol{\xi}}e^{\boldsymbol{\eta}} = e^{\boldsymbol{\xi}+\boldsymbol{\eta}+\frac{1}{2}[\boldsymbol{\xi},\boldsymbol{\eta}]+\frac{1}{12}([\boldsymbol{\xi},[\boldsymbol{\xi},\boldsymbol{\eta}]]+[[\boldsymbol{\xi},\boldsymbol{\eta}],\boldsymbol{\eta}])+\cdots}$, donde los puntos que siguen representan una expresión adicional en conmutadores múltiples de $\boldsymbol{\xi}$ y $\boldsymbol{\eta}$, i.e., un elemento del álgebra de Lie generada por $\boldsymbol{\xi}$ y $\boldsymbol{\eta}$.

14.7. De forma algo más precisa, podemos escoger coordenadas $x^2, x^3, \ldots, x^n$ constantes a lo largo de la curva, con $x^1 = t$; entonces $\boldsymbol{\xi} = \partial/\partial t$, a lo largo de la curva. Es simplemente el teorema de Taylor (§6.4) que nos dice que la receta anterior da $e^{t\boldsymbol{\xi}}(\Phi)$.

14.8. Análoga a la exponenciación $e^{t\boldsymbol{\xi}}$ de $\boldsymbol{\xi}$ que da el valor de una cantidad escalar Φ a distancia finita, existe una expresión correspondiente con $\pounds_{\boldsymbol{\xi}}$ en lugar de $\boldsymbol{\xi}$ para obtener un tensor $\boldsymbol{Q}$ a distancia finita, medida frente a un sistema de referencia «arrastrado».

14.9. Véanse Schouten (1954) y Penrose y Rindler (1984), p. 202.

Sección 14.7

14.10. En algunos libros de matemáticas se ha utilizado el término «semirriemanniana» para el caso indefinido (véase O'Neill, 1983), pero creo que «pseudorriemanniana» es una terminología más adecuada.

14.11. Una forma habitual de dar significado a esta expresión es introducir un parámetro, digamos u, a lo largo de la curva y escribir $ds = (ds/du)du$. La cantidad ds/du es una función ordinaria de u, expresada en términos de dx^a/du.

14.12. Esta «localidad» puede entenderse en el sentido siguiente. Para cada punto p de $\mathcal{M}$ existe una exponenciación (§14.6) de una pequeña constante no nula múltiplo de κ que lleva algún conjunto abierto que

contiene a p a algún otro conjunto abierto en $\mathcal{M}$ con una estructura métrica idéntica.

Sección 14.8

14.13. Aquí, «idénticos» se refiere al hecho de que cada uno de ellos puede aplicarse en el otro de forma tal que las estructuras simplécticas se corresponden.

15

Fibrados y conexiones gauge

15.1. Algunas motivaciones físicas para los fibrados

Las herramientas introducidas en los capítulos 14 y 15 son suficientes para el tratamiento de la relatividad general de Einstein y para los espacios de fases de la mecánica clásica. Sin embargo, una buena parte de la moderna teoría de interacciones entre partículas depende de una generalización de la noción concreta de «conexión» (o derivada covariante) que se ha introducido en §14.3, generalización que se conoce como *conexión gauge*. Básicamente, nuestra noción original de derivada covariante se basaba en lo que entendemos por transporte paralelo de un vector a lo largo de una curva en nuestra variedad $\mathcal{M}$ (§14.2). Conociendo el transporte paralelo para vectores, podemos extenderlo unívocamente al transporte de cualquier cantidad tensorial (§14.3). Ahora bien, vectores y tensores son cantidades que se refieren a los espacios tangentes en puntos de $\mathcal{M}$ (véanse §12.3, §14.1 y la Fig. 12.6). Pero una conexión gauge se refiere al «transporte paralelo» de ciertas cantidades de particular interés físico que es mejor considerar referidas a cierto tipo de «espacio» distinto del espacio tangente en un punto p de $\mathcal{M}$, aunque en cierto sentido debe seguir considerándose «localizado en el punto p».

Para aclarar un poco lo que se necesita aquí, recordemos de §§12.3,8 que una vez que tenemos un espacio vectorial —aquí, el espacio de vectores tangentes en un punto— podemos construir su (espacio de covectores) dual y todos los diversos espacios de vectores $[^p_q]$-valentes. Así, en un sentido claro, los espacios de $[^p_q]$-tensores (incluyendo los espacios cotangentes, pues los covectores son $[^0_1]$-tensores) «no son

nada nuevo», una vez que tenemos los espacios tangentes T_p en puntos p. (Un comentario similar sería aplicable —al menos según mi modo de ver las cosas— a los espacios de *espinores* en p; véase §11.3. Otros podrían adoptar una actitud diferente respecto a los espinores; pero estas perspectivas alternativas no nos interesan aquí.) Los espacios que necesitamos para las teorías gauge de interacciones entre partículas (distintas de la gravedad) son diferentes de estos (y por ello *son* algo nuevo), y es mejor considerar que se refieren a un tipo de dimensión «espacial» que se añade a las del espacio y tiempo ordinarios. Estas dimensiones «espaciales» extras suelen conocerse como dimensiones *internas*, de modo que el movimiento a lo largo de una de estas «direcciones internas» no nos aleja del punto espaciotemporal en el que estamos situados.

Para dar un sentido geométrico a esa idea necesitamos la noción de un *fibrado*. Esta es una noción matemática muy precisa, y llegaremos a ella en §15.2. Se ha encontrado útil en matemáticas puras[1] mucho antes de que los físicos se dieran cuenta de que algunas de las nociones importantes que habían estado utilizando anteriormente tenían que ser entendidas realmente en términos de fibrados. En años posteriores, los físicos teóricos se han familiarizado con los conceptos matemáticos requeridos y los han incorporado en sus teorías. Sin embargo, en algunas teorías modernas estas nociones están presentadas de una forma modificada, según la cual se considera que el propio espaciotiempo adquiere dimensiones extras.

En realidad, en muchos (¿o en casi todos?) de los intentos actuales de encontrar un marco más profundo para la física fundamental (por ejemplo, supergravedad o teoría de cuerdas), la propia noción de espaciotiempo se extiende a una dimensionalidad más alta. Las «dimensiones internas» son entonces resultado de la presencia de dichas dimensiones espaciales extras, que se colocan esencialmente en pie de igualdad con las del espacio y tiempo ordinarios. El «espaciotiempo» resultante adquiere así más dimensiones que las cuatro estándar. Estas ideas se remontan aproximadamente a 1919, cuando Theodor Kaluza y Oskar Klein propusieron una ampliación de la relatividad general de Einstein en la que el número de dimensiones espaciotemporales se incrementa de 4 a 5. La dimensión extra hace posible que la maravillosa teoría del

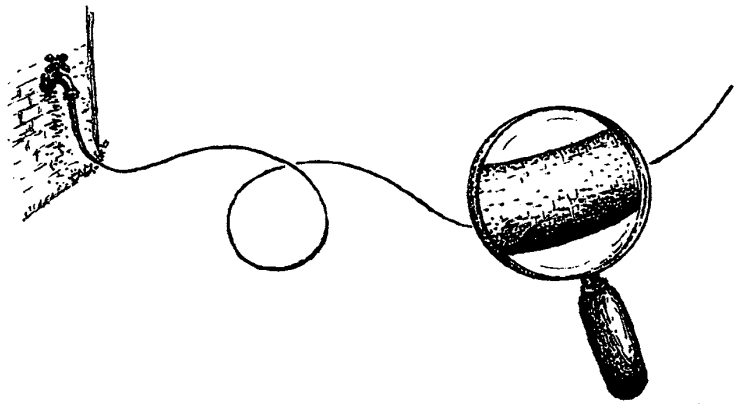


Fig. 15.1. La analogía de una manguera. Vista a gran escala, parece 1-dimensional, pero cuando se examina más minuciosamente se ve que es una superficie 2-dimensional. Análogamente, según la idea de Kaluza-Klein, podría haber «pequeñas» dimensiones espaciales extra inobservadas en una escala ordinaria.

electromagnetismo de Maxwell (véanse §§19.2,4) sea incorporada en cierto sentido dentro de una «descripción geométrica espaciotemporal». No obstante, esta «quinta dimensión» tiene que considerarse «enrollada en un bucle minúsculo», de modo que no somos directamente conscientes de ella como dimensión espacial ordinaria.

Es habitual hacer la analogía con una manguera (véase la Fig. 15.1), con la que se representa una modificación tipo Kaluza-Klein de un universo 1-dimensional. Cuando miramos en una escala grande, la manguera parece 1-dimensional: la dimensión de su longitud. Pero cuando se examina más de cerca, nos damos cuenta de que la superficie de la manguera es realmente 2-dimensional, con la dimensión extra apretadamente enrollada en una escala mucho más pequeña que la longitud de la manguera. Esto debe tomarse como una analogía directa de nuestra percepción de un espaciotiempo *físico* 4-dimensional dentro de un «espaciotiempo» *total* de Kaluza-Klein 5-dimensional. El 5-espacio de Kaluza-Klein va a ser el análogo directo de la 2-superficie de la manguera, y el espaciotiempo que percibimos es el análogo directo de la apariencia 1-dimensional de la manguera.

En muchos aspectos, esta es una idea atractiva e ingeniosa. Los de-

fensores de muchas de las teorías físicas modernas (tales como la supergravedad y la teoría de cuerdas que veremos en el capítulo 31) se sienten realmente impulsados a considerar versiones de la idea de Kaluza-Klein con dimensiones aún más altas (entre las más populares se encuentran las versiones con una dimensionalidad total de 26, 11, 10). En tales teorías se entiende que mediante la idea de conexión gauge pueden introducirse otras interacciones además de las electromagnéticas, como veremos enseguida.

Sin embargo, hay que destacar que la idea de Kaluza-Klein sigue siendo especulativa. Las «dimensiones internas» de las que dependen las presentes teorías gauge convencionales de las interacciones entre partículas no deben considerarse equiparables a las dimensiones espaciotemporales ordinarias, y por ello no surgen de un esquema de tipo Kaluza-Klein. Un tema de especulación interesante es el de si es razonable considerar que las dimensiones internas de las teorías gauge actuales surgen en definitiva de este tipo de «espaciotiempo extendido» (tipo Kaluza-Klein) en algún sentido significativo.[2] Más adelante volveré a plantear esta cuestión (§31.4).

En lugar de considerar estas dimensiones internas como parte de un espaciotiempo de dimensiones más altas, será más conveniente pensar que nos proporcionan lo que se denomina un *haz de fibras* (o simplemente un *fibrado*) sobre el espaciotiempo. Esta es una noción importante que resulta fundamental para las modernas teorías gauge de las interacciones entre partículas. Imaginemos que «sobre» cada punto del espaciotiempo hay otro espacio, denominado una *fibra*. La fibra consiste en todas las dimensiones internas, de acuerdo con la imagen física mencionada antes. Pero el concepto de fibrado tiene aplicaciones mucho más generales que esto, de modo que será mejor que no nos atemos necesariamente a este tipo de interpretación física, al menos por el momento.

15.2. La idea matemática de un fibrado

Un *fibrado* $\mathcal{B}$ es una variedad con cierta estructura, que está definida en términos de otras dos variedades $\mathcal{M}$ y $\mathcal{V}$, de las que $\mathcal{M}$ se denomina *espacio base* (que, en la mayoría de las aplicaciones físicas, es el propio espaciotiempo), y $\mathcal{V}$ se denomina *fibra* (el espacio interno en la mayoría de las aplicaciones físicas). El propio fibrado $\mathcal{B}$ puede considerarse compuesto de una familia completa de fibras $\mathcal{V}$; de hecho, está constituido como un «valor de $\mathcal{M}$ en $\mathcal{V}$s» (véase la Fig. 15.2). El tipo más simple de fibrado es lo que se denomina un *espacio producto*. Este sería un fibrado *trivial* o «no retorcido», pero son más interesantes los fibrados *retorcidos*. Enseguida daré algunos ejemplos de ambos. Es importante que el espacio $\mathcal{V}$ tenga también ciertas *simetrías*, pues es la presencia de estas simetrías la que da libertad para el *retorcimiento* que hace interesante el concepto de fibrado. El grupo $\mathcal{G}$ de simetrías de $\mathcal{V}$ en el que estamos interesados se denomina el *grupo* del fibrado $\mathcal{B}$. A menudo decimos que $\mathcal{B}$ es un $\mathcal{G}$-*fibrado* sobre $\mathcal{M}$. En muchas situaciones se toma como $\mathcal{V}$ un espacio vectorial, en cuyo caso diremos que el fibrado es un *fibrado vectorial*. Entonces, el grupo $\mathcal{G}$ es el grupo lineal general de la dimensión relevante, o un subgrupo del mismo (véanse §§13.3,6-10).

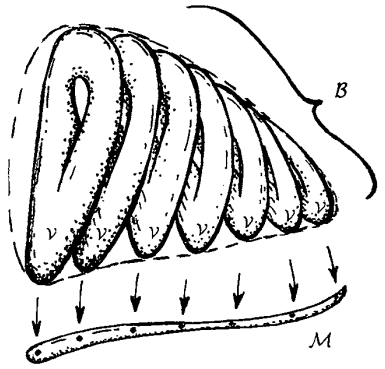


Fig. 15.2. Un fibrado $\mathcal{B}$ con espacio base $\mathcal{M}$ y fibra $\mathcal{V}$ puede considerarse constituido como un «valor de $\mathcal{M}$ en $\mathcal{V}$s». La proyección canónica de $\mathcal{B}$ en $\mathcal{M}$ puede verse como el colapso de cada fibra $\mathcal{V}$ en un solo punto.

No tenemos que pensar en $\mathcal{M}$ como una *parte* de $\mathcal{B}$ (i.e., $\mathcal{M}$ no está dentro de $\mathcal{B}$); en su lugar, $\mathcal{B}$ debe verse como un espacio independiente de $\mathcal{M}$, que tendemos a considerar como si en cierto sentido estuviera *por encima* del espacio base $\mathcal{M}$. Existen muchas copias de la fibra $\mathcal{V}$ en el fibrado $\mathcal{B}$, pues hay una copia entera de $\mathcal{V}$ por encima de cada punto de $\mathcal{M}$. Todas las copias de las fibras son disjuntas (i.e., no hay dos que se intersecten), y en conjunto forman el fi-

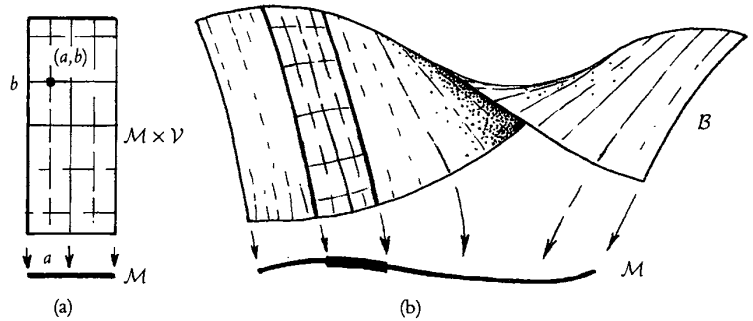


Fig. 15.3. (a) El caso particular de un fibrado «trivial», que es el espacio producto $\mathcal{M}\times\mathcal{V}$ de $\mathcal{M}$ por $\mathcal{V}$. Los puntos de $\mathcal{M}\times\mathcal{V}$ pueden interpretarse como pares de elementos (a, b), con a en $\mathcal{M}$ y b en $\mathcal{V}$. (b) El fibrado «retorcido» general $\mathcal{B}$, sobre $\mathcal{M}$, con fibra $\mathcal{V}$, se parece a $\mathcal{M}\times\mathcal{V}$ localmente, i.e., la parte de $\mathcal{B}$ sobre cualquier región abierta suficientemente pequeña de $\mathcal{M}$ es idéntica a la parte de $\mathcal{M}\times\mathcal{V}$ sobre la misma región de $\mathcal{M}$. Pero las fibras se retuercen, de modo que $\mathcal{B}$ no es globalmente el mismo que $\mathcal{M}\times\mathcal{V}$.

brado $\mathcal{B}$. La forma de considerar $\mathcal{M}$ en relación con $\mathcal{B}$ es como un *espacio cociente* del fibrado $\mathcal{B}$ por la familia de fibras $\mathcal{V}$. Es decir, cada punto de $\mathcal{M}$ corresponde exactamente a una copia de $\mathcal{V}$ individual e independiente. Existe una aplicación continua de $\mathcal{B}$ en $\mathcal{M}$, denominada la *proyección canónica* de $\mathcal{B}$ en $\mathcal{M}$, que colapsa cada fibra entera $\mathcal{V}$ en el punto concreto de $\mathcal{M}$ por encima del cual está (véase la Fig. 15.2).

El *espacio producto* de $\mathcal{M}$ por $\mathcal{V}$ (el fibrado trivial de $\mathcal{V}$ sobre $\mathcal{M}$) se escribe $\mathcal{M}\times\mathcal{V}$. Los puntos de $\mathcal{M}\times\mathcal{V}$ son los *pares* de elementos (a, b), donde a pertenece a $\mathcal{M}$ y b pertenece a $\mathcal{V}$; véase la Fig. 15.3a. (Ya hemos visto la misma idea aplicada a grupos en §13.2.).[3] Una fibra «retorcida» $\mathcal{B}$ más general sobre $\mathcal{M}$ se parece a $\mathcal{M}\times\mathcal{V}$ *localmente*, en el sentido de que la parte de $\mathcal{B}$ que reside sobre cualquier región abierta y suficientemente pequeña de $\mathcal{M}$ tiene una estructura idéntica a la de la parte de $\mathcal{M}\times\mathcal{V}$ que reside sobre la misma región abierta de $\mathcal{M}$.Véase la Fig. 15.3b. Pero a medida que nos movemos a lo largo de $\mathcal{M}$, las fibras superiores pueden retorcerse, de modo que, en conjunto, $\mathcal{B}$ es *diferente* (a menudo topológicamente diferente) de $\mathcal{M}\times\mathcal{V}$. La dimensión

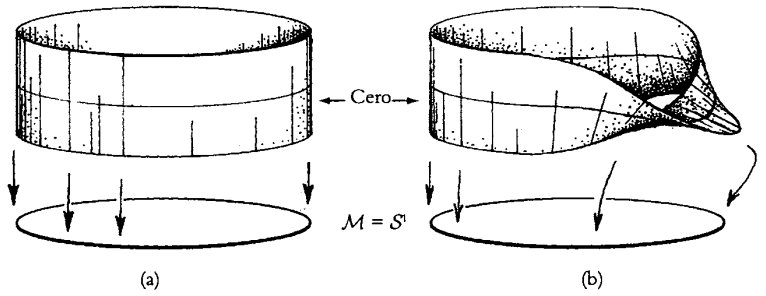


Fig. 15.4. Para entender cómo puede darse este giro, consideremos el caso en que $\mathcal{M}$ es un círculo S^1 y la fibra $\mathcal{V}$ es un espacio vectorial 1-dimensional (i.e., un espacio modelado sobre $\mathbb{R}$, pero donde solo está marcado el origen 0, y ningún otro valor (tal como el elemento identidad 1). (a) El caso trivial $\mathcal{M}\times\mathcal{V}$, que es aquí un cilindro ordinario 2-dimensional. (b) En el caso retorcido, obtenemos una cinta de Moebius (como en la Fig. 12.15).

de $\mathcal{B}$ es siempre la suma de las dimensiones de $\mathcal{M}$ y de $\mathcal{V}$ con independencia del retorcimiento.[15.1]

Todo esto puede ser confuso, de modo que para hacernos una idea de a qué se parece un fibrado pondré un ejemplo. En primer lugar, consideremos que nuestro espacio $\mathcal{M}$ es un círculo S^1, y la fibra $\mathcal{V}$ es un espacio vectorial 1-dimensional (que podemos representar como una copia de la recta real $\mathbb{R}$, con el origen 0 marcado). Dicho fibrado se denomina un *fibrado línea* (real) sobre S^1. Ahora $\mathcal{M}\times\mathcal{V}$ es un cilindro 2-dimensional; véase la Fig. 15.4a. ¿Cómo podemos construir un fibrado *retorcido* $\mathcal{B}$ sobre $\mathcal{M}$ con fibra $\mathcal{V}$? Podemos tomar una *cinta de Moebius* (véase la Fig. 15.4b). Veamos por qué esto es un fibrado, «localmente» igual al cilindro. Podemos obtener una región adecuadamente «local» del espacio base S^1 eliminando un punto p de S^1. Esto rompe el círculo base para dar un segmento[4] simplemente conexo[5] S^1-p, y la parte de $\mathcal{B}$ que está por encima de dicho segmento es precisamente la misma que la parte del *cilindro* que está por encima de S^1-p. La distinción entre el fibrado de Moebius $\mathcal{B}$ y el cilindro aparece solo

[15.1] Explique por qué la dimensión de $\mathcal{M}\times\mathcal{V}$ es la suma de las dimensiones de $\mathcal{M}$ y $\mathcal{V}$.

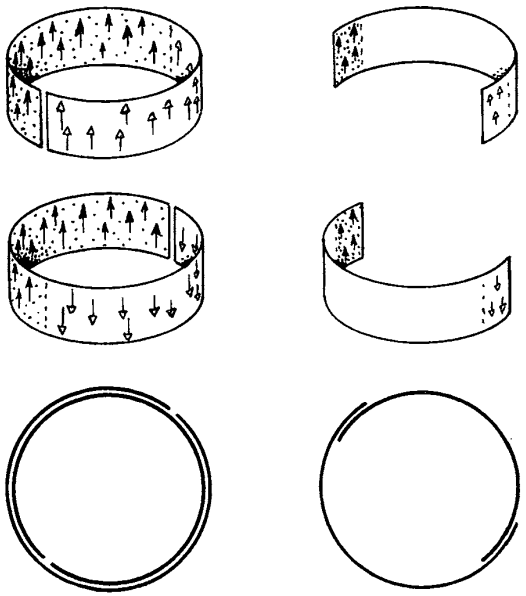


Fig. 15.5. (a) Podemos crear una región adecuadamente «local» (simplemente conexa) de la base S^1 eliminando un punto p de ella, siendo solo un producto la parte de la fibra sobre $S^1 - p$. Lo mismo se aplica a la parte de $\mathcal{B}$ sobre $S^1 - q$, donde q es un punto diferente de S^1. Obtenemos un cilindro si podemos empalmar las dos partes de $\mathcal{B}$ directamente, pero obtenemos el fibrado de Moebius, como se ha ilustrado antes, si aplicamos una reflexión arriba/abajo (una simetría de $\mathcal{V}$) a una de las dos porciones empalmadas. (b) La cinta de Moebius resultante es un poco más obvia si reducimos el tamaño de las dos partes de S^1, de modo que solo haya pequeñas regiones de solapamiento.

cuando miramos lo que hay por encima del S^1 *entero*. Podemos imaginar que S^1 está compuesto de dos de tales regiones, a saber, $S^1 - p$ y $S^1 - q$, donde p y q son dos puntos distintos de S^1; entonces podemos componer la totalidad de $\mathcal{B}$ a partir de dos regiones correspondientes, cada una de las cuales es un fibrado trivial sobre una de las regiones individuales de S^1. Es al «pegar» estas dos regiones de fibrados triviales cuando aparece el «giro» en el fibrado de Moebius (Fig. 15.5). De hecho, se ve claramente que lo que aparece con tan solo un giro es una cinta de Moebius, si reducimos el tamaño de nuestras cartas en S^1, como

se indica en la Fig. 15.5b, reducción que no supone diferencia para la estructura de $\mathcal{B}$.

Es importante darse cuenta de que la posibilidad de este giro es el resultado de una *simetría* particular que posee la fibra $\mathcal{V}$, a saber, la que invierte el signo de los elementos del espacio vectorial 1-dimensional $\mathcal{V}$. (Esto es $v \mapsto -v$, para cada v en $\mathcal{V}$.) Esta operación conserva la estructura de $\mathcal{V}$ como espacio vectorial. Deberíamos notar que esta operación no es realmente una simetría del sistema de los números reales $\mathbb{R}$. De hecho, $\mathbb{R}$ no posee simetrías en absoluto. (Por ejemplo, el número 1 es ciertamente diferente de -1, y $x \mapsto -x$ *no* es una simetría de $\mathbb{R}$, puesto que no conserva la estructura multiplicativa de $\mathbb{R}$.)[15.2] Por eso se toma $\mathcal{V}$ como un espacio vectorial real 1-dimensional antes que solo *como* la propia recta real $\mathbb{R}$. A veces decimos que $\mathcal{V}$ está *modelado* sobre la recta real. Enseguida veremos cómo otras simetrías de la fibra proporcionan oportunidades para otros tipos de giro.

15.3. Secciones transversales de fibrados

Una forma de caracterizar la diferencia entre el cilindro y el fibrado de Moebius es en términos de lo que se denominan *secciones transversales* (o simplemente *secciones*) de un fibrado. Desde un punto de vista geométrico, consideramos una sección transversal de un fibrado $\mathcal{B}$ sobre $\mathcal{M}$ como una imagen continua de $\mathcal{M}$ en $\mathcal{B}$ que corta a cada fibra individual en un único punto (véase la Fig. 15.6a). Llamamos a esto un «elevador» del espacio base $\mathcal{M}$ dentro del fibrado. Nótese que, si ejecutamos la aplicación que eleva $\mathcal{M}$ a una sección transversal de $\mathcal{B}$ y lo hacemos seguir de la proyección canónica, obtenemos la aplicación identidad de $\mathcal{M}$ en sí mismo (es decir, cada punto de $\mathcal{M}$ se aplica de nuevo en sí mismo).

En el caso de un fibrado trivial $\mathcal{M} \times \mathcal{V}$, las secciones transversales pueden interpretarse simplemente como las funciones continuas sobre el espacio base $\mathcal{M}$ que toman valores en el espacio $\mathcal{V}$ (i.e., son aplicaciones continuas de $\mathcal{M}$ en $\mathcal{V}$). Así, una sección transversal de $\mathcal{M} \times \mathcal{V}$

[15.2] Explique esto.

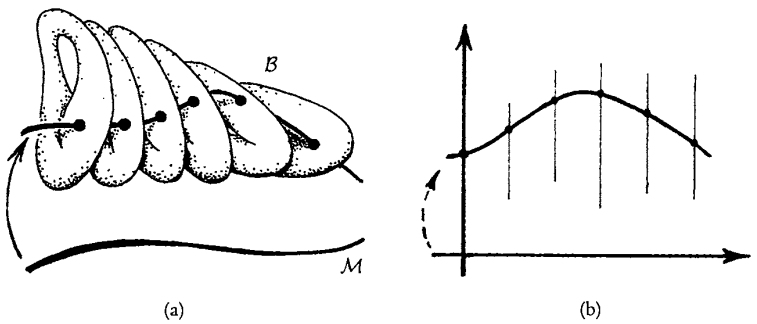


Fig. 15.6. (a) Una sección transversal (o sección) de un fibrado $\mathcal{B}$ es una imagen continua de $\mathcal{M}$ en $\mathcal{B}$ que corta cada fibra individual en un solo punto. (b) Esto generaliza la idea ordinaria de la gráfica de una función.

asigna[6] de una forma continua un punto de $\mathcal{V}$ a cada punto de $\mathcal{M}$. Esto es similar a la idea ordinaria de la *gráfica* de una función que se ilustra en la Fig. 15.6b. Con más generalidad, en el caso de un fibrado retorcido $\mathcal{B}$, cualquier sección transversal de $\mathcal{B}$ define una noción de «función retorcida» que es más general que la idea ordinaria de una función.

Volvamos a nuestro ejemplo particular de §15.2. En el caso del cilindro (fibrado producto $\mathcal{M} \times \mathcal{V}$), nuestras secciones transversales pueden representarse simplemente como curvas que se enrollan una vez alrededor del cilindro, cortando cada fibra solo una vez (Fig. 15.7a). Puesto que el fibrado es precisamente un espacio producto, podemos pensar en cada fibra como si fuera solo una copia de la recta real, y así es posible asignar de manera consistente coordenadas con valores reales a las fibras. La coordenada de valor 0 en cada fibra señala la *sección cero* de «puntos marcados» que representan los ceros de los espacios vectoriales $\mathcal{V}$. Una sección transversal general proporciona una función continua con valores reales en el círculo (siendo la «altura» sobre la sección cero el valor de la función en cada punto del círculo). Evidentemente hay muchas secciones transversales que no cortan la sección cero (las funciones que no se anulan en S^1). Por ejemplo, podemos escoger una sección del cilindro que es paralela a la sección cero aunque no coincidente con ella. Esto representa una función *constante* no nula en el círculo.

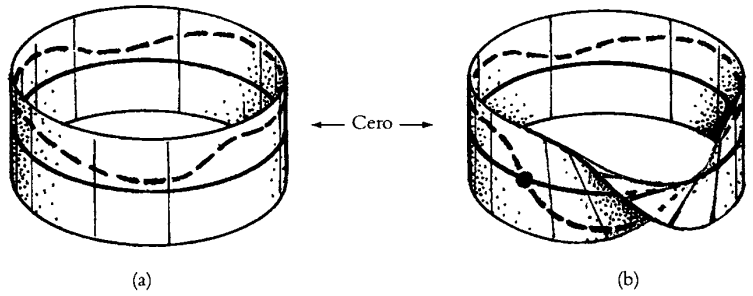


Fig. 15.7. Una sección transversal de un fibrado línea sobre S^1 es un lazo que da una vuelta, cortando cada fibra solo una vez. (a) Cilindro: hay secciones que no cortan en ninguna parte la sección 0. (b) Fibrado de Moebius: toda sección corta la sección 0.

Sin embargo, cuando consideramos el fibrado de Moebius $\mathcal{B}$, nos damos cuenta de que las cosas son muy diferentes. Para el lector no debería ser difícil aceptar que ahora cada sección transversal de $\mathcal{B}$ debe cortar la sección cero (Fig. 15.7b). (La noción de sección cero aún se aplica, puesto que $\mathcal{V}$ es un espacio vectorial, con su cero «marcado».) Esta diferencia cualitativa respecto al caso anterior deja claro que $\mathcal{B}$ debe ser topológicamente distinto de $\mathcal{M} \times \mathcal{V}$. Para ser un poco más concretos, podemos empezar a asignar coordenadas reales a las diversas fibras $\mathcal{V}$, igual que antes, pero tenemos que adoptar un convenio según el cual en cierto punto del círculo debe haber un «salto» de signo $(x \mapsto -x)$, de modo que una sección transversal de $\mathcal{B}$ corresponde a una función con valor real en el círculo que sería continua excepto que cambia de signo cuando se recorre el círculo por completo. Cualquier sección transversal semejante debe tomar el valor cero en alguna parte.[15.3]

En este ejemplo, la naturaleza de la familia de secciones transversales es suficiente para distinguir el fibrado de Moebius del cilindro. Un examen de la familia de secciones transversales conduce a menudo a una forma útil de distinguir varios fibrados diferentes sobre el mismo espacio base $\mathcal{M}$. La distinción entre el fibrado de Moebius y el espacio

[15.3] Explique en detalle este argumento, utilizando la construcción de **B** a partir de dos cartas, tal como se ha indicado antes.

producto (cilindro) es, no obstante, un poco menos extrema que en el caso de otros ejemplos de fibrados. ¡A veces un fibrado no tiene secciones transversales en absoluto! Consideremos a continuación uno de tales ejemplos particularmente importante y famoso.

15.4. El fibrado de Clifford

¡En este ejemplo, nos vamos a poner serios! El espacio base $\mathcal{M}$ va a ser una esfera 2-dimensional S^2 y la variedad fibrada $\mathcal{B}$ resulta ser una 3-esfera S^3. Las fibras $\mathcal{V}$ son círculos S^1 («1-esferas»). Esto se conoce habitualmente como la *fibración de Hopf* de S^3, una construcción topológica apuntada por Heinz Hopf (1931). Pero el procedimiento de Hopf se basaba de manera explícita (con la referencia debida) en una construcción geométrica anterior de «paralelos de Clifford», que se debe a William Clifford (1873). Llamaré *fibrado de Clifford* a la S^3 fibrada geométricamente de este modo.

La forma más reveladora de obtener el fibrado de Clifford es considerar primero el espacio $\mathbb{C}^2$ de pares de números complejos (w, z). (Aquí, lo relevante de la estructura de $\mathbb{C}^2$ es que se trata simplemente de un espacio vectorial complejo 2-dimensional; véase §12.9.) Nuestro espacio fibrado $B(= S^3)$ debe considerarse como la 3-esfera unidad S^3 en $\mathbb{C}^2$, definida por la ecuación (véase la parte final de §10.1)

$$|w|^2 + |z|^2 = 1.$$

Esto representa la ecuación real $u^2 + v^2 + x^2 + y^2 = 1$, la ecuación de una 3-esfera, donde $w = u + iv$ y $z = x + iy$ son las expresiones respectivas de w y z en términos de sus partes real e imaginaria. (Ello está en analogía directa con la ecuación de una 2-esfera ordinaria $x^2 + y^2 + z^2 = 1$ en el 3-espacio euclídeo con coordenadas cartesianas reales x, y, z.)

Para obtener la fibración vamos a considerar la familia de *líneas rectas complejas* que pasan por el origen (i.e., los subespacios vectoriales complejos 1-dimensionales de $\mathbb{C}^2$). Cada una de estas rectas viene dada por una ecuación de la forma

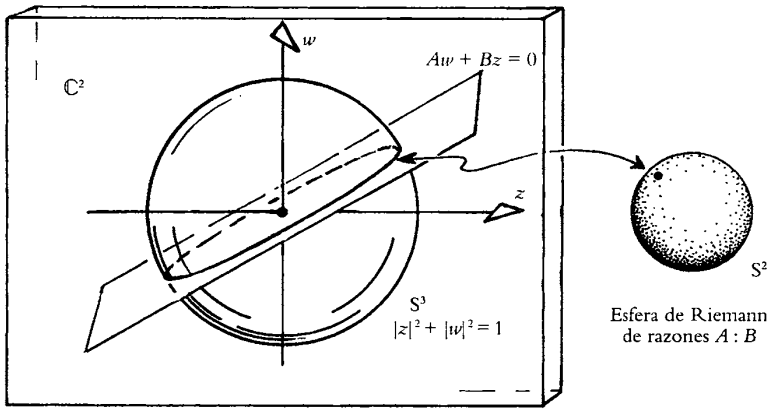


Fig. 15.8. El fibrado de Clifford. Tómese $\mathbb{C}^2$ con coordenadas (w, z), que contiene la 3-esfera $\mathcal{B} = S^3$ dada por $|w|^2 + |z|^2 = 1$. Cada fibra $\mathcal{V} = S^1$ es el círculo unidad en una línea recta compleja que pasa por el origen $Aw + Bz$ (subespacio vectorial 1-dimensional complejo de $\mathbb{C}^2$), y está determinada por la razón $A : B$. La esfera de Riemann S^2 de tales razones es el espacio base $\mathcal{B}$.

$$Aw + Bz = 0,$$

donde A y B son números complejos (al menos uno de los cuales es distinto de cero). Al ser un espacio vectorial 1-dimensional complejo, esta línea es una copia del plano complejo, y corta a S^3 en un círculo S^1, que podemos considerar como el círculo unidad en dicho plano (Fig. 15.8). Estos círculos van a ser nuestras fibras $\mathcal{V} = S^1$. Las líneas diferentes solo pueden cortarse en el origen, de modo que dos S^1 distintas no pueden tener un punto en común. Así pues, esta familia de círculos S^1 constituyen realmente fibras que dan a S^3 una estructura de fibrado.

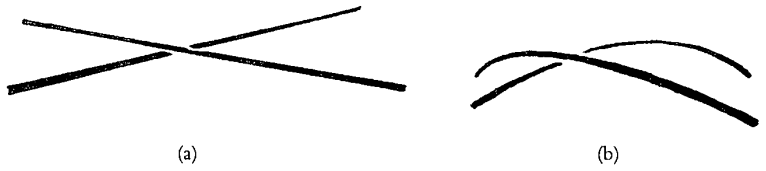
¿Cuál es el espacio base $\mathcal{M}$? Evidentemente, obtenemos la misma recta $Aw + Bz = 0$ si multiplicamos A y B por el mismo número complejo distinto de cero, de modo que es realmente la razón $A : B$ la que distingue unas rectas de otras. Uno u otro de los números A, B puede ser cero, pero no ambos a la vez. El espacio de tales razones es la *esfera de Riemann* descrita con detalle en §8.3. Tenemos que identificar así el espacio base $\mathcal{M}$ de nuestro fibrado como esta esfera de

Riemann S^2. Por consiguiente, podemos ver que S^3 puede considerarse como un S^1-fibrado sobre S^2. (No debemos esperar una relación semejante para otras dimensiones, donde fibrado, espacio base y fibra son esferas.) Sin embargo, resulta que S^7 puede considerarse como un S^3-fibrado sobre S^4, como puede obtenerse (con cuidado) reemplazando los números complejos w y z en el argumento anterior por cuaterniones;[15.4] también S^{15} puede considerarse como un S^7-fibrado sobre S^8, donde w y z son ahora reemplazados por octoniones (véanse §11.2 y §16.2; pero esto no funciona para ninguna otra esfera de dimensiones más altas.[7]

Esta familia de círculos en S^3, denominados *paralelos de Clifford*, es particularmente interesante. Los círculos, que son círculos máximos, se retuercen uno alrededor del otro, permaneciendo siempre a la misma distancia (por ello se conocen como «paralelos»). Dos cualesquiera de los círculos están enlazados, de modo que son *oblicuos* (no coesféricos). En el 3-espacio euclídeo, las líneas rectas que son oblicuas (no coplanares) tienen la propiedad de que se alejan unas de otras a medida que se prolongan hacia el infinito. Sin embargo, la 3-esfera tiene curvatura positiva, de modo que los círculos de Clifford, que son geodésicas en S^3, tienen una tendencia compensadora a curvarse unos hacia otros de acuerdo con el efecto de desviación geodésica considerado en §14.5 (véase la Fig. 14.12). Estos dos efectos se cancelan exactamente en el caso de los paralelos de Clifford; véase la Fig. 15.9. Para obtener una imagen de la familia de paralelos de Clifford podemos proyectar S^3 estereográficamente desde su «polo sur» en un 3-espacio euclídeo ecuatorial, en exacta analogía con la correspondiente proyección estereográfica de S^2 sobre el espacio euclídeo que adoptamos en nuestro estudio de la esfera de Riemann en §8.3 (véase la Fig. 8.7). Como sucede con la proyección estereográfica de S^2, los círculos en S^3 se aplican en círculos en el 3-espacio euclídeo bajo esta proyección. Véase la Fig. 33.15 para una imagen de la familia de círculos de Clifford proyectados. Esta configuración tuvo una importancia seminal para la teoría de twistores,[8] y la geometría relevante se describirá en §33.6.

Antes he afirmado que este fibrado particular de Clifford no po-

[15.4] Desarrolle este argumento. ¿Puede ver como tratar el caso S^{15}?



(a) (b)

Fig. 15.9. (a) En un 3-espacio euclídeo las líneas rectas oblicuas se alejan cada vez más una de otra. (b) En S^3 la curvatura positiva proporciona una tendencia compensatoria a curvar las geodésicas (círculos máximos) unas hacia otras (por desviación geodésica; véase la Fig. 14.12). En el caso de los paralelos de Clifford, la compensación es exacta.

seería ninguna sección transversal. ¿Cómo debemos entender esto? En primer lugar habría que señalar que el «giro» en el fibrado de Clifford debe su existencia al hecho de que las fibras-círculo poseen una simetría exacta dada por las rotaciones del círculo (el grupo O(2) o, de forma equivalente, U(1); véase el ejercicio [13.59]). No podemos identificar cada una de estas fibras con un círculo dado específicamente, tal como el círculo unidad en el plano complejo $\mathbb{C}$. Si pudiéramos hacerlo, entonces podríamos escoger de manera consistente algún punto específico en el círculo (por ejemplo, el punto «1» en el círculo unidad en $\mathbb{C}$) y con ello obtener una sección transversal del fibrado de Clifford. La no existencia de secciones transversales es posible porque los círculos de Clifford solo están modelados sobre el círculo unidad en $\mathbb{C}$, y no identificados con él.

Por supuesto, esto por sí mismo no nos dice por qué el fibrado de Clifford no tiene secciones transversales continuas. Para entender esto, será útil mirar el fibrado de Clifford de otra manera. De hecho, resulta que cada punto de nuestra esfera S^3 puede interpretarse como un vector «espinorial» de longitud unidad tangente a S^2 en uno de sus puntos.[15.5] Recordemos de §11.3 que un objeto espinorial es una cantidad que cuando experimenta una rotación completa de 2π se convierte en el negativo de lo que era originalmente. De acuerdo con el enunciado anterior, una sección transversal de nuestro fibrado $\mathcal{B}(= S^3)$

[15.5] Demuéstrelo. *Sugerencia*: Tome como vector tangente $u\partial/\partial v - v\partial/\partial u + x\partial/\partial y - y\partial/\partial x$.

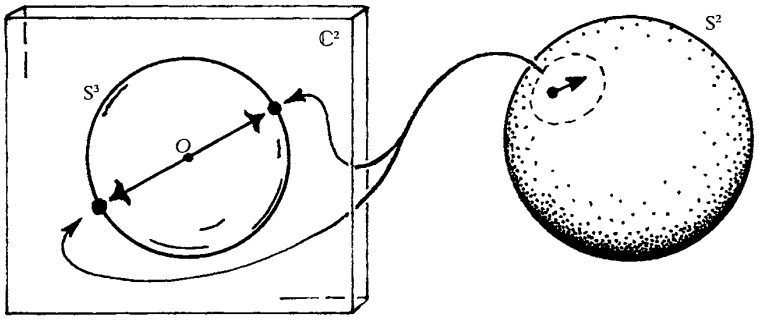


Fig. 15.10. El fibrado $\mathcal{B}'$ de vectores unidad tangentes a S^2 es una ligera modificación del fibrado de Clifford, en el que se identifican los puntos antipodales de S^3. Sin esta identificación, obtenemos S^3 como el fibrado (de Clifford) $\mathcal{B}$ de vectores *espinoriales* tangentes a S^2. Las fibras de $\mathcal{B}'$ siguen siendo círculos, pero cada fibra-círculo de $\mathcal{B}$ se enrolla dos veces alrededor de cada fibra-círculo de $\mathcal{B}'$.

representaría un campo continuo de tales vectores unidad espinoriales en $\mathcal{M}(= S^2)$. Ahora bien, es un hecho topológico bien conocido que no hay ningún campo continuo de vectores tangentes unidad ordinarios en S^2. (¡Ese es el problema de peinar el pelo de un «perro esférico»! Es imposible que los pelos queden lisos de una forma continua en toda la esfera.) Hacer «espinoriales» estas direcciones no ayuda, de modo que tampoco puede existir un campo continuo de vectores tangentes espinoriales. De ahí que nuestro fibrado $\mathcal{B}(= S^3$ no tenga secciones transversales.

Esto merece alguna discusión adicional, pues hay mucho más que aprender de este ejemplo. En primer lugar, podemos obtener el fibrado real $\mathcal{B}'$ de vectores unidad tangentes a S^2 modificando ligeramente el fibrado de Clifford antes descrito. Puesto que cualquier vector tangente unidad tiene solo dos manifestaciones como un objeto espinorial (siendo una la «negativa» de la otra), debemos identificarlas si queremos pasar del vector espinorial al vector ordinario. Lo que esto significa, en términos del fibrado de Clifford $\mathcal{B}(= S^3)$, es que dos puntos de S^3 deben ser identificados para dar un único punto[9] del haz $\mathcal{B}'$ de vectores unitarios en S^3. Los pares de puntos de S^3 que deben identificarse son los *puntos antipodales* en la 3-esfera.Véase la Fig. 15.10. Las

fibras de $\mathcal{B}'$ siguen siendo círculos. Lo que sucede es simplemente que cada fibra-círculo de $\mathcal{B}(= S^3)$ «se enrolla dos veces» alrededor de cada fibra-círculo de $\mathcal{B}'$. Cada punto de $\mathcal{B}'$ representa ahora un punto de S^2 con un vector tangente unidad en dicho punto. De hecho, el espacio $\mathcal{B}'$ es topológicamente idéntico al espacio $\mathcal{R}$ que encontramos en §12.1, y que representa las diferentes orientaciones espaciales de un objeto (tal como el libro considerado en §11.3) en el 3-espacio euclídeo. Esto se hace evidente si pensamos que nuestro «objeto» es la esfera S^2 con una flecha (vector tangente unidad) marcada en uno de sus puntos. Esta flecha marcada fijará por completo la orientación espacial de la esfera.

15.5. Fibrados vectoriales complejos, fibrados (co)tangentes

Una ligera ampliación de la idea que hay tras el fibrado de Clifford (y también de $\mathcal{B}'$) nos ofrece un buen ejemplo de un *fibrado vectorial complejo*, en este caso un fibrado al que llamaré $\mathcal{B}^{\mathbb{C}}$ (o correspondientemente $\mathcal{B}'^{\mathbb{C}}$). Cada una de las rectas $Aw + Bz = 0$ es en sí misma un espacio vectorial complejo 1-dimensional. (La recta entera consiste en la familia de múltiplos de un único vector (w, z) por números complejos λ, de modo que (w, z) se multiplica para dar $(\lambda w, \lambda z)$.) Consideremos ahora este 1-espacio vectorial complejo como nuestra fibra $\mathcal{V}$. La esfera de Riemann S^2 es nuestro espacio base $\mathcal{M}$, igual que antes.

No obstante, debemos hacer algo más para obtener el fibrado vectorial complejo correcto $\mathcal{B}^{\mathbb{C}}$. En $\mathbb{C}^2$ las diferentes fibras no son disjuntas, pues todas ellas tienen en común el origen $(0, 0)$. Así pues, para obtener $\mathcal{B}^{\mathbb{C}}$ debemos modificar $\mathbb{C}^2$ reemplazando el origen por una copia de la esfera de Riemann entera ($\mathbb{CP}^1$; véase §15.6), de modo que en lugar de tener solo un cero, tenemos todo un equivalente de ceros por la esfera de Riemann, uno por cada fibra, lo que da lugar a la *sección cero* del fibrado (véase la Fig. 15.11). Este procedimiento se conoce como *ampliar* el origen de $\mathbb{C}^2$ (una idea importante para la geometría algebraica, la teoría de variedades complejas, la teoría de cuerdas, la teoría de twistores y muchas otras áreas). Puesto que ahora admitimos el cero en las fibras, existen secciones transversales continuas de $\mathcal{B}$. Resulta que

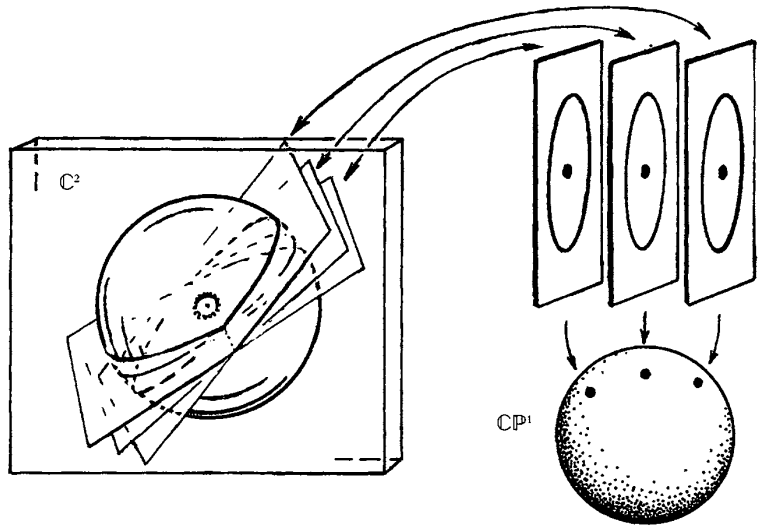


Fig. 15.11. Tomando la línea entera $Aw + Bz = 0$ (un plano complejo), en lugar de solo su círculo unidad, obtenemos un ejemplo de un fibrado lineal complejo $\mathcal{B}^{\mathbb{C}}$, siendo ahora la fibra $\mathcal{V}$ un espacio vectorial 1-dimensional complejo. La esfera de Riemann $S^2 = \mathbb{CP}^1$ (también una variedad compleja; véanse §8.3 y §15.6) sigue siendo el espacio base $\mathcal{M}$. Pero para hacer disjuntas las diferentes fibras, debemos «ampliar» el origen (0, 0) reemplazándolo por una esfera de Riemann entera, lo que nos da el valor en 0s de una esfera de Riemann.

estas secciones transversales representan los *campos de espinores* en S^2. Un «espinor» en un punto de S^2 va a representarse no solo como un «vector tangente unidad espinorial» en un punto de S^2, sino que ahora el vector puede ser «ampliado o reducido de escala» por un número real positivo, o puede admitirse que se haga cero. Puede demostrarse que tales «espinores» posibles en un punto de S^2 nos ofrecen un espacio vectorial 2-complejo dimensional.[10,[15.6]]

El fibrado entero $\mathcal{B}^{\mathbb{C}}$ es una estructura compleja (i.e., holomorfa) —de hecho, se denomina un fibrado *línea* complejo, porque las fibras son líneas 1-complejo dimensionales. Es un objeto holomorfo por-

[15.6] ¿Por qué cada uno de tales campos espinoriales toma el valor cero al menos en un punto de S^2?

que su construcción viene dada enteramente en términos de nociones holomorfas.[15.7] En particular, el espacio base es una curva compleja —la esfera de Riemann (véase §8.3)— y las fibras son espacios vectoriales complejos 1-dimensionales. Por consiguiente, existe también otra noción de sección transversal que tiene relevancia aquí, a saber, la de una sección transversal *holomorfa*. Una sección transversal holomorfa es una sección transversal de un fibrado complejo que es ella misma una subvariedad compleja del fibrado (lo que significa que está dada localmente por ecuaciones holomorfas). A veces, en el caso de un fibrado línea complejo, una sección transversal semejante se conoce como función holomorfa *retorcida* en el espacio base. Tales objetos tienen considerable importancia en muchas áreas de las matemáticas puras y de la física matemática.[11] También desempeñan un papel especial en la teoría de twistores (véase §33.8). Las secciones holomorfas constituyen una familia firmemente controlada pero importante. En el caso de $\mathcal{B}^{\mathbb{C}}$, resulta que *no* hay secciones holomorfas (globales) distintas de la sección *cero* (i.e., cero en todas partes).

Con una modificación menor de esta construcción (correspondiente al paso de $\mathcal{B}$ a $\mathcal{B}'$), obtenemos campos vectoriales, en lugar de campos espinoriales, en S^2. El fibrado apropiado $\mathcal{B}'^{\mathbb{C}}$ puede interpretarse de nuevo como un fibrado vectorial complejo; de hecho, es lo que se denomina el *cuadrado* del fibrado vectorial $\mathcal{B}^{\mathbb{C}}$. Está construido exactamente de la misma forma que $\mathcal{B}'^{\mathbb{C}}$, excepto que ahora *identificamos* cada punto (w, z) con su punto «antipodal» $(-w, -z)$, estando ahora dada la multiplicación de (w, z) por el número complejo λ por $(\lambda^{1/2}w, \lambda^{1/2}z)$ (en lugar de $(\lambda w, \lambda z)$).

Para terminar esta sección, debería señalar que el fibrado $\mathcal{B}'^{\mathbb{C}}$ puede reinterpretarse vagamente en términos reales como lo que se denomina el *fibrado tangente* $T(S^2)$ de S^2. El fibrado tangente $T(\mathcal{M})$ de una variedad general $\mathcal{M}$ es el espacio cada uno de cuyos puntos representa un punto de $\mathcal{M}$ junto con un vector tangente a $\mathcal{M}$ en dicho punto. Véase la Fig. 15.12a.[15.8] Una sección transversal de $T(\mathcal{M})$ representa

[15.7] Explique esto en detalle.

[15.8] Demuestre que $\mathcal{B}'^{\mathbb{C}}$, interpretado como un fibrado real sobre S^2, es en realidad el mismo que $T(S^2)$. *Sugerencia*: Reexamine el ejercicio [15.5].

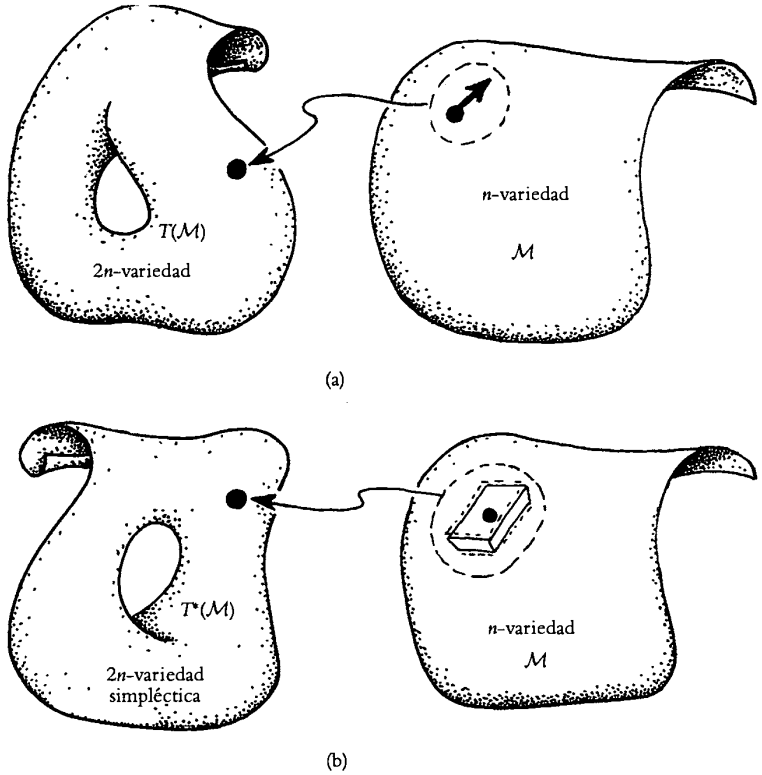


Fig. 15.12. (a) Para una variedad general $\mathcal{M}$, cada punto de su fibrado tangente $T(\mathcal{M})$ representa un punto de $\mathcal{M}$ junto con un vector tangente a $\mathcal{M}$ allí. Una sección transversal de $T(\mathcal{M})$ representa un campo vectorial en $\mathcal{M}$. (b) El fibrado cotangente $T^*(\mathcal{M})$ es similar, pero con covectores en lugar de vectores. Los fibrados cotangentes son siempre variedades simplécticas.

un campo vectorial en $\mathcal{M}$. Una noción quizá de importancia física aún mayor es la de *fibrado cotangente* $\mathrm{T}^*(\mathcal{M})$ de una variedad $\mathcal{M}$, cada uno de cuyos puntos representa un punto $\mathcal{M}$, junto con un covector en dicho punto (Fig. 15.12b). En el capítulo 20 vislumbraremos algo de la importancia de estas ideas. Las secciones transversales de $\mathrm{T}^*(\mathcal{M})$ representan campos de covectores en $\mathcal{M}$. Resulta que los fibrados cotangentes son siempre variedades simplécticas (véanse §14.8 y §§20.2,4),

un hecho de considerable importancia para la mecánica clásica. También podemos definir correspondientemente varios tipos de fibrados tensoriales. Un campo tensorial puede interpretarse como una sección transversal de uno de estos fibrados.

15.6. Espacios proyectivos

Otra noción importante relacionada con un espacio vectorial general es la de un *espacio proyectivo*. El propio espacio vectorial es «casi» un fibrado sobre el espacio proyectivo. Si eliminamos el origen del espacio vectorial, entonces obtenemos un fibrado sobre el espacio proyectivo, siendo la fibra una recta de la que se ha eliminado el origen; alternativamente, como sucede con el ejemplo particular de $\mathcal{B}^{\mathbb{C}}$ que se ha dado en §15.5, podemos «ampliar» el origen del espacio vectorial. (Volveré sobre este tema más adelante.) Los espacios proyectivos tienen una considerable importancia en matemáticas y desempeñan un papel importante en la geometría de la mecánica cuántica (véanse §21.9 y §22.9), así como en la teoría de twistores (§33.5). Por lo tanto, es conveniente que haga un breve comentario sobre dichos espacios.

La idea de un espacio proyectivo parece tener su origen en el estudio de la perspectiva en el dibujo y la pintura, dentro del contexto de la geometría euclídea. Recordemos que en el plano euclídeo dos rectas distintas se cortan siempre, a menos que sean paralelas. Sin embargo, si en una hoja de papel vertical dibujamos una imagen de un par de rectas paralelas que se alejan en la distancia en un plano horizontal (por ejemplo, los bordes de una carretera recta), observamos que en el dibujo los bordes parecen cortarse en un «punto de fuga» en el horizonte (véase la Fig. 15.13). La geometría proyectiva toma en serio estos puntos de fuga, añadiendo al plano euclídeo «puntos en el infinito» que hacen posible que las rectas paralelas se corten en tales puntos adicionales.

Existen muchos teoremas sobre rectas en el 3-espacio euclídeo ordinario que son difíciles de enunciar debido a que tienen que hacerse excepciones para rectas paralelas. En la Fig. 15.14 muestro dos ejemplos notables, a saber, los teoremas de Pappos[12] (descubierto a finales del siglo III a.C.) y de Desargues (hallado en 1636). En cada caso, el

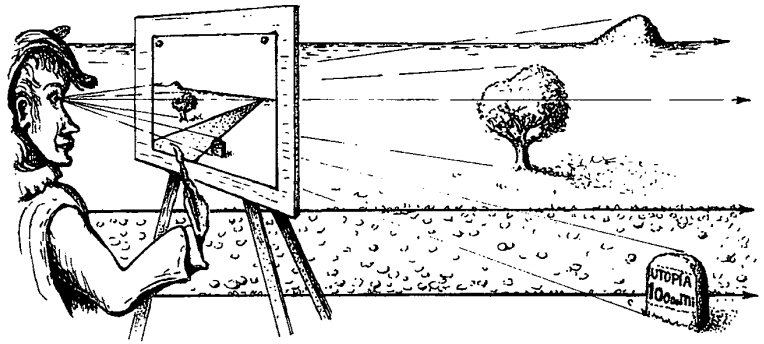
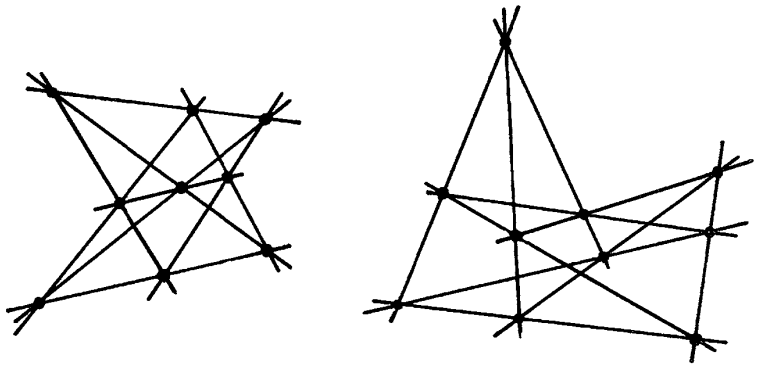


Fig. 15.13. La geometría proyectiva añade «puntos en el infinito» al plano euclídeo permitiendo que en ellos se corten las líneas paralelas. En la imagen del artista, pintada sobre un lienzo vertical, un par de líneas paralelas horizontales que se alejan en la distancia —los bordes de una carretera horizontal recta— parecen cortarse en un «punto de fuga» en el horizonte.

teorema (que estoy enunciando de forma «inversa») afirma que si todas las líneas rectas mostradas en el diagrama (9 líneas en el caso de Pappos y 10 en el de Desargues) se cruzan de tres en tres en todos salvo uno de los puntos marcados con círculos negros (9 círculos negros en total en el caso de Pappos y 10 en total en el caso de Desargues), entonces el trío de rectas que se muestran cortándose en el círculo negro restante tienen, en efecto, un punto en común. Sin embargo, enunciados de este modo, estos teoremas son ciertos solo si consideramos que un triplete de líneas mutuamente *paralelas* cuenta como si tuviera un punto en común, a saber, un «punto en el infinito». Con esta interpretación, los teoremas continúan siendo ciertos cuando las rectas son paralelas. También siguen siendo ciertos incluso si una de las rectas yace por ejemplo en el infinito. Así pues, los teoremas de Pappos y Desargues son más propiamente teoremas en geometría proyectiva que en geometría euclídea.

¿Cómo construimos un espacio proyectivo n-dimensional $\mathbb{P}^n$? La forma más inmediata consiste en tomar un espacio vectorial $(n+1)$-dimensional $\mathbf{V}^{n+1}$, y considerar nuestro espacio $\mathbb{P}^n$ como el espacio de subespacios vectoriales 1-dimensionales de $\mathbf{V}^{n+1}$. (Estos subespacios vectoriales 1-dimensionales son las rectas que pasan por el origen de $\mathbf{V}^{n+1}$.)



(a) (b)

Fig. 15. 14. Configuraciones de dos famosos teoremas de la geometría proyectiva plana: (a) el de Pappos, con nueve líneas y nueve puntos marcados, y (b) el de Desargues, con diez líneas y diez puntos marcados. En ambos casos, la afirmación es que si todos salvo uno de los puntos marcados es la intersección de un trío de líneas, entonces el restante punto marcado también lo es.

Una línea recta en $\mathbb{P}^n$ (que es en sí misma un ejemplo de un $\mathbb{P}^1$) está dada por un subespacio 2-dimensional de $\mathbf{V}^{n+1}$ (un plano que pasa por el origen), y los puntos colineales de $\mathbb{P}^n$ aparecen como líneas que yacen en dicho plano (Fig. 15.15). Existen también subespacios planos de $\mathbb{P}^n$ de dimensiones más altas, y los espacios proyectivos $\mathbb{P}^r$ están contenidos en $\mathbb{P}^n$ $(r < n)$. Cada $\mathbb{P}^r$ corresponde a un subespacio vectorial $(r+1)$-dimensional de $\mathbf{V}^{n+1}$.

Esta construcción (en el caso $n = 2$) formaliza los procedimientos de la perspectiva en la representación pictórica; en efecto, podemos considerar que el ojo del artista está situado en el origen O del espacio vectorial $\mathbf{V}^3$, que representa al espacio 3-euclídeo que rodea al artista. Un rayo de luz que pasa por O (el ojo del artista) es visto por el artista como un único punto. Así, el «campo de visión» del artista, tomado como la totalidad de tales rayos de luz, puede considerarse como un plano proyectivo $\mathbb{P}^2$ (véase de nuevo la Fig. 15.15). Cualquier línea recta (que no pase por O) en el espacio que percibe el artista corresponde al plano que une dicha línea con O, de acuerdo con la definición de una «línea recta» en $\mathbb{P}^2$ que se ha dado antes.

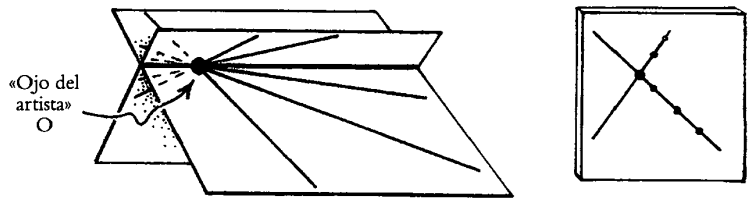


Fig. 15.15. Para construir un espacio proyectivo n-dimensional $\mathbb{P}^n$, tomemos un espacio vectorial $(n+1)$-dimensional $\mathbf{V}^{n+1}$, y consideremos $\mathbb{P}^n$ como el espacio de los subespacios vectoriales 1-dimensionales de $\mathbf{V}^{n+1}$ (líneas que pasan por el origen de $\mathbf{V}^{n+1}$). Una línea recta en $\mathbb{P}^n$ está dada por un subespacio 2-dimensional de $\mathbf{V}^{n+1}$ (plano a través del origen), y los puntos colineales de $\mathbb{P}^n$ surgen como líneas que pasan por O en un plano semejante. Esto se aplica tanto al caso real ($\mathbb{RP}^n$) como al caso complejo ($\mathbb{CP}^n$). La geometría de $\mathbb{RP}^2$ formaliza los procedimientos de perspectiva en representación pictórica: consideremos que el ojo del artista está en el origen O de $\mathbf{V}^3$, tomando $\mathbf{V}^3$ como el 3-espacio euclídeo ambiente del artista. Un rayo de luz a través de O es visto por el artista como un solo punto. Lo que el artista representa como una «línea recta» ($\mathbb{RP}^1$ en $\mathbb{RP}^2$) (en cualquier elección particular del lienzo del artista) corresponde en realidad al plano ($\mathbf{V}^2$)que une dicha línea con O. Los pares de planos que pasan por O se cortan siempre, incluso cuando unen líneas paralelas en $\mathbf{V}^3$ a O. (Por ejemplo, las dos líneas frontera inferiores en la imagen de la izquierda representan el papel de los bordes de la carretera de la Fig. 15.13).

Imaginemos que el artista pinta una imagen aproximada de la escena que percibe en un lienzo que coincide con un plano concreto (que no pasa por O). Cualquier plano semejante recogerá solo una parte del $\mathbb{P}^2$ entero. Ciertamente no cortará aquellos rayos de luz que sean paralelos al mismo. Pero varios de estos planos proporcionarán un «mosaico» adecuado que cubre la totalidad de $\mathbb{P}^2$ (tres planos bastarán).[13,[15.9]] Líneas paralelas en dicho plano se representarán en otro plano como líneas con un *punto de fuga* común.

Podemos considerar o bien espacios proyectivos reales, $\mathbb{RP}^n$, o bien complejos, $\mathbb{CP}^n$. Ya hemos considerado un ejemplo de un espacio proyectivo complejo, a saber, la *esfera de Riemann*, que es $\mathbb{CP}^1$. Recordemos

[15.9] Explique cómo hacerlo. *Sugerencia*: Piense en coordenadas cartesianas (x, y, z). Tome dos en un instante, con el lienzo dado por la tercera establecida a la unidad.

que la esfera de Riemann aparece como el espacio de razones de pares de números complejos (w, z), no ambos nulos, que es el espacio de las rectas complejas que pasan por el origen en $\mathbb{C}^2$ (véase la Fig. 15.8). Con más generalidad, a cualquier espacio proyectivo se le pueden asignar lo que se denominan coordenadas *homogéneas*. Estas son las coordenadas $z^0, z^1, z^2, \ldots, z^n$ para el espacio vectorial $(n+1)$-dimensional $\mathbf{V}^{n+1}$, a partir del cual surge $\mathbb{P}^n$, pero las «coordenadas homogéneas» para $\mathbb{P}^n$ son las n *razones* independientes

$$z^0 : z^1 : z^2 : \ldots : z^n$$

(donde al menos una de las zs es distinta de cero), en lugar de los valores individuales de las propias zs.[15.10] Si todas las z^r son reales, entonces estas coordenadas describen $\mathbb{RP}^n$, y el espacio $\mathbf{V}^{n+1}$ puede identificarse con $\mathbb{R}^{n+1}$ (espacio de $n+1$ números reales; véase §12.2). Si todas son complejas, entonces describen $\mathbb{CP}^n$, y el espacio $\mathbf{V}^{n+1}$ puede identificarse con $\mathbb{C}^{n+1}$ (espacio de $n+1$ números complejos; véase §12.9).

Puesto que excluimos el punto $O = (0, 0, \ldots, 0)$ de las coordenadas homogéneas admisibles, el origen de $\mathbb{R}^{n+1}$ o $\mathbb{C}^{n+1}$ se omite[14] (para dar $\mathbb{R}^{n+1} - O$ o $\mathbb{C}^{n+1} - O$) cuando lo consideramos como un fibrado sobre $\mathbb{RP}^n$ o sobre $\mathbb{CP}^n$, respectivamente. Por consiguiente, también debe eliminarse el origen de la fibra. En el caso real, esto divide la fibra en dos piezas (pero eso no significa que el fibrado se divida en dos piezas; de hecho, $\mathbb{R}^{n+1} - O$ es conexo, cuando $n > 0$).[15.11] En el caso complejo, la fibra es $\mathbb{C} - O$ (a menudo escrita $\mathbb{C}^*$), que es conexa. En cualquiera de los casos, podemos preferir reinstaurar el origen en la fibra, de modo que obtenemos un fibrado vectorial. Pero si lo hacemos, esto equivale a algo más que colocar de nuevo el origen en $\mathbb{R}^{n+1}$ o $\mathbb{C}^{n+1}$. Como sucede con el caso particular de $\mathbb{C}^2$, que se ha comentado antes, debemos volver a colocar el origen en cada fibra por separado, de modo que el origen es «ampliado». El espacio fibrado se convierte en $\mathbb{R}^{n+1}$ con un $\mathbb{RP}^n$ insertado en lugar de O, o $\mathbb{C}^{n+1}$ con un $\mathbb{CP}^n$ en lugar de O.

[15.10] Explique por qué hay n razones independientes. Encuentre $n+1$ conjuntos de n coordenadas ordinarias (construidas a partir de las zs) para $n+1$ cartas coordenadas que, juntas, cubran $\mathbb{P}^n$.

[15.11] Explique esta geometría demostrando que el fibrado $\mathbb{R}^{n+1} - O$ sobre $\mathbb{RP}^n$ puede entenderse como la composición del fibrado $\mathbb{R}^{n+1} - O$ sobre S^n (siendo la fibra, $\mathbb{R}^+$, los reales positivos) y de S^2 como una cubierta doble de $\mathbb{RP}^n$.

En el caso complejo, también podemos considerar la $(2n + 1)$-esfera unidad S^{2n+1} en $\mathbb{C}^{n+1}$, exactamente igual que se ha hecho en el caso particular $n = 1$ al construir el fibrado de Clifford. Cada fibra corta S^{2n+1} en un círculo S^1, de modo que ahora obtenemos S^{2n+1} como un S^1-fibrado sobre $\mathbb{CP}^n$. Esta estructura subyace en la geometría de la mecánica cuántica —aunque no es muy frecuente que este bello hecho geométrico repercuta en el pensamiento de los físicos cuánticos—, donde veremos que el espacio de los estados cuánticos físicamente distintos, para un sistema de $(n + 1)$-estados, es un $\mathbb{CP}^n$. Además, existe una cantidad conocida como *fase*, que normalmente se considera que es un número complejo de módulo unidad ($e^{i\theta}$, con θ real; véase §5.3), aunque en realidad es un número complejo *retorcido* de módulo unidad.[15] Volveremos a estas cuestiones al final de este capítulo, y cuando profundicemos en serio en la mecánica cuántica en los capítulos 21 y 22 (véanse §21.9 y §22.9).

15.7. No trivialidad en una conexión fibrada

¡Tan solo he llevado al lector en un viaje relámpago por algunos conceptos importantes acerca de fibrados o relacionados con ellos! Parte de la geometría y la topología implicadas es bastante intrincada, de modo que el lector no debería asustarse si todo esto parece un poco desconcertante. Volvamos ahora a algo mucho más sencillo, en el sentido de que no necesitamos muchas dimensiones (¡al menos al principio!) para explorar la idea. Aunque mi siguiente ejemplo de un fibrado es realmente muy simple, expresa una sutileza importante implícita en la noción de fibrado que no hemos visto antes. En todos los fibrados que se han considerado hasta ahora, la no trivialidad del fibrado se revelaba en un aspecto topológico de la geometría, al tener el «giro» un carácter topológico. Sin embargo, es perfectamente posible que un fibrado sea no trivial en un sentido importante, a pesar de ser topológicamente trivial.

Volvamos a nuestro ejemplo original, donde el espacio base $\mathcal{M}$ es *un círculo* ordinario S^1 y la fibra $\mathcal{V}$ es un espacio vectorial real 1-dimensional. Construiremos ahora nuestro fibrado $\mathcal{B}$ de una forma algo

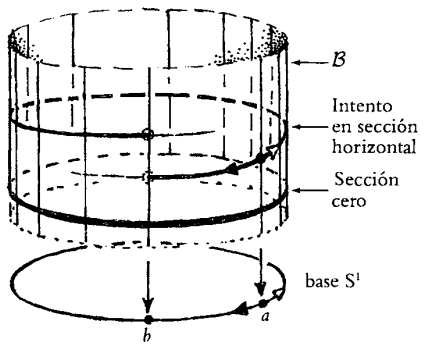


Fig. 15.16. Un fibrado de línea $\mathcal{B}$ «tensado» sobre $\mathcal{M}$, utilizando una simetría diferente de la fibra $\mathcal{V}$ de la de las Figs. 15.4, 15.5 y 15.7 (donde $\mathcal{V}$ sigue siendo un espacio real 1-dimensional $\mathbf{V}^1$), a saber, una dilatación en un factor positivo (aquí 2). La topología es precisamente la del cilindro $S^1 \times \mathcal{R}$, pero hay una «tensión» que puede reconocerse en términos de una *conexión* en $\mathcal{B}$. Esta conexión define una noción local de «horizontal», para curvas en $\mathcal{B}$. Pero consideremos dos caminos de a a b en la base, el camino directo (flecha negra) y el indirecto (flecha blanca). Cuando llegamos a b encontramos una discrepancia (en un factor de 2), que indica que la noción de «horizontal» aquí es dependiente del camino.

diferente al simple «darle la vuelta» a la fibra $\mathcal{V}$, cuando circunnavegamos $\mathcal{M}$, que nos ha dado el fibrado de Moebius. En lugar de ello, *estirémoslo* en un factor 2. Esto se muestra en la Fig. 15.16. Esto explota una simetría de un espacio vectorial real 1-dimensional diferente de la simetría «salto» $\boldsymbol{v} \mapsto -\boldsymbol{v}$ utilizada en el fibrado de Moebius. La transformación «estiramiento» $\boldsymbol{v} \mapsto 2\boldsymbol{v}$ conserva también la estructura de espacio vectorial de $\mathcal{V}$. Ahora la topología del fibrado no es la cuestión. Topológicamente, tenemos solo un cilindro $S^1 \times \mathbb{R}$, igual que en nuestro primer ejemplo de la Fig. 15.4a, pero ahora hay un tipo diferente de «tensión» en el fibrado, que podemos reconocer en términos de un tipo apropiado de *conexión* en el mismo.

Nuestro tipo previo de conexión, expuesto en el capítulo 14, estaba relacionado con una noción de «paralelismo» para vectores tangentes transportados a lo largo de curvas en la variedad $\mathcal{M}$. La forma de ver esto en el presente contexto es pensar en términos del fibrado tangente $T(\mathcal{M})$ de $\mathcal{M}$. Puesto que un punto de $T(\mathcal{M})$ representa un vector $\boldsymbol{v}$

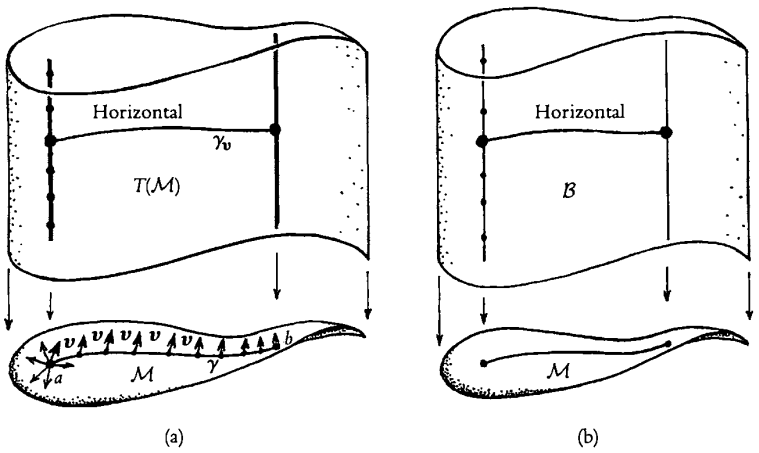


Fig. 15.17. Tipos de conexión en una variedad general $\mathcal{M}$ comparados. (a) La noción original (§14.3), que define una noción de «paralelo» para vectores tangentes transportados a lo largo de curvas en $\mathcal{M}$, se describe en términos del fibrado tangente $T(\mathcal{M})$ de $\mathcal{M}$ (Fig. 15.12a). Un vector tangente particular $\boldsymbol{v}$ en un punto a de $\mathcal{M}$ se representa en $T(\mathcal{M})$ por un punto particular de la fibra sobre a. Una curva «horizontal» γ_v en $T(\mathcal{M})$ desde este punto representa el transporte paralelo de $\boldsymbol{v}$ a lo largo de una curva γ en $\mathcal{M}$. (b) La misma idea se aplica a un fibrado $\mathcal{B}$ sobre $\mathcal{M}$, distinto de $T(\mathcal{M})$, donde el «transporte constante» en $\mathcal{M}$ se define a partir de una noción de «horizontal» en $\mathcal{B}$.

tangente a $\mathcal{M}$ en un punto a de $\mathcal{M}$, el transporte de $\boldsymbol{v}$ a lo largo de una curva γ en $\mathcal{M}$ estará representado por una curva γ_v en $T(\mathcal{M})$. Véase la Fig. 15.17a. Tener una idea de lo que significa «paralelo» para el transporte de v es equivalente a tener una noción de «horizontal» para la curva γ_v en el fibrado (puesto que mantener γ_v «horizontal» en el fibrado equivale a mantener $\boldsymbol{v}$ «constante» a lo largo de γ en la base). Véase la Fig. 15.19b. Aquí la idea consiste en generalizar esta noción de modo que se aplique a otros fibrados que no sean el fibrado tangente; véase la Fig. 15.17b. Ya hemos visto en el capítulo 14 los inicios de dicha generalización, puesto que hemos ampliado la noción de conexión, de modo que se aplica a entidades distintas de vectores tangentes, a saber, a covectores y a $[^p_q]$-tensores en general. Sin embargo, como se ha señalado en §15.1, este es un tipo muy limitado de generalización, porque la extensión de la conexión desde vectores a estas entidades de otro tipo está

unívocamente prescrita, sin que quede ninguna libertad adicional (esencialmente porque el fibrado cotangente y los fibrados tensoriales están completamente determinados por el fibrado tangente). Para un fibrado general sobre $\mathcal{M}$, no tiene por qué haber asociación con el fibrado tangente, de modo que la forma en que actúa la conexión sobre dicho fibrado puede ser especificada independientemente de la forma en que actúa sobre vectores tangentes. Para un fibrado sobre $\mathcal{M}$ que está desasociado de $T(\mathcal{M})$, no resulta tan apropiado hablar en términos de un «paralelismo», porque la noción (local) de «paralela» es algo que se refiere a *direcciones*, lo que básicamente significa direcciones de vectores tangentes. En consecuencia, es más habitual referirse a una «constancia» local para la cantidad que está descrita por el fibrado más que al «paralelismo» que se refiere a los vectores tangentes descritos por $T(\mathcal{M})$. Semejante noción local de «constancia» —i.e., de «horizontalidad» en el fibrado— proporciona la estructura conocida como una *conexión fibrada*.

Volvamos ahora a nuestro fibrado «tensado» $\mathcal{B}$, sobre el círculo S^1, tal como se representa en la Fig. 15.16. Consideremos una parte de $\mathcal{B}$ que es «trivial» en el sentido de que está situada por encima de cierta región «topológicamente trivial» de S^1; supongamos que es la parte $\mathcal{B}_p$, que está sobre el segmento conexo $S^1 - p$ (como en la Fig. 15.5), donde p es un punto de S^1. Consideraremos $\mathcal{B}_p$ como el espacio *producto* $(S^1 - p) \times \mathbb{R}$, y nuestra conexión fibrada tiene que proporcionar la noción de *constancia* de una sección transversal que pueda tomarse como constancia en el sentido ordinario de una función con valores reales en $S^1 - p$. Así, en la Fig. 15.18 encontramos las secciones constantes representadas como líneas horizontales reales en $\mathcal{B}_p$. Lo mismo se aplica a una segunda región $\mathcal{B}_q$, con $q \neq p$, donde el fibrado entero está compuesto pegando estas dos cartas. Sin embargo, en el pegado hay un estiramiento relativo en un factor 2 entre la región derecha del pegado y la región izquierda (donde la región derecha se muestra incluyendo un estiramiento en un factor 2). Así pues, una sección (no nula) que permanece localmente horizontal diferirá en un factor 2 cuando se recorra por completo el espacio S^1 (véase la Fig. 15.16). En consecuencia, el fibrado $\mathcal{B}$ no tiene secciones transversales (aparte de la sección cero) que sean localmente horizontales según nuestra conexión fibrada específica.

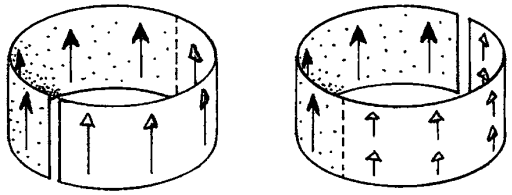


Fig. 15.18. Consideremos una parte $\mathcal{B}_p$ de $\mathcal{B}$ (de la Fig. 15.16) que permanece sobre una región «trivial» $S^1 - p$ de S^1, y análogamente para $\mathcal{B}_p$, igual que en la Fig. 15.5a. Tengamos en cuenta que «horizontal» en cada camino significa horizontal en el sentido ordinario. Sin embargo, al pegar hay una dilatación relativa en un factor 2 entre una región de pegado y la otra (ilustrada a la derecha). Esto proporciona la conexión que se ilustra en la Fig. 15.16.

Podemos considerar esta situación de un modo ligeramente diferente. Imaginemos una curva γ en el espacio base S^1 que empieza en un punto a y termina en b, e imaginemos el «transporte constante», de una función fibrovaluada en S^1, desde a a b. Es decir, buscamos una curva en $\mathcal{B}$ que es localmente una sección transversal horizontal por encima de esta curva. Véase la Fig. 15.16. Ahora bien, hay más de una curva de a a b en el espacio base; si seguimos una, obtenemos una respuesta para el valor final en b diferente de la respuesta que obtenemos cuando seguimos otra. La noción de transporte constante que hemos definido es *dependiente del camino*.

Esto no es exactamente lo mismo que la dependencia del camino que encontramos para nuestra conexión en la fibra tangente ∇ que hemos visto en el capítulo 13. En efecto, en ese caso había una dependencia del camino local que se daba incluso para lazos infinitesimales, y se manifestaba en la curvatura de la conexión. En el caso de nuestro fibrado «tensado» $\mathcal{B}$, la dependencia de camino es de un carácter global. Por supuesto, no hay posibilidad de una dependencia de camino *local* en este ejemplo, ya que el espacio base es 1-dimensional. Pero este ejemplo muestra de paso que es posible tener dependencia de camino globalmente incluso cuando no hay ninguna presente localmente.

15.8. Curvatura fibrada

Sin embargo, podemos modificar nuestro ejemplo para obtener un fibrado sobre un espacio *2-dimensional*, dentro del cual escogemos un círculo concreto que representa a nuestro S^1 original. Consideremos, por conveniencia, que nuestro S^1 es el círculo unidad en el plano complejo, de modo que supondremos que el espacio base $\mathcal{M}^{\mathbb{C}}$, de nuestro nuevo fibrado $\mathcal{B}^{\mathbb{C}}$, está dado por $\mathcal{M}^{\mathbb{C}} = \mathbb{C}$. Véase la Fig. 15.19. Las fibras van a seguir siendo copias de la recta real $\mathbb{R}$. Veamos cómo podemos extender nuestra conexión fibrada a este espacio.

Si no tuviera que haber «tensión» en nuestro nuevo fibrado $\mathcal{B}^{\mathbb{C}}$, entonces podríamos considerar que esta conexión está dada por diferenciación directa con respecto a las coordenadas estándar $(z, \bar{z})$ para el plano complejo $\mathcal{M}^{\mathbb{C}}$. Entonces, la «constancia» de una sección transversal Φ (una función de valor real de z y $\bar{z}$) podría considerarse simplemente como constancia en el sentido ordinario, a saber, $\partial\Phi/\partial z = 0$ (de donde también $\partial\Phi/\partial\bar{z} = 0$, puesto que Φ es real). Cuando introducimos «tensión» en la conexión fibrada, podemos hacerlo modificando el operador $\partial/\partial z$ para convertirlo en un nuevo operador ∇ donde

$$\nabla = \frac{\partial f}{\partial x} - A,$$

siendo la cantidad A una función suave compleja (no necesariamente holomorfa) de z que «opera» por multiplicación (escalar). El operador ∇ actúa sobre cantidades como Φ. Topológicamente, nuestro fibrado $\mathcal{B}^{\mathbb{C}}$ va a ser simplemente el fibrado trivial $\mathbb{C} \times \mathbb{R}$, de modo que podemos utilizar coordenadas globales (z, Φ) para $\mathcal{B}^{\mathbb{C}}$, con z complejo y Φ real.

Una sección transversal de $\mathcal{B}^{\mathbb{C}}$ está determinada por Φ dada como función de z:

$$\Phi = \Phi(z, \bar{z})$$

(la aparición de $\bar{z}$ indica la falta de holomorficidad; véase §10.5). Para que la sección transversal sea *constante* (i.e., horizontal), exigimos $\nabla\Phi = 0$ (de donde $\bar{\nabla}\Phi = 0$ también, puesto que Φ es real), i.e.,

$$\frac{\partial\Phi}{\partial z} = A\Phi.$$

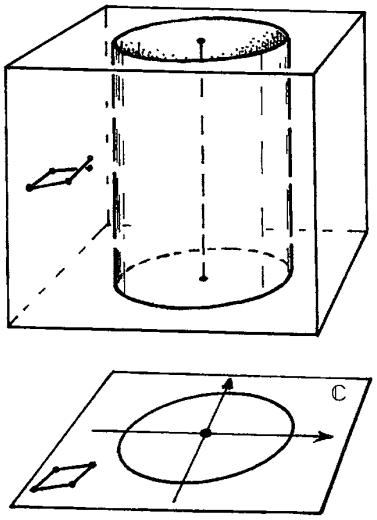


Fig. 15.19. Para obtener una dependencia de camino local (con curvatura) en nuestro fibrado (ahora $\mathcal{B}^{\mathbb{C}}$), necesitamos al menos dos dimensiones en la base $\mathcal{M}^{\mathbb{C}}$, ahora tomada como plano complejo $\mathbb{C}$, donde el S^1 de la Fig. 15.16 es su círculo unidad. Las fibras van a seguir siendo V^1 (i.e., modeladas sobre la recta real $\mathbb{R}$). Utilizando z como una coordenada compleja para $\mathbb{C} = \mathcal{M}^{\mathbb{C}}$, utilizamos la conexión explícita $\nabla = \partial/\partial z - A$, donde A es una función suave compleja de z. Cuando A es holomorfa, la curvatura del fibrado se anula, pero si $A = ikz$ (con k adecuada), obtenemos el fibrado tensado de la Fig. 15.16 para la parte por encima del círculo unidad. La curvatura del fibrado se manifiesta en la falta de cierre de un polígono horizontal sobre un pequeño paralelogramo en $\mathcal{M}^{\mathbb{C}}$.

Si A es holomorfa, entonces no hay ningún problema para resolver esta ecuación, puesto que una expresión de la forma $\Phi = e^{(B+\bar{B})}$ servirá perfectamente, donde $B = \int A dz$.[15.12] Sin embargo, en el caso general, con una A no holomorfa, no tendemos a obtener soluciones no nulas, debido a la relación del conmutador

$$\nabla\bar{\nabla} - \bar{\nabla}\nabla = \frac{\partial A}{\partial \bar{z}} - \frac{\partial \bar{A}}{\partial z}$$

que actúa sobre Φ.[15.13] (El segundo miembro da un número que multiplica a Φ y que no se anula en general, aunque el primer miembro

[15.12] Compruébelo.
[15.13] Verifique esta fórmula.

aniquila cualquier solución real de la ecuación $\partial\Phi/\partial z = A\Phi$.) Este conmutador sirve para definir una *curvatura* para ∇, dada por la parte imaginaria de $\partial A/\partial\bar{z}$, curvatura que mide el grado local de «tensión» en el fibrado.

Haciendo una elección específica de A, para la cual este conmutador toma un valor constante no nulo, tal como $A = ik\bar{z}$ para una constante real adecuada k, podemos obtener un «factor de estiramiento», cuando recorremos un lazo cerrado en $\mathcal{M}^{\mathbb{C}}$, que es simplemente proporcional al área del lazo. Esto se aplica en particular al círculo unidad S^1, de modo que podemos reproducir nuestro fibrado «tensado» original $\mathcal{B}$ sobre S^1 tomando solo esa parte del fibrado que yace por encima de este S^1. Tomando un valor apropiado de k, obtenemos el exigido «estiramiento en un factor 2» sobre el círculo unidad.[15.14]

Este conmutador es el análogo directo del conmutador de operadores ∇_a que hemos visto en §14.4, y que da lugar a la torsión y la curvatura. También podemos suponer que la torsión es nula. (La torsión tiene que ver con la actuación de la conexión sobre vectores tangentes, y no tiene interés para nosotros en relación con aquellos fibrados, como el aquí considerado, que no están asociados con el fibrado tangente.) Para un espacio base n-dimensional $\mathcal{M}$, tenemos cantidades exactamente como las ∇_a y ∇_X del capítulo 14, salvo que ahora actúan sobre cantidades del fibrado.[16] Cuando formamos sus conmutadores de forma apropiada, extraemos la curvatura de la conexión fibrada. Cuando esta curvatura se anula, entonces tenemos muchas secciones localmente constantes del fibrado; en caso contrario, tropezamos con obstáculos para encontrar tales secciones, i.e., encontramos una dependencia de camino local de la conexión. La curvatura describe esta dependencia de camino en el nivel infinitesimal. Esto se ilustra en la Fig. 15.19.

En términos de índices, la conexión se suele expresar en un sistema de coordenadas como un operador de la forma

$$\nabla_a = \frac{\partial}{\partial x^a} - A_a,$$

[15.14] Confirme las afirmaciones de este párrafo, encontrando el valor explícito de k que da este factor 2 requerido.

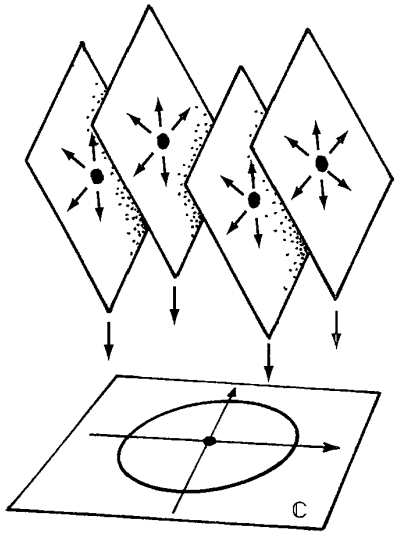


Fig. 15.20. Podemos hacer también la fibra en un espacio vectorial 1-dimensional complejo, correspondiendo la «dilatación» a la multiplicación por un número real.

donde puede considerarse que la cantidad A_a tiene algunos «índices del fibrado» suprimidos. Podemos utilizar letras griegas para estos[17] (suponiendo que estemos interesados en un fibrado vectorial, se aplicarán ideas tensoriales), y entonces la cantidad A_a aparece como $A_a{}^{\mu}{}_{\lambda}$. (Para la expresión con todos los índices habría una «δ^{μ}_{λ}» multiplicando los otros dos términos.) La *curvatura* del fibrado sería una cantidad

$$F_{ab}{}^{\mu}{}_{\lambda},$$

donde el par antisimétrico de índices *ab* se refiere a las direcciones 2-plano tangentes en $\mathcal{M}$, exactamente de la misma forma que en el caso del tensor de curvatura que teníamos antes, pero ahora los índices λ y μ se refieren a las direcciones en la fibra (y normalmente se suprimen en la mayoría de los tratamientos). Hay también un análogo directo de la (segunda) identidad de Bianchi (véase §14.4). (El uso de coordenadas complejas en el ejemplo específico de $\mathcal{B}^{\mathbb{C}}$ era solo una conveniencia, y podía haberse utilizado una notación de índices, igual que en el caso *n*-dimensional.)

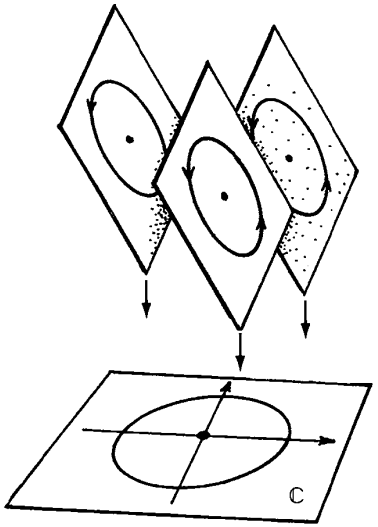


Fig. 15.21. Alternativamente, podemos imponer una «dilatación compleja» en su lugar, tal como la multiplicación por una fase compleja ($e^{i\theta}$, con θ real), de modo que el grupo del fibrado sea ahora U(1), el grupo multiplicativo de dichos números complejos.

Debería señalarse que en muchos casos de fibrados la simetría relevante implicada en la construcción del fibrado no tiene por qué coincidir por completo con la simetría de la fibra. Por ejemplo, en el ejemplo del haz «tensado» $\mathcal{B}$ sobre S^1, o $\mathcal{B}^{\mathbb{C}}$ sobre $\mathbb{C}$, podríamos considerar que la fibra 1-dimensional se ensancha en un espacio vectorial real 2-dimensional, donde el «estiramiento» de la fibra se representa como una expansión uniforme del 2-espacio vectorial. También podríamos dotar a este 2-espacio vectorial real con la estructura adicional que haga de él un espacio vectorial *complejo* 1-dimensional, correspondiendo el «estiramiento» a una multiplicación por un número real (Fig. 15.20). Esto nos lleva a considerar lo que sucede cuando, en su lugar, imponemos un «estiramiento complejo». Un caso particular de este sería una multiplicación por un número complejo de módulo unidad ($\times e^{i\theta}$, con θ real), que proporcionaría una rotación antes que un estiramiento real (Fig. 15.21) (que es el tipo de objeto que está implicado en el fibrado de Clifford del que se ha hablado antes). En este caso, el gru-

po implicado es U(1), el grupo multiplicativo de números complejos unimodulares (véase §13.9). Las conexiones fibradas con este grupo de simetría U(1) son de particular importancia en física, puesto que describen interacciones electromagnéticas, como veremos en §19.4. La esencia de un fibrado semejante se captura si la fibra se toma modelada precisamente sobre el círculo unidad S^1, antes que en el plano complejo entero $\mathbb{C}$. Esto es en cierto sentido más «económico», puesto que el resto del plano es simplemente «transportado» con el círculo y no proporciona ninguna información extra. De todas formas podría sacarse alguna ventaja de utilizar el plano complejo como fibra, puesto que el fibrado se convierte entonces en un fibrado vectorial (complejo).[18]

Más adelante veremos la potencia de estas ideas en relación con las modernas teorías de las fuerzas físicas. En su aspecto como «conexiones gauge», las conexiones fibradas son, de hecho, un ingrediente clave, y ciertos campos físicos emergen como las curvaturas de dichas conexiones (siendo el electromagnetismo de Maxwell el ejemplo arquetípico). Hemos visto cuán esencial es para esta idea que tengamos fibras que posean una *simetría exacta*. Esto plantea cuestiones fundamentales acerca del origen de tales simetrías, y cuáles son realmente. Más adelante volveré a esta importante cuestión, muy especialmente en los capítulos 28, 31 y 34.

Notas

Sección 15.1

15.1. Véase, por ejemplo, Steenrod (1951). Parece que uno de los primeros físicos en darse cuenta, alrededor de 1967, de que la noción del físico de una «teoría gauge» está realmente relacionada con una conexión en un fibrado fue Andrzej Trautman; véase Trautman (1970) (también Penrose *et al.*, 1997, p. A4).

15.2. De hecho, las dimensiones espaciotemporales extras (espacios de Calabi-Yau; véase §31.14) de la teoría de cuerdas no deben considerarse directamente como las «fibras» de un fibrado. Dichas fibras serían espacios de ciertos campos espinoriales en los espacios de Calabi-Yau.

Sección 15.2

15.3. Se requiere más información para una definición completa de espacio producto, de modo que las nociones de topología y suavidad estén correctamente definidas para $\mathcal{M} \times \mathcal{V}$. Cuando puede asignarse una medida de volumen a cada $\mathcal{M}$ y $\mathcal{V}$, entonces el volumen de $\mathcal{M} \times \mathcal{V}$ es el producto de los volúmenes de $\mathcal{M}$ y $\mathcal{V}$. Me distraería si ahora entrara en los detalles de estas materias, aun si, técnicamente hablando, fueran necesarios. Para una referencia adecuada, véanse Kelley (1965), Lefshetz (1949) o Munkres (1954).

15.4. Por simplicidad notacional, estoy adoptando un (moderado) abuso de notación al escribir «$S^1 - p$» para el espacio que consiste en S^1 pero con el punto p eliminado. Los puristas escribirían «$S^1 - \{p\}$», o más probablemente «$S^1\{p\}$» (véase la nota 9.13). La «diferencia» expresada en estas notaciones es entre dos *conjuntos*, y $\{p\}$ denota el conjunto cuyo único elemento es el punto p.

15.5. Véase §12.1 para el significado general de «simplemente conexo».

Sección 15.3

15.6. Normalmente los matemáticos puros suelen ser respetuosos con la gramática, pero muchos de ellos han adoptado la costumbre de utilizar la expresión «asociado a» cuando parecen tener la sensación que «asociado con» no tiene un significado lo bastante específico. No entiendo por qué no utilizan en su lugar la expresión perfectamente gramatical «asignado a». En mi opinión, «asociado a» es bastante peor que otro abuso del lenguaje común por parte de los matemáticos, a saber, «de acuerdo con» (que confieso haber utilizado yo mismo en ocasiones), puesto que la expresión «de acuerdo con si o no», que es lo que representa, es un poco larga.

Sección 15.4

15.7. Véase Adams y Atiyah (1966).

15.8. Véanse Penrose (1987a) y Penrose y Rindler (1986).

15.9. Decimos que $\mathcal{B}$ es un *espacio recubrimiento* de $\mathcal{B}'$. De hecho, $\mathcal{B}$ es lo que se denomina el espacio de recubrimiento *universal* de $\mathcal{B}'$. Al ser simplemente conexo, no puede recubrirse más.

Sección 15.5

15.10. Esta descripción geométrica de los 2-espinores se discute con cierto detalle en Penrose y Rindler (1984), capítulo 1.

15.11. Por ejemplo, en §9.5 la separación de funciones (de una variable real) en partes de frecuencia positiva y negativa (crucial para la teoría cuántica de campos) se analizaba en términos de extensiones a funciones holomorfas; pero el lector puede recordar cierta dificultad en relación con las funciones constantes. Tal cuestión se aclara mucho cuando permitimos que estas sean funciones holomorfas retorcidas y tiene relevancia para la teoría de twistores en §§33.8,10.

Sección 15.6

15.12. Utilizo aquí el nombre griego, aunque es algo más habitual la versión latina Pappus.

15.13. No sería irrazonable adoptar la postura de que el campo de visión del artista sea considerado con más propiedad como una *esfera* S^2, en lugar de $\mathbb{P}^2$, donde tomamos los rayos de luz dirigidos que pasan por O como el campo de visión del artista, en lugar de los no dirigidos que he venido utilizando (implícitamente) en el texto. La esfera es precisamente una cubierta doble del plano proyectivo y el único problema con ella, como algo que proporciona una «geometría», en este contexto, es que pares de «líneas» (a saber, círculos máximos) se intersectan en pares de puntos y no en puntos simples. El artista necesitaría cuatro lienzos en lugar de tres para cubrir la esfera S^2.

15.14. Véase la nota 15.5.

15.15. Este hecho tiene relevancia para una importante e intrigante cuestión mecanocuántica conocida como «fase de Berry» (véanse Berry, 1984, 1985; Simon, 1983; Aharonov y Anandan, 1987; Shankar, 1994, y Woodhouse, 1991, pp. 225-249), que tiene en cuenta el hecho de que no sabemos si «1» está en el círculo unidad, i.e., si dicho «número» es un elemento de una S^1-fibra, en este caso S^{2n+1} sobre $\mathbb{CP}^n$.

Sección 15.8

15.16. En el caso de ∇_a, también necesitamos que actúe sobre vectores (co)tangentes, de modo que $\boldsymbol{\nabla}_a$ pueda operar sobre cantidades con índices espaciotemporales, para poder dar significado al conmutador $\nabla_{[a}\nabla_{b]}$. En el caso de $\underset{x}{\boldsymbol{\nabla}}$, podemos utilizar la expresión de conmutadores $\underset{L}{\boldsymbol{\nabla}}\underset{M}{\boldsymbol{\nabla}} - \underset{M}{\boldsymbol{\nabla}}\underset{L}{\boldsymbol{\nabla}} - \underset{[L,M]}{\boldsymbol{\nabla}}$, que no requiere esto.

15.17. Este tipo de notación de índices para índices del fibrado está desarrollada explícitamente en Penrose y Rindler (1984), capítulo 5.

15.18. Por otro lado, cuando la fibra es el círculo unidad, el fibrado se convierte en un ejemplo de un *fibrado principal* que tiene ventajas en otros contextos. Un fibrado principal es uno en el que la fibra $\mathcal{V}$ está realmente modelada sobre el grupo $\mathcal{G}$ de sus propias simetrías. En términos generales, $\mathcal{G}$ y $\mathcal{V}$ son lo «mismo» para un fibrado principal, pero donde más correctamente $\mathcal{V}$ es $\mathcal{G}$, pero donde uno «olvida» cuál es el elemento identidad de $\mathcal{G}$; por consiguiente, $\mathcal{V}$ es un espacio *afín* (no necesariamente abeliano), de acuerdo con §14.1 y los ejercicios [14.1] y [14.2].

16

La escalera del infinito

16.1. Campos finitos

Parece que una característica universal de las matemáticas que normalmente se consideran subyacentes en el funcionamiento de nuestro universo físico es su dependencia fundamental del *infinito*. Ya los antiguos griegos, antes incluso de haberse visto obligados a hacer consideraciones sobre el sistema de los números reales, habían llegado a habituarse al uso de los números racionales (véase §3.1). El sistema de los racionales no solo es infinito en cuanto que tiene la capacidad de permitir cantidades infinitamente grandes (una propiedad compartida con los propios números naturales), sino que también permite un inagotable grado de refinamiento en una escala infinitamente pequeña. Hay quienes se sienten molestos con estos dos aspectos del infinito. Preferirían un universo que fuera, por una parte, finito en extensión y, por otra, solo finitamente divisible, de modo que una discretización fundamental podría empezar a emerger en los niveles más minúsculos.

Aunque una postura semejante debe considerarse poco convencional, no es intrínsecamente inconsistente. De hecho, ha existido una escuela de pensamiento para la cual el papel en apariencia básico del sistema de números reales $\mathbb{R}$ es tan solo un tipo de aproximación a un «verdadero» sistema de números físicos que solo tiene un número finito de elementos. (Un enfoque de este tipo es el que ha seguido en particular Y. Ahmavaara [1965] y algunos colaboradores; véase §33.1.) ¿Cómo podemos dar sentido a semejante sistema de números finito? Los ejemplos más simples son aquellos construidos a partir de los enteros, «reduciéndolos a *módulo p*», donde p es cualquier número primo.

(Recordemos que los números primos son los números naturales 2, 3, 5, 7, 11, 13, 17, ... que no tienen más divisores que ellos mismos y la unidad, y donde el propio 1 no se considera como primo.) Para reducir los enteros módulo p consideramos que dos enteros son *equivalentes* si su diferencia es un múltiplo de p, es decir,

$$a \equiv b \pmod{p}$$

si y solo si

$$a - b = kp \quad \text{(para algún entero } k\text{)}.$$

De acuerdo con esta receta, los enteros caen exactamente en p «clases de equivalencia» (para la noción de clase de equivalencia, véase el prefacio), de modo que a y b pertenecen a la misma clase siempre que $a \equiv b$. Estas clases se consideran los elementos del *campo finito* $\mathbb{F}_p$, y hay exactamente p de tales elementos. (Estoy utilizando el término «campo» con el sentido que le dan los algebristas. Este no debería confundirse con los «campos» en una variedad, tales como los campos vectoriales o tensoriales, ni con un campo físico tal como el electromagnético. El campo de un algebrista es un *anillo de división conmutativo*; véase §11.1.) Las reglas ordinarias de adición, sustracción, multiplicación (conmutativa) y división son válidas para los elementos de $\mathbb{F}_p$.[16.1] Sin embargo, tenemos la curiosa propiedad adicional de que si sumamos p elementos idénticos, obtenemos siempre *cero* (y, por supuesto, el propio número p tiene que contar como «cero»).

Nótese que, tal como se acaba de describir $\mathbb{F}_p$, sus propios elementos están definidos como «conjuntos infinitos de enteros», puesto que las propias «clases de equivalencia» son conjuntos infinitos, tales como la clase de equivalencia particular $\{\ldots, -7, -2, 3, 8, 13, \ldots\}$ que define el elemento de $\mathbb{F}_5$ ($p = 5$) que denotaríamos por «3». Así pues, ¡hemos apelado al *infinito* para definir las cantidades mismas que constituyen nuestro sistema de números finito! Este es un ejemplo de la forma en que los matemáticos suelen ofrecer una receta rigurosa para una entidad definiéndola en términos de conjuntos infinitos. Es el mismo pro-

[16.1] Demuestre cómo funcionan estas reglas, explicando por qué p tiene que ser primo.

cedimiento mediante «clases de equivalencia» que está implicado en la definición de fracciones, como he mencionado en el prefacio a propósito de la «simplificación» que la amiga de mi madre encontraba tan confusa. Imagino que para alguien convencido de que el sistema de números $\mathbb{F}_p$ (para un p apropiado) está «realmente» enraizado en la naturaleza, el procedimiento de «clases de equivalencia» sería solo una «conveniencia» del matemático, dirigida a proporcionar algún tipo de receta rigurosa en términos de los procedimientos infinitos más familiares (históricamente). De hecho, no necesitamos apelar a conjuntos infinitos de enteros; lo que sucede es que este es el procedimiento más sistemático. En cualquiera de los casos podríamos alternativamente hacer un listado de todas las operaciones, puesto que estas son finitas en número.

Consideremos más detalladamente el caso $p = 5$ a modo de ejemplo. Podemos etiquetar los elementos de $\mathbb{F}_5$ mediante los símbolos estándar 0, 1, 2, 3, 4, y tenemos las tablas de suma y multiplicación

+	0	1	2	3	4
0	0	1	2	3	4
1	1	2	3	4	0
2	2	3	4	0	1
3	3	4	0	1	2
4	4	0	1	2	3

×	0	1	2	3	4
0	0	0	0	0	0
1	0	1	2	3	4
2	0	2	4	1	3
3	0	3	1	4	2
4	0	4	3	2	1

y notamos que cada elemento no nulo tiene un inverso multiplicativo:

$$1^{-1} = 1, \quad 2^{-1} = 3, \quad 3^{-1} = 2, \quad 4^{-1} = 4,$$

en el sentido de que $2 \times 3 \equiv 1 \pmod 5$, etc. (a partir de ahora, utilizaré «=» en lugar de «≡», cuando trabaje dentro de un sistema de números finito concreto).

Existen también otros campos finitos $\mathbb{F}_q$, construidos de una forma algo más elaborada, en los que el número total de elementos es una *potencia* de un primo: $q = p^n$. Permítanme dar el ejemplo más sencillo, a saber, el caso $q = 4 = 2^2$. Aquí podemos etiquetar los diferentes elementos como $0, 1, \omega, \omega^2$, donde $\omega^3 = 1$ y cada elemento x está sujeto a $x + x = 0$. Esto amplía ligeramente el grupo multiplicativo de números complejos $1, \omega, \omega^2$ que son raíces cúbicas de la unidad (descritas en §5.4 y mencionadas en §5.5 como descripción de la «quarkedad» de las partículas que interaccionan fuertemente). Para ampliar este grupo multiplicativo a $\mathbb{F}_4$, añadimos un cero «0» y proporcionamos una operación «suma» para la que $x + x = 0$.[16.2] En el caso general $\mathbb{F}_{p^n}$, tendríamos $x + x + \ldots + x = 0$, donde el número de xs en la suma es p.

16.2. ¿Una geometría finita o una geometría infinita para la física?

No está claro si tales objetos tienen realmente un papel significativo en física, aunque la idea ha sido reavivada de vez en cuando. Si $\mathbb{F}_q$ tuviera que tomar el lugar del sistema de números reales en cualquier sentido significativo, entonces p tendría que ser realmente muy grande (de modo que los «$x + x + \ldots + x = 0$» no se manifestarían como una seria discrepancia en el comportamiento observado). En mi opinión, una teoría física que dependa de un modo fundamental de cierto número primo absurdamente enorme sería una teoría mucho más complicada (e improbable) que una que pueda depender de una simple noción de infinito. De todas formas, tiene algún interés seguir trabajando en estos temas. De hecho, gran parte de la geometría sobrevive cuando las coordenadas se dan como elementos de un $\mathbb{F}_q$. Más cuidados necesitan las ideas del cálculo infinitesimal; de todas formas, muchas de ellas también sobreviven.

Resulta instructivo (y divertido) ver cómo funciona la *geometría proyectiva* con un número total finito de puntos y en consecuencia, podemos explorar los n-espacios proyectivos $\mathbb{P}^n(\mathbb{F}_q)$ sobre el campo $\mathbb{F}_q$. Encontramos que $\mathbb{P}^n(\mathbb{F}_q)$ tiene exactamente $1 + q + q^2 + \ldots + q^n = (q^{n+1} - 1)/(q - 1)$ puntos diferentes.[16.3] Los *planos* proyectivos $\mathbb{P}^2(\mathbb{F}_q)$ son particularmente fascinantes porque puede darse una construcción muy elegante para ellos. Esta puede describirse de la siguiente forma.

[16.2] Haga tablas completas de suma y multiplicación para $\mathbb{F}_4$ y compruebe que las leyes del álgebra funcionan (donde suponemos que $1 + \omega + \omega^2 = 0$).
[16.3] Demuéstrelo.

Tomemos un círculo hecho de algún material apropiado tal como cartulina y pinchémoslo en su centro con un alfiler que lo una a otro trozo fijo de cartulina de base, de modo que pueda rotar libremente. Marquemos $1 + q + q^2$ puntos igualmente espaciados alrededor de la circunferencia en la cartulina de base, etiquetándolos, en sentido contrario a las agujas del reloj, mediante los números $0, 1, 2, \ldots, q(1 + q)$. En el disco rotatorio, marquemos $1 + q$ puntos especiales en ciertas posiciones cuidadosamente escogidas. Estas posiciones serán tales que, para cualquier selección de dos de los puntos marcados en la base, hay exactamente una posición del disco para la que los dos puntos seleccionados coinciden con dos de estos puntos especiales en el disco. Otra forma de decirlo es la siguiente: si $a_0, a_1, \ldots, a_q$ son las distancias sucesivas a lo largo de la circunferencia entre estos puntos especiales, tomados cíclicamente (donde la distancia a lo largo de la circunferencia entre puntos marcados sucesivos en el círculo de base se toma como distancia *unidad*), entonces cada distancia $1, 2, 3, \ldots, q + 1$ puede representarse unívocamente como una suma de una colección de *as* cíclicamente sucesivos. En la Fig. 16.1 he representado los discos para $q = 2, 3, 4$ y 5, para los que $a_0, \ldots, a_q$ pueden tomarse como 1, 2, 4; 1, 2, 6, 4; 1, 3, 10, 2, 5; 1, 2, 7, 4, 12, 5, respectivamente.[16.4] En los casos $q = 7, 8, 9, 11, 13$ y 16 podemos hacer discos mágicos definidos por 1, 2, 10, 19, 4, 7, 9, 5; 1, 2, 4, 8, 16, 5, 18, 9, 10; 1, 2, 6, 18, 22, 7, 5, 16, 4, 10; 1, 2, 13, 7, 5, 14, 34, 6, 4, 33, 18, 17, 21, 8; 1, 2, 4, 8, 16, 32, 27, 26, 11, 9, 45, 13, 10, 29, 5, 17, 18, respectivamente. Un teorema matemático dice que existen discos mágicos para todo $\mathbb{P}^2(\mathbb{F}_q)$ (con q una potencia de un número primo).[1] El lector quizá encuentre divertido comprobar varios casos de los teoremas de Pappos y Desargues (véanse §15.6 y la Fig. 15.14).[2] (¡Tome $q > 2$, para tener puntos suficientes para una configuración no degenerada!) En la Fig. 16.2 se ilustran dos ejemplos (Desargues, para $q = 3$ y Pappos, para $q = 5$).

El caso más simple $q = 2$ tiene un interés particular que procede de

[16.4] Muestre cómo construir nuevos discos en los casos $q = 3,5$, empezando en un punto concreto marcado en uno de los discos que he dado y multiplicando luego cada una de las distancias angulares a los otros puntos marcados por algún entero fijo. ¿Por qué funciona esto?

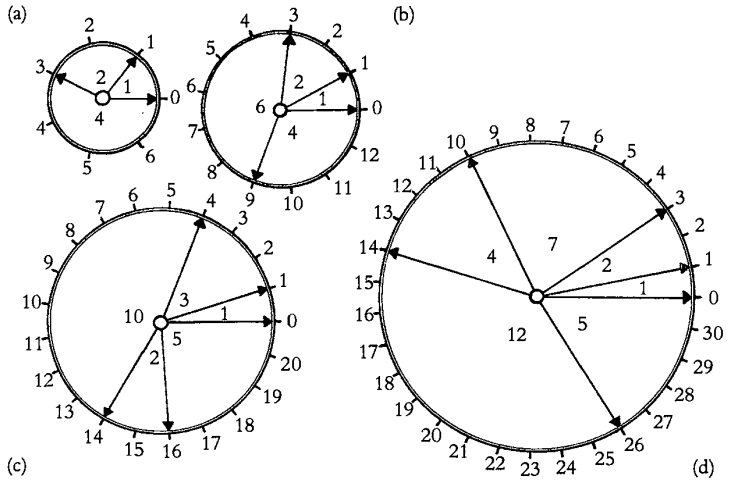


Fig. 16.1. «Discos mágicos» para planos proyectivos finitos $\mathbb{P}^2(\mathbb{F}_q)$ (siendo q una potencia de un número primo). Los $1 + q + q^2$ puntos se representan como numerales sucesivos $0, 1, 2 \ldots, q(1 + q)$ colocados equidistantemente alrededor de un círculo de fondo. Se une un disco circular que rota libremente, con flechas que etiquetan $1 + q$ lugares particulares: los puntos de una línea en $\mathbb{P}^2(\mathbb{F}_q)$. Estos son tales que para cada par de numerales distintos hay exactamente una posición del disco en la que las flechas apuntan a ellos. Se muestran discos mágicos para (a) $q = 2$; (b) $q = 3$; (c) $q = 4 = 2^2$, y (d) $q = 5$.

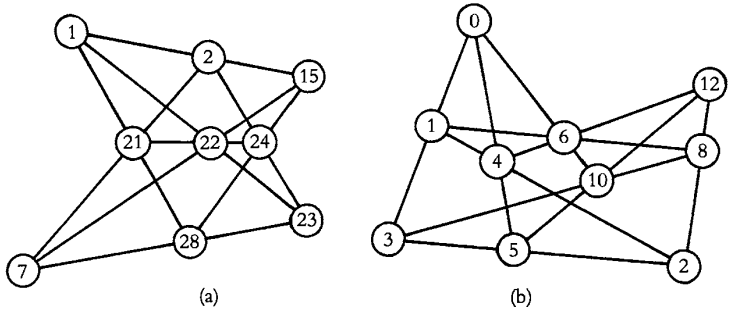


Fig. 16.2. Versiones en geometría finita de los teoremas de la Fig. 5.14. (a) Pappos (con $q = 5$) y (b) Desargues (con $q = 3$), ilustrados por el uso respectivo de los discos mostrados en las Figs. 16.1d y 16.1b.

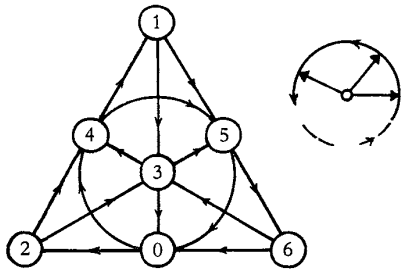


Fig. 16.3. El plano de Fano $\mathbb{P}^2(\mathbb{F}_2)$, con 7 puntos y 7 líneas (contando el círculo como una «línea recta») numeradas según la Fig. 16.1a. Esto da la tabla de multiplicación para los elementos base $\mathbf{i}_0, \mathbf{i}_1, \mathbf{i}_2, \ldots, \mathbf{i}_6$ del álgebra de división de octoniones, donde las flechas proporcionan el ordenamiento cíclico que da un signo «+».

otras direcciones.[16.5] Este plano, con 7 puntos, se denomina *plano de Fano*, y se muestra en la Fig. 16.3, donde el círculo central cuenta como una «línea recta». Aunque su alcance como geometría es bastante limitado, desempeña un papel importante de un tipo diferente al proporcionar la ley de multiplicación para *octoniones* (véanse §11.2 y §15.4). El plano de Fano tiene 7 puntos en él y cada punto debe ser asociado con uno de los elementos generadores $\mathbf{i}_0, \mathbf{i}_1, \mathbf{i}_2, \ldots, \mathbf{i}_6$ del álgebra de octoniones. Cada uno de estos debe satisfacer $\mathbf{i}_r^2 = -1$. Para encontrar el producto de dos elementos generadores *distintos*, simplemente encontramos la recta en el plano de Fano que une los puntos que los representan, y entonces el punto restante en la recta es el punto que representa al producto (salvo un signo) de los otros dos. Para esto, la imagen del plano de Fano no es suficiente, puesto que también hay que determinar el signo del producto. Podemos encontrar este signo si volvemos a la descripción dada por el disco, que se muestra en la Fig. 16.1a, o utilizando las (equivalentes) configuraciones de flechas (interpretadas cíclicamente) de la Fig. 16.3. Asignemos un ordenamiento cíclico a los puntos marcados en el disco, por ejemplo, en sentido contrario a las agujas del reloj.

[16.5] El campo finito $\mathbb{F}_8$ tiene elementos $0, 1, \varepsilon, \varepsilon^2, \varepsilon^3, \varepsilon^4, \varepsilon^5, \varepsilon^6$, donde $\varepsilon^7 = 1$ y $1 + 1 = 0$. Demuestre que o bien (1) hay una identidad de la forma $\varepsilon^a + \varepsilon^b + \varepsilon^c = 0$ siempre que a, b y c sean números en el círculo de fondo de la Fig. 16.1a que puede alinearse con los tres puntos del disco, o de lo contrario (2) es válido lo mismo, pero con ε^3 en lugar de ε (i.e., $\varepsilon^{3a} + \varepsilon^{3b} + \varepsilon^{3c} = 0$).

Entonces tenemos $\mathbf{i}_x\mathbf{i}_y = \mathbf{i}_z$, si el ordenamiento cíclico de $\mathbf{i}_x, \mathbf{i}_y, \mathbf{i}_z$ coincide con el asignado por el disco, e $\mathbf{i}_x\mathbf{i}_y = -\mathbf{i}_z$ en caso contrario. En particular, tenemos $\mathbf{i}_0\,\mathbf{i}_1 = \mathbf{i}_3 = -\mathbf{i}_1\mathbf{i}_0$, $\mathbf{i}_0\mathbf{i}_2 = \mathbf{i}_6$, $\mathbf{i}_1\mathbf{i}_6 = -\mathbf{i}_6$, $\mathbf{i}_1\mathbf{i}_6 = -\mathbf{i}_1$, etc.[16.6]

Aunque hay una gran elegancia en estas estructuras geométricas y algebraicas, parecen tener poco contacto obvio con el funcionamiento del mundo físico. Tal vez esto no debería sorprendernos si adoptamos el punto de vista expresado en la Fig. 1.3, en §1.4; pues las matemáticas que tienen una relevancia directa para las leyes físicas que gobiernan nuestro universo no son sino una minúscula parte del mundo platónico en su conjunto, o así podría parecerlo, hasta donde nuestro conocimiento actual nos ha llevado. Es posible que, a medida que nuestro conocimiento se haga más profundo en el futuro, se encuentren papeles importantes para estructuras tan elegantes como las geometrías finitas o para el álgebra de octoniones. Pero tal como están las cosas, todavía no se ha presentado un argumento convincente.[3] Parece que la sola elegancia matemática está lejos de ser suficiente (véase también §34.9). Esto debería enseñarnos a ser cautos en nuestra búsqueda de los principios subyacentes en las leyes del universo.

Abstengámonos de tales flirteos con estas atractivas estructuras finitas y volvamos a la impresionante riqueza matemática que es inherente al infinito. Como preliminar, habría que señalar que estructuras infinitas (tales como la totalidad de los números naturales $\mathbb{N}$) podrían formar parte de algún formalismo matemático dirigido a una descripción de la realidad, mientras no se pretenda que tales estructuras infinitas tengan una interpretación física directa como entidades físicas infinitas (o infinitesimales). Por ejemplo, se han hecho intentos por desarrollar un esquema en el que la discretización (y, de hecho, la finitud) aparece en el nivel más pequeño, mientras que sigue existiendo la capacidad para describir estructuras indefinidamente (e incluso infinitamente) grandes. Esto se aplica en particular a ciertas viejas ideas mías para construir el espacio de una forma finita, utilizando la teoría de redes de espín que describiré brevemente en §32.6, y que depende del

[16.6] Demuestre que el «asociado» $a(bc) - (ab)c$ es antisimétrico en a, b, c cuando son elementos generadores, y deduzca que esto (y por ello también $a(ab) = a^2b$) es válido para *todos* los elementos. *Sugerencia*: Utilice la Fig. 16.3 y la simetría completa del plano de Fano.

hecho de que, según la mecánica cuántica estándar, la medida del espín de un objeto viene dada por un múltiplo número natural de cierta cantidad fija ($\frac{1}{2}\hbar$). En realidad, como he mencionado en §3.3, en los primeros días de la mecánica cuántica hubo grandes esperanzas, no satisfechas por los desarrollos futuros, de que la teoría cuántica llevara a la física a una imagen del mundo en la que realmente hubiera discretización en los niveles más minúsculos. Tal como resultaron las cosas, en las teorías exitosas de nuestros días consideramos el espaciotiempo continuo incluso cuando están involucrados conceptos cuánticos, y las ideas que implican discretización a pequeña escala deben considerarse «poco convencionales» (§33.1). El continuo sigue siendo un protagonista esencial incluso en aquellas teorías que intentan aplicar las ideas de la mecánica cuántica a la propia estructura del espacio y el tiempo. Esto se aplica, en particular, a la teoría de variables de lazo de Ashtekar-Rovelli-Smolin-Jacobson, en la que ideas discretas (combinatorias) tales como las de la teoría de nudos y enlaces desempeñan realmente papeles clave, y en las que las redes de espín entran también en la estructura básica. (Veremos algo de este notable esquema en el capítulo 32, y en §33.1 encontraremos brevemente otras ideas relativas a un «espaciotiempo discreto».)

Así pues, parece, al menos por el momento, que tenemos que tomar en serio el uso del infinito, especialmente en su papel en la descripción matemática del continuo físico. Pero ¿qué tipo de infinito necesitamos aquí? En §3.2 he descrito brevemente el método de las «cortaduras de Dedekind» para construir el sistema de los números reales en términos de conjuntos infinitos de números racionales. De hecho, este es un gran paso, que implica una noción de infinito que supera en mucho a la que está implicada en los propios números racionales. Será importante para nosotros que abordemos aquí esta cuestión. De hecho, como demostró en 1874 el gran matemático danés-ruso-alemán Georg Cantor, ¡existen *diferentes tamaños* de infinitos! La infinitud de los números naturales es en realidad el menor de estos tamaños, y diferentes infinitos se suceden sin cesar en escalas cada vez mayores. Tratemos de captar una idea de las originales y fundamentales ideas de Cantor.

16.3. Diferentes tamaños de infinito

El primer ingrediente clave de la revolución de Cantor es la idea de una correspondencia 1–1 (i.e., uno a uno).[4] Decimos que dos conjuntos tienen la misma *cardinalidad* (lo que en lenguaje ordinario significa que tienen «el mismo número de elementos») si es posible establecer una correspondencia entre los elementos de un conjunto y los elementos del otro conjunto, uno a uno, de modo que no haya ningún elemento de ninguno de los conjuntos que no tome parte en la correspondencia. Es evidente que este procedimiento da la respuesta correcta («mismo número de elementos») para conjuntos finitos (i.e., conjuntos con un número finito $1, 2, 3, 4, \ldots$ de miembros, o incluso 0 elementos, en cuyo caso exigimos que la correspondencia sea vacía). Pero en el caso de conjuntos infinitos hay una característica nueva (ya advertida en 1638 por el gran físico y astrónomo Galileo Galilei)[5] y es que un conjunto infinito tiene la misma cardinalidad que uno de sus subconjuntos propios (donde «propio» significa distinto del conjunto entero).

Veamos esto en el caso del conjunto $\mathbb{N}$ de los números naturales

$$\mathbb{N} = \{0, 1, 2, 3, 4, 5, \ldots\}.$$

Ahora, si eliminamos el 0 de este conjunto,[6] encontramos un nuevo conjunto $\mathbb{N} - 0$ que tiene claramente la misma cardinalidad que $\mathbb{N}$, porque podemos establecer la correspondencia 1–1, en la que al elemento r de $\mathbb{N}$ se le hace corresponder el elemento $r + 1$ en $\mathbb{N} - 0$. Alternativamente, podemos tomar el ejemplo de Galileo y ver que el conjunto de *cuadrados* $\{0, 1, 4, 9, 16, 25, \ldots\}$ también debe tener la misma cardinalidad que $\mathbb{N}$, pese al hecho de que, en un sentido bien definido, los cuadrados constituyen una proporción cada vez más pequeña de los números naturales en conjunto. Asimismo, podemos ver que el conjunto $\mathbb{Z}$ de todos los enteros es de nuevo de esta misma cardinalidad. Puede verse esto si consideramos el ordenamiento de $\mathbb{Z}$ dado por

$$\{0, 1, -1, 2, -2, 3, -3, 4, -4, \ldots\},$$

que podemos emparejar simplemente con los elementos $\{0, 1, 2, 3, 4, 5, 6, 7, 8, \ldots\}$ del conjunto $\mathbb{N}$. Más sorprendente es el hecho de que la

cardinalidad de los números *racionales* sea también la misma que la cardinalidad de $\mathbb{N}$. Hay muchas maneras de verlo directamente,[16.7],[16.8] pero antes de demostrar esto en detalle veamos cómo este ejemplo concreto cae en el marco general de la maravillosa teoría de Cantor de los números cardinales infinitos.

En primer lugar, ¿qué es un *número cardinal*? Básicamente, es el «número» de elementos en cierto conjunto, y consideramos que dos conjuntos tienen el «mismo número de elementos» si y solo si pueden ponerse en correspondencia 1–1 entre sí. Podríamos tratar de ser más precisos utilizando la idea de «clases de equivalencia» (empleada antes para definir $\mathbb{F}_p$ para un número primo p; véase también el prefacio) y decir que el número cardinal α de cierto conjunto A *es* la clase de equivalencia de todos los conjuntos con la misma cardinalidad de A. De hecho, el lógico Gottlob Frege trató de hacer precisamente esto en 1884, pero resulta que hay dificultades fundamentales con conceptos abiertos tales como «todos los conjuntos», puesto que con ellos pueden surgir graves contradicciones (como veremos en §16.5). Para evitar tales contradicciones, parece necesario poner algunas limitaciones al tamaño del «universo de conjuntos posibles». Pronto tendré que hacer algunos comentarios sobre esta preocupante cuestión. Por el momento, la evitaremos refugiándonos en una postura que he adoptado antes (como se ha mencionado en el prefacio en relación con la definición por «clases de equivalencia» de los números racionales). Consideraremos que los cardinales son simples entidades matemáticas (¡habitantes del mundo de Platón!) que pueden ser abstraídas de la noción de equivalencia 1–1 entre conjuntos. Nos permitimos decir que el conjunto A «tiene cardinalidad α», o que «tiene α elementos», siempre que seamos consistentes y digamos que el conjunto B también «tiene cardinalidad α», o que «tiene α elementos», si y solo si A y B pueden ponerse en correspondencia 1–1. Nótese que todos los números naturales pueden

[16.7] Vea si puede proporcionar un procedimiento explícito semejante, encontrando una manera sistemática de ordenar todas las fracciones. Puede encontrar útil el resultado del ejercicio [16.8].

[16.8] Demuestre que la función $1/2((a + b)^2 + 3a + b)$ proporciona explícitamente una correspondencia 1–1 entre los números naturales y los pares (a, b) de números naturales.

considerarse como números cardinales en este sentido, y esto está mucho más cerca de la noción intuitiva de lo que «es» un número natural que la definición de «ordinal» ($0 = \{\}, 1 = \{0\}, 2 = \{0, \{0\}\}, 3 = \{0, \{0\}, \{0, \{0\}\}\}, \ldots$) dada en §3.4. Los números naturales son, de hecho, los cardinales *finitos* (en el sentido en que los cardinales *infinitos* son las cardinalidades de aquellos conjuntos, como el $\mathbb{N}$ anterior, que contienen subconjuntos propios de la misma cardinalidad que sí mismos).

A continuación, podemos establecer relaciones entre números cardinales. Decimos que el cardinal α es *menor o igual que* el cardinal β, y escribimos

$$\alpha \leq \beta$$

(o equivalentemente $\beta \geq \alpha$) si los elementos de un conjunto A con cardinalidad α pueden ponerse en correspondencia 1–1 con los elementos de algún subconjunto (no necesariamente un subconjunto propio) de elementos de un conjunto B, con cardinalidad β. Debería ser evidente que si $\alpha \leq \beta$ y $\beta \leq \gamma$, entonces $\alpha \leq \gamma$.[16.9] Uno de los bellos resultados de la teoría de los números cardinales es que si

$$\alpha \leq \beta \text{ y } \beta \leq \alpha,$$

entonces,

$$\alpha = \beta,$$

lo que significa que existe una correspondencia 1–1 entre A y B.[16.10] Podemos preguntar si existen pares de cardinales α y β para los que no se satisface ninguna de las relaciones $\alpha \leq \beta, \beta \leq \alpha$. Tales cardinales serían *no comparables*. De hecho, se sigue de la hipótesis conocida como *axioma de elección* (mencionada brevemente en §1.3), según la cual no existen cardinales no comparables.

El axioma de elección afirma que si tenemos un conjunto A, cuyos miembros son todos conjuntos no vacíos, entonces existe un conjunto

[16.9] Demuéstrelo en detalle.

[16.10] Pruébelo. Esbozo: existe una aplicación 1–1 b que lleva A a cierto subconjunto $bA(= b(A))$ de B, y una aplicación 1–1 a que lleva B a cierto subconjunto aB de A; considere la aplicación de A en B que utiliza b para aplicar A–aB en bA–baB y abA–$abaB$ en $babA$–$babaB$, etc., y que utiliza a^{-1} para aplicar aB–abA en B–bA y $abaB$–$ababA$ en baB–$babA$, etc. y resuelva qué hacer con el resto de A y B.

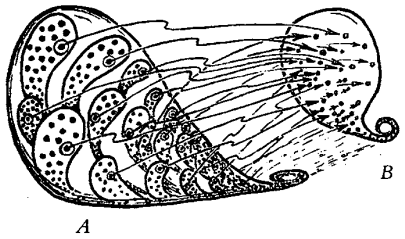


Fig. 16.4. El axioma de elección afirma que, para cualquier conjunto A, cuyos miembros todos son conjuntos no vacíos, existe un conjunto B que contiene exactamente un elemento de cada uno de los conjuntos pertenecientes a A.

B que contiene exactamente un elemento de cada uno de los conjuntos que pertenecen a A. A primera vista, podría parecer que el axioma de elección está afirmando algo obvio (véase la Fig. 16.4). Sin embargo, no está del todo libre de controversia el que el axioma de elección *debiera* ser aceptado como algo de validez universal. Mi posición sobre ello es de cautela. El problema con este axioma es que se trata de una pura afirmación de «existencia», sin ningún indicio de una regla mediante la cual pudiera especificarse el conjunto B. De hecho, tiene varias consecuencias alarmantes. Una de ellas es el teorema de Banach-Tarski,[7] una de cuyas versiones dice que la esfera unidad ordinaria en el 3-espacio euclídeo puede dividirse en cinco piezas con la propiedad de que, mediante movimientos euclídeos simplemente (i.e., traslaciones y rotaciones), estas piezas pueden recomponerse ¡para formar *dos* esferas unidad completas! Las «piezas», por supuesto, no son cuerpos sólidos, sino colecciones intrincadas de puntos, y están definidas de una forma muy no constructiva, pues su «existencia» solo se afirma mediante el uso del axioma de elección.

Déjenme hacer una lista, sin demostración, de algunas propiedades muy básicas de los números cardinales. En primer lugar, el símbolo «$\leq$» da el significado normal (véase la nota 3.1) cuando se aplica a los números naturales (los cardinales *finitos*). Además, cualquier número natural es menor o igual ($\leq$) que cualquier número cardinal infinito, y por supuesto es *estrictamente* menor, i.e., menor que ($<$) y no igual a este. Supongamos ahora que $\beta \leq \alpha$, con α infinito; entonces (en claro contraste con lo que nos es familiar para los números finitos), la cardinalidad de la *unión* $A \cup B$ es simplemente la mayor de las dos, a saber α, y la cardinalidad del *producto* $A \times B$ es también α. (Hemos visto ejemplos del producto antes, por ejemplo, §13.2 y §15.2. El conjunto

$A \times B$ consiste en todos los *pares* (a, b), donde a se toma de A y b se toma de B. Para conjuntos finitos, la cardinalidad de su producto, como conjuntos, es el producto numérico ordinario de sus cardinalidades, que en el caso de conjuntos finitos con más de un miembro es siempre mayor que la cardinalidad de cada uno de ellos por separado.) Esto no parece llevarnos muy lejos si queremos encontrar infinitos que sean mayores que los que ya tenemos. Parece que estamos «atascados» en α.

Veremos cómo «desatascarnos» en la próxima sección. Por el momento, sin embargo, podemos ver que lo que hemos hecho antes es al menos suficiente para demostrarnos que el número de números racionales es el mismo que el número de números naturales. Siguiendo a Cantor, utilicemos el símbolo $\aleph_0$ («aleph cero») para la cardinalidad de los números naturales $\mathbb{N}$, que, como hemos visto antes, es la misma que la cardinalidad de los enteros $\mathbb{Z}$. De hecho, el número infinito $\aleph_0$ es el menor de los cardinales infinitos. Ahora bien, ¿cuál es la cardinalidad ρ de los racionales? Cualquier número racional puede escribirse (de muchas maneras) de la forma a/b, donde a y b son enteros. Escogiendo una de estas maneras (por ejemplo, con «términos mínimos») para cada racional, hemos encontrado una correspondencia 1–1 entre el conjunto de racionales y un *subconjunto* del conjunto $\mathbb{N} \times \mathbb{N}$. Por consiguiente, ρ es menor o igual que la cardinalidad de $\mathbb{N} \times \mathbb{N}$. Pero por lo anterior (o por aplicación directa del ejercicio [16.8]), la cardinalidad de $\mathbb{N} \times \mathbb{N}$ es la misma que la cardinalidad de $\mathbb{N}$, a saber, $\aleph_0$. Así pues, $\rho \leq \aleph_0$. Pero los enteros están contenidos en los racionales, de modo que $\aleph_0 \leq \rho$. De ahí que $\rho = \aleph_0$.

16.4. El corte diagonal de Cantor

Llegamos ahora al primer logro sorprendente de Cantor, a saber, su demostración de que realmente existen infinitos estrictamente mayores que $\aleph_0$, y que la cardinalidad del conjunto $\mathbb{R}$ de los números reales es uno de estos infinitos. Daré este resultado como un caso particular del resultado más general de Cantor

$$\alpha < 2^\alpha,$$

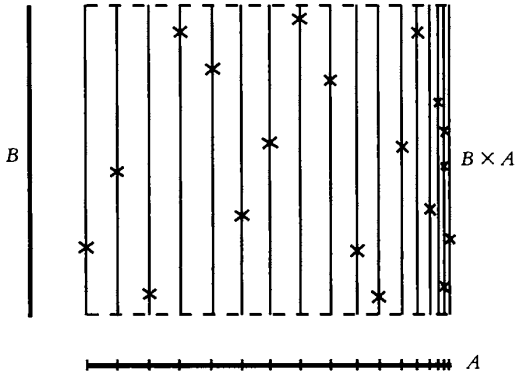


Fig. 16.5. Para conjuntos generales A, B, el conjunto de todas las aplicaciones de A en B se denota por B^A (véase también la Fig. 6.1). A cada elemento de A se le asigna un elemento particular de B. Esto proporciona una sección transversal de $B \times A$, considerada como un fibrado sobre A (como en la Fig. 15.6a), excepto que no hay ninguna noción de continuidad implicada.

donde $\alpha < \beta$ significa $\alpha \leq \beta$ y $\alpha \neq \beta$ (y, por supuesto, también podemos escribir $\alpha < \beta$ como $\beta > \alpha$). La notable demostración de Cantor de este resultado (y el resultado mismo) constituye una de las hazañas más originales e influyentes en el conjunto de las matemáticas. Pese a todo, es lo suficientemente simple como para que pueda darla aquí en su totalidad.

Antes de nada, debería explicar la notación. Si tenemos dos conjuntos A y B, entonces el conjunto B^A es el *conjunto de todas las aplicaciones de* A en B. ¿Cuál es la lógica para este uso de la notación? Consideremos el conjunto A extendido ante nosotros, estando cada elemento de A representado por un «punto». Entonces, para representar un elemento de B^A, colocamos un elemento de B en cada uno de estos puntos. Esto es una aplicación de A en B porque proporciona una asignación de un elemento de B a cada elemento de A (véase la Fig. 16.5). La razón para la «notación exponencial» B^A es que cuando aplicamos este procedimiento a conjuntos finitos, digamos a un conjunto A, con a elementos, y un conjunto B, con b elementos, entonces el número total de modos de asignar un elemento de B a cada elemento de A es realmente b^a. (Existen b modos para el primer miembro de A; existen b modos para el segundo; existen b modos para el tercero, y así sucesivamente para cada uno de los a miembros de A. El número total

es, por lo tanto, $b \times b \times b \times \ldots \times b$, siendo a el número de bs en el producto, de modo que es exactamente b^a.) La notación de Cantor es

$$\beta^\alpha$$

para la cardinalidad de B^A, donde β y α son las cardinalidades respectivas de B y A.

Esto adquiere un significado particular cuando $\beta = 2$. Aquí podemos tomar B como un conjunto de dos elementos que consideraremos que son las etiquetas «dentro» y «fuera». Cada elemento de B^A es así una asignación de «dentro» o de «fuera» a cada elemento de A. Tal asignación equivale a escoger un *subconjunto* de A (a saber, el subconjunto de elementos «dentro»). Así pues, B^A es en este caso precisamente el conjunto de subconjuntos de A (y con frecuencia denotamos este conjunto de subconjuntos de A por 2^A). En consecuencia,

2^α es el número total de subconjuntos de cualquier conjunto con α elementos.

Ahora viene la sorprendente demostración de Cantor. Esta procede de acuerdo con la tradición clásica de la antigua Grecia de «demostración por contradicción» (§2.6 y §3.1). En primer lugar, tratemos de suponer que $\alpha = 2^\alpha$, de modo que hay una correspondencia 1–1 entre cierto conjunto A y su conjunto de subconjuntos 2^A. Entonces, cada elemento a de A estará asociado con un subconjunto concreto $S(a)$ de A, bajo esta correspondencia. Podemos esperar que a veces el conjunto $S(a)$ contenga el propio a como un miembro y otras veces no lo haga. Consideremos la colección de todos los elementos a para los que $S(a)$ no contiene a. Esta colección será un particular subconjunto Q de A (que admitimos que pueda ser el conjunto vacío o la totalidad de A si es necesario). Bajo la supuesta correspondencia 1–1, debemos tener $Q = S(q)$, para algún q en S. Ahora planteamos la pregunta: ¿está q en Q o no lo está? Supongamos primero que no lo está. Entonces, q debe pertenecer a la colección de elementos de A que acabamos de señalar como el subconjunto Q, de modo que q debe pertenecer a Q después de todo: una contradicción. Esto nos deja con la hipótesis alternativa, a saber, que q está en Q. Pero, entonces, q no puede pertenecer a la colección que hemos denominado Q, de modo que no perte-

nece a Q después de todo: de nuevo una contradicción. Por lo tanto, concluimos que nuestra supuesta correspondencia 1–1 entre A y 2^A no puede existir.

Por último, necesitamos demostrar que $\alpha \leq 2^\alpha$, i.e., que existe una correspondencia 1–1 entre A y algún subconjunto de 2^A. Esto se consigue utilizando simplemente la correspondencia 1–1 que asigna cada elemento a de A al subconjunto particular de A que contiene precisamente el elemento a y ninguno más. Así pues, hemos establecido $\alpha < 2^\alpha$, como se exigía, al haber demostrado $\alpha \leq 2^\alpha$ pero $\alpha \neq 2^\alpha$.

Aunque este argumento puede ser algo confuso (y cualquier lector confundido se preocupará de estudiarlo de nuevo), es extraordinariamente «elemental» en el sentido de que no apela a ideas matemáticas que exijan un conocimiento propio de expertos. A la vista de ello, es muy notable que sus implicaciones tienen un extraordinario alcance. No solo nos permite ver que existen esencialmente más números reales que números naturales, sino que también demuestra que no existe límite a la enormidad de los números infinitos posibles. Además, en una forma ligeramente modificada, el argumento demuestra que no hay un modo computacional de decidir si una computación general llegará a término (Turing), y una consecuencia relacionada es el famoso *teorema de la incompletitud* de Gödel que demuestra que ningún conjunto de reglas matemáticas fiables preasignadas puede englobar todos los procedimientos mediante los cuales se establecen las verdades matemáticas. En la próxima sección trataré de dar una idea de cómo se obtienen estos resultados.

Para concluir esta sección, veamos por qué el resultado anterior establece realmente el primer y extraordinario gran paso de Cantor concerniente al infinito, a saber, que existen realmente muchos más números reales que números naturales, pese al hecho de que hay exactamente tantas fracciones como números naturales. (¡Este gran paso estableció que hay realmente una teoría no trivial del infinito!) Esto se seguirá si podemos ver que la cardinalidad de los reales, normalmente denotada por C, es realmente igual a $2^{\aleph_0}$:

$$\mathsf{C} = 2^{\aleph_0}.$$

Entonces, por el argumento anterior, $\mathsf{C} > \aleph_0$, como se requería.

Existen muchas formas de ver que $\mathsf{C} = 2^{\aleph_0}$. Para demostrar que $2^{\aleph_0} \leq \mathsf{C}$ (que es realmente todo lo que ahora necesitamos para $\mathsf{C} > \aleph_0$), es suficiente con establecer que existe una correspondencia 1–1 entre $2^{\mathbb{N}}$ y algún *subconjunto* de $\mathbb{R}$. Podemos pensar que cada elemento de $2^{\mathbb{N}}$ es una asignación de 0 o de 1 («fuera» o «dentro») a cada número natural, i.e., uno de tales elementos puede considerarse como una secuencia infinita, tal como

100110001011101...

(Este elemento particular de $2^{\mathbb{N}}$ asigna 1 al número natural 0, asigna 0 al número natural 1, asigna 0 al número natural 2, asigna 1 al número natural 3, etc.) Ahora podríamos tratar de leer esta secuencia completa de dígitos como el desarrollo binario de cierto número real, donde consideramos que hay una coma decimal situada en el extremo izquierdo. Por desgracia, esto no funciona del todo, debido al irritante hecho de que existe una ambigüedad en algunas de tales representaciones, a saber, aquellas que acaban con una secuencia infinita que consta enteramente de 0s o enteramente de 1s.[16.11] Podemos sortear esta dificultad mediante varios artificios burdos. Uno de estos consistiría en intercalar entre los dígitos binarios otro dígito, por ejemplo el 3, para obtener

.31303031313030303130313131303 1...,

y leer entonces este número como la expresión decimal ordinaria de cierto número real. En consecuencia, hemos establecido realmente una correspondencia 1–1 entre $2^{\mathbb{N}}$ y cierto subconjunto de $\mathbb{R}$ (a saber, el subconjunto cuyos desarrollos decimales tienen esta singular forma intercalada). De ahí que $2^{\aleph_0} \leq \mathsf{C}$ (y ahora obtenemos $\mathsf{C} > \aleph_0$), como se requería.

Para deducir que $\mathsf{C} = 2^{\aleph_0}$, tenemos que ser capaces de demostrar que $\mathsf{C} \leq 2^{\aleph_0}$. Ahora todo número real estrictamente entre 0 y 1 tiene un desarrollo binario (como se ha considerado antes), aunque a veces de forma redundante; así pues, dicho conjunto concreto de reales tiene cardinalidad $\leq 2^{\aleph_0}$. Existen muchas funciones simples que llevan este

[16.11] Explíquelo.

intervalo a la totalidad de $\mathbb{R}$,[16.12] lo que establece que $\mathsf{C} \leq 2^{\aleph_0}$, y de ahí $\mathsf{C} = 2^{\aleph_0}$, como se requería.

La versión original que dio Cantor del argumento era algo diferente de la que se ha presentado aquí, aunque las ideas esenciales eran las mismas. Su versión original era también una demostración por contradicción, aunque más directa. Se imaginaba una hipotética correspondencia 1–1 entre $\mathbb{N}$ y los números reales comprendidos estrictamente entre 0 y 1, y se presentaba como un listado vertical de todos los números reales, cada uno de ellos escrito en desarrollo decimal. Se obtenía una contradicción con la hipótesis de que la lista es completa por un «argumento diagonal» mediante el cual se construye un nuevo número real que no está en la lista: para ello se desciende por la diagonal principal de la matriz, partiendo de la esquina superior izquierda, y se escribe un dígito que difiera del n-ésimo dígito del n-ésimo número real de la lista. (Hay muchas exposiciones divulgativas de esto; véase, por ejemplo, la versión que se da en el capítulo 3 de mi libro *La nueva mente del emperador*).[16.13] Este tipo general de argumento (incluido el que he utilizado al principio de esta sección para demostrar $\alpha < 2^{\alpha}$) se suele conocer como el «corte diagonal» de Cantor.

16.5. Enigmas en los fundamentos de las matemáticas

Como se ha comentado antes, la cardinalidad, $2^{\aleph_0}$, del continuo (i.e., de $\mathbb{R}$) se suele denotar por la letra C. A Cantor le hubiera gustado poder etiquetarla «$\aleph_1$», por lo que él entendía el «siguiente cardinal más pequeño» después de $\aleph_0$. Intentó demostrar, aunque fracasó, que $2^{\aleph_0} = \aleph_1$; de hecho, la afirmación «$2^{\aleph_0} = \aleph_1$», conocida como la *hipótesis del continuo*, se convirtió en una famosa cuestión no resuelta durante muchos años después de que Cantor la propusiera. Aún sigue sin estar resuelta, en un sentido «absoluto». Kurt Gödel y Paul Cohen demostraron que la hipótesis del continuo (y también el axioma de elección) no es de-

[16.12] Muestre una. *Sugerencia:* Considere la Fig. 9.8, por ejemplo.
[16.13] Explique por qué este es esencialmente el mismo argumento que el que he dado aquí, en el caso $\alpha = \aleph_0$ para demostrar $\alpha < 2^{\alpha}$.

cidible por medio de la teoría de conjuntos estándar. Sin embargo, debido al teorema de la incompletitud de Gödel, que comentaré enseguida, y varias cuestiones relacionadas, esto no resuelve propiamente la cuestión de la *verdad* de la hipótesis del continuo. Sigue siendo posible que métodos de demostración más poderosos que los de la teoría de conjuntos estándar puedan ser capaces de decidir la verdad o falsedad de la hipótesis del continuo; por otra parte, podría darse el caso de que su verdad o falsedad fuera una cuestión subjetiva dependiente del punto de vista que se adopte.[8] Esta cuestión ha sido mencionada en §1.3, pero en relación con el axioma de elección, y no con la hipótesis del continuo.

Vemos que la relación $\alpha < 2^{\alpha}$ nos dice que no puede haber ningún infinito máximo; pues si se propusiera que algún número cardinal Ω fuera el máximo, entonces el número cardinal 2^{Ω} resultaría aún mayor. Este hecho (y el argumento de Cantor que lo establece) ha tenido consecuencias trascendentales para los fundamentos de las matemáticas. En particular, el filósofo Bertrand Russell, que previamente era de la opinión de que debía haber un número cardinal máximo (a saber, el de la clase de todas las clases), había desconfiado de la conclusión de Cantor; pero cambió de opinión alrededor de 1903, tras estudiarla en detalle. En efecto, aplicó el argumento de Cantor al «conjunto de todos los conjuntos», ¡lo que le llevó inmediatamente a la ahora famosa «paradoja de Russell»!

Esta paradoja procede del siguiente modo. Consideremos el conjunto $\mathcal{R}$ que consiste en «todos los conjuntos que no son miembros de sí mismos». (Por el momento, no importa si uno está dispuesto a creer que un conjunto pueda ser miembro de sí mismo. Si ningún conjunto pertenece a sí mismo, entonces $\mathcal{R}$ es el conjunto de *todos* los conjuntos.) Planteemos la pregunta: ¿qué pasa con el propio $\mathcal{R}$? ¿Es $\mathcal{R}$ un miembro de sí mismo? Supongamos que lo es. Entonces, puesto que pertenece al conjunto $\mathcal{R}$ de conjuntos que *no* son miembros de sí mismos, no pertenece a sí mismo después de todo: ¡una contradicción! La hipótesis alternativa es que no pertenece a sí mismo. Pero, entonces, debe ser un miembro de la familia de los conjuntos que no son miembros de sí mismos, a saber, el conjunto $\mathcal{R}$. Así pues, $\mathcal{R}$ pertenece a $\mathcal{R}$, lo que contradice la hipótesis de que no pertenece a sí mismo. ¡Lo cual es una clara contradicción!

Puede advertirse que esto es simplemente lo que sucede con la demostración de Cantor $\alpha < 2^{\alpha}$, si se aplica al caso en que se toma α como el «conjunto de todos los conjuntos».[16.14] De hecho, así es como Russell llegó a su paradoja.[9] Lo que el argumento está demostrando realmente es que no hay nada semejante al «conjunto de todos los conjuntos». (De hecho, Cantor ya era consciente de esto, y conocía la «paradoja de Russell unos años antes que el propio Russell.)[10] Puede parecer extraño que algo tan sencillo como el «conjunto de todos los conjuntos» sea un concepto prohibido. Se podría imaginar que cualquier propuesta para un conjunto debería ser perfectamente aceptable si existe una regla bien definida que nos diga cuándo algo pertenece a aquel y cuándo no. Aquí parece que ciertamente existe tal regla, a saber, ¡que *todo* conjunto está en él! La trampa parece residir en que estamos admitiendo el mismo estatus para esta magnífica colección que el que damos a cada uno de sus miembros, a saber, llamar simplemente «conjunto» a ambos tipos de colecciones. El argumento general depende de que tengamos una idea clara de qué es realmente un conjunto. Y una vez que tengamos tal idea, surge la pregunta: ¿debe realmente contarse como un conjunto la propia colección de todos estos objetos? Lo que Cantor y Russell nos han dicho es que ¡la respuesta a esta pregunta debe ser no!

De hecho, la forma en que los matemáticos han llegado a entender esta situación aparentemente paradójica consiste en imaginar que se ha hecho algún tipo de distinción entre «conjuntos» y «clases». (Pensemos en las clases como grandes objetos rebeldes de los que no se supone que se afilian a clubes, mientras que los conjuntos se consideran siempre suficientemente respetables para que sí lo hagan.) Hablando con crudeza, a una colección de conjuntos cualesquiera se le podría considerar como una totalidad, y a semejante colección se la denominaría una *clase*. Algunas clases son lo bastante respetables para que a ellas mismas se las considere como conjuntos, pero otras clases se consideran «demasiado grandes» o «demasiado revueltas» para que cuenten como conjuntos. Por otra parte, no estamos autorizados necesariamente para reunir *clases* para formar entidades mayores. Así pues, el «conjunto de

[16.14] Demuestre que es esto lo que sucede.

todos los conjuntos» no está permitido (ni está permitida la «clase de todas las clases»), pero la «clase de todos los conjuntos» sí se considera legítima. Cantor denotó a esta clase suprema por Ω, y le atribuyó un significado casi deístico. No se nos permite formar clases más grandes que Ω. El problema con «2^{Ω}» residiría en que implica «reunir» todas las diferentes «subclases» de Ω, muchas de las cuales no son propiamente conjuntos, de modo que esto no está permitido.

En todo ello hay algo que parece bastante insatisfactorio. Tengo que confesar que yo mismo me siento insatisfecho. Este procedimiento podría ser razonable si hubiera un criterio claro y tajante que nos dijera cuándo una clase está realmente cualificada para ser un conjunto. Sin embargo, la «distinción» parece estar hecha a menudo de un modo muy circular. Se estima que una clase es un conjunto si y solo si puede ser ella misma un miembro de alguna otra clase, lo que ¡me parece una petición de principio! El problema está en que no hay ningún lugar obvio en donde trazar la línea divisoria. Una vez que se ha trazado una línea, no pasa mucho tiempo sin que parezca que la línea se ha trazado de una forma muy restringida. No parece que haya ninguna razón para no incluir algunas clases mayores (o más rebeldes) en nuestro club de conjuntos. Por supuesto, se tiene que evitar una contradicción flagrante; pero resulta que cuanto más liberales son las reglas de admisión al club de conjuntos, más potentes son los métodos de demostración matemática que proporciona el concepto de conjunto. ¡Pero si abrimos una rendija suficientemente ancha en la puerta de este club, se producirá el desastre —¡CONTRADICCIÓN!—, ¡y el edificio entero se vendrá abajo! El trazado de la línea es uno de los procedimientos más difíciles y delicados de las matemáticas.[11]

Muchos matemáticos preferirían retroceder ante un liberalismo tan extremo, adoptando incluso un enfoque «constructivista» rígidamente conservador, según el cual solo se admite un conjunto si existe una construcción directa que nos permita decir cuándo un elemento pertenece al conjunto y cuándo no. ¡Ciertamente, el axioma de elección sería un criterio inadmisible bajo reglas tan estrictas! Pero el caso es que estos conservadores extremos no son más inmunes al corte diagonal de Cantor que los liberales extremos. Tratemos de ver dónde está la dificultad.

16.6. Las máquinas de Turing y el teorema de Gödel

En primer lugar, necesitamos una noción de lo que significa «construir» algo en matemáticas. Es mejor que restrinjamos la atención a subconjuntos del conjunto $\mathbb{N}$ de los números naturales, al menos para nuestras consideraciones presentes. ¿Podemos pedir que tales subconjuntos estén definidos «constructivamente»? Por fortuna, tenemos a nuestra disposición una noción maravillosa introducida por varios lógicos[12] del primer tercio del siglo XX y establecida sobre una base clara por Alan Turing en 1936. Se trata de la noción de *computabilidad*; y puesto que los computadores electrónicos se han hecho ahora tan familiares para nosotros, bastará probablemente con que me refiera a las acciones de estos dispositivos físicos en lugar de presentar las ideas relevantes en términos de alguna formulación matemática precisa. En términos aproximados, una *computación* (o *algoritmo*) es lo que realizaría un computador idealizado, donde «idealizado» significa que puede seguir funcionando sin «gastarse» durante un período de tiempo indefinido, que nunca comete errores, y que tiene un espacio de almacenamiento ilimitado. Matemáticamente, una entidad semejante es lo que, en efecto, se denomina una *máquina de Turing*.[13]

Cualquier máquina de Turing particular $\boldsymbol{T}$ corresponde a una computación específica que puede llevarse a cabo sobre los números naturales. La acción de $\boldsymbol{T}$ sobre el número natural concreto n se escribe $\boldsymbol{T}(n)$, y normalmente realizamos esta acción para dar algún (otro) número natural m:

$$\boldsymbol{T}(n) = m.$$

Ahora una máquina de Turing podría tener la propiedad de que se queda «bloqueada» (o «entra en un bucle») porque la computación que está ejecutando no termina nunca. Diré que una máquina de Turing es *defectuosa* si no termina su ejecución cuando se aplica a algún número natural n. La llamaré *efectiva* si, por el contrario, siempre termina, cualquiera que sea el número que se le presenta.

Un ejemplo de una máquina de Turing $\boldsymbol{T}$ que nunca termina (defectuosa) sería la que, cuando se le presenta n, trata de encontrar el número natural más pequeño que no es la suma de n cuadrados ($0^2 = 0$

incluido). Encontramos $\boldsymbol{T}(0) = 1$, $\boldsymbol{T}(1) = 2$, $\boldsymbol{T}(2) = 3$, $\boldsymbol{T}(3) = 7$ (que significa, por ejemplo, en el caso de la última ecuación: «7 es el número más pequeño que no es la suma de 3 cuadrados»),[16.15] pero cuando $\boldsymbol{T}$ se aplica a 4, sigue computando para siempre, tratando de encontrar un número que no sea la suma de cuatro cuadrados. La causa de que nuestra máquina se cuelgue es un famoso teorema que se debe al gran matemático franco-italiano del siglo XVIII Joseph L. Lagrange, que fue capaz de demostrar que todo número natural es la suma de cuatro cuadrados. (Más adelante, Lagrange tendrá una gran importancia para nosotros en un contexto diferente, ¡muy especialmente en los capítulos 20 y 26!)

Cada máquina de Turing individual (ya sea defectuosa o efectiva) tiene cierta «tabla de instrucciones» que caracteriza el algoritmo concreto que ejecuta esta máquina de Turing concreta. Dicha tabla de instrucciones puede especificarse completamente mediante un «código», que podemos escribir como una secuencia de dígitos. Podemos reinterpretar esta secuencia como un número natural t; así pues, t codifica el «programa» que permite que la máquina lleve a cabo su algoritmo concreto. La máquina de Turing que está codificada por el número natural t será denotada por $\boldsymbol{T}_t$. Quizá la codificación no funcione para todos los números naturales t; si no lo hace por alguna razón, entonces podemos referirnos a $\boldsymbol{T}_t$ como «defectuosa», además de aquellos casos recién considerados en los que la máquina no se detiene cuando se aplica a algún n. Las únicas máquinas de Turing efectivas $\boldsymbol{T}_t$ son aquellas que dan una respuesta al cabo de un tiempo finito, cuando se aplican a cualquier n individual.

Una de las hazañas fundamentales de Turing consistió en darse cuenta de que es posible especificar una única máquina de Turing, llamada una máquina de Turing *universal* $\boldsymbol{U}$, que puede imitar la acción de cualquier máquina de Turing. Todo lo que se necesita es que $\boldsymbol{U}$ actúe primero sobre el número natural t, que especifica la máquina de Turing concreta $\boldsymbol{T}_t$ que tiene que ser imitada, después de lo cual $\boldsymbol{U}$ actúa sobre el número n, de modo que puede proceder a evaluar $\boldsymbol{T}_t(n)$. (Los mo-

[16.15] Dé una descripción aproximada de cómo podría realizarse nuestro algoritmo y explique estos valores concretos.

dernos computadores de propósito general son, en esencia, tan solo máquinas de Turing universales.) Escribiré esta acción combinada en la forma $U(t, n)$, de modo que

$$U(t, n) = T_t(n).$$

Sin embargo, deberíamos tener en cuenta que se supone que las máquinas de Turing definidas aquí actúan solo sobre un único número natural, y no sobre un par tal como (t, n). Pero no es difícil codificar un par de números naturales como un único número natural, como hemos visto antes (por ejemplo, en el ejercicio [16.8]). La propia máquina U estará definida por algún número natural, digamos u, de modo que tenemos

$$U = T_u.$$

¿Cómo podemos decir si una máquina de Turing es efectiva o defectuosa? ¿Podemos encontrar algún algoritmo para tomar esta decisión? Uno de los logros importantes de Turing consistió en demostrar que ¡la respuesta a esta pregunta es «no»! La demostración es una aplicación del corte diagonal de Cantor. Consideraremos el conjunto $\mathbb{N}$, como antes, pero ahora en lugar de considerar *todos* los subconjuntos de $\mathbb{N}$, consideramos solo aquellos subconjuntos para los cuales es una cuestión *computacional* el decidir si un elemento está o no en el conjunto. (Estos no pueden ser todos los subconjuntos de $\mathbb{N}$, puesto que el número de computaciones diferentes es solo $\aleph_0$, mientras que el número de todos los subconjuntos de $\mathbb{N}$ es $\mathbf{C}$.) Tales conjuntos definidos computacionalmente se denominan *recursivos*. De hecho, cualquier subconjunto recursivo de $\mathbb{N}$ está definido por la salida de una máquina de Turing efectiva T, del tipo particular que solo da 0 o 1. Si $T(n) = 1$, entonces n es un miembro del conjunto recursivo definido por T («dentro»), mientras que si $T(n) = 0$, entonces n no es un miembro («fuera»). Ahora aplicamos el argumento de Cantor igual que antes, pero solo a los subconjuntos *recursivos* de $\mathbb{N}$. El argumento nos dice enseguida que el conjunto de los números naturales t para los que T_t es efectiva no puede ser recursivo. ¡No hay algoritmo, aplicable a todas las máquinas de Turing T, que nos diga si T es o no defectuosa!

Vale la pena detenerse en este razonamiento un poco más detalla-

damente. Lo que el argumento de Turing/Cantor demuestra en realidad es que el conjunto de los t para los que T_t es efectiva no es siquiera *recursivamente numerable*. ¿Qué es un subconjunto de $\mathbb{N}$ recursivamente numerable? Es un conjunto de números naturales para los que existe una máquina de Turing efectiva T que, con el tiempo, genera cada miembro (posiblemente más de una vez) de dicho conjunto cuando se aplica a 0, 1, 2, 3, 4, ... sucesivamente. (Es decir, m es un miembro del conjunto si y solo si $m = T(n)$ para algún número natural n.) Un subconjunto S de $\mathbb{N}$ es recursivo si y solo si es recursivamente numerable y su *complemento* $\mathbb{N} - S$ es también recursivamente numerable.[16.16] La supuesta correspondencia 1–1 con la que el argumento de Turing/Cantor obtiene una contradicción es una numeración recursiva de las máquinas de Turing efectivas. Una pequeña reflexión nos dice que lo que hemos aprendido es que no existe ningún algoritmo general que nos diga cuándo la acción de una máquina de Turing $T_t(n)$ *no llegará* a detenerse.

Lo que esto nos dice en definitiva es que a pesar de las esperanzas que se pudieran haber tenido en una postura de «conservadurismo extremo», en la que los únicos conjuntos aceptables serían aquellos —los conjuntos recursivos— cuya admisión está determinada por reglas computacionales tajantes, este punto de vista nos conduce inmediatamente a tener que considerar conjuntos que son no recursivos. Este punto de vista tropieza incluso con la dificultad fundamental de que no existe ningún modo computacional de decidir *en general* si dos conjuntos recursivos son el mismo o son conjuntos diferentes, si están definidos por dos máquinas de Turing efectivas diferentes T_t y T_s.[16.17] Además, este tipo de problema reaparece una y otra vez en diferentes niveles, cuando tratamos de restringir nuestra noción de «conjunto» con un punto de vista demasiado conservador. Siempre nos vemos llevados a considerar clases que no pertenecen a nuestra familia de conjuntos previamente admitidos.

Estas cuestiones están estrechamente relacionadas con el famoso teo-

[16.16] Demuéstrelo.

[16.17] ¿Puede ver por qué es así? *Sugerencia*: En el caso de la acción de una máquina de Turing arbitraria T aplicada a n, podemos considerar una máquina de Turing efectiva Q que tiene la propiedad de que $Q(r) = 0$ si T aplicada a n no se ha detenido al cabo de r pasos computacionales y $Q(r) = 1$ si lo ha hecho. Tome la suma mod 2 de $Q(n)$ y $T_t(n)$ para obtener $T_s(n)$.

rema de Kurt Gödel. Este estaba interesado en la cuestión de los métodos de demostración que están disponibles para los matemáticos. Desde comienzos del siglo XX, y durante muchos años después, los matemáticos habían intentado evitar las paradojas (tales como la paradoja de Russell) que surgían cuando se hacía un uso excesivamente liberal de la teoría de conjuntos, para lo que introdujeron la idea de un *sistema matemático formal*, de acuerdo con el cual había que establecer una serie de reglas absolutamente precisas acerca de qué líneas de razonamiento iban a contar como demostración matemática. Lo que Gödel demostró es que este programa no funcionaría. Demostró, en efecto, que si estamos dispuestos a aceptar que las reglas de un sistema formal semejante $\boldsymbol{F}$ deben ser fiables y darnos solo conclusiones matemáticamente correctas, entonces también debemos aceptar como correcto cierto enunciado matemático preciso $G(\boldsymbol{F})$, aunque concluyendo que $G(\boldsymbol{F})$ no es demostrable por los métodos de $\boldsymbol{F}$ solamente. Así, Gödel nos demuestra cómo trascender cualquier $\boldsymbol{F}$ en el que estamos dispuestos a confiar.

Existe la idea equivocada de que el teorema de Gödel nos dice que existen «proposiciones matemáticas indemostrables», y que esto implica que existen regiones del «mundo platónico» de verdades matemáticas (véase §1.4) que, en principio, nos son inaccesibles. Esto está muy lejos de la conclusión que deberíamos sacar del teorema de Gödel. Lo que Gödel realmente nos dice es que cualesquiera que sean las reglas de demostración que hayamos establecido por adelantado, si ya aceptamos que dichas reglas son dignas de confianza (i.e., que no nos permiten deducir falsedades) y no son demasiado limitadas, entonces disponemos de un nuevo medio de acceso a ciertas verdades matemáticas para cuya deducción aquellas reglas particulares no son lo bastante potentes.

El resultado de Gödel se sigue del de Turing (aunque históricamente las cosas sucedieron al revés). ¿Cómo funciona esto? El punto importante acerca de un sistema formal es que no se necesita ningún juicio matemático adicional para comprobar si se han aplicado correctamente las reglas de $\boldsymbol{F}$. Decidir la corrección de una demostración matemática de acuerdo con $\boldsymbol{F}$ tiene que ser una cuestión completamente computacional. Observamos que, para cualquier $\boldsymbol{F}$, el conjunto

de teoremas matemáticos que pueden ser demostrados utilizando sus reglas es necesaria y recursivamente numerable.

Ahora algunos enunciados matemáticos bien conocidos pueden expresarse de la forma «la acción de tal y cual máquina de Turing no termina». Ya hemos visto un ejemplo, a saber, el teorema de Lagrange según el cual todo número natural es la suma de cuatro cuadrados. Otro ejemplo aún más famoso es el «último teorema de Fermat», demostrado a finales del siglo XX por Andrew Wiles (§1.3).[14] Otro ejemplo (aunque no resuelto) es la bien conocida «conjetura de Goldbach», según la cual todo número par mayor que dos es la suma de dos números primos. Los enunciados de esta naturaleza son conocidos por los lógicos matemáticos como Π_1-*sentencias*. Se sigue ahora inmediatamente del anterior argumento de Turing que la familia de Π_1-sentencias verdaderas constituye un conjunto no recursivamente numerable. Así pues, existen Π_1-sentencias verdaderas que no pueden obtenerse a partir de las reglas de $\boldsymbol{F}$ (donde suponemos que $\boldsymbol{F}$ es fiable). Esta es la forma básica del teorema de Gödel. De hecho, examinando los detalles de este con más atención, podemos afinar el argumento de modo que obtenemos la versión del mismo que hemos establecido antes, y obtener una Π_1-sentencia específica $G(\boldsymbol{F})$, que, si creemos que $\boldsymbol{F}$ proporciona solo Π_1-sentencias verdaderas, debe escapar de la red arrojada por $\boldsymbol{F}$ a pesar del hecho notable de que debemos concluir que $G(\boldsymbol{F})$ es también una Π_1-sentencia verdadera.[16.18]

16.7. Tamaños de infinitos en física

Por último, veamos qué relación guardan estas cuestiones del infinito y la constructibilidad con las matemáticas de nuestros capítulos anteriores y con nuestra comprensión actual de la física. A la vista de la estrecha relación entre matemáticas y física, es quizá notable que cuestiones de una importancia tan básica en matemáticas como la teoría de conjuntos transfinitos y de la computabilidad solo hayan tenido por el momento un impacto limitado sobre nuestra descripción del mundo físi-

[16.18] Vea si puede establecerlo.

co. Mi opinión personal es que con el tiempo descubriremos que las cuestiones de computabilidad tienen una profunda relevancia para la teoría física futura,[15] aunque por el momento estas ideas se hayan utilizado muy poco en física matemática.[16]

Con respecto al *tamaño* de los infinitos que han encontrado valor, resulta bastante sorprendente que solo una pequeña parte de la teoría física parezca necesitar ir más allá de $\mathsf{C}(= 2^{\aleph_0})$, la cardinalidad del sistema de los números reales $\mathbb{R}$. La cardinalidad del campo complejo $\mathbb{C}$ es la misma que la de $\mathbb{R}$ (a saber, C), puesto que $\mathbb{C}$ es simplemente $\mathbb{R} \times \mathbb{R}$ (pares de números reales) con ciertas leyes de suma y multiplicación definidas en él. Análogamente, los espacios vectoriales y las variedades que hemos considerado están construidos a partir de familias de puntos a los que se les pueden asignar coordenadas a partir de un $\mathbb{R} \times \mathbb{R} \times \ldots \times \mathbb{R}$ (o $\mathbb{C} \times \mathbb{C} \times \ldots \times \mathbb{C}$) o a partir de cartas coordenada finitas (o infinitas numerables, i.e., en número $\aleph_0$), y una vez más la cardinalidad es C.

¿Qué sucede con las familias de funciones en tales espacios? Si consideramos, pongamos por caso, la familia de todas las funciones con valores reales en cierto espacio con C puntos, entonces encontramos, a partir de las consideraciones anteriores, que la familia tiene C^{C} miembros (al ser aplicaciones de un espacio de C elementos en un espacio de C elementos). Esto es ciertamente mayor que C. De hecho, $\mathsf{C}^{\mathsf{C}} = 2^{\mathsf{C}}$. (Esto se sigue de que cada elemento de $\mathbb{R}^{\mathbb{R}}$ puede ser reinterpretado como un elemento particular de $2^{\mathbb{R}\times\mathbb{R}}$, a saber, como una sección transversal (normalmente lejos de ser continua) del fibrado $\mathbb{R} \times \mathbb{R}$, y la cardinalidad de $\mathbb{R} \times \mathbb{R}$ es C.) Sin embargo, las funciones (o los campos tensoriales, o las conexiones) reales (o complejas) *continuas* en una variedad son solo C en número, porque una función continua está determinada una vez que se conocen sus valores en el conjunto de puntos con coordenadas racionales. El número de estos es precisamente $\mathsf{C}^{\aleph_0}$, puesto que el número de puntos con coordenadas racionales es precisamente $\aleph_0$. Pero $\mathsf{C}^{\aleph_0} = (2^{\aleph_0})^{\aleph_0} = 2^{\aleph_0\times\aleph_0} = 2^{\aleph_0} = \mathsf{C}$.[16.19] En §§6.4,6 hemos considerado ciertas generalizaciones de las funciones continuas, que llevan a la gran generalización conocida como *hiperfunciones* (§9.7). Sin embargo, una vez más el número de estas no es mayor que C, puesto que están definidas por pares de funciones holomorfas (cada una de ellas en número C).

[16.19] Explique por qué $(A^B)^C$ puede identificarse con $A^{B\times C}$, para conjuntos A, B, C.

En §22.3 veremos que la teoría cuántica requiere el uso de ciertos espacios, conocidos como *espacios de Hilbert*, que pueden tener infinitas dimensiones. Sin embargo, aunque estos espacios concretos de dimensión infinita difieren de manera significativa de los espacios de dimensión finita, no hay en ellos más funciones continuas que en el caso de dimensión finita, y de nuevo tenemos que el número total es C. La mejor apuesta para llegar más alto que esto está relacionada con la formulación de integrales de camino de la teoría cuántica de campos (como se expondrá en §26.6), cuando se considera un espacio de curvas de apariencia incontrolada (o de configuraciones de campos físicos de apariencia incontrolada) en el espaciotiempo. Sin embargo, parece que seguimos obteniendo C para el número total, puesto que, a pesar de su incontrolabilidad, hay un resto suficiente de continuidad en estas estructuras.

La noción de cardinalidad no parece suficientemente refinada para recoger el concepto apropiado de *tamaño* para los espacios que encontramos en física. Casi todos los espacios de importancia tienen simplemente C puntos en ellos. Sin embargo, existe una gran diferencia entre los «tamaños» de dichos espacios, que en primera instancia consideramos simplemente como la dimensión del espacio vectorial o variedad $\mathcal{M}$ en consideración. Esta dimensión de $\mathcal{M}$ puede ser un número natural (por ejemplo, 4, en el caso del espaciotiempo ordinario, o 6×10^{19}, en el caso del espacio de fases considerado en §12.1), o podría ser infinito, tal como sucede con (la mayoría de) los espacios de estados de Hilbert que aparecen en mecánica cuántica. Desde el punto de vista matemático, el espacio de Hilbert más simple de dimensión infinita es el espacio de sucesiones $(z_1, z_2, z_3, \ldots)$ de números complejos para las cuales la suma infinita $|z_1|^2 + |z_2|^2 + |z_3|^2 + \ldots$ converge. En el caso de un espacio de Hilbert de dimensión infinita, es más apropiado pensar que esta dimensionalidad es $\aleph_0$. (Existen varios puntos sutiles acerca de esto, pero es mejor que no entremos en ello ahora.) Para un espacio n-real dimensional, diré que tiene «∞^n» puntos (lo que expresa

que este continuo de puntos está organizado en una disposición n-dimensional). En el caso de dimensión infinita, me referiré a esto como «∞^{∞}» puntos.

También estamos interesados en los espacios de varios tipos de campos definidos en $\mathcal{M}$. Normalmente estos se toman suaves, aunque a veces son más generales (por ejemplo, distribuciones) y entran dentro del ámbito de la teoría de hiperfunciones (véase §9.7). Pueden estar sujetos a ecuaciones diferenciales (en derivadas parciales) que restringen su libertad. Si no están restringidos de este modo, entonces cuentan como «funciones de n variables», para una $\mathcal{M}$ n-dimensional (donde $n = 4$ para el espaciotiempo estándar). En cada punto el campo puede tener k componentes independientes. Entonces diré que la libertad en el campo es $\infty^{k\infty^n}$. La explicación para esta notación[17] está en que los campos pueden considerarse (cruda y localmente) como aplicaciones de un espacio con ∞^n puntos en un espacio con ∞^k puntos, y nos aprovecharemos de la relación notacional (formal)

$$(\infty^k)^{\infty^n} = \infty^{k\infty^n}.$$

Cuando los campos están restringidos por ecuaciones en derivadas parciales apropiadas, entonces puede darse el caso de que estén completamente determinados por las *condiciones iniciales* de los campos (véase §27.1 en particular), es decir, por algunos datos subsidiarios del campo especificados en cierto espacio $\mathcal{S}$ de menor dimensión, por ejemplo, de q dimensiones. Si los datos pueden expresarse *libremente* en $\mathcal{S}$ (lo que significa, básicamente, que no están sujetos a *ligaduras*, que son ecuaciones diferenciales o algebraicas que los datos tendrían que satisfacer en $\mathcal{S}$), y si estos datos constan de r-componentes independientes en cada punto de $\mathcal{S}$, entonces diré que la libertad en el campo es $\infty^{r\infty^q}$. En muchos casos, no es nada fácil encontrar r y q, pero lo importante es que parecen ser cantidades *invariantes*, independientes de cómo puedan reexpresarse los campos en términos de otras cantidades equivalentes.[18] Estas cuestiones cobrarán gran importancia más adelante (cf. §23.2 y §§31.10-12, 15-17).

Notas

Sección 16.2

16.1. Véanse Stephenson (1972), §7; Howie (1989), pp. 269-271, y Hirschfeld (1998); los discos mágicos son equivalentes a lo que se denominan *conjuntos con diferencia perfecta.*

16.2. Al parecer se desconoce si existen discos mágicos (que no surgen necesariamente de un $\mathbb{P}^3(\mathbb{F}_q)$) para los que el teorema de Desargues (o, de manera equivalente, el de Pappos) siempre falla —o, en realidad, si existen planos proyectivos finitos no desarguianos (equivalentemente, no pappianos).

16.3. De todas formas, a veces se ha sugerido un papel físico para los octoniones (véanse, por ejemplo, Gürsey y Tze, 1996; Dixon, 1994; Manogue y Dray, 1999, y Dray y Manogue, 1999); pero hay dificultades fundamentales para la construcción de una «mecánica cuántica octoniónica» general (Adler, 1995), y la situación con respecto a una «mecánica cuántica cuaterniónica» es solo un poco más positiva. Otro sistema de números, sugerido en una ocasión como candidato para un papel físico importante, es el de los «números p-ádicos». Estos constituyen sistemas de números a los que se aplican las reglas del cálculo infinitesimal, y pueden expresarse como números reales ordinarios desarrollados de forma decimal, salvo que los dígitos representan 0, 1, 2, 3, ..., $p-1$ (donde p se escoge primo) y se permite que sean infinitos *en el sentido contrario* de lo que sucede en el caso de los decimales ordinarios (y no necesitamos signo menos). Por ejemplo,

..... 24033200411,3104

representa un número 5-ádico concreto. Las reglas para sumar y multiplicar son las mismas que para la aritmética p-aria «ordinaria» (en la que el símbolo «10» representa el primo p, etc.). Véanse Mahler (1981), Gouvea (1993), Brekke y Frend (1993), Vladimirov y Volovich (1989) y Pitkäenen (1995).

Sección 16.3

16.4. La terminología matemática moderna llama a esto un isomorfismo. Hay otras palabras tales como «endomorfismo», «epimorfismo» y «monomorfismo» (o, simplemente, «morfismo») que los matemáticos tienden a utilizar en un contexto general para caracterizar aplicaciones en-

tre un conjunto o estructura y otro. Prefiero evitar este tipo de terminología en este libro, pues creo que para acostumbrarse a esto se necesita un esfuerzo excesivo.

16.5. Para algunas elucubraciones aun anteriores de esta naturaleza, véase Moore (1990), capítulo 3.

16.6. Recordemos de la nota 15.5 que he estado dispuesto a adoptar un abuso de notación donde $\mathbb{N} - 0$ representa el conjunto de números naturales distintos de cero. Aquí se da la ironía de que si se adoptara el aparentemente «más correcto» $\mathbb{N} - \{0\}$, aunque adoptando también los procedimientos de §3.4, donde $0 = 1$, ¡deberíamos acabar en el incluso más confuso «$\mathbb{N} - 1$» para el conjunto bajo consideración!

16.7. Véanse Wagon (1985) y para una exposición divulgativa, Runde (2002).

Sección 16.5

16.8. Comentarios similares se aplican a la *hipótesis del continuo generalizada* de Cantor: $2^{\aleph_\alpha} = \aleph_{\alpha+1}$ (donde α es ahora un «número ordinal», cuya definición no he discutido aquí), y estos comentarios se aplican también al axioma de elección.

16.9. Véase Russell (1903), p. 362, segunda nota final [en la edición de 1937].

16.10. Véase Van Heijenoort (1967), p. 114.

16.11. Véase Woodin (2001) para un nuevo enfoque de estas materias. Para referencias generales sobre los fundamentos de las matemáticas, véanse Abian (1965) y Wilder (1965).

Sección 16.6

16.12. Los precursores de Turing fueron, en general, Alonzo Church, Haskell B. Curry, Stephen Kleene, Kurt Gödel y Emil Post; véase Gandy (1988).

16.13. Para una descripción detallada de una máquina de Turing, véanse Penrose (1989), capítulo 2; por ejemplo, Davis (1978), o la referencia original, Turing (1937).

16.14. Véanse Singh (1997) y Wiles (1995).

Sección 16.7

16.15. Véase Penrose (1989, 1994, 1997a).

16.16. Véanse Komar (1964) y Geroch y Hartle (1986), §34.7.

16.17. Debo esta útil notación a John A. Wheeler; véase Wheeler (1960), p. 67.

16.18. Véase Cartan (1945), especialmente §§68,69, pp. 75 y 76 (edición original). Hay que tener cuidado en garantizar que la cantidad r en $\infty^{r\infty^s}$ se cuenta correctamente. Dos sistemas pueden ser equivalentes, pero tener valores r que parecen diferir a primera vista. Sin embargo, no puede haber ambigüedad en la determinación del valor de q. El tratamiento moderno riguroso de estas cuestiones hace las cosas más claras; viene dado en términos de la teoría de *fibrados en chorro* (véase Bryant *et al.*, 1991). Puede mencionarse que existe un refinamiento de la notación de Wheeler (véase Penrose, 2003), donde, por ejemplo, $\infty^{2\infty^2} + 3^{\infty^1} + 5$ significa «los campos dependen de 2 funciones de 2 variables, 3 funciones de una variable, y 5 constantes». Nos vemos así llevados a considerar expresiones como $\infty^{p(\infty)}$, donde p denota un polinomio con coeficientes enteros no negativos.

17

Espaciotiempo

17.1. El espaciotiempo de la física aristotélica

De ahora en adelante nuestra atención se volverá desde las consideraciones básicamente matemáticas que nos han ocupado en los capítulos anteriores a las imágenes reales del mundo físico a las que nos han llevado la teoría y la observación. Empecemos tratando de entender ese escenario en el que parecen tener lugar todos los fenómenos del universo físico: el *espaciotiempo*. ¡Encontraremos que esta noción desempeña un papel vital en casi todo lo que sigue en este libro!

Antes de nada, debemos preguntarnos: ¿por qué «espaciotiempo»?[1] ¿Qué hay de erróneo en considerar espacio y tiempo por separado, en lugar de intentar unificar estas dos nociones aparentemente muy diferentes? Pese a lo que parece ser una percepción común sobre esta cuestión, y pese al soberbio uso que hizo Einstein de esta idea en su formulación de la teoría de la relatividad general, el espaciotiempo no fue una idea original de Einstein ni, según parece, él se mostró muy entusiasmado cuando oyó hablar de ella por primera vez. Además, si echamos una ojeada retrospectivamente a las magníficas intuiciones «relativistas» más antiguas de Galileo y Newton, observamos que también ellos pudieron, en principio, haber sacado grandes ventajas de la perspectiva espaciotemporal.

Para entender esto, retrocedamos mucho más en la historia y tratemos de ver qué tipo de estructura espaciotemporal hubiera sido adecuada para el armazón dinámico de Aristóteles y sus contemporáneos. En la física aristotélica existe una noción de 3-espacio euclídeo $\mathbb{E}^3$ para representar el espacio físico, y los puntos de este espacio mantienen su

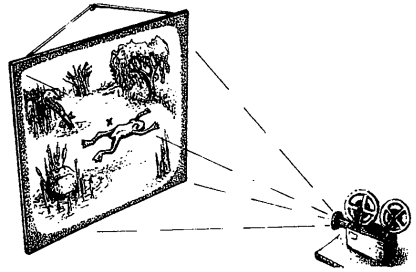


Fig. 17.1. ¿Es el movimiento físico similar al que se percibe en una pantalla de cine? Un punto concreto de la pantalla (aquí marcado «×») retiene su identidad independientemente de qué movimiento se proyecte sobre él.

identidad de un instante al siguiente. Esto se debe a que, en el esquema aristotélico, el estado de reposo es dinámicamente preferido a todos los demás estados de movimiento. Adoptamos la postura de que un punto espacial concreto, en un instante de tiempo, es el *mismo* punto espacial, en un instante de tiempo posterior, si una partícula situada en dicho punto se mantiene en reposo al pasar de un instante al siguiente. Nuestra imagen de la realidad es como la pantalla de una sala de cine, donde un punto concreto de la pantalla retiene su identidad independientemente de qué tipo de movimientos vigorosos pudieran proyectarse en ella. Véase la Fig. 17.1.

También el tiempo se representa como un espacio euclídeo, aunque uno bastante trivial, a saber, el espacio 1-dimensional $\mathbb{E}^1$. Así pues, consideramos el tiempo, tanto como el espacio físico, como una «geometría euclídea», en lugar de ser solo una copia de la recta real $\mathbb{R}$. Esto se debe a que $\mathbb{R}$ tiene un elemento privilegiado 0 que representaría el «cero» de tiempos, mientras que en nuestra visión dinámica «aristotélica» no tiene que haber ningún origen privilegiado. (En esto estoy adoptando una visión idealizada de lo que pudiera llamarse «dinámica aristotélica», o «física aristotélica», ¡y no adopto ningún punto de vista respecto a lo que el Aristóteles *real* pudiera haber pensado!)[2] Si existiera un «origen de tiempos» privilegiado, podría concebirse que las leyes dinámicas cambian a medida que transcurre el tiempo a partir de dicho origen privilegiado. Sin un origen privilegiado, las leyes deben seguir siendo las mismas todo el tiempo, porque no hay ningún *parámetro temporal* privilegiado del que dichas leyes puedan depender.

De modo análogo, estoy adoptando el punto de vista de que no existe ningún origen espacial privilegiado, y que el espacio se prolonga indefinidamente en todas direcciones, con completa uniformidad en las leyes dinámicas (una vez más, ¡con independencia de lo que pudiera haber creído el Aristóteles real!). En la geometría euclídea, ya sea 1-dimensional o 3-dimensional, existe una noción de *distancia*. En el caso espacial 3-dimensional, esta va a ser la distancia euclídea ordinaria (medida en metros, o en pies, pongamos por caso); en el caso 1-dimensional, esta distancia es el intervalo de tiempo ordinario (medido, digamos, en segundos).

En la física aristotélica —y en los esquemas dinámicos posteriores de Galileo y Newton— existe una noción absoluta de *simultaneidad* temporal. Así, de acuerdo con estos esquemas dinámicos, tiene un significado absoluto decir que el tiempo aquí, *en este preciso instante*, mientras escribo esto sentado en el despacho de mi casa en Oxford, es «el mismo tiempo» en el que tiene lugar cierto suceso en la galaxia Andrómeda (por ejemplo, la explosión de una supernova). Para volver a nuestra analogía con la pantalla de un cine, podemos preguntar si dos imágenes proyectadas que ocurren en dos lugares muy separados de la pantalla están teniendo lugar simultáneamente o no. La respuesta en este caso es evidente. Los sucesos deben considerarse simultáneos si y solo si ocurren en el mismo fotograma proyectado. Así pues, no solo tenemos una noción clara de si dos sucesos (temporalmente separados) ocurren o no en la misma localización espacial de la pantalla, sino que también tenemos una noción clara de si dos sucesos (espacialmente separados) ocurren o no al mismo tiempo. Además, si las localizaciones espaciales de los dos sucesos son diferentes, tenemos una noción clara de la *distancia* entre ellos, ya ocurran o no al mismo tiempo (i.e., la distancia medida a lo largo de la pantalla); asimismo, si los tiempos de los dos sucesos son diferentes, tenemos una clara noción del *intervalo temporal* entre ellos, ocurran o no en el mismo lugar.

Lo que esto nos dice es que en nuestro esquema aristotélico es apropiado considerar el *espaciotiempo* como el simple producto

$$\mathcal{A} = \mathbb{E}^1 \times \mathbb{E}^3,$$

que llamaré *espaciotiempo aristotélico*. Este es simplemente el espacio de pares $(t, \mathbf{x})$, donde t es un elemento de $\mathbb{E}^1$, un «instante de tiempo», y $\mathbf{x}$ es un elemento de $\mathbb{E}^3$, un «punto en el espacio». (Véase la Fig. 17.2.) Para dos puntos diferentes de $\mathbb{E}^1 \times \mathbb{E}^3$, digamos $(t, \mathbf{x})$ y $(t', \mathbf{x}')$ —i.e., dos *sucesos* diferentes—, tenemos una noción bien definida de su separación espacial, a saber, la distancia entre los puntos $\mathbf{x}$ y $\mathbf{x}'$ de $\mathbb{E}^3$, y tenemos también una noción bien definida de su diferencia temporal, a saber, la separación entre t y t' medida en $\mathbb{E}^1$. En particular, sabemos si dos sucesos ocurren o no en el mismo lugar (anulación del desplazamiento espacial) y si tienen lugar o no al mismo tiempo (anulación de la diferencia temporal).

17.2. El espaciotiempo para la relatividad galileana

Veamos ahora qué noción de espaciotiempo es apropiada para el esquema dinámico introducido por Galileo en 1638. Queremos incorporar el *principio de relatividad galileana* en nuestra imagen espaciotemporal. Tratemos de recordar lo que afirma este principio. No hay mejor manera de hacerlo que citar al propio Galileo (según una traducción de Stillman Drake[3] que doy aquí solo de forma abreviada; y recomiendo encarecidamente un análisis de la cita entera para quienes tengan acceso a ella):

> Encerraos con un amigo en la cabina principal bajo la cubierta de un barco grande, y llevad con vosotros moscas, mariposas, y otros pequeños animales voladores... colgad una botella que se vacíe gota a gota en un

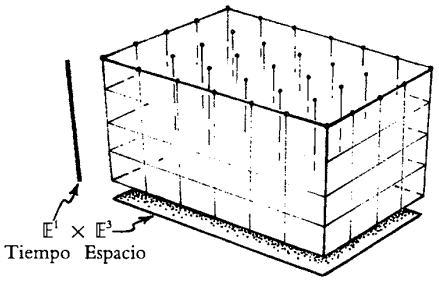


Fig. 17.2. El espaciotiempo aristotélico $\mathcal{A} = \mathbb{E}^1 \times \mathbb{E}^3$ es el espacio de pares $(t, \mathbf{x})$, donde t («tiempo») recorre un 1-espacio euclídeo $\mathbb{E}^1$, y $\mathbf{x}$ («punto en el espacio») recorre un 3-espacio euclídeo $\mathbb{E}^3$.

> amplio recipiente colocado por debajo de la misma... haced que el barco vaya con la velocidad que queráis, siempre que el movimiento sea uniforme y no haya fluctuaciones en un sentido u otro. ...Las gotas caerán... en el recipiente inferior sin desviarse hacia la popa, aunque el barco haya avanzado mientras las gotas están en el aire... las mariposas y las moscas seguirán su vuelo por igual hacia cada lado, y no sucederá que se concentren en la popa, como si se cansaran de seguir el curso del barco...

Lo que Galileo nos enseña es que las leyes dinámicas son exactamente las mismas cuando se refieren a cualquier sistema de referencia en movimiento uniforme. (Este era un ingrediente esencial de su entusiasta aceptación del esquema copernicano, en el que se permite que la Tierra esté en movimiento sin que advirtamos directamente dicho movimiento, en contraposición a su estatus necesariamente estacionario que tenía en el marco aristotélico anterior.) No hay nada que distinga la física del estado de reposo de la del movimiento uniforme. Teniendo en cuenta lo que se ha señalado antes, lo que esto nos dice es que no tiene significado dinámico decir que un punto particular en el espacio es, o no es, el mismo punto que un punto escogido en el espacio en un instante posterior. En otras palabras, ¡nuestra analogía con la pantalla del cine es inadecuada! No hay un espacio de fondo —una «pantalla»— que permanece fijo mientras el tiempo pasa. No podemos decir significativamente que un punto particular p en el espacio (por ejemplo, el punto de la señal de admiración en el teclado de mi ordenador portátil) es, o no es, el *mismo* punto del espacio que era hace un minuto. Para abordar esta cuestión con más fuerza, consideremos la rotación de la Tierra. Según este movimiento, un punto fijo en la superficie de la Tierra (pongamos por caso en la latitud de Oxford) se habrá movido unos dieciséis kilómetros durante el minuto en cuestión. En consecuencia, el punto p que yo acababa de seleccionar estará situado ahora en algún lugar en las proximidades de la vecina ciudad de Witney, o más allá. ¡Pero esperen! No he tomado en consideración el movimiento de la Tierra alrededor del Sol. Si lo hago, entonces encuentro que p estará ahora aproximadamente cien veces más lejos, pero en dirección contraria (porque es un poco después del mediodía, y la superficie de la Tierra, aquí, se mueve ahora en dirección contraria a su movimiento alrededor del Sol), y la Tierra se habrá alejado de p a tal extremo que p está ahora ¡más allá del límite de la atmósfera terrestre! Pero ¿no debería haber tenido en cuenta el movimiento del Sol alrededor del centro de nuestra Vía Láctea? ¿Y qué pasa con el «movimiento propio» de la galaxia misma dentro del grupo local? ¿O con el movimiento del grupo local respecto al centro del cúmulo Virgo, o del cúmulo Virgo respecto al cúmulo Coma, o del cúmulo Coma respecto al «Gran Atractor»? (§27.11.)

Evidentemente deberíamos tomar en serio a Galileo. No se puede atribuir ningún significado a la idea de que un punto concreto del espacio dentro de un minuto debe considerarse el *mismo* punto del espacio que el que yo he escogido. En la dinámica galileana no tenemos solamente un 3-espacio euclídeo $\mathbb{E}^3$ como escenario para las acciones del mundo físico que evoluciona con el tiempo, sino que tenemos un $\mathbb{E}^3$ *diferente* para cada instante de tiempo, sin ninguna identificación natural entre estos diversos $\mathbb{E}^3$.

Puede resultar alarmante que nuestra misma noción de espacio físico parezca ser algo que se evapora por completo a cada instante que pasa y reaparece como un espacio completamente diferente cuando

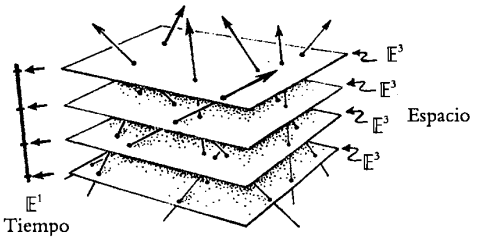


Fig. 17.3. El espaciotiempo galileano $\mathcal{G}$ es un fibrado con espacio base $\mathbb{E}^1$ y fibra $\mathbb{E}^3$, de modo que no se da identificación punto a punto entre diferentes fibras $\mathbb{E}^3$ (no hay espacio absoluto), mientras que a cada suceso espaciotemporal se le asigna un tiempo vía la proyección canónica (tiempo absoluto). (Compárese con la Fig. 15.2, pero la proyección canónica a la base se muestra aquí en horizontal.) Las historias de partículas (líneas de universo) son secciones transversales del fibrado (compárese con la Fig. 15.6a), estando mostrados aquí los movimientos inerciales de partículas como lo que especifica la estructura de $\mathcal{G}$, esto es, líneas de universo «rectas».

llega el instante siguiente. Pero las matemáticas del capítulo 15 vienen ahora en nuestra ayuda, pues esta situación es precisamente lo que hemos estudiado allí. El *espaciotiempo galileano* $\mathcal{G}$ no es un espacio producto $\mathbb{E}^1 \times \mathbb{E}^3$; ¡es un *espacio fibrado*[4] con espacio base $\mathbb{E}^1$ y fibra $\mathbb{E}^3$! En un fibrado no hay identificación puntual entre una fibra y la siguiente; de todas formas, las fibras encajan para formar una totalidad conexa. A cada suceso espaciotemporal se le asigna de forma natural un *tiempo*, como un elemento particular de un «espacio-reloj» específico $\mathbb{E}^1$, pero no hay ninguna asignación natural de una localización espacial en un «espacio de localización» específico $\mathbb{E}^3$. En el lenguaje de fibrados de §15.2, esta asignación natural de un tiempo se consigue mediante la *proyección canónica* de $\mathcal{G}$ en $\mathbb{E}^1$. (Véase la Fig. 17.3; compárese también con la Fig. 15.2.)

17.3. La dinámica newtoniana en términos del espaciotiempo

Toda esta imagen «fibrada» del espaciotiempo está muy bien, pero ¿cómo vamos a expresar la dinámica de Galileo-Newton en términos de la misma? No es sorprendente que cuando Newton llegó a formular sus leyes de la dinámica se viera llevado a una descripción en la que parecía favorecer una noción de «espacio absoluto». De hecho, Newton era, al menos inicialmente, tan relativista galileano como el propio Galileo. Esto queda claro por el hecho de que en su formulación original de las leyes del movimiento establecía explícitamente como ley fundamental el principio de relatividad de Galileo (según el cual las leyes de la física deberían ser ciegas ante un cambio de un sistema de referencia en movimiento uniforme a otro, al ser absoluta la noción de *tiempo* como se manifiesta en la imagen anterior del espaciotiempo galileano $\mathcal{G}$).

Originalmente, Newton había propuesto cinco (o seis) leyes, de las que la ley 4 era el principio galileano,[5] pero más tarde las simplificó, en sus *Principia*, hasta quedarse en las tres «leyes de Newton» que ahora nos son familiares. En efecto, se había dado cuenta de que estas eran suficientes para derivar todas las demás. Para hacer preciso el marco conceptual para sus leyes, necesitaba adoptar un «espacio absoluto» con

respecto al cual se describirían sus movimientos. Si la noción de «espacio fibrado» hubiera existido en aquella época (lo que desde luego es una posibilidad inverosímil), entonces habría sido concebible que Newton formulara sus leyes de un modo que es completamente «invariante galileano». Pero sin dicha noción es difícil ver cómo podría haber procedido Newton sin introducir algún concepto de «espacio absoluto», que es lo que hizo en realidad.

¿Qué tipo de estructura debemos asignar a nuestro «espaciotiempo galileano» $\mathcal{G}$? Sería, por supuesto, demasiado fuerte dotar a nuestro espacio fibrado $\mathcal{G}$ de una conexión fibrada (§15.7).[17.1] Lo que debemos hacer en su lugar es dotarle de algo que esté de acuerdo con la *primera ley de Newton*. Esta ley afirma que el movimiento de una partícula sobre la que no actúa ninguna fuerza debe ser uniforme y en línea recta. Esto se denomina *movimiento inercial*. En términos espaciotemporales, el movimiento (i.e., «historia») de cualquier partícula, ya sea movimiento inercial o no, está representado por una curva, denominada *línea de universo* de la partícula. De hecho, en nuestro espaciotiempo galileano, las líneas de universo deben ser siempre *secciones transversales* del fibrado galileano; véanse §15.3[17.2] y la Fig. 17.3. La noción de «uniforme y en línea recta», en términos espaciales ordinarios (un movimiento inercial), se interpreta simplemente como «recto», en términos espaciotemporales. Así pues, el fibrado galileano $\mathcal{G}$ debe tener una estructura que codifica la noción de «rectitud» de líneas de universo. Una manera de decirlo es establecer que $\mathcal{G}$ es un espacio *afín* (§14.1) en el que la estructura afín, cuando se restringe a fibras individuales $\mathbb{E}^3$, coincide con la estructura afín euclídea de cada $\mathbb{E}^3$. Otra forma es especificar simplemente la ∞^6 familia de líneas rectas que reside de forma natural en $\mathbb{E}^1 \times \mathbb{E}^3$ (los movimientos uniformes «aristotélicos») y considerar que estas proporcionan la estructura de «línea recta» del fibrado galileano, aunque «olvidando» la estructura producto real del espaciotiempo aristotélico $\mathcal{A}$. (Recordemos que ∞^6 significa una familia 6-dimensional; véase §16.7.) Pero otra manera consiste en afirmar que el espaciotiempo galileano, considerado como

[17.1] ¿Por qué?
[17.2] Explique la razón para esto.

una variedad, posee una conexión que tiene curvatura y torsión nulas (que es muy diferente de poseer una conexión fibrada, cuando se considera como un fibrado sobre $\mathbb{E}^1$).[17.3]

De hecho, este tercer punto de vista es el más satisfactorio, pues permite la generalización que necesitaremos en §§17.5,9 para describir la gravitación de acuerdo con las ideas de Einstein. Teniendo una conexión definida en $\mathcal{G}$, disponemos de una noción de *geodésica* (§14.5), y estas geodésicas (aparte de las que son simplemente líneas rectas en los $\mathbb{E}^3$s individuales) definen los *movimientos inerciales* de Newton. También podemos considerar líneas de universo que no sean geodésicas. En términos espaciales ordinarios, estas representan movimientos de partículas que se aceleran. La magnitud real de dicha aceleración viene medida, en términos espaciotemporales, como una curvatura de la línea de universo.[17.4] Según la *segunda ley de Newton*, esta aceleración es igual a la fuerza total que actúa sobre la partícula dividida por la masa de esta. (Esta es la «$f = ma$» de Newton en la forma $a = f \div m$, siendo a la aceleración de la partícula, m su masa y f la fuerza total que actúa sobre ella.) Así pues, la curvatura de una línea de universo, para una partícula de masa dada, da una medida directa de la fuerza total que actúa sobre la partícula.

En mecánica newtoniana estándar, la fuerza total sobre una partícula es la *suma* (vectorial) de las contribuciones de todas las demás partículas (Fig. 17.4a). En cualquier $\mathbb{E}^3$ particular (es decir, en cualquier instante), la contribución a la fuerza sobre una partícula actúa a lo largo de la recta que une las dos que yacen en ese $\mathbb{E}^3$ particular. Es decir, actúa *simultáneamente* entre las dos partículas. (Véase la Fig. 17.4b.) *La tercera ley de Newton* afirma que la fuerza ejercida por una de estas partículas sobre la otra es siempre de la misma magnitud y de sentido opuesto a la fuerza que esta última ejerce sobre la primera. Además, para cada tipo de fuerza diferente existe una *ley de fuerzas* que nos dice cuál debe ser la magnitud de la fuerza en función de la distancia espa-

[17.3] Explique estas tres formas con más detalle, demostrando por qué todas dan la misma estructura.

[17.4] Trate de escribir una expresión para dicha curvatura en términos de la conexión ∇. ¿Qué condición de normalización se necesita (si se necesita alguna) para los vectores tangentes?

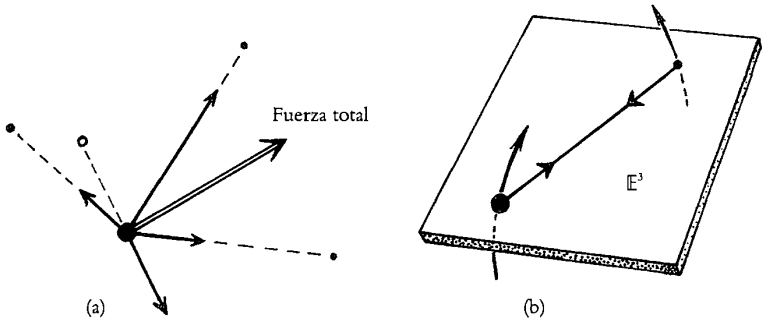


Fig. 17.4. (a) Fuerza newtoniana: en cualquier instante la fuerza total sobre una partícula (flecha doble) es la suma vectorial de las contribuciones (atractivas o repulsivas) de todas las demás partículas. (b) Dos líneas de universo de dos partículas y la fuerza entre ellas actuando «instantáneamente» en una línea que une las dos partículas en cualquier instante dentro del $\mathbb{E}^1$ particular definido por el instante. La tercera ley de Newton afirma que la fuerza que una ejerce sobre la otra es igual en magnitud y de sentido contrario a la fuerza que la segunda ejerce sobre la primera.

cial entre las partículas, y qué parámetros deben utilizarse para cada tipo de partícula, lo que describe la escala global de dicha fuerza. En el caso particular de la gravedad, esta función es la inversa del cuadrado de la distancia, y la escala global viene dada por cierta constante G, llamada constante gravitatoria de Newton, multiplicada por el producto de las dos masas implicadas. En términos de símbolos, obtenemos la bien conocida fórmula de Newton según la cual la fuerza de atracción que sobre una partícula de masa m ejerce otra partícula de masa M situada a una distancia r de ella es

$$\frac{GmM}{r^2}.$$

Es notable cómo a partir de estos simples ingredientes surge una teoría de extraordinaria potencia y versatilidad, que puede utilizarse con gran precisión para describir el comportamiento de cuerpos macroscópicos (y, para muchas consideraciones básicas, también cuerpos microscópicos), siempre que sus velocidades sean significativamente menores que la de la luz. En el caso de la gravedad, el acuerdo entre la teoría y las observaciones es especialmente evidente, debido a las muy detalladas observaciones de los movimientos planetarios en nuestro sis-

tema solar. Hoy día se sabe que la teoría de Newton es exacta hasta una parte en 10^7, lo que es una hazaña realmente impresionante, sobre todo si se tiene en cuenta que la exactitud de los datos de los que tuvo que partir Newton era solo de una diezmilésima de esto (una parte en 10^3).

17.4. El principio de equivalencia

Pese a esta extraordinaria precisión, y al hecho de que la gran teoría de Newton no tuvo prácticamente rival durante casi doscientos cincuenta años, en la actualidad sabemos que esta teoría no es exacta; además, para mejorar el esquema de Newton se necesitó la perspectiva mucho más profunda y revolucionaria de Einstein con respecto a la naturaleza de la gravitación. No obstante, esta perspectiva por sí misma no altera en absoluto la teoría de Newton en lo que respecta a cualquier consecuencia observacional. Los cambios se dan solo cuando se combina la perspectiva de Einstein con otras consideraciones relacionadas con la finitud de la velocidad de la luz y las ideas de la relatividad especial, que se describirán en §§17.6-8. La combinación completa, que da lugar a la relatividad *general* de Einstein, se verá en términos cualitativos en §17.9 y con todo detalle en §§19.6-8.

¿Cuál es entonces la perspectiva más profunda de Einstein? Es la comprensión de la importancia fundamental del *principio de equivalencia*. ¿Qué es el principio de equivalencia? La idea esencial se remonta (¡una vez más!) al gran Galileo (a finales del siglo XVI, aunque tuvo precursores, a saber, Simon Stevin en 1586, e incluso otros anteriores, tales como Ioannes Philiponos, en el siglo V o VI). Recordemos el (supuesto) experimento de Galileo, que consistió en dejar caer dos piedras, una grande y otra pequeña, desde lo alto de la torre inclinada de Pisa (Fig. 17.5a). La gran intuición de Galileo era que las dos caerían a la vez, suponiendo que pudieran despreciarse los efectos de la resistencia del aire. Dejara o no caer realmente las piedras desde la torre inclinada, realizó otros experimentos que le convencieron de esta conclusión.

El primer punto que se debe señalar aquí es que esta es una pro-

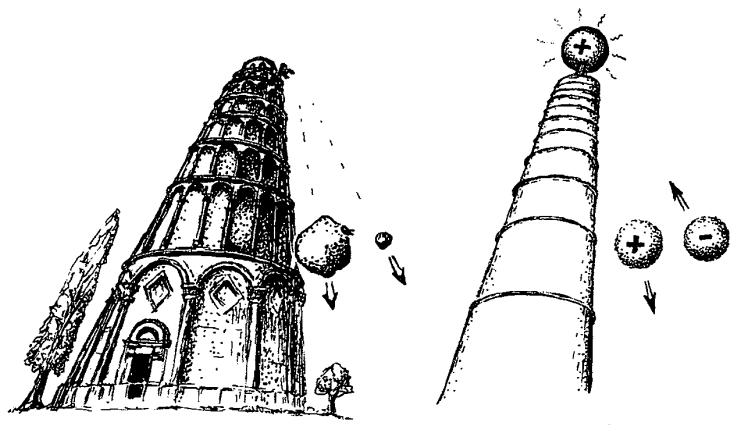


Fig. 17.5. (a) Experimento (supuesto) de Galileo. Se dejan caer dos piedras, una grande y otra pequeña, desde lo alto de la torre inclinada de Pisa. La idea de Galileo era que si pueden ignorarse los efectos de la resistencia del aire, las dos caerían al mismo tiempo. (b) Bolas de médula (de la misma masa pequeña) con cargas de signo opuesto en un campo eléctrico dirigidas hacia el suelo. Una carga «caería» hacia abajo, pero la otra subiría.

piedad particular del campo *gravitatorio*, y no se espera de ninguna otra fuerza que actúe sobre cuerpos. La propiedad de la gravedad de la que depende la intuición de Galileo es el hecho de que la intensidad de la fuerza gravitatoria ejercida por un campo gravitatorio sobre un cuerpo es proporcional a la *masa* de dicho cuerpo, mientras que la resistencia al movimiento (la cantidad m que aparece en la segunda ley de Newton) es también la masa. Es útil distinguir estas dos nociones de masa y llamar *masa gravitatoria* a la primera y *masa inercial* a la segunda. (También podríamos distinguir la masa gravitatoria *pasiva* de la *activa*. La masa pasiva es la contribución m en la fórmula de Newton de la inversa del cuadrado de la distancia GmM/r^2, cuando consideramos la fuerza gravitatoria sobre la partícula m debida a la partícula M. Cuando consideramos la fuerza sobre la partícula M debida a la partícula m, entonces la masa m aparece en su papel *activo*. Pero la tercera ley de Newton decreta que las masas activa y pasiva son iguales, de modo que no voy a distinguirlas aquí.)[6] Así pues, la intuición de Galileo depende de

la igualdad (o, más correctamente, la proporcionalidad) de las masas gravitatoria e inerte.

Desde la perspectiva del esquema dinámico global de Newton, podría parecer que el hecho de que las masas inercial y gravitatoria sean iguales es un golpe de suerte de la naturaleza. Si el campo no fuera gravitatorio, sino, pongamos por caso, un campo eléctrico, entonces el resultado sería completamente diferente. El análogo eléctrico a la masa gravitatoria pasiva es la *carga eléctrica*, mientras que la masa inercial (i.e., la resistencia a la aceleración) es exactamente la misma que en el caso gravitatorio (i.e., sigue siendo la m de la segunda ley de Newton $f = ma$). La diferencia es obvia si en lugar del par de piedras de Galileo se toma un par de bolas de médula de la misma masa pequeña pero con cargas opuestas. En un campo eléctrico dirigido hacia el suelo, una carga «caerá» hacia abajo, pero la otra subirá hacia arriba: ¡una aceleración en la dirección exactamente opuesta! (véase la Fig. 17.5b). Esto puede ocurrir porque la carga eléctrica de un cuerpo no guarda relación con su masa inercial, incluso hasta el punto de que su signo puede ser diferente. La idea de Galileo no se aplica a las fuerzas eléctricas; es una característica especial de la gravedad.

¿Por qué a esta característica de la gravedad se le llama «el principio de equivalencia»? La «equivalencia» se refiere al hecho de que un campo gravitatorio uniforme es equivalente a una aceleración. El efecto resulta muy familiar en un vuelo aéreo, donde podemos hacernos una idea completamente equivocada de dónde está «abajo» cuando nos encontramos en el interior de un avión que está realizando un movimiento acelerado (que podría ser simplemente un cambio de dirección). Los efectos de la aceleración y del campo gravitatorio de la Tierra no pueden distinguirse de forma sencilla por lo que «sentimos» dentro del avión, y los dos efectos en dos direcciones diferentes pueden sumarse para dar una sensación de dónde «debería estar» abajo, que (quizá para nuestra sorpresa cuando miremos por la ventanilla) puede ser muy diferente de la dirección hacia abajo real.

Para ver por qué esta equivalencia entre aceleración y los efectos de la gravedad es precisamente la idea de Galileo que se ha descrito antes, consideremos de nuevo sus piedras en caída mientras descienden desde lo alto de la torre inclinada. Imaginemos un insecto situado en una

de las piedras y que mira a la otra. Para el insecto, parece que la otra piedra se cierne inmóvil, como si no hubiera ningún campo gravitatorio en absoluto. (véase la Fig. 17.6a.) La aceleración que el insecto comparte mientras cae con las piedras cancela el campo gravitatorio, y es como si la gravedad estuviera totalmente ausente, hasta que las piedras y el insecto chocan contra el suelo y la experiencia «ingrávida»[7] termina abruptamente.

Estamos familiarizados con las imágenes de astronautas que tienen también experiencias «ingrávidas», aunque evitan el difícil y abrupto final de nuestro insecto al estar en órbita alrededor de la Tierra (Fig. 17.6b) (¡o en un avión que sale de su picado en el momento oportuno!). Una vez más, ellos están simplemente en caída libre, como el insecto, pero en una trayectoria escogida de forma más sensata. El hecho de que la gravedad pueda ser neutralizada de este modo por la aceleración (mediante el uso del principio de equivalencia) es una consecuencia directa del hecho de que la masa gravitatoria (pasiva) es la misma que (o proporcional a) la masa inercial, el hecho mismo que subyace en la gran intuición de Galileo.

Si vamos a tomar en serio el principio de equivalencia, entonces debemos adoptar una postura diferente de la que hemos adoptado en §17.3 con respecto a lo que debería contar como un «movimiento inercial». Previamente, un movimiento inercial se distinguía como el

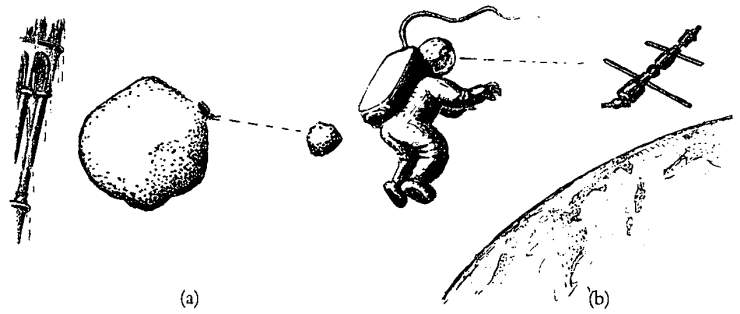


Fig. 17.6. (a) Para un insecto que repta en una piedra de la Fig. 17.5a, la otra piedra parece cernirse sin movimiento, como si el campo gravitatorio estuviese ausente. (b) Análogamente, un astronauta que orbita libremente se siente libre de gravedad, y la estación espacial parece cernirse sin movimiento, pese a la obvia presencia de la Tierra.

tipo de movimiento que se da cuando una partícula está sometida a una fuerza externa total nula. Pero con la gravedad tenemos una dificultad. Debido al principio de equivalencia no hay un modo local de decir si una fuerza gravitatoria está actuando o si lo que se «siente» como una fuerza gravitatoria puede ser solo el efecto de una aceleración. Además, como sucede con nuestro insecto en la piedra o nuestro astronauta en órbita, la fuerza gravitatoria puede ser eliminada simplemente cayendo en caída libre. Y puesto que podemos eliminar de esta forma la fuerza gravitatoria, debemos adoptar una actitud diferente hacia ella. Esta fue la visión profundamente nueva de Einstein: considerar que los *movimientos inerciales* son los movimientos que siguen las partículas cuando las fuerzas totales de origen *no* gravitatorio que actúan sobre ellas son nulas, de modo que las partículas deben estar en caída libre en el campo gravitatorio (y así la fuerza gravitatoria *efectiva* se reduce también a cero). Así pues, la trayectoria de caída de nuestro insecto y el movimiento orbital de nuestro astronauta alrededor de la Tierra deben contar como movimientos inerciales. Por el contrario, alguien que solo esté de pie en el suelo no está realizando ningún movimiento inercial, en el esquema einsteniano, porque permanecer quieto en un campo gravitatorio no es un movimiento de caída libre. Para Newton, eso hubiera contado como inercial, porque el «estado de reposo» debe contar siempre como «inercial» en el esquema newtoniano. La fuerza gravitatoria que actúa sobre la persona está compensada por la fuerza hacia arriba ejercida por el suelo, pero no son cero por separado como requiere Einstein. Por otra parte, los movimientos del insecto o del astronauta no son inerciales de acuerdo con Newton.

17.5. El «espaciotiempo newtoniano» de Cartan

¿Cómo incorporamos la noción de Einstein de un movimiento «inercial» en la estructura del espaciotiempo? Para dar un paso en la dirección de la teoría de Einstein completa, será útil proceder primero a cierta reformulación de la teoría gravitatoria de Newton de acuerdo con la perspectiva de Einstein. Como se ha mencionado al principio de §17.4, esto no representa realmente un cambio en la teoría de Newton, sino que proporciona meramente una descripción diferente de la misma. Al hacer esto me estoy tomando otra libertad con la historia, ya que tal reformulación fue propuesta por el destacado geómetra y algebrista Élie Cartan —cuya importante influencia sobre la teoría de los grupos continuos se ha señalado en el capítulo 13 (y recordemos también §12.5)— unos seis años después de que Einstein hubiera establecido su revolucionario punto de vista.

En términos generales, en el esquema de Cartan son los movimientos inerciales en este sentido einsteniano, antes que en el sentido newtoniano, los que proporcionan las líneas de universo «rectas» del espaciotiempo. En caso contrario, la geometría es similar a la galileana de §17.2. Voy a llamarlo el espaciotiempo *newtoniano* $\mathcal{N}$, puesto que el campo gravitatorio newtoniano va a estar completamente codificado en su estructura. (Quizá debería haberlo llamado «cartaniano», pero esta es una palabra horrible. En cualquier caso, Aristóteles no conocía los espacios producto, ¡ni Galileo conocía los espacios fibrados!)

El espaciotiempo $\mathcal{N}$ va a ser un fibrado con espacio base $\mathbb{E}^1$ y fibra $\mathbb{E}^3$, igual que en el caso de nuestro anterior espaciotiempo galileano $\mathcal{G}$. Pero ahora va a haber algún tipo de estructura en $\mathcal{N}$ diferente de la de $\mathcal{G}$, porque la familia de líneas de universo «rectas» que representan movimientos inerciales son diferentes; véase la Fig. 17.7a. Al menos son esencialmente diferentes en todos los casos excepto en aquellos en los que el campo gravitatorio puede ser eliminado por completo mediante una elección del sistema de referencia global en caída libre. Tal excepción sería un campo gravitatorio newtoniano que es absolutamente constante (tanto en magnitud como en dirección) en todo el espacio, aunque puede variar en el tiempo. A un observador en caída libre en dicho campo no le parecería que hay campo en absoluto.[17.5] En tal caso, la estructura de $\mathcal{N}$ sería la misma que la de $\mathcal{G}$ (Fig. 17.7b,c). Pero la mayoría de los campos gravitatorios cuentan como «esencialmente diferentes» de la ausencia de un campo gravitatorio. ¿Podemos ver por qué? ¿Podemos reconocer cuándo es la estructura de $\mathcal{N}$ diferente de la de $\mathcal{G}$? Llegaremos a esto enseguida.

[17.5] Encuentre una transformación explícita de $\mathbf{x}$ como función de t que hace esto para un campo gravitatorio newtoniano $\mathbf{F}(t)$ que es constante en el espacio en cualquier instante, pero variable en el tiempo tanto en intensidad como en dirección.

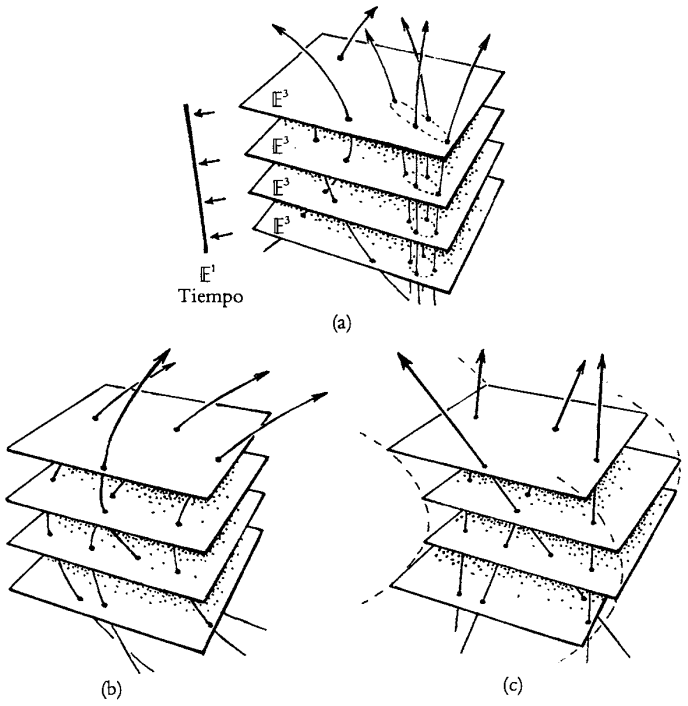


Fig. 17.7. (a) El espaciotiempo de Newton-Cartan $\mathcal{N}$, como el caso galileano particular $\mathcal{G}$, es un fibrado con espacio base $\mathbb{E}^1$ y fibra $\mathbb{E}^3$. Su estructura la proporciona la familia de movimientos «inerciales», en el sentido de Einstein de caída libre bajo gravedad. (b) El caso especial de un campo gravitatorio newtoniano constante en todo el espacio. (c) Su estructura es completamente equivalente a la de $\mathcal{G}$, como puede verse «deslizando» horizontalmente las fibras $\mathbb{E}^3$ hasta que todas las líneas de universo de caída libre son rectas.

La idea consiste en que la variedad $\mathcal{N}$ va a poseer una conexión, como sucedía en el caso particular $\mathcal{G}$. Las geodésicas de esta conexión, ∇ (véase §14.5), van a ser las líneas de universo «rectas» que representan movimientos inerciales en el sentido einsteniano. Esta conexión será libre de torsión (§14.4), pero en general tendrá curvatura (§14.4). Es la presencia de esta curvatura la que hace a algunos campos gravitatorios «esencialmente diferentes» de la ausencia de campo gravitatorio, en contraste con el campo espacialmente constante que se acaba

de considerar. Tratemos de entender el significado físico de esta curvatura.

Imaginemos a un astronauta, Albert, a quien nos referiremos como «A», en caída libre en el espacio ligeramente por encima de la atmósfera terrestre. Es útil considerar que A está precisamente en el momento en que empieza a caer hacia la superficie de la Tierra, pero no importa realmente cuál sea la velocidad del astronauta; es su aceleración, y la aceleración de sus partículas vecinas, la que nos interesa. A podría estar en una órbita segura y no necesita estar cayendo hacia el suelo. Imaginemos que hay una esfera de partículas que rodea a A, e inicialmente en reposo con respecto a A. Ahora, en términos newtonianos ordinarios, las diversas partículas en esta esfera estarán acelerándose hacia el centro de la Tierra, E, en diversas direcciones ligeramente diferentes (puesto que la dirección de E diferirá ligeramente para las diferentes partículas) y la magnitud de esta aceleración también variará (puesto que variará la distancia a E). Nos interesarán las aceleraciones *relativas*, comparadas con la aceleración del astronauta A, puesto que estamos interesados en lo que un observador inercial (en el sentido einsteniano) —en este caso A— observará que está sucediendo a las partículas inerciales próximas. La situación se ilustra en la Fig. 17.8a. Aquellas partículas que están desplazadas horizontalmente respecto a A se acelerarán hacia E en direcciones que apuntan ligeramente hacia dentro con relación a la aceleración de A, debido a la distancia finita al centro de la Tierra, mientras que aquellas partículas que están desplazadas verticalmente respecto a A se acelerarán ligeramente hacia fuera con relación a A porque el campo gravitatorio decrece con la distancia a E. En consecuencia, la esfera de partículas se distorsionará. De hecho, esta distorsión, para partículas próximas, transformará la esfera en un elipsoide de revolución, un elipsoide (prolato) que tiene su eje mayor (el eje de simetría) a lo largo de la recta AE. Además, la distorsión inicial de la esfera será un elipsoide cuyo volumen es igual al de la esfera.[17.6] Esta última propiedad es característica de la ley de la inversa del cuadrado de la gravedad newtoniana, un hecho notable que tendrá im-

[17.6] Deduzca estas diversas propiedades, dejando claro mediante el uso de la notación $O(\ldots)$ a qué orden se pretende que sean válidas estas afirmaciones.

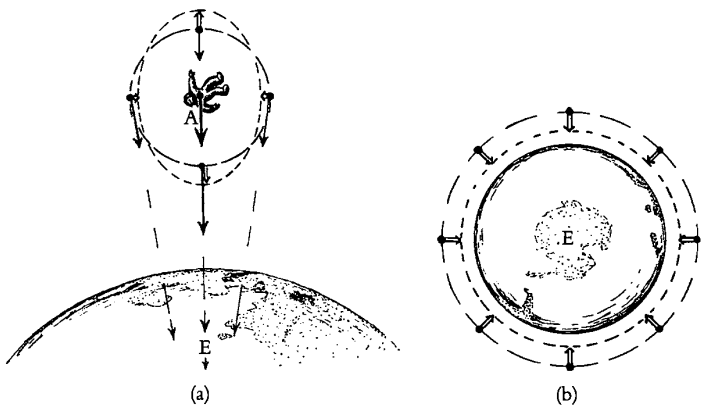


Fig. 17.8. (a) Efecto de marea. El astronauta A (Albert) rodeado por una esfera de partículas vecinas inicialmente en reposo con respecto a A. En términos newtonianos, estas tienen una aceleración hacia el centro de la Tierra E, que varía ligeramente en dirección y magnitud (flechas simples). Restando la aceleración de A de cada una de ellas obtenemos las aceleraciones relativas a A (flechas dobles); esta aceleración relativa está dirigida ligeramente hacia dentro para aquellas partículas desplazadas horizontalmente de A, pero ligeramente hacia fuera para las desplazadas verticalmente de A. En consecuencia, la esfera se distorsiona para dar un elipsoide (prolato) de revolución, con eje de simetría en la dirección AE. La distorsión inicial conserva el volumen. (b) Movamos ahora A al centro de la Tierra E y movamos la esfera de partículas hasta que rodee a E justo sobre la atmósfera. La aceleración (relativa a A = E) es hacia dentro en toda la esfera, con una aceleración de reducción de volumen inicial $4\pi GM$, donde M es la masa total rodeada.

portancia para nosotros cuando lleguemos a la relatividad general de Einstein propiamente dicha. Debería advertirse que este efecto de conservación de volumen solo tiene validez inicialmente, cuando las partículas parten del reposo respecto a A; de todas formas, con esta salvedad, es una característica general de los campos gravitatorios newtonianos, cuando A está en una región vacía. (Por el contrario, la simetría rotacional del elipsoide es un accidente de la simetría de la geometría concreta considerada aquí.)

Ahora bien, ¿cómo tenemos que considerar todo esto en términos de nuestra imagen espaciotemporal $\mathcal{N}$? En la Fig. 17.9a he tratado de indicar qué aspecto tendría esta situación para las líneas de universo

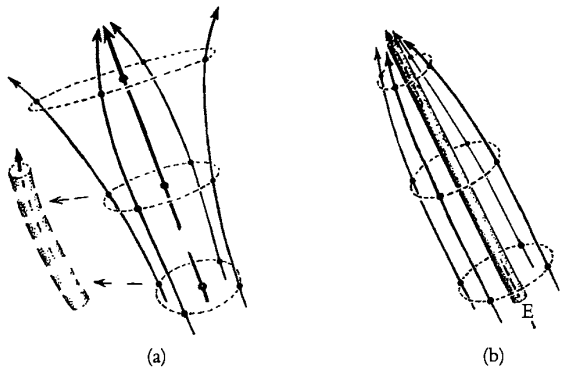


Fig. 17.9. Versiones espaciotemporales de la Fig. 17.8 (en la imagen $\mathcal{N}$ de Newton-Cartan de la Fig. 17.7), en términos de la distorsión relativa de geodésicas vecinas. (a) Desviación geodésica en espacio vacío (básicamente, la curvatura de Weyl de §19.7) vista en las líneas de universo de A y las partículas que la rodean (se ha suprimido una dirección espacial), tal como podría inducirse a partir del campo gravitatorio de un cuerpo próximo E. (b) La correspondiente aceleración hacia dentro (básicamente, la curvatura de Ricci) debida a la densidad de masa dentro del haz de geodésicas.

de A y las partículas que le rodean. (Por supuesto, he tenido que prescindir de una dimensión espacial porque ¡es difícil representar una geometría genuinamente 4-dimensional! Afortunadamente, dos dimensiones espaciales son suficientes aquí para transmitir la idea esencial.) Nótese que la distorsión de la esfera de partículas (representada como un círculo de partículas) aparece a causa de la desviación geodésica de las geodésicas que están en la vecindad de la línea de universo geodésica de A. En §14.5 he indicado por qué esta desviación geodésica es, de hecho, una medida de la *curvatura* $\boldsymbol{R}$ de la conexión $\boldsymbol{\nabla}$.

En términos físicos newtonianos, el efecto de distorsión que acabo de describir es lo que se denomina el efecto *de marea* de la gravedad. La razón de esta terminología se hace evidente si dejamos que E y A intercambien sus papeles, de modo que ahora consideramos que A está en el centro de la Tierra pero con la Luna (o quizá el Sol) localizada en E. Consideremos que la esfera de partículas es la superficie de los océanos de la Tierra, de modo que vemos que hay un efecto de distorsión debido al campo gravitatorio no uniforme de la Luna (o del

Sol).[17.7] Esta distorsión es realmente la causa de las mareas oceánicas, de modo que la terminología «efecto de marea» para esta manifestación física directa de la curvatura espaciotemporal es realmente apropiada.

De hecho, en la situación que se acaba de considerar el efecto de la Luna (o del Sol) sobre las aceleraciones relativas de las partículas en la superficie de la Tierra es solo una pequeña corrección al efecto gravitatorio principal sobre dichas partículas, a saber, la atracción gravitatoria de la propia Tierra. Por supuesto, esta es hacia dentro, a saber, en la dirección del centro de la Tierra (ahora el punto A en nuestra descripción espacial; véase la Fig. 17.8b), medido desde la localización individual de cada partícula. Si ahora se toma la esfera de partículas alrededor de la Tierra, inmediatamente por encima de la atmósfera terrestre (de modo que podemos ignorar la presión del aire), entonces las partículas estarán en caída libre (movimiento inercial einsteniano) hacia dentro en toda la esfera. En lugar de una distorsión de la forma esférica para dar un elipsoide inicialmente del mismo volumen, ahora tenemos una *reducción de volumen* más que una distorsión. En el caso general, ambos efectos podrían estar presentes. En el espacio vacío hay solo distorsión y no reducción de volumen inicial; cuando la esfera rodea la materia, existe una reducción de volumen inicial que es proporcional a la masa total rodeada. Si esta masa es M, entonces la «tasa» inicial de reducción de volumen (como medida de la aceleración hacia dentro) es, de hecho,

$$4\pi GM,$$

donde G es la constante gravitatoria de Newton.[17.8],[17.9]

De hecho, como demostró Cartan, es posible reformular por completo la teoría gravitatoria de Newton en términos de condiciones matemáticas sobre la conexión ∇, que básicamente son ecuaciones sobre la curvatura $\boldsymbol{R}$ que proporcionan una expresión matemática precisa de los requisitos antes esbozados, y que relacionan la densidad de materia ρ (masa por unidad de volumen espacial) con la parte «reductora de volumen» de $\boldsymbol{R}$. No daré en detalle la descripción de Cartan para esto porque no es necesaria para nuestras consideraciones posteriores, pues la teoría completa de Einstein es, en cierto sentido, más simple. Sin embargo, la idea misma es importante para nosotros, no solo porque nos lleva suavemente a la teoría de Einstein, sino también porque nos ayudará en nuestras consideraciones posteriores en el capítulo 30 (§30.11), concernientes a los profundos enigmas que nos presenta la teoría cuántica y a su posible solución.

[17.7] Demuestre que esta distorsión de marea es proporcional de mr^{-3}, donde m es la masa del cuerpo gravitante (considerado puntual) y r es su distancia. El Sol y la Luna muestran discos en la Tierra de un tamaño angular aproximadamente igual, pero la distorsión de marea de la Luna sobre los océanos de la Tierra es aproximadamente cinco veces la debida al Sol. ¿Qué nos dice esto sobre sus densidades relativas?

[17.8] Establezca este resultado, suponiendo que toda la masa está concentrada en el centro de la esfera.

[17.9] Demuestre que este resultado sigue siendo cierto con toda generalidad, independientemente de cuán grande sea o qué forma tenga la corteza circundante de partículas estacionarias, y de cuál sea la distribución de masa.

17.6. La velocidad finita y fija de la luz

En nuestra exposición anterior hemos considerado dos aspectos fundamentales de la relatividad general de Einstein, a saber, el principio de *relatividad*, que nos dice que las leyes de la física son ciegas a la distinción entre reposo y movimiento uniforme, y el principio de *equivalencia*, que nos dice de qué forma sutil deben modificarse estas ideas para englobar el campo gravitatorio. Ahora debemos dirigirnos al tercer ingrediente fundamental de la teoría de Einstein, que tiene que ver con la finitud de la velocidad de la luz. Es un hecho notable que estos tres ingredientes básicos puedan remontarse a Galileo; en efecto, parece que fue también Galileo el primero que tuvo una expectativa clara de que la luz debería viajar con velocidad finita, hasta el punto de que intentó medir dicha velocidad. El método que propuso (en 1638), que implica la sincronización de destellos de linternas entre colinas distantes, era, como sabemos hoy día, demasiado tosco. Pero él no tenía forma alguna de anticipar la extraordinaria rapidez a la que realmente viaja la luz.

Parece que tanto Galileo como Newton[8] tenían poderosas sospechas respecto a un profundo papel que conecta la naturaleza de la luz con las fuerzas que mantienen la materia unida. Pero la comprensión adecuada de estas ideas tuvo que esperar hasta el siglo xx, cuando se

reveló la verdadera naturaleza de las fuerzas químicas y de las fuerzas que mantienen unidos a los átomos individuales. Ahora sabemos que tales fuerzas tienen un origen fundamentalmente *electromagnético* (que concierne a la implicación del campo electromagnético con partículas cargadas) y que la teoría del electromagnetismo es también la teoría de la luz. Para entender los átomos y la química se necesitan otros ingredientes procedentes de la teoría cuántica, pero las ecuaciones básicas que describen el electromagnetismo y la luz fueron propuestas en 1865 por el gran físico escocés James Clerk Maxwell, que había sido inspirado por los magníficos descubrimientos experimentales de Michael Faraday, unos treinta años antes. Más adelante volveremos a la teoría de Maxwell (§19.2), pero su importancia inmediata para lo que ahora nos ocupa es que requiere que la velocidad de la luz tenga un valor fijo y definido, que normalmente se conoce como «c», y que en unidades ordinarias es de aproximadamente 3×10^8 metros por segundo.

Sin embargo, esto nos presenta un enigma, si queremos conservar el principio de relatividad. El sentido común nos diría que si se mide que la velocidad de la luz toma el valor concreto c en el sistema de reposo de un observador, entonces un segundo observador que se mueva a una velocidad muy alta con respecto al primero medirá que la luz viaja a una velocidad diferente, aumentada o disminuida, según sea el movimiento del segundo observador. Pero el principio de relatividad exigiría que las leyes físicas del segundo observador —que definen en particular la velocidad de la luz que percibe el segundo observador— deberían ser idénticas a las del primer observador. Esta aparente contradicción entre la constancia de la velocidad de la luz y el principio de relatividad condujo a Einstein —como, de hecho, había llevado previamente al físico holandés Hendrick Antoon Lorentz y muy en especial al matemático francés Henri Poincaré— a un punto de vista notable por el que el principio de relatividad del movimiento puede hacerse compatible con la constancia de una velocidad finita de la luz.

¿Cómo funciona esto? Sería natural que creyéramos que existe un conflicto irresoluble entre los requisitos de (a) una teoría, tal como la de Maxwell, en la que existe una velocidad absoluta para la luz, y (b) un principio de relatividad, según el cual las leyes físicas parecen las mismas con independencia de la velocidad del sistema de referencia

utilizado para su descripción. Pues ¿no podría hacerse que el sistema de referencia se moviera con una velocidad que se acercara o incluso superara a la de la luz? Y según este sistema, ¿no es cierto que la velocidad aparente de la luz no podría seguir siendo la misma que era antes? Esta indudable paradoja no aparece en una teoría, tal como la originalmente preferida por Newton (y yo diría que también por Galileo), en la que la luz se comporta como *partículas* cuya velocidad depende de la velocidad de la fuente. En consecuencia, Galileo y Newton podían seguir viviendo cómodamente con un principio de relatividad. Pero semejante imagen de la naturaleza de la luz había entrado en conflicto con la observación a lo largo de los años, como era el caso de las observaciones de estrellas dobles lejanas que mostraban que la velocidad de la luz era *independiente* de la de su fuente.[9] Por el contrario, la teoría de Maxwell había ganado fuerza, no solo por el poderoso apoyo que obtuvo de la observación (muy especialmente en los experimentos de Heinrich Hertz en 1888), sino también por la naturaleza convincente y unificadora de la propia teoría, por la que las leyes que gobiernan los campos eléctricos, los campos magnéticos y la luz están todas subsumidas en un esquema matemático de notable elegancia y simplicidad. En la teoría de Maxwell, la luz toma la forma de ondas, no de partículas; y debemos enfrentarnos al hecho de que, en esta teoría, hay realmente una velocidad fija a la que deben viajar las ondas luminosas.

17.7. Conos de luz

El punto de vista geométrico-espaciotemporal nos proporciona una ruta particularmente clara hacia la solución de la paradoja que presenta el conflicto entre la teoría de Maxwell y el principio de relatividad. Como he comentado antes, este punto de vista espaciotemporal no fue el que Einstein adoptó originalmente (ni fue el punto de vista de Lorentz, ni siquiera, al parecer, el de Poincaré). Pero, mirando en retrospectiva, podemos ver la potencia de este enfoque. Por el momento, ignoremos la gravedad y las sutilezas y complicaciones asociadas que proporciona el principio de equivalencia. Empezaremos con una

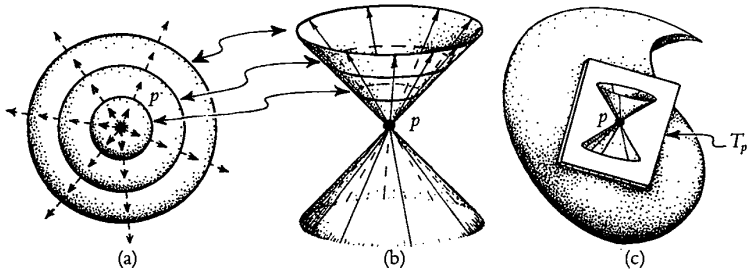


Fig. 17.10. El cono de luz especifica la velocidad fundamental de la luz. Historias de fotones que pasan por un punto espaciotemporal (suceso) p. (a) En términos puramente espaciales, el cono de luz (futuro) es una esfera que se expande hacia fuera desde p (frentes de onda). (b) En el espaciotiempo, las historias de fotones que pasan por p barren el cono de luz en p. (c) Puesto que más tarde consideraremos espaciotiempos curvos, es mejor pensar en el cono —frecuentemente denominado el cono nulo en p— como una estructura local en el espaciotiempo, i.e., en el espacio tangente T_p en p.

pizarra en blanco o más bien con una 4-variedad real sin características distintivas. Queremos ver qué podría significar el decir que existe una velocidad fundamental, que va a ser la velocidad de la luz. En cualquier punto (i.e., «suceso») p en el espaciotiempo podemos concebir la familia de todos los diferentes rayos de luz que pasan por p, en todas las diferentes direcciones espaciales. La descripción espaciotemporal es una familia de líneas de universo que pasan por p. Véase la Fig. 17.10a,b.

Será conveniente referirse a dichas líneas de universo como «historias de fotones» que pasan por p, aunque la teoría de Maxwell considera la luz como un efecto ondulatorio. Este no es realmente un conflicto importante, por diversas razones. En la teoría de Maxwell uno puede considerar un «fotón» como un paquete minúsculo de perturbación electromagnética de muy alta frecuencia, y este se comportará, adecuadamente para nuestros propósitos, como una pequeña partícula que viaja a la velocidad de la luz. (Alternativamente, podríamos pensar en términos de «frentes de onda» o de lo que los matemáticos llaman «bicaracterísticas», o quizá prefiramos apelar a la teoría cuántica, según la cual también puede considerarse que la luz consiste en «partículas» que en realidad se conocen como «fotones».)

En la vecindad de p, la familia de historias de fotones que pasan por p, tal como se representa en la Fig. 17.10b, describe un cono en el espaciotiempo conocido como el *cono de luz* en p. Considerar fundamental la velocidad de la luz es equivalente en términos espaciotemporales a considerar fundamentales los conos de luz. De hecho, desde el punto de vista que es apropiado para la geometría de variedades (véanse los capítulos 12 y 14), suele ser mejor considerar el cono de luz como una estructura en el espacio tangente T_p en p (véase la Fig. 17.10c). (Después de todo, estamos interesados en velocidades en p, y una velocidad es algo que está definido en el espacio tangente.) Con frecuencia se utiliza el término *cono nulo* para esta estructura del *espacio tangente* —y esta es mi preferencia personal—, reservando el término «cono de luz» para el lugar geométrico real en el espaciotiempo que es barrido por los rayos de luz que pasan por un punto p. Nótese que el cono de luz (o el cono nulo) tiene dos partes: el cono *pasado* y el cono *futuro*. Podemos considerar que el cono pasado representa la historia de un destello luminoso que está implosionando hacia p, de modo que toda la luz converge simultáneamente en el suceso p; en correspondencia, el cono futuro representa la historia del destello luminoso de una explosión que tiene lugar en el suceso p. Véase la Fig. 17.11.

¿Cómo vamos a dar una descripción matemática del cono nulo en

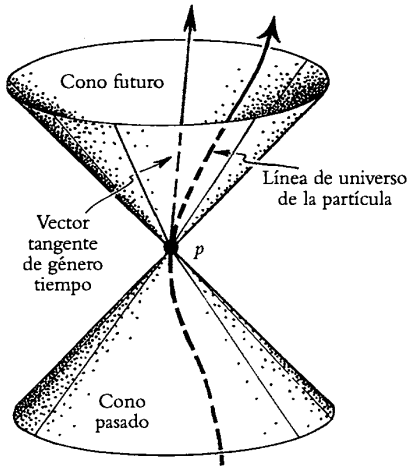


Fig. 17.11. El cono pasado y el cono futuro. El cono nulo pasado (de vectores nulos pasados) se refiere a la luz que implosiona en p de la misma manera que el cono futuro (de vectores nulos futuros) se refiere a la luz que se origina en p. La línea de universo de cualquier partícula masiva en p tiene un vector tangente que es de género tiempo (futuro), y por lo tanto yace dentro del cono nulo (futuro) en p.

p? Los capítulos 13 y 14 nos han proporcionado la base. Exigimos que la velocidad de la luz sea la misma en todas las direcciones en p, de modo que un instante después de un destello de luz la configuración espacial que rodea al punto aparece como una esfera antes que como alguna otra forma ovoide.[10] Al referirme a «un instante», quiero decir realmente que estas consideraciones se aplican a la vecindad infinitesimal temporal (tanto como espacial) de p, de modo que es legítimo considerar que esto se está refiriendo realmente a estructuras en el espacio tangente en p. Decir que el cono nulo parece «esférico» es decir simplemente que el cono está dado por una ecuación en el espacio tangente que es *cuadrática*. Esto significa que esta ecuación toma la forma

$$g_{ab}v^a v^b = 0,$$

donde g_{ab} es la forma de índices de cierto $\left[{}^0_2\right]$-tensor $\boldsymbol{g}$ simétrico y no singular, de signatura lorentziana (§13.8).[17.10] El término «nulo» en «cono nulo» se refiere al hecho de que el vector $\boldsymbol{v}$ tiene *longitud cero* ($|\boldsymbol{v}|^2 = 0$) con respecto a la (pseudo)métrica $\boldsymbol{g}$.

En esta etapa lo que nos interesa de $\boldsymbol{g}$ es solo su papel para definir los conos nulos, de acuerdo con la ecuación anterior. Si multiplicamos $\boldsymbol{g}$ por un número real diferente de cero, obtenemos exactamente el mismo cono nulo que antes (véanse también §27.12 y §33.3). Dentro de poco exigiremos que $\boldsymbol{g}$ desempeñe el papel físico adicional de proporcionar la métrica espaciotemporal, y para esto exigiremos el factor de escala apropiado; pero, por el momento, lo que nos interesa es solo la familia de conos nulos, uno en cada punto espaciotemporal. Para poder asegurar que la velocidad de la luz es constante, adoptamos la postura de que tiene sentido considerar que los conos nulos en sucesos diferentes son *paralelos* entre sí, puesto que «velocidad» en términos espaciales se refiere a la «pendiente» en términos espaciotemporales. Esto nos lleva a la imagen del espaciotiempo representada en la Fig. 17.12.

[17.10] Explique por qué.

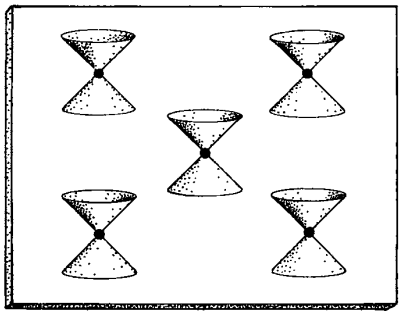


Fig. 17.12. El espacio $\mathbb{M}$ de Minkowski es plano y sus conos nulos están dispuestos uniformemente; aquí se muestran todos paralelos.

17.8. El abandono del tiempo absoluto

Podemos preguntar ahora si sería apropiado imponer, además, la estructura fibrada del espaciotiempo galileano $\mathcal{G}$. En otras palabras, ¿podemos incluir una noción de *tiempo absoluto* en nuestra imagen? Esto nos llevaría a una imagen como la de la Fig. 17.13. Las secciones $\mathbb{E}^3$ en el espaciotiempo nos darán un *elemento 3-plano* en cada espacio tangente T_p, además del cono nulo, como se muestra en la Fig. 17.13. Pero, como explicaré con más detalle en el próximo capítulo, $\boldsymbol{g}$ determina una noción de *ortogonalidad* que significa que ahora hay una dirección temporal privilegiada en cada suceso p (el complemento ortogonal con respecto a $\boldsymbol{g}$ de dicho 3-elemento plano), y esta dirección temporal privilegiada nos da un *estado de reposo* privilegiado en cada suceso. ¡Hemos perdido el principio de relatividad!

En términos más prosaicos, este argumento está expresando simplemente la noción de «sentido común» de que si existe una velocidad de la luz absoluta, entonces existe un «estado de reposo» privilegiado con respecto al cual la velocidad parece ser la misma en todas las direcciones. Lo que es menos obvio es que este conflicto surge *solamente* si tratamos de mantener la noción de un tiempo absoluto (o, al menos, un 3-espacio privilegiado en cada T_p). Debería quedar claro cómo debemos proceder. La noción de un tiempo absoluto (y, por consiguiente, de la estructura fibrada de $\mathcal{G}$ y $\mathcal{N}$) debe ser abandonada. En el estadio de sofisticación al que hemos llegado, esto no debería sorprendernos especialmente. Ya hemos visto que el espacio absoluto debe ser aban-

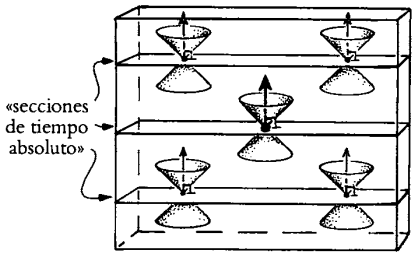


Fig. 17.13. Una noción de tiempo absoluto introducida en $\mathbb{M}$ especificaría una familia de secciones $\mathbb{E}^3$ que cortan a $\mathbb{M}$ y con ello un elemento plano local en cada suceso. Pero cada cono nulo define una (pseudo)métrica $\boldsymbol{g}$, salvo la proporcionalidad, cuya noción de ortogonalidad determina un estado de reposo.

donado tan pronto como se adopta seriamente un principio de relatividad siquiera galileano (aunque esta percepción no suele tener un reconocimiento tan amplio como debería). De modo que, por ahora, la aceptación del hecho de que el tiempo no es un concepto absoluto, igual que el espacio no es un concepto absoluto, no debería parecer tan revolucionaria como podríamos haber pensado.

Así pues, debemos despedirnos de las secciones $\mathbb{E}^3$ en el espaciotiempo, y aceptar que la única razón de que tengamos un tiempo absoluto tan firmemente incrustado en nuestro pensamiento es que la velocidad de la luz es extraordinariamente grande comparada con las velocidades que nos son familiares. En la Fig. 17.14 he vuelto a dibujar parte de la Fig. 17.13, con una razón de escala horizontal/vertical que es un poco más próxima a la que sería adecuada para las unidades normales que solemos utilizar en la vida cotidiana. Pero solo está un poco más próxima, puesto que debemos tener en cuenta que en unidades ordinarias, segundos para el tiempo y metros para la distancia, pongamos por caso, encontramos que la velocidad de la luz está dada por

$$c = 299\ 792\ 458 \text{ metros/segundo},$$

¡que es realmente un valor exacto![11] Puesto que nuestros diagramas (y nuestras fórmulas) espaciotemporales parecen tan complicados en unidades convencionales, una práctica común al trabajar en la teoría de la relatividad es la de utilizar unidades para las que $c = 1$. Esto significa que si escogemos un *segundo* como nuestra unidad de tiempo, entonces debemos utilizar un *segundo luz* (i.e., 299792458 metros) para nuestra unidad de distancia; si utilizamos el *año* como nuestra unidad de tiempo, entonces utilizamos el *año luz* (aproximadamente $9{,}45 \times 10^{15}$ me-

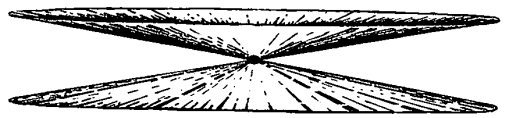


Fig. 17.14. El cono nulo redibujado, de modo que las escalas espacial y temporal estén ligeramente más próximas a las de la experiencia normal.

tros) como la unidad de distancia; si queremos utilizar un metro como nuestra medida de distancia, entonces debemos utilizar para nuestra medida de tiempo algo como 3 1/3 nanosegundos, etc.

La imagen espaciotemporal de la Fig. 17.12 fue introducida por primera vez por Hermann Minkowski (1864-1909), que era un matemático extraordinariamente bueno y original. Casualmente, él fue también uno de los profesores de Einstein en el ETH, el Instituto Federal de Tecnología de Zurich, a finales de la última década del siglo XIX. De hecho, la idea misma del espaciotiempo es de Minkowski,[12] que ya en 1908 escribía: «En lo sucesivo el espacio por sí solo, y el tiempo por sí solo están condenados a desvanecerse en meras sombras, y solo un tipo de unión entre ambos conservará una realidad independiente». En mi opinión, la teoría de la relatividad especial no estaba aún completa, pese a las maravillosas intuiciones físicas de Einstein y las excelentes contribuciones de Lorentz y Poincaré, hasta que Minkowski aportó su punto de vista fundamental y revolucionario: el *espaciotiempo*.

Para completar el punto de vista de Minkowski con relación a la geometría subyacente en la relatividad especial, y definir con ello el *espaciotiempo minkowskiano* $\mathbb{M}$, debemos fijar el *escalamiento* de $\boldsymbol{g}$, de modo que proporcione una medida de «longitud» a lo largo de las líneas de universo. Esto se aplica a curvas en $\mathbb{M}$ a las que llamamos *de género tiempo*, lo que significa que sus tangentes yacen siempre dentro de los conos nulos (véanse las Figs. 17.11 y 17.15a) y, de acuerdo con la teoría, son líneas de universo posibles para partículas masivas ordinarias. Esta «longitud» es realmente un *tiempo* y mide el tiempo real τ que registraría un reloj (ideal), entre dos puntos A y B en la curva, de acuerdo con la fórmula (véanse §13.8 y §14.7)

$$\tau = \int_A^B \mathrm{d}s, \quad \text{donde } \mathrm{d}s = (g_{ab}\mathrm{d}x^a\mathrm{d}x^b)^{\frac{1}{2}}.$$

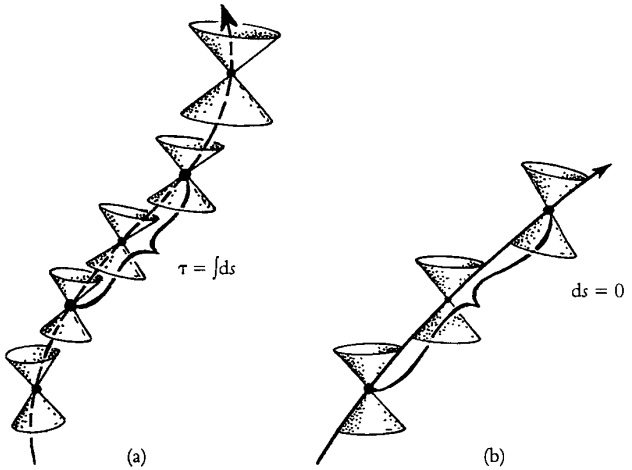


Fig. 17.15. (a) La línea de universo de una partícula masiva es una curva de género tiempo, de modo que sus tangentes están siempre dentro de los conos nulos locales, lo que da $ds^2 = g_{ab}dx^a dx^b$ positiva. La cantidad $ds = (g_{ab}dx^a dx^b)^{1/2}$ mide el intervalo de tiempo infinitesimal a lo largo de la curva, de modo que la «longitud», $\tau = \int ds$, es el tiempo medido por un reloj ideal transportado por la partícula entre dos sucesos de la curva. (b) En el caso de una partícula sin masa (por ejemplo, un fotón), las líneas de universo tienen tangentes en los conos nulos (línea de universo nula), de modo que el intervalo de tiempo $\tau = \int ds$ se anula siempre.

Para esto exigimos que la elección de la métrica espaciotemporal $\boldsymbol{g}$ tenga signatura $+---$ (que es mi elección preferida, antes que $+++-$, que prefieren otras personas por diferentes razones). Los fotones tienen líneas de universo que se denominan *nulas* (o *de género luz*), pues tienen tangentes que están *sobre* los conos nulos (Fig. 17.15b). En consecuencia, el «tiempo» que experimenta un fotón (si realmente un fotón pudiera tener experiencias), ¡tiene que ser cero!

En mi exposición anterior he preferido resaltar la estructura de conos nulos del espaciotiempo, más incluso que su métrica. En ciertos aspectos, los conos nulos son realmente más fundamentales que la métrica. En particular, determinan las propiedades de causalidad del espaciotiempo. Como acabamos de ver, las líneas de universo de las partículas materiales están obligadas a yacer dentro de los conos, y los rayos de luz tienen líneas de universo a lo largo de estos. A ninguna partícu-

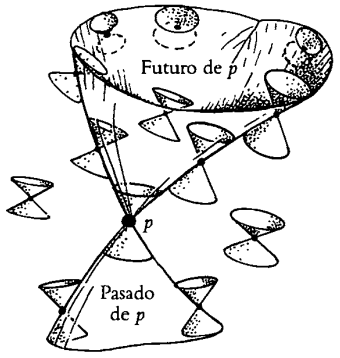


Fig. 17.16. El futuro de p es la región que puede ser alcanzada por curvas de género tiempo futuro que parten de p. Se indica un caso de espaciotiempo curvo (véase la Fig. 17.17). La frontera de esta región (suave en todo punto) es tangente a los conos de luz. Las señales, ya sean transportadas por partículas masivas o por fotones sin masa, llegan a puntos dentro de esta región o en su frontera. El pasado de p se define de forma similar.

la física se le permite tener una línea de universo de *género espacio*, i.e., una línea fuera de sus conos de luz asociados.[13] Si pensamos que las señales reales son transmitidas por partículas materiales o fotones, entonces observamos que ninguna señal semejante puede superar las restricciones impuestas por los conos nulos. Si consideramos un punto p en $\mathbb{M}$, entonces observamos que la región que yace sobre o dentro de su cono de luz futuro consiste en todos los sucesos que, en principio, pueden recibir una señal desde p. Análogamente, los puntos de $\mathbb{M}$ que yacen sobre o dentro del cono de luz pasado de p son precisamente aquellos sucesos que, en principio, pueden enviar una señal al punto p; véase la Fig. 17.16. La situación es similar cuando consideramos campos que se propagan e incluso efectos mecanocuánticos (aunque pueden surgir algunas situaciones extrañamente enigmáticas con lo que se denomina *entrelazamiento cuántico* —o *«cuanlazamiento»*—, como veremos en §23.10). Los conos nulos definen, de hecho, la estructura de *causalidad* de $\mathbb{M}$: ningún cuerpo material o señal puede viajar más rápido que la luz; está necesariamente obligado a estar dentro de (o sobre) los conos de luz.

¿Qué pasa con el principio de relatividad? En §18.2 veremos que la notable geometría de Minkowski tiene precisamente un grupo de simetría tan grande como el que tiene el espaciotiempo $\mathcal{G}$ de la física galileana. No solo están todos los puntos de $\mathbb{M}$ en pie de igualdad sino que todas las velocidades posibles (direcciones de género tiempo que apuntan al futuro) están también en pie de igualdad entre sí. Todo esto

se explicará más extensamente en §18.2. ¡El principio de relatividad sigue siendo tan válido para $\mathbb{M}$ como para $\mathcal{G}$!

17.9. El espaciotiempo de la relatividad general de Einstein

Finalmente, llegamos al espaciotiempo *einsteniano* $\mathcal{E}$ de la relatividad general. Básicamente aplicamos la misma generalización al $\mathbb{M}$ de Minkowski que hemos aplicado previamente al $\mathcal{G}$ de Galileo, cuando hemos obtenido el espaciotiempo $\mathcal{N}$ de Newton(-Cartan). En lugar de tener la disposición uniforme de conos nulos que se muestra en la Fig. 17.12, ahora tenemos la disposición de aspecto más irregular de la Fig. 17.17. Una vez más, tenemos una métrica $\boldsymbol{g}$ lorentziana $(+---)$ cuya interpretación física es definir el tiempo medido por un reloj ideal, de acuerdo precisamente con la misma fórmula que para $\mathbb{M}$, aunque ahora $\boldsymbol{g}$ es una métrica más general sin la uniformidad que es la característica de la métrica de $\mathbb{M}$.

La estructura de conos nulos definida por esta $\boldsymbol{g}$ especifica la estructura de causalidad de los $\mathcal{E}$s, como sucedía con el espacio de Minkowski $\mathbb{M}$. Localmente, las diferencias son pequeñas, pero las cosas pueden ser más elaboradas cuando examinamos la estructura de causalidad global de un complicado espaciotiempo einsteniano $\mathcal{E}$. Surge una situación extrema cuando tenemos lo que se conoce como *violación de causalidad*, en la que pueden darse «curvas cerradas de género tiempo» y se hace posible enviar una señal desde algún suceso ¡al pasado de ese

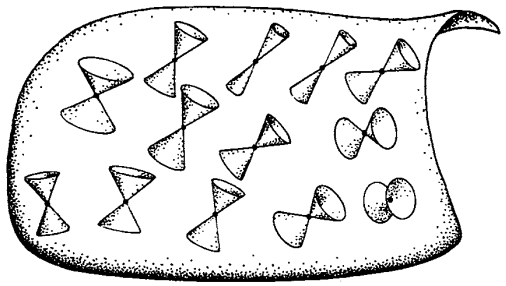


Fig. 17.17. El espaciotiempo einsteniano $\mathcal{E}$ de la relatividad general. Esta generalización del $\mathbb{M}$ de Minkowski es similar al paso de $\mathcal{G}$ a $\mathcal{N}$ (Figs. 17.12, 17.3 y 17.7a, respectivamente). Como sucede con $\mathbb{M}$, la pseudométrica lorentziana $\boldsymbol{g}$ $(+---)$ define la medida física del tiempo.

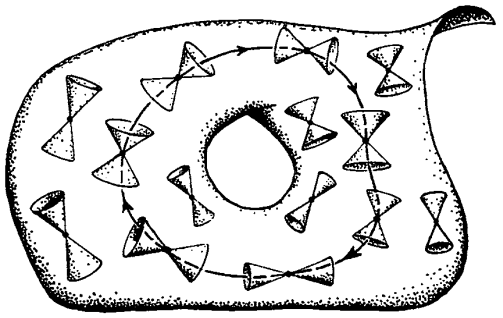


Fig. 17.18. La estructura de causalidad de $\mathcal{E}$ está determinada por $\boldsymbol{g}$ (como sucede con $\mathbb{M}$; véase la Fig. 17.16), de modo que hipotéticamente podrían darse situaciones no físicas extremas con «curvas cerradas de género tiempo», lo que permitiría que regresasen desde el pasado señales dirigidas al futuro.

mismo suceso! Véase la Fig. 17.18. Tales situaciones suelen descartarse como «no físicas», y descartarlas sería mi propia posición para un espaciotiempo clásicamente aceptable. Pese a todo, algunos físicos adoptan un punto de vista mucho más relajado sobre la cuestión[14] y están dispuestos a admitir la posibilidad del *viaje en el tiempo* que tales curvas cerradas de género tiempo permitirían. (Véase §30.6 para una exposición de estas cuestiones.) Por otra parte, pueden aparecer estructuras de causalidad menos extremas —aunque ciertamente algo exóticas— en algunos espaciotiempos interesantes de gran relevancia para la astrofísica moderna, a saber, las que representan agujeros negros. Estos serán considerados en §27.8.

En §14.7 encontrábamos el hecho de que una (pseudo)métrica $\boldsymbol{g}$ determina una única conexión $\boldsymbol{\nabla}$ libre de torsión para la que $\boldsymbol{\nabla g} = 0$, de modo que esto también se aplicará aquí. Este es un hecho notable. Nos dice que el concepto de Einstein de movimiento inercial está completamente determinado por la métrica espaciotemporal. Esto es completamente diferente de la situación con el espaciotiempo newtoniano de Cartan, donde la «$\boldsymbol{\nabla}$» tenía que ser especificada, además de las nociones métricas. La ventaja aquí es que la métrica $\boldsymbol{g}$ es ahora no degenerada, de modo que $\boldsymbol{\nabla}$ está completamente determinada por ella. De hecho, las *geodésicas* de género tiempo de $\boldsymbol{\nabla}$ (movimientos inerciales) están fijadas por la propiedad de que sean (localmente) las curvas que *maximizan* lo que se denomina el *tiempo propio*. Este tiempo propio es simplemente la longitud, medida a lo largo de la línea de universo, y es lo que mide un reloj ideal que tenga dicha línea de universo. (Esto

es curiosamente «opuesto» a la noción de geodésica como «cuerda estirada» en una superficie riemanniana ordinaria con una métrica definida positiva; véase §14.7. En §18.3 veremos que esta maximización del tiempo propio para la línea de universo no acelerada es básicamente una expresión de la «paradoja del reloj» de la teoría de la relatividad.)

La conexión ∇ tiene un tensor de curvatura $\boldsymbol{R}$, cuya interpretación física es básicamente la misma que la que se ha dado antes en el caso de $\mathcal{N}$. Lo que distingue localmente al $\mathbb{M}$ de Minkowski de la relatividad *especial*, del $\mathcal{E}$ de Einstein, de la relatividad general, es que $\boldsymbol{R} = 0$ para $\mathbb{M}$. En el próximo capítulo exploraremos con más detalle la geometría *lorentziana*, y en el siguiente veremos que las ecuaciones de campo de Einstein son la codificación natural, dentro de la estructura de las $\mathcal{E}$s, de la «tasa de reducción de volumen» $4\pi GM$ mencionada hacia el final de §17.5. También empezaremos a ser testigos de la extraordinaria potencia, belleza y precisión de la revolucionaria teoría de Einstein.

Notas

Sección 17.1

17.1. Aunque en el pasado he sido partidario de escribir «espacio-tiempo», con guión intercalado, creo que hay pasajes en este libro donde causaría confusiones en la terminología. En consecuencia, aquí adopto «espaciotiempo» de forma consistente.

17.2. Parece que Aristóteles pudo muy bien haber tenido dificultades con la noción de un espacio físico infinito, como se requiere, si la geometría euclídea $\mathbb{E}^3$ tiene que ofrecer una descripción aproximada de la geometría espacial, pero sus ideas con respecto al tiempo pueden haber sido más acordes con el $\mathbb{E}^1$ de la imagen $\mathbb{E}^1 \times \mathbb{E}^3$. Véase Moore (1990), capítulo 2.

Sección 17.2

17.3. Véase Drake (1953), pp. 186-187.

17.4. Véanse Trautman (1965), Arnol'd (1978) y Penrose (1968).

Sección 17.3

17.5. Esto era en su fragmento manuscrito *De motu corporum in mediis regulariter cedentibus*, un precursor de los *Principia*, escrito en 1684. Véase también Penrose (1987c), p. 49.

Sección 17.4

17.6. Pero véase Bondi (1957).

17.7. ¡Ahora en Rusia hay «oportunidades turísticas» de experiencias de este tipo para seres humanos, en aviones y en vuelos parabólicos!

Sección 17.6

17.8. Véase Drake (1957), p. 278, concerniente a un comentario de Galileo en *El ensayador*; véanse también Newton (1730), Query 30, y Penrose (1987c), p.23.

17.9. Véase De Sitter (1913).

Sección 17.7

17.10. Existe la peliaguda cuestión de cómo se distingue una «esfera» de un «elipsoide», puesto que las distancias pueden recalibrarse en direcciones diferentes, de modo que cualquier «elipsoide» parezca esférico. La cuestión no es aquí realmente importante. Sin embargo, lo que las recalibraciones no pueden hacer es que un ovoide no elipsoidal parezca esférico, al menos con recalibraciones «suaves». Tales ovoides darían lugar a un *espacio de Finsler*, que no tiene la agradable simetría local de las estructuras (pseudo)riemannianas de la teoría de la relatividad.

Sección 17.8

17.11. El lector podría muy bien sentirse intrigado por el hecho de que la velocidad de la luz sea un número entero exacto cuando se mide en metros por segundo. Esto no es una casualidad, sino un mero reflejo del hecho de que las medidas muy precisas de distancias son ahora mucho más difíciles de establecer que las medidas precisas de tiempo. En consecuencia, el patrón más exacto para el metro se *define* convenientemente de modo que haya 299792458 de ellos para la distancia recorrida por la luz en un segundo patrón, lo que da un valor para el metro que encaja de forma muy aproximada con el ahora insuficientemente preciso metro patrón de París.

17.12. Véase Minkowski (1952). Esta es una traducción de la conferencia que

pronunció Minkowski en la L Asamblea de Científicos Naturales y Médicos Alemanes, en Colonia, el 21 de septiembre de 1908.

17.13. Algunos físicos han jugado con la idea de «partículas» hipotéticas conocidas como *taquiones* que tendrían líneas de universo de género espacio (de modo que viajan más rápidas que la luz). Véase Bilaniuk y Sudarshan (1969); para una referencia más técnica, véase Sudarshan y Dhar (1968). Es difícil desarrollar algo parecido a una teoría consistente en la que estén presentes los taquiones, y normalmente se cree que tales entidades no existen.

Sección 17.9

17.14. Véanse, por ejemplo, Novikov (2001) y Davies (2003).

18

Geometría minkowskiana

18.1. Los 4-espacios euclídeo y minkowskiano

Las geometrías de los 2-espacios y 3-espacios euclídeos nos son muy familiares. Además, la generalización a una geometría euclídea 4-dimensional $\mathbb{E}^4$ no es difícil de hacer en principio, aunque para ello no se pueda apelar directamente a la «intuición visual». Sin embargo, es evidente que existen muchas configuraciones 4-dimensionales muy bellas, ¡o que seguramente serían muy bellas si pudiéramos verlas realmente! Una de las más sencillas (!) de estas configuraciones es la estructura de paralelos de Clifford en la 3-esfera, cuando consideramos que esta esfera está situada en $\mathbb{E}^4$. (Por supuesto, aquí sí podemos valernos un poco de la visualización porque S^3 es solamente 3-dimensional, y su proyección estereográfica, tal como se presenta en la Fig. 33.15, nos da alguna idea de la configuración de Clifford real. Si fuéramos capaces de «ver» realmente dicha configuración como parte de $\mathbb{E}^4$, podríamos hacernos una idea del aspecto que realmente tiene la estructura de 2-espacio vectorial complejo de $\mathbb{C}^2$;[1] véanse §15.4 y la Fig. 15.8.) El espacio de Minkowski $\mathbb{M}$ es en muchos aspectos muy similar a $\mathbb{E}^4$, pero existen algunas diferencias importantes en las que vamos a entrar.

Desde el punto de vista algebraico el tratamiento de $\mathbb{E}^4$ es muy similar al tratamiento en coordenadas del 3-espacio «ordinario» $\mathbb{E}^3$. Todo lo que se necesita es una coordenada cartesiana más, w, además de las coordenadas estándar x, y y z. La distancia s en $\mathbb{E}^4$ entre los puntos (w, x, y, z) y (w', x', y', z') viene dada por la relación pitagórica

$$s^2 = (w - w')^2 + (x - x')^2 + (y - y')^2 + (z - z')^2.$$

Si consideramos que (w, x, y, z) y (w', x', y', z') están solo «infinitesimalmente» desplazados uno respecto de otro, y escribimos formalmente $(\mathrm{d}w, \mathrm{d}x, \mathrm{d}y, \mathrm{d}z)$ para la diferencia $(w', x', y', z') - (w, x, y, z)$, i.e.,[2]

$$w' = w + \mathrm{d}w, x' = x + \mathrm{d}x, y' = y + \mathrm{d}y, z' = z + \mathrm{d}z,$$

entonces, encontramos

$$\mathrm{d}s^2 = \mathrm{d}w^2 + \mathrm{d}x^2 + \mathrm{d}y^2 + \mathrm{d}z^2.$$

La longitud de una curva en $\mathbb{E}^4$ viene dada por la misma fórmula que en $\mathbb{E}^3$, a saber $\int \mathrm{d}s$ (tomando el signo positivo para $\mathrm{d}s$).

Ahora la geometría del espaciotiempo de Minkowski $\mathbb{M}$ está muy próxima a esta, siendo los signos la única diferencia. Muchos de los que trabajan en este campo prefieren concentrarse en la pseudométrica con signatura $(+\ +\ +\ -)$

$$\mathrm{d}\ell^2 = -\mathrm{d}t^2 + \mathrm{d}x^2 + \mathrm{d}y^2 + \mathrm{d}z^2,$$

porque resulta conveniente cuando se considera geometría espacial, ya que la cantidad representada arriba por «$\mathrm{d}\ell^2$» es positiva para desplazamientos *de género espacio* (i.e., desplazamientos que no están ni dentro ni sobre los conos de luz nulos futuros o pasados; véase la Fig. 18.1). Pero la cantidad «$\mathrm{d}s^2$» definida por la signatura $(+\ -\ -\ -)$

$$\mathrm{d}s^2 = \mathrm{d}t^2 - \mathrm{d}x^2 - \mathrm{d}y^2 - \mathrm{d}z^2$$

tiene un significado físico más directo, ya que es positiva a lo largo de curvas *de género tiempo* que son las líneas de universo admisibles para partículas masivas, y la integral $\int \mathrm{d}s$ (con $\mathrm{d}s > 0$) es directamente interpretable como el *tiempo físico* real medido por un reloj ideal que tiene a dicha curva como línea de universo. Utilizaré *esta* signatura $(+\ -\ -\ -)$ para mi elección del tensor (pseudo)métrico $\boldsymbol{g}$, con forma de índices g_{ab}, de modo que la expresión puede escribirse en forma de índices (véase §13.8)

$$\mathrm{d}s^2 = g_{ab}\mathrm{d}x^a\,\mathrm{d}x^b.$$

No obstante, deberíamos recordar de §17.8 que, a diferencia del caso de una partícula masiva, $\int \mathrm{d}s$ es cero para una línea de universo de un fotón (de modo que puntos no coincidentes en la línea de universo pueden estar a «distancia cero»). Esto también sería cierto para cualquier otra partícula que viaje a la velocidad de la luz. El tiempo «experimentado» por dicha partícula sería siempre cero, ¡no importa cuán lejos viaje! Esto está permitido debido a la naturaleza no definida positiva (lorentziana) de g_{ab}.

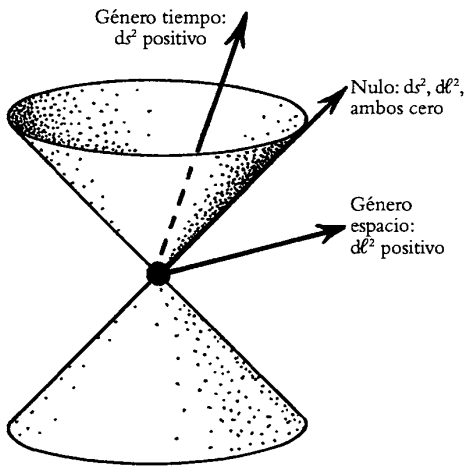


Fig. 18.1. En el espacio de Minkowski $\mathbb{M}$, la métrica $\mathrm{d}\ell^2$ proporciona una medida de (distancia)^2 espacial para desplazamientos de género espacio (ni sobre ni dentro de conos nulos futuros o pasados). Para desplazamientos de género tiempo (dentro del cono nulo), $\mathrm{d}s^2$ proporciona una medida del (intervalo)^2 temporal, donde $\int \mathrm{d}s$ es un tiempo físico tal como lo mide un reloj ideal. Para un desplazamiento nulo (a lo largo del cono nulo), ambos $\mathrm{d}\ell^2$ y $\mathrm{d}s^2$ dan cero.

En los primeros días de la teoría de la relatividad había una tendencia a resaltar la proximidad de la geometría de $\mathbb{M}$ a la de $\mathbb{E}^4$ considerando la coordenada temporal t como puramente imaginaria:

$$t = \mathrm{i}w,$$

lo que hace que la forma «$\mathrm{d}\ell^2$» de la métrica de Minkowski parezca exactamente la misma que la $\mathrm{d}s^2$ de $\mathbb{E}^4$. Por supuesto, las apariencias son algo ilusorias debido a que por la condición de «realidad» el tiempo se mide en unidades puramente imaginarias mientras que las coordenadas espaciales utilizan unidades reales normales. Además, en un sistema de referencia en movimiento las condiciones de realidad se complican porque las coordenadas reales e imaginarias se mezclan. De hecho, hay una tendencia moderna a hacer algo muy similar a esto, con varios disfraces diferentes en nombre de lo que se denomina «teoría cuántica de

campos euclídea». Más adelante, en §28.9, expondré mis razones para sentirme menos que satisfecho con este procedimiento (al menos si se considera un ingrediente clave en una aproximación a una nueva teoría física fundamental, como a veces se hace; el artificio se utiliza también como un «truco» para obtener soluciones a preguntas en teoría cuántica de campos, y para esto puede desempeñar realmente un papel honesto y valioso).

En lugar de adoptar un procedimiento como este que a mí, al menos, me parece tan poco natural, «tiremos la casa por la ventana» y permitamos que *todas* nuestras coordenadas sean complejas (véase la Fig. 18.2). Entonces no hay diferencia entre las diferentes signaturas, pues nuestras coordenadas *complejas* ω, ξ, η, ζ se refieren ahora al espacio complejo $\mathbb{C}^4$, que podemos considerar como la *complexificación* $\mathbb{CE}^4$ de $\mathbb{E}^4$. Como espacio afín complejo —véase §14.1—, este es el mismo que la complexificación $\mathbb{CM}$ de $\mathbb{M}$. Más aún, cada 4-espacio complejo $\mathbb{CE}^4$ y $\mathbb{CM}$ tiene una métrica compleja $\mathbb{C}g$ *plana* (curvatura nula) completamente equivalente. Dicha métrica puede tomarse como $ds^2 = d\omega^2 + d\xi^2 + d\eta^2 + d\zeta^2$, donde $\mathbb{E}^4$ es el subespacio real de $\mathbb{CM}$ para el que todos los ω, ξ, η, ζ son reales y $\mathbb{M}$ es el subespacio para el que ω es real pero ξ, η, ζ son imaginarios puros. El subespacio real minkowskiano alternativo $\tilde{\mathbb{M}}$, dado cuando ω es imaginario puro pero ξ, η, ζ son reales, tiene su «ds^2» que da la anterior versión «$d\ell^2$» de la métrica de Minkowski. Los tres subespacios $\mathbb{E}^4$, $\mathbb{M}$ y $\tilde{\mathbb{M}}$ se denominan *secciones reales* (alternativas) de $\mathbb{CE}^4$. Podemos distinguir una cualquiera de estas si dotamos a $\mathbb{CE}^4$ de una operación de *conjugación compleja* C que es involutiva (i.e., $C^2 = 1$) y que deja solo la sección real escogida invariante punto a punto.[18.1]

[18.1] Encuentre **C** explícitamente para cada uno de los tres casos $\mathbb{E}^4$, $\mathbb{M}$ y $\tilde{\mathbb{M}}$. *Sugerencia*: Considere cómo va a actuar **C** sobre ω, ξ, η y ζ. Modifique la conjugación compleja estándar con signos en los casos $\mathbb{M}$ y $\tilde{\mathbb{M}}$.

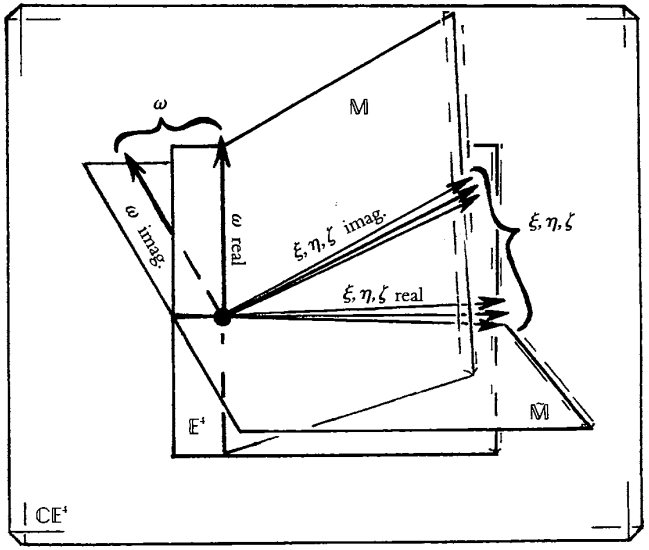


Fig. 18.2. El espacio euclídeo complejo $\mathbb{CE}^4$ tiene una métrica (holomorfa) compleja $ds^2 = d\omega^2 + d\xi^2 + d\eta^2 + d\zeta^2$ en coordenadas cartesianas complejas $(\omega, \xi, \eta, \zeta)$. El 4-espacio euclídeo $\mathbb{E}^4$ es la «sección real» para la que ω, ξ, η, ζ son todas reales. El espacio-tiempo de Minkowski $\mathbb{M}$, con la métrica $ds^2 + - - -$, es una sección real diferente, siendo ω real y ξ, η, ζ imaginarios puros. Obtenemos otra sección real lorentziana $\tilde{\mathbb{M}}$ tomando ω imaginario puro y ξ, η, ζ reales, donde la ds^2 inducida da ahora la versión «$d\ell^2$» $+ + + -$ de la métrica de Minkowski.

18.2. Los grupos de simetría del espacio de Minkowski

El grupo de simetrías de $\mathbb{E}^4$ (i.e., su grupo de movimientos euclídeos) es 10-dimensional, puesto que (i) el grupo de simetría para el que el origen está fijo es el grupo 6-dimensional O(4) (puesto que $n(n-1)/2 = 6$, cuando $n = 4$; véase §13.8), y (ii) existe un grupo de simetría 4-dimensional de traslaciones del origen (véase la Fig. 18.3a). Cuando complejificamos $\mathbb{E}^4$ a $\mathbb{CE}^4$, obtenemos un grupo 10-*complejo* dimensional evidentemente, puesto que si escribimos cualquiera de los movimientos euclídeos reales de $\mathbb{E}^4$ como una fórmula algebraica en términos de las coordenadas, todo lo que tenemos que hacer es permitir que

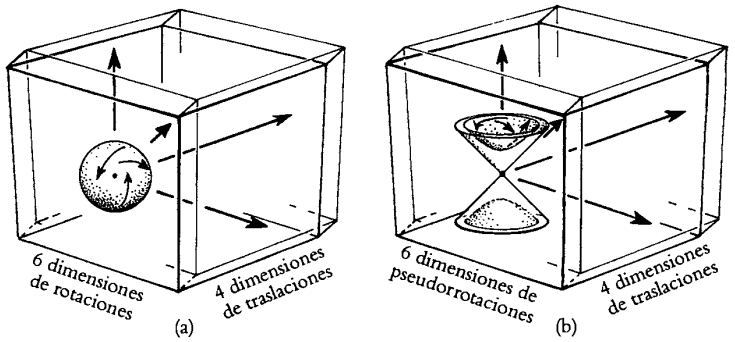


Fig. 18.3. (a) El grupo de movimientos euclídeos de $\mathbb{E}^4$ es 10-dimensional, siendo el grupo de simetría con origen fijo el grupo de rotación 6-dimensional $O(4)$ y el grupo de traslaciones del origen, 4-dimensional. (b) Para las simetrías de $\mathbb{M}$, obtenemos el grupo de Lorentz 6-dimensional $O(1, 3)$ (o $O(3, 1)$) para origen fijo y 4 dimensiones de traslaciones, lo que da el grupo de simetría de Poincaré 10-dimensional.

todas las cantidades que aparecen en la fórmula (coordenadas y coeficientes) sean complejas en lugar de reales, y obtenemos un movimiento complejo correspondiente de $\mathbb{CE}^4$. Puesto que el primero conserva la métrica, también lo hará el segundo. Además, todos los movimientos continuos de $\mathbb{CE}^4$ en sí mismo que conservan la métrica complejificada $\mathbb{C}\boldsymbol{g}$ son de esta naturaleza.[18.2]

Resulta ahora muy plausible, aunque no del todo obvio en esta fase, que el grupo tendría la misma dimensión, a saber 10 (pero ahora dimensiones *reales*), si particularizamos a una diferente «sección real» de $\mathbb{C}^4$, tal como aquella para la que las coordenadas $(\omega, \xi, \eta, \zeta)$ tienen la condición de realidad de que ω es imaginario puro y ξ, η, ζ son reales (signatura $+ + + -$) o para las que ω es real y ξ, η, ζ son imaginarios puros (signatura $(+ - - -)$; véase la Fig. 18.2. La parte traslacional sigue siendo obviamente 4-dimensional. De hecho, esta parte nos dice que el grupo es *transitivo* en $\mathbb{M}$, lo que significa que cualquier punto dado de $\mathbb{M}$ puede ser enviado a cualquier otro punto dado de $\mathbb{M}$ por algún elemento del grupo, como era el caso para $\mathbb{E}^4$. Pero ¿qué pasa con el grupo de Lorentz (O(3, 1) u O(1, 3))? ¿Cómo podemos ver que este es

[18.2] ¿Puede ver por qué?

«exactamente tan 6-dimensional» como lo es O(4)? De hecho, el grupo de Lorentz *es* 6-dimensional (véase la Fig. 18.3b.) La forma más general de verlo es examinar el álgebra de Lie —véase §13.6— y comprobar que sigue funcionando con los cambios de signo menores exigidos.[18.3] Más adelante (§18.5) veremos una muy notable forma alternativa de considerar O(1, 3), relacionándola con el grupo de simetría de la esfera de Riemann.

El grupo de simetría completo 10-dimensional del espacio de Minkowski $\mathbb{M}$ se denomina *grupo de Poincaré*, en reconocimiento de los logros del destacado matemático francés Henri Poincaré (1854-1912) al desarrollar la estructura matemática esencial de la relatividad especial en los años comprendidos entre 1898 y 1905, independientemente de la aportación fundamental de Einstein en 1905.[3] El grupo de Poincaré es importante en física relativista, especialmente en física de partículas y en teoría cuántica de campos (capítulos 25 y 26). Resulta que, según las reglas de la mecánica cuántica, las partículas individuales corresponden a *representaciones* (§§13.6,7) del grupo de Poincaré, donde los valores de su masa y espín determinan las representaciones concretas (§22.12).

Es, en esencia, la exhaustividad de este grupo lo que nos permite afirmar que el principio de relatividad sigue siendo válido para $\mathbb{M}$, incluso si tenemos una velocidad de la luz fija (§§17.6,8). En primer lugar, vemos que todos los puntos del espaciotiempo $\mathbb{M}$ están en pie de igualdad debido a la naturaleza transitiva del subgrupo de traslaciones. Además, tenemos una completa simetría de rotación espacial (3 dimensiones). Esto nos deja 3 dimensiones más para expresar el hecho de que hay una completa libertad para pasar de una velocidad ($< c$) a cualquier otra, y la estructura global sigue siendo la misma —¡lo que básicamente es el principio de relatividad de $\mathbb{M}$!—. De manera algo más formal, lo que el principio de relatividad afirma es que el grupo de Poincaré actúa transitivamente sobre el *fibrado de direcciones futuras de género tiempo* en $\mathbb{M}$.[4] Estas son las direcciones que apuntan hacia los interiores de los conos nulos futuros, que son las posibles direcciones tan-

[18.3] Confírmelo en este caso examinando explícitamente las matrices 4×4 del álgebra de Lie.

gentes a las líneas de universo de observadores.[18.4] No obstante, hay que señalar que esto solo funciona debido a que hemos abandonado la familia de «secciones simultáneas» en el espaciotiempo galileano o newtoniano. Conservarlas hubiera reducido la simetría respecto a un punto espaciotemporal a la del O(3) 3-dimensional, sin dejar ninguna libertad para pasar de una velocidad a otra.

18.3. Ortogonalidad lorentziana; la «paradoja del reloj»

Este punto de vista considera a $\mathbb{M}$ como tan solo una «sección real» del espacio complejo $\mathbb{CE}^4$ (o $\mathbb{C}^4$), aunque una sección con un carácter diferente del propio $\mathbb{E}^4$. Se trata de un punto de vista muy conveniente, siempre que podamos adoptar la actitud mental correcta. Por ejemplo, en el $\mathbb{E}^4$ euclídeo tenemos una noción de «ortogonal» (que significa «a ángulos rectos»). Esto se traslada directamente a $\mathbb{CE}^4$ por el proceso de «complexificación».[5] Sin embargo, hay ciertos tipos de propiedades de las que cabe esperar que serán un poco diferentes después de aplicar este procedimiento. Por ejemplo, encontramos que en $\mathbb{CE}^4$ una dirección puede ahora ser ortogonal a sí misma, que es algo que ciertamente no puede suceder en $\mathbb{E}^4$. Sin embargo, esta característica persiste cuando regresamos a nuestra nueva sección real, el $\mathbb{M}$ lorentziano. Así pues, retenemos una noción de ortogonalidad en $\mathbb{M}$, pero vemos que existen direcciones reales que son ortogonales a sí mismas, siendo estas las direcciones *nulas* que apuntan a lo largo de las líneas de universo de los fotones (véase *infra*).

Podemos llevar más lejos esta noción de ortogonalidad y considerar el *complemento ortogonal* $\boldsymbol{\eta}^\perp$ de un elemento r-plano $\boldsymbol{\eta}$ en un punto p. Este es el elemento $(4-r)$-plano $\boldsymbol{\eta}^\perp$ de todas las direcciones en p que son ortogonales a todas las direcciones dentro de $\boldsymbol{\eta}$ en p. Así, el complemento ortogonal de un elemento de línea es un elemento 3-plano, el complemento ortogonal de un elemento 2-plano es otro elemento 2-plano, y el complemento ortogonal de un elemento 3-plano es un elemento de línea. En cada caso, tomar de nuevo el complemento ortogonal nos devolvería al elemento del que hemos partido; en otras palabras $(\boldsymbol{\eta}^\perp)^\perp = \boldsymbol{\eta}$. Recordemos que en §13.9 y §14.7 hemos considerado las operaciones de bajar y subir índices, en una magnitud vectorial o tensorial, con g_{ab} o g^{ab}. Cuando se aplica al simple r-vector o una simple $(4-r)$-forma que representa un elemento 3-superficial, de acuerdo con §§12.4,7 (por ejemplo, $\boldsymbol{\eta}_{ab} \mapsto \boldsymbol{\eta}^{ab} = \boldsymbol{\eta}_{cd}g^{ac}g^{bd}$; $\boldsymbol{\eta}^{ab} \mapsto \boldsymbol{\eta}_{ab} = \boldsymbol{\eta}^{cd}g_{ac}g_{bd}$), esta operación de subir y bajar índices corresponde a pasar al complemento ortogonal; véase también §19.2.

En $\mathbb{E}^4$, el complemento ortogonal de un elemento 3-plano $\boldsymbol{\eta}$, por ejemplo, es un elemento de línea $\boldsymbol{\eta}^\perp$ (normal a $\boldsymbol{\eta}$) que nunca está contenido en $\boldsymbol{\eta}$; véase la Fig. 18.4. Pero como en la Fig. 18.2 podemos pasar a la complexificación $\mathbb{CE}^4$ y de ahí a la diferente sección real $\mathbb{M}$. Este es precisamente el procedimiento al que apelábamos, en efecto, en el capítulo anterior (§17.8), cuando buscábamos el complemento ortogonal de una sección temporal (elemento 3-plano de género espacio) en un punto p para encontrar una dirección de género tiempo («estado de reposo») que nos mostraba que no puede mantenerse un principio de relatividad si queremos tener a la vez una velocidad finita de la

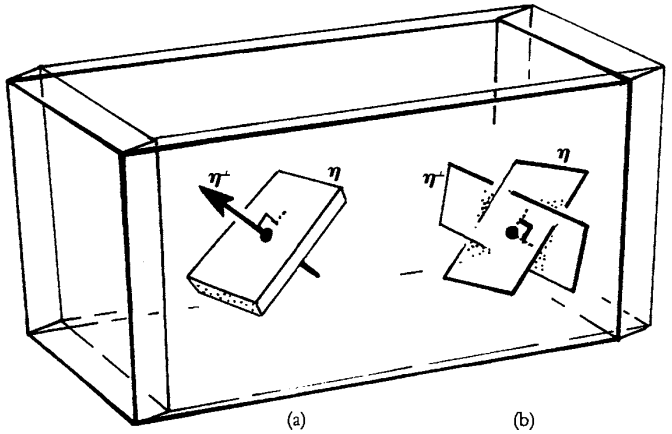


Fig. 18.4. En $\mathbb{E}^4$, un elemento r-plano $\boldsymbol{\eta}$ en un punto p tiene un complemento ortogonal $\boldsymbol{\eta}^\perp$ que es un elemento $(4-r)$-plano, donde $\boldsymbol{\eta}$ y $\boldsymbol{\eta}^\perp$ nunca tienen una dirección en común. (a) En particular, si $\boldsymbol{\eta}$ es un elemento 3-plano, entonces $\boldsymbol{\eta}^\perp$ es la dirección normal al mismo. (b) Si $\boldsymbol{\eta}$ es un elemento 2-plano, entonces $\boldsymbol{\eta}^\perp$ es otro elemento plano.

☎ [18.4] Explique más detalladamente esta acción del grupo de Poincaré.

luz y un tiempo absoluto (véase la Fig. 17.13).[18.5] No obstante, leamos ahora esto en la dirección contraria. Consideremos un observador inercial en un suceso concreto p en $\mathbb{M}$. Supongamos que la línea de universo del observador tiene alguna dirección (de género tiempo) $\boldsymbol{\tau}$ en p. Entonces el 3-espacio $\boldsymbol{\tau}^{\perp}$ representa la familia de direcciones «puramente espaciales» en p para dicho observador, i.e., aquellos sucesos vecinos que el observador estima que son simultáneos con p.

Mi intención no es explicar aquí todos los detalles de la teoría de la relatividad especial ni ver por qué, en particular, esta es una noción razonable de «simultáneo». Para esto, el lector debe remitirse a varios textos excelentes.[6] No obstante, habría que señalar que esta noción de simultaneidad depende en realidad de la velocidad del observador. En la geometría euclídea, el complemento ortogonal de una dirección en el espacio cambiará cuando cambia dicha dirección (Fig. 18.5a). En correspondencia, en geometría lorentziana el complemento ortogonal también cambiará cuando cambie la dirección (i.e., la velocidad del observador). La única diferencia está en que el cambio inclina al complemento ortogonal en sentido contrario a lo que sucede en el caso euclídeo (véase la Fig. 18.5b) y, por consiguiente, es posible que el complemento ortogonal de una dirección *contenga* a dicha dirección (véase la Fig. 18.5c), como se ha comentado antes, siendo esto lo que sucede para una dirección *nula* (i.e., a lo largo del cono de luz).

Al pasar de $\mathbb{E}^4$ a $\mathbb{M}$ hay también cambios relacionados con las desigualdades. El más espectacular de estos contiene la esencia de la denominada «paradoja del reloj» (o «paradoja de los gemelos») de la relatividad especial. A algunos lectores quizá les resultará familiar esta «paradoja»: se refiere a un viajero espacial que toma un cohete que se dirige a un planeta lejano a una velocidad próxima a la de la luz, y luego regresa para descubrir que en la Tierra han transcurrido muchos siglos, mientras que el viajero solo ha envejecido algunos años. Como ha resaltado Bondi (1964, 1967), si aceptamos que el paso del tiempo, registrado por un reloj en movimiento, es realmente una especie de «longitud de arco» medida a lo largo de una línea de universo, entonces el fenómeno no es más enigmático que el hecho de que la distancia entre dos puntos en el espacio euclídeo depende del camino a lo largo del cual se mide dicha distancia. Ambos se miden por la misma fórmula, a saber $\int ds$, pero en el caso euclídeo el camino recto representa la minimización de la distancia medida entre dos puntos extremos fijos, mientras que en el caso minkowskiano resulta que el camino recto, i.e., *inercial*, representa la *maximización* del tiempo medido entre dos sucesos extremos fijos (véase también §17.9).

La desigualdad básica, de la que dimana todo esto, es la que se denomina *desigualdad triangular* de la geometría euclídea ordinaria. Si ABC es un triángulo euclídeo cualquiera, entonces las longitudes de sus lados satisfacen

$$AB + BC \geq AC,$$

y la igualdad solo es válida en el caso degenerado en que A, B y C son colineales (véase la Fig. 18.6a). Por supuesto, las cosas son simétricas, y no importa qué lado escojamos como lado AC. En la geometría lorentziana solo obtenemos una desigualdad triangular consistente cuando todos los lados son de género tiempo, y ahora debemos tener cuidado

[18.5] (i) ¿En qué circunstancias es posible que un elemento 3-plano $\boldsymbol{\eta}$ contenga a su normal $\boldsymbol{\eta}^{\perp}$, en $\mathbb{M}$? (ii) Demuestre que existen dos familias distintas de 2-planos que son los complementos ortogonales de sí mismos en $\mathbb{CE}^4$, pero ninguna de estas familias sobrevive en $\mathbb{M}$. (Estos 2-planos complejos denominados «autodual» y «antiautodual» tendrán considerable importancia más adelante; véanse §32.2 y §33.11).

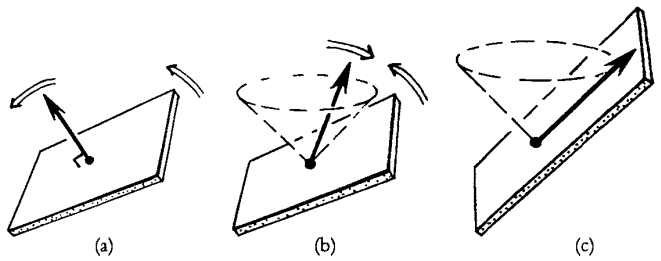


Fig. 18.5. (a) En la 4-geometría euclídea, si una dirección rota también lo hace su complemento ortogonal 3-plano. (b) Esto es cierto también en la 4-geometría lorentziana, pero en el caso de una dirección de género tiempo la pendiente del complemento ortogonal 3-plano (direcciones espaciales de «simultaneidad») se mueve en sentido inverso; (c) en consecuencia, si la dirección se hace nula, el complemento ortogonal contiene realmente dicha dirección.

en ordenar las cosas adecuadamente, de modo que AB, BC y AC están todos dirigidos hacia el futuro (véase la Fig. 18.6b). Nuestra desigualdad está ahora invertida:

$$AB + BC \leq AC,$$

y de nuevo la igualdad es válida solo cuando A, B y C son colineales, i.e., están en la línea de universo de una partícula inercial. La interpretación de esto es precisamente la denominada «paradoja del reloj». La línea de universo del viajero espacial es la trayectoria quebrada ABC, mientras que los habitantes de la Tierra tienen AC como línea de universo. Vemos que, según la desigualdad, el reloj del viajero espacial registra realmente un tiempo total transcurrido más corto que el que registran los relojes en la Tierra.

A algunas personas les preocupa que en esta descripción no se tome en cuenta adecuadamente la aceleración del cohete, y de hecho he idealizado las cosas de modo que parece que el astronauta está sometido a una aceleración impulsiva (i.e., infinita) en el suceso B (¡lo que resultaría fatal!). Sin embargo, esta cuestión es fácilmente tratable solo suavizando las esquinas del triángulo, como se indica en la Fig. 18.6d. La diferencia de tiempo no se ve muy afectada, como es obvio en la situación correspondiente para el triángulo euclídeo «suavizado» que se muestra en la Fig. 18.6c. Se suele argumentar que sería necesario pasar a la *relatividad general* de Einstein para tratar la aceleración, pero esto es erróneo. La respuesta para los tiempos del reloj se obtiene utilizando la fórmula $\int ds$ (con $ds > 0$) en ambas teorías. Al astronauta se le permite acelerar en la relatividad especial, igual que en la relatividad general. La diferencia reside tan solo en la métrica que realmente se está utilizando para evaluar la cantidad ds; i.e., depende del g_{ij} real. Estaremos trabajando en relatividad especial siempre que esta métrica sea la métrica plana de la geometría de Minkowski $\mathbb{M}$. Desde el punto de vista físico, esto significa que pueden despreciarse los campos gravitatorios. Cuando necesitemos tener en cuenta el campo gravitatorio, deberemos introducir la métrica curva de la relatividad general de Einstein. Esto se examinará con más detalle en el próximo capítulo.

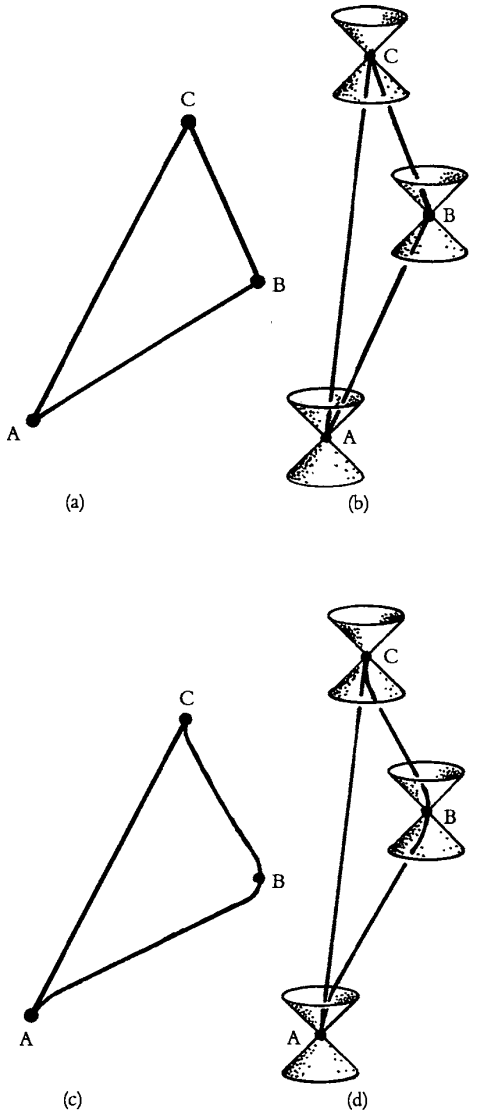


Fig. 18.6. (a) La desigualdad del triángulo euclídeo AB + + BC ≥ AC, donde la igualdad solo es válida en el caso degenerado en que A, B, C son colineales. (b) En geometría lorentziana, con AB, BC, AC de género tiempo futuro, la desigualdad se invierte: AB + + BC ≤ AC, con la igualdad válida solo cuando A, B, C están en la línea de universo de una partícula inercial. Esto ilustra la «paradoja del reloj» de la relatividad especial, donde un viajero espacial con línea de universo ABC experimenta un intervalo de tiempo más corto que el de los habitantes de la Tierra AC. (c) «Suavizar» las esquinas de un triángulo euclídeo supone poca diferencia para las longitudes de los lados, y el camino recto sigue siendo el más corto. (d) Análogamente, hacer las aceleraciones finitas («suavizando» las esquinas) supone poca diferencia para los tiempos, y el camino (inercial) recto sigue siendo el más largo.

18.4. Geometría hiperbólica en el espacio de Minkowski

Consideremos algunos otros aspectos de la geometría minkowskiana y su relación con la de Euclides. En la geometría euclídea, el lugar geométrico de los puntos situados a una distancia fija a de un punto fijo O es una superficie esférica. En $\mathbb{E}^4$, por supuesto, esta es una 3-esfera S^3. ¿Qué sucede en $\mathbb{M}$? Ahora hay dos situaciones que se deben considerar, dependiendo de si tomamos a como un número real (digamos positivo) o puramente imaginario (donde estoy adoptando mi signatura preferida $+---$; en caso contrario, los papeles se invertirían); véase la Fig. 18.7, que ilustra ambos casos.

El caso de a imaginario no nos interesa particularmente aquí. Por ello, supongamos $a > 0$ (el caso $a < 0$ es equivalente). Ahora nuestra «esfera» consta de dos piezas, una de las cuales tiene «forma de cuenco», $\mathcal{H}^+$, y yace dentro del cono de luz futuro, y la otra, $\mathcal{H}^-$, con «forma de colina», que yace dentro del cono de luz pasado. Nos concentraremos en $\mathcal{H}^+$ (el espacio $\mathcal{H}^-$ es similar). ¿Cuál es la métrica intrínseca en $\mathcal{H}^+$? Ciertamente hereda una métrica, inducida en el mismo por su inmersión en $\mathbb{M}$. (La longitud de una curva en $\mathcal{H}^+$, por ejemplo, se define simplemente considerándola como una curva en $\mathbb{M}$.) De hecho, para este caso, la $d\ell^2$ (con signatura $+++-$) es la mejor medida, puesto que las direcciones a lo largo de $\mathcal{H}^+$ son de género espacio. Podemos hacer una buena conjetura respecto a la métrica de $\mathcal{H}^+$ porque es esencialmente una «esfera» de algún tipo, pero con un «signo cambiado». ¿Cuál puede ser? Recordemos las elucubraciones de Johann Lambert, en 1786, sobre la posibilidad de construir una geometría en la que se violara el quinto postulado de Euclides. Él consideraba que una esfera de radio imaginario proporcionaría una geometría tal, siempre que algo semejante tuviera realmente un sentido consistente. De hecho, la construcción de $\mathcal{H}^+$ que acabamos de dar proporciona precisamente un espacio semejante —un modelo de geometría hiperbólica— pero ahora es 3-dimensional. Para obtener el plano no euclídeo de Lambert (el plano hiperbólico), todo lo que tenemos que hacer es prescindir de una de las dimensiones espaciales en lo que se ha descrito más arriba. En cada caso las «líneas rectas hiperbólicas» (geodésicas) son simplemente intersecciones de $\mathcal{H}^+$ con 2-planos que pasan por O (Fig. 18.8).

Por supuesto, resulta extravagante imaginar que Lambert pudiera

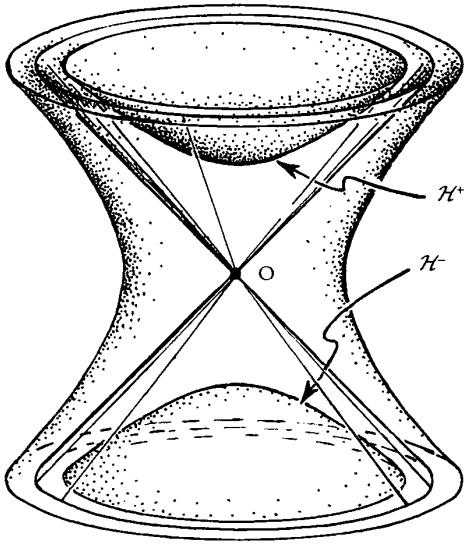


Fig. 18.7. «Esferas» en $\mathbb{M}$, como los lugares geométricos de puntos a una distancia de Minkowski dada a a un punto fijo O. Si $a > 0$, obtenemos dos piezas «hiperbólicas», la pieza «en forma de cuenco» $\mathcal{H}^+$ (dentro del cono de luz futuro) y la pieza «en forma de colina» $\mathcal{H}^-$ (dentro del cono de luz pasado). Para a imaginaria (o con a real y la signatura $+++-$ para $d\ell^2$), obtenemos un hiperboloide de una hoja, con separación de género espacio de O.

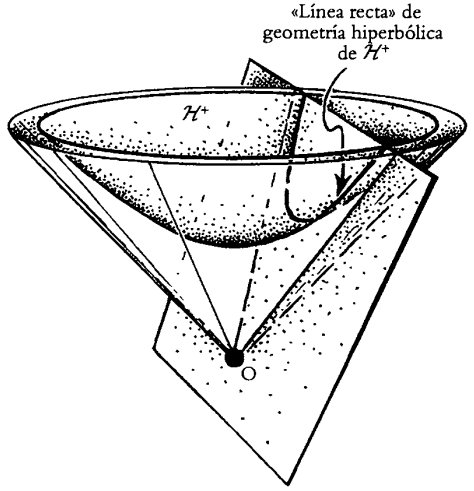


Fig. 18.8. Una «línea recta hiperbólica» (geodésica) en $\mathcal{H}^+$ es la intersección con $\mathcal{H}^+$ de un 2-plano que pasa por O. (Se ilustra el caso 2-dimensional, pero es análogo para un $\mathcal{H}^+$ 3-dimensional.)

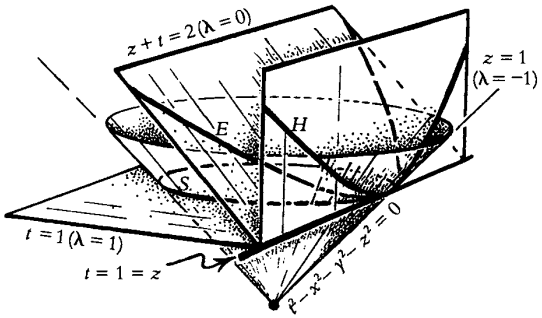


Fig. 18.9. Secciones del cono de luz $t^2 - x^2 - y^2 - z^2 = 0$, por 3-planos $(z + t) + \lambda(t - z) = 2$, que pasan por el 2-plano $t = 1 = z$. Se ha suprimido la coordenada y, de modo que las dimensiones aparecen reducidas en uno. Cuando $\lambda > 0$, la sección S tiene una métrica $d\ell^2$ que es 2-esfera, ilustrada por el caso horizontal $\lambda = 1$. Cuando $\lambda = 0$, obtenemos la métrica $d\ell^2$ euclídea plana de la sección paraboloide E. Cuando $\lambda < 0$, obtenemos una métrica $d\ell^2$ hiperbólica, ilustrada por la sección hiperbólica vertical H, en el caso $\lambda = -1$.

haber guardado en el fondo de su mente algo parecido a esta construcción. De todas formas, ilustra en cierto modo la consistencia interna de las ideas de este tipo, en las que las signaturas pueden ser «conmutadas» y las cantidades reales hechas imaginarias y las imaginarias hechas reales. Esto es algo acerca de lo que Lambert podría haber tenido intuiciones muy encomiables. Quizá sea instructivo examinar la Fig. 18.9. En ella he dibujado un cono de luz $t^2 - x^2 - y^2 - z^2 = 0$ (con y suprimida), para el 4-espacio de Minkowski $\mathbb{M}$, con coordenadas (t, x, y, z), y he tomado una familia de secciones del cono por los planos

$$z + t + \lambda(t - z) = 2,$$

para varios valores de λ, todos ellos tomados a través de un plano concreto $t = 1 = z$. Esta intersección es 2-dimensional (al ser el propio cono 3-dimensional), y resulta que para cada valor positivo de λ, la métrica de esta 2-superficie es exactamente la de una esfera de radio $\lambda^{-1/2} = 1/\sqrt{\lambda}$ (con respecto a la métrica $d\ell^2$). Cuando $\lambda = 0$, obtenemos la métrica de un plano euclídeo ordinario. (Esta intersección no se ve «plana», sino que en su lugar parece «paraboloide»; de todas formas, su

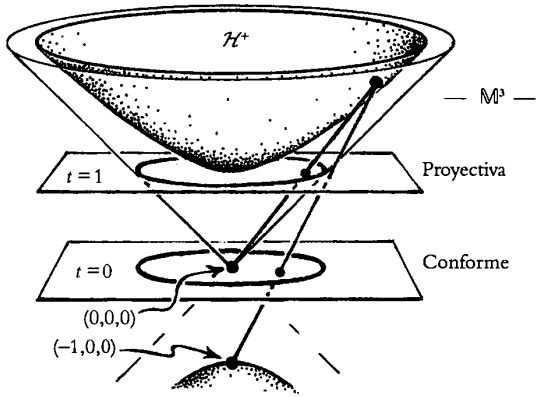


Fig. 18.10. En el 3-espacio de Minkowski $\mathbb{M}^3$, la 2-geometría hiperbólica de $\mathcal{H}^+$ (dada por $t^2 - x^2 - y^2 = 1$) se relaciona directamente con las representaciones conforme y proyectiva de Beltrami (ilustradas en las Figs. 2.11 y 2.16 respectivamente —grabado de M. C. Escher y su versión distorsionada—. El modelo proyectivo de Beltrami («Klein») se obtiene proyectando $\mathcal{H}^+$ desde el origen (0,0,0) al interior del círculo unidad en el plano $t = 1$. El modelo conforme de Beltrami («Poincaré») se obtiene proyectando $\mathcal{H}^+$ desde $(-1, 0, 0)$ en el interior del círculo unidad en $t = 0$. (Véase también la geometría de Beltrami de la Fig. 2.17.) La construcción análoga funciona asimismo para la 3-geometría hiperbólica en $\mathbb{M}$.

métrica intrínseca es realmente plana.)[18.6] Cuando λ se hace negativo, la intersección es la esfera de Lambert de radio imaginario $(= 1/\sqrt{\lambda})$. Realmente tiene una métrica intrínseca (a partir de $d\ell^2$) de geometría hiperbólica. De este modo, vemos que la intuición tentativa de Lambert, según la cual podrían tener sentido esferas de radio imaginario, estaba perfectamente justificada aunque siglos por delante de su tiempo.

La construcción para la geometría hiperbólica como la «pseudoesfera» $\mathcal{H}^+$ puede relacionarse directamente con las representaciones conforme y proyectiva de Beltrami que se han descrito (en el caso 2-dimensional) en §§2.4,5. En la Fig. 18.10 he ilustrado la forma en que ambas pueden obtenerse directamente a partir de $\mathcal{H}^+$, mostrando explí-

[18.6] Demuestre todo esto. *Sugerencia*: Resulta útil utilizar coordenadas x, y y w, donde $w = (t - z - 1/\lambda)\sqrt{\lambda} = (1 - t - z)/\sqrt{\lambda}$.

citamente el caso 2-dimensional de pseudoesferas en el 3-espacio de Minkowski $\mathbb{M}^3$ (con coordenadas t, x, y). Considerando que $\mathcal{H}^+$ tiene ecuación $t^2 - x^2 - y^2 = 1$, obtenemos la representación «Klein» (i.e., proyectiva) de Beltrami proyectándolo desde el origen $(0, 0, 0)$ en el plano $t = 1$, y obtenemos la representación «Poincaré» (i.e., conforme) de Beltrami proyectándolo desde el «polo sur» $(-1, 0, 0)$ en el «plano ecuatorial» $t = 0$ (i.e, «proyección estereográfica»; véanse §8.3 y la Fig. 8.7).[18.7]

Nótese que las *direcciones* de género tiempo futuras se representan por los puntos de $\mathcal{H}^+$ (donde, por claridad, tomo $a = 1$). Estos son simplemente las velocidades posibles de una partícula con masa. Así pues, $\mathcal{H}^+$ puede considerarse como un *espacio de velocidades* en la teoría de la relatividad. (Recordemos que esta cuestión se ha planteado al final de §2.7.) Uno de los aspectos de la relatividad que la gente suele encontrar más perturbador es que no se pueden sumar velocidades al modo usual. Así, en particular, si una nave espacial viajara en una dirección a velocidad $\frac{3}{4}c$, con respecto a la Tierra, y lanzara un misil en la misma dirección espacial a una velocidad $\frac{3}{4}c$, con respecto a la nave, entonces el misil viajaría solo a $\frac{24}{25}c$, con respecto a la Tierra, y no a la velocidad superlumínica $\left(\frac{3}{4} + \frac{3}{4}\right)c = \frac{3}{2}c$. (Aquí c es la velocidad de la luz, reintroducida solo por claridad; las unidades se escogen de modo que $c = 1$.) Esto se comprende ahora como un efecto de sumar *longitudes* en la geometría hiperbólica (véase la Fig. 18.11).[18.8]

Para apreciarlo, tenemos que entender la interpretación física de esta «longitud» hiperbólica. De hecho, es una magnitud, conocida

[18.7] Demuestre por qué las líneas rectas hiperbólicas se representan por rectas en el caso «Klein» y por círculos que cortan la frontera ortogonalmente en el caso «Poincaré», indicando mediante el uso de un «salto de signatura» por qué este segundo caso es, de hecho, conforme.

[18.8] Utilice un argumento de «salto de signatura» para ver por qué sumar longitudes en geometría hiperbólica debería dar lugar a la fórmula de adición que se está utilizando aquí, a saber $(u + v)c/(1 + uv)$, para «sumar» velocidades uc, vc en la misma dirección espacial. Considere la suma de longitudes de arco alrededor de un círculo o una esfera, siendo la «velocidad» correspondiente a cada longitud de arco la tangente del ángulo que subtiende en el centro.

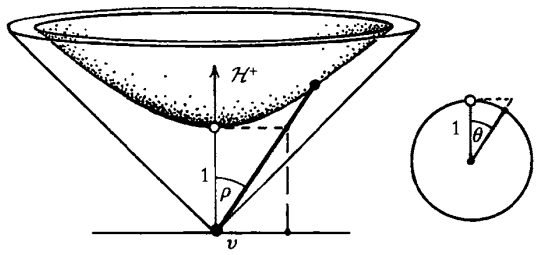


Fig. 18.11. El espacio de velocidades en la teoría de la relatividad es el espacio hiperbólico (unidad) $\mathcal{H}^+$, donde la rapidez ρ ($= \tanh^{-1}v$) mide la distancia hiperbólica a lo largo de $\mathcal{H}^+$ (la velocidad de la luz $c = 1$ corresponde a ρ infinita). Esto es equivalente (por «salto de signatura») a que la distancia a lo largo del círculo unidad sea el ángulo θ subtendido en su centro.

como la *rapidez*, para la que utilizaré la letra griega ρ, definida en términos de la velocidad v por las fórmulas (representadas gráficamente en la Fig. 18.12)

$$\rho = \frac{1}{2}\log\frac{1+v}{1-v}, \quad \text{i.e.,} \quad v = \frac{e^{\rho} - e^{-\rho}}{e^{\rho} + e^{-\rho}}$$

(la expresión derecha es lo que se denomina la «tangente hiperbólica» de ρ, escrita «tanhρ»). La rapidez es simplemente la medida de «distancia» en el espacio hiperbólico $\mathcal{H}^+$ (escogida para tener pseudorradio unidad —véanse §§2.4,6— puesto que $a = 1$). Para velocidades v pequeñas comparadas con la de la luz, la rapidez es lo mismo que v.[18.9] Nótese que la frontera, en el grabado de Escher mostrado en la Fig.

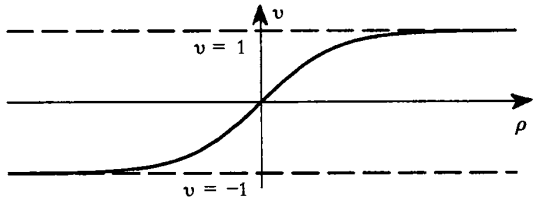


Fig. 18.12. La gráfica de la velocidad v (con $c = 1$) en términos de la rapidez ρ definida por $\rho = \frac{1}{2}\log\{(1 + v)/(1 - v)\}$, i.e., $v = (e^{\rho} - e^{-\rho})/(e^{\rho} + e^{-\rho}) = \tanh\rho$.

[18.9] Justifique esta afirmación; demuestre la equivalencia de las dos fórmulas mostradas antes.

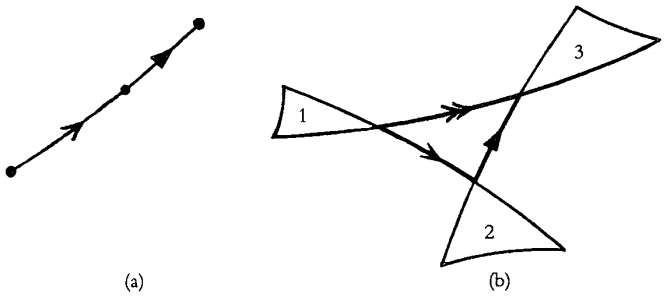


Fig. 18.13. Composición de velocidades relativistas en el espacio de velocidad hiperbólico $\mathcal{H}^+$. (a) Para velocidades en la misma dirección, simplemente sumamos las rapideces. (b) Para velocidades en direcciones diferentes, utilizamos una ley del triángulo para componerlas, donde las longitudes hiperbólicas de los lados son la mitad de sus rapideces respectivas. (Compárese con la Fig. 11.4b, que describe la composición de rotaciones ordinarias en el 3-espacio, cuya demostración es la misma.)

2.11 y que describe el infinito para la geometría hiperbólica ($\rho = \infty$), representa la velocidad límite inalcanzable c (= 1).

La composición de velocidades en la misma dirección se describe simplemente sumando sus rapideces (i.e., sumando longitudes hiperbólicas); véase la Fig. 18.13a. Podemos componer velocidades en direcciones diferentes simplemente utilizando el procedimiento dado para rotaciones ordinarias en §11.4, como se ilustra en la Fig. 11.4 (con «signatura conmutada» de manera adecuada). Aquí utilizamos una *ley triangular* hiperbólica, aplicada a las dos velocidades que se tienen que componer, donde cada una está representada por un segmento hiperbólico cuya longitud hiperbólica es exactamente la mitad de la rapidez que representa (correspondiente al hecho de que las longitudes de arco en la Fig. 11.4 son exactamente la mitad del ángulo que se está rotando); véase la Fig. 18.13b.

18.5. La esfera celeste como una esfera de Riemann

A continuación echemos una ojeada a la geometría interna de la «frontera en el infinito» para la geometría hiperbólica $\mathcal{H}^+$, donde debemos dejar claro que estamos interesados en el espaciotiempo de Minkowski 4-dimensional completo, de modo que esta frontera es ahora una esfera S^2 en lugar de un círculo (S^1) que hemos encontrado como frontera del grabado de Escher de la Fig. 2.11. Cada punto de esta esfera representa una dirección a lo largo del propio cono de luz nulo, que representa la velocidad de la luz límite que es inalcanzable para las partículas masivas. Sin embargo, estas velocidades límite son alcanzables para las partículas *sin masa*; de hecho, estas son las únicas velocidades permitidas para las partículas sin masa en vuelo libre. Afortunadamente, los fotones son partículas sin masa, y podemos *ver* los fotones. Si miramos al cielo en una noche despejada veremos lo que parece una cúpula hemisférica por encima de nosotros, salpicada de miríadas de estrellas. De hecho, nos estamos representando de forma realista la familia de los rayos de luz que constituyen el cono de luz centrado en el suceso O que está ocupado por nuestro ojo en el momento en que percibimos el escenario celeste. En realidad, estamos percibiendo solo la mitad de los rayos del cono de luz, pero si imaginamos que nos encontramos en el espacio exterior, con una completa visión de la esfera celeste que nos rodea, entonces tendremos una imagen mejor de la esfera de rayos que constituyen el cono de luz entero de O. Quizá resulta más fácil imaginar que esta esfera representa el cono *pasado* de O, puesto que nos interesamos en la luz que entra en el ojo y no en la que sale de él. Pero los rayos luminosos, en el sentido de líneas rectas nulas, se extienden en ambas direcciones, desde el pasado al futuro, de modo que podría considerarse que la esfera celeste representa solo esta familia S de rayos luminosos *completos* que pasan por O. (Véase también §33.2.)

Por supuesto, el espacio $\mathcal{S}$ es topológicamente una 2-esfera, pero ¿tiene alguna estructura particular digna de mención? Podríamos imaginarlo dotado de una métrica y considerarlo como un espacio riemanniano 2-dimensional. La forma más obvia sería tomar una sección del cono de luz, por ejemplo un corte por el 3-plano espacial $t = -1$, para obtener (a partir de la ecuación del cono $t^2 - x^2 - y^2 - z^2 = 0$) la esfera métrica de radio unidad $x^2 + y^2 + z^2 = 1$ como representación de $\mathcal{S}$. Alternativamente, podríamos seccionar el cono por $t = 1$, y de nuevo obtenemos una esfera de radio unidad, estando ambas relacionadas por la aplicación antipodal (que conserva esta métrica). Pero no hay nada especial en estas formas particulares de seccionar el cono, a menos que discriminemos la línea de universo de algún observador

particular que pasa por O y utilicemos la coordenada t de dicho observador. Si otro observador, que podría estar viajando a una velocidad alta con respecto al primero, se encuentra con el mismo suceso O, puede haber cierta distorsión entre el mapa del cielo celeste que hace un observador y el mapa que hace el otro.

En realidad, *existe* un tipo de distorsión que se debe al efecto conocido como *aberración estelar*, que fue observada por James Bradley en 1725. Según este efecto, la posición aparente de una estrella en la esfera celeste se desplaza ligeramente con las estaciones, debido al hecho de que la velocidad de la Tierra es diferente cuando está en diferentes lugares en su órbita alrededor del Sol. Este efecto es similar al que suelen observar los automovilistas cuando viajan a gran velocidad bajo la lluvia. Para quienes están dentro del automóvil parece que la lluvia viene directamente de frente, mientras que desde la perspectiva de un observador en reposo en el suelo la lluvia cae en dirección vertical. Este efecto se debe al hecho de que la velocidad finita de la lluvia debe componerse de la forma adecuada con la velocidad del automóvil para que pueda discernirse el efecto relativo observado. De hecho, en esta situación estamos considerando que la velocidad del automóvil es mucho mayor que la de la lluvia, de modo que el efecto aparente principal procede del movimiento del automóvil. En el caso de la estrella, por el contrario, la variación en la velocidad orbital de la Tierra es mucho menor que la velocidad de la luz de la estrella cuando viaja hacia nosotros. En consecuencia, la variación estacional en la posición aparente de la estrella es muy pequeña (aproximadamente 20 segundos de arco). En cualquier caso, el efecto está presente y representa una distorsión de la esfera celeste dependiente de la velocidad, lo que nos dice que no podemos considerar que esta esfera tenga una estructura métrica natural, independiente de la velocidad del observador.

La cuestión que estoy planteando aquí es si hay alguna bonita estructura matemática en $\mathcal{S}$, más débil que una estructura métrica, que se conserva al pasar del mapa celeste que hace un observador al mapa que hace otro, cuando ambos coinciden en el suceso O cruzándose a gran velocidad relativa. De hecho, *sí* existe tal estructura, y curiosamente es precisamente la estructura que hemos estudiado antes en §§8.2,3, cuando hemos considerado la esfera de Riemann. Recordemos que la esfera de Riemann posee una estructura conforme, de modo que aunque no tenga asignada ninguna métrica particular y no hay definida una noción de distancia entre puntos o de longitudes asignadas a curvas, existe una noción absoluta de *ángulo* definido entre curvas en la esfera. Cualquier transformación admisible de la esfera de Riemann en sí misma debe conservar esta noción de ángulo. En consecuencia, las formas pequeñas (infinitesimales) se conservan en estas transformaciones, aunque sus tamaños pueden cambiar. Además, círculos de cualquier tamaño en la esfera se transforman de nuevo en círculos. Esta es, de hecho, la estructura propia de la esfera celeste $\mathcal{S}$. Por consiguiente, cualquier figura circular de estrellas, tal como la percibe un observador, debe ser también percibida como circular por cualquier otro.[18.10] Esto sugiere que un etiquetado conveniente de las estrellas del cielo podría consistir en ¡asignar un número complejo a cada una de ellas (admitiendo también ∞)! No tengo noticia de que dicha propuesta haya sido

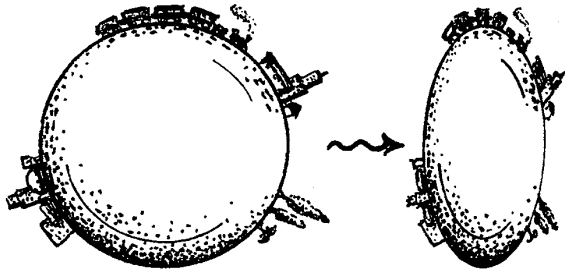


Fig. 18.14. «Efecto de achatamiento» de FitzGerald-Lorentz. Un planeta esférico se mueve hacia la derecha a velocidad v (próxima a la de la luz) con respecto a un sistema de referencia fijo. En ese sistema se describiría como si estuviera achatada en un factor $(1 - v^2/c^2)^{1/2}$ en su dirección de movimiento.

[18.10] Trate de reconstruir en detalle un argumento ingenioso para esto, que se debe al muy original e influyente teórico relativista irlandés John L. Synge y que ¡no requiere cálculos! El argumento procede aproximadamente como sigue. Considere la configuración geométrica consistente en el cono de luz pasado $\mathcal{C}$ de un suceso O y un 3-plano P (de género tiempo) que pasa por O. Sea Σ la intersección de $\mathcal{C}$ y P. Describa la «historia», a medida que avanza el tiempo, de las descripciones espaciales respectivas de $\mathcal{C}$, P y Σ, de acuerdo con algún sistema de referencia minkowskiano concreto. Explique por qué cualquier observador en O ve Σ como un círculo y, además, que esta construcción geométrica caracteriza, de una forma independiente del sistema, a aquellos haces de rayos que se le aparecen a un observador como un círculo.

adoptada en astronomía, pero el uso de un parámetro complejo semejante, denominado «coordenada estereográfica», relacionada con los ángulos polares esféricos estándar (§2.11, Fig. 2.16) por la fórmula $\zeta = e^{i\varphi}\cot\frac{1}{2}\theta$,[18.11] es común en la teoría de la relatividad general.[7]

Esta propiedad puede parecer sorprendente, sobre todo para quienes estén familiarizados con la contracción de FitzGerald-Lorentz en la que se considera que una esfera que se mueve rápidamente con velocidad v sufre un achatamiento en la dirección de su movimiento en un factor $\gamma = \sqrt{(1 - (v^2/c^2))}$; véase la Fig. 18.14. (No he discutido explícitamente este efecto de achatamiento. Aparece cuando consideramos la descripción espacial de un objeto en movimiento, y puede encontrarse en la mayoría de las exposiciones estándar de la teoría de la relatividad).[8,[18.12]] Imaginemos que la esfera pasa en dirección horizontal a una velocidad próxima a la de la luz. Es fácil imaginar que este achatamiento debería ser perceptible para un observador que esté en reposo en el suelo. Por el principio de relatividad, el efecto debería ser idéntico al que percibiría un observador que se moviera a velocidad v en dirección opuesta mientras la esfera permanece en reposo. Pero para un observador en reposo que ve una esfera en reposo, la esfera es percibida ciertamente como un objeto con perfil circular. ¡Esto parecería contradecir la afirmación hecha en el párrafo anterior según la cual «lo que se percibe como un círculo va a algo que se percibe como un círculo»! De hecho, no hay contradicción, porque este «efecto de achatamiento» de FitzGerald-Lorentz no es directamente observable. Esto se sigue de una consideración detallada de las longitudes de camino de la luz que llega a un observador con respecto al cual la esfera está en movimiento.Véase la Fig. 18.15. La luz que procede de la parte trasera de la esfera llega al observador desde un punto más lejano que la que procede de la parte frontal de la esfera.[9,[18.13]]

[18.11] Deduzca esta fórmula.

[18.12] Trate de obtener esta fórmula utilizando las ideas anteriores de la geometría espaciotemporal.

[18.13] Desarrolle este argumento en detalle para demostrar por qué el achatamiento de FitzGerald-Lorentz compensa exactamente el efecto que aparece de la diferencia de la longitud del camino. Demuestre que para un diámetro angular pequeño el efecto aparente es una rotación de la esfera, y no un achatamiento.

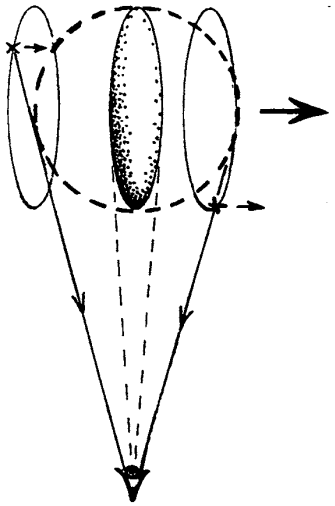


Fig. 18.15. El achatamiento de FitzGerald-Lorentz no es directamente visible porque lo que para un observador parece ser la parte trasera de la esfera implica una longitud de camino mayor que lo que parece ser la parte frontal de la esfera (pues la parte trasera se aleja del camino de la luz y la parte frontal se mueve hacia él). Por consiguiente, el borde trasero aparente se refiere a una posición de la esfera anterior a la que se refiere el borde frontal, por lo que la imagen sugiere un estiramiento compensatorio en la dirección del movimiento.

18.6. Energía y momento (angular) newtonianos

Hay un último aspecto de la geometría minkowskiana que quiero examinar en este capítulo. Concierne a las importantes cuestiones de la *energía*, el *momento* y el *momento angular* en la teoría de la relatividad. Llegaremos a ello enseguida, en §18.7, pero antes debería hacer algunos comentarios sobre estos conceptos esenciales en la teoría newtoniana, pues no los he introducido antes en este libro. La importancia vital de estas magnitudes se debe a que son cosas con un significado bien definido en la teoría newtoniana, y que se *conservan* —para un sistema sobre el que no actúan fuerzas externas— en el sentido de que la energía, el momento y el momento angular total son constantes en el tiempo.

La energía de un sistema puede considerarse compuesta de dos partes, a saber, la energía *cinética* (i.e., energía de movimiento) y la energía *potencial* (la energía almacenada en las fuerzas entre partículas). La energía cinética de una partícula (sin estructura), en la teoría newtoniana, viene dada por la expresión

$$\frac{1}{2}mv^2,$$

donde m es la masa de la partícula y v es su velocidad. Para obtener la energía cinética total, simplemente sumamos las energías cinéticas de todas las partículas individuales (aunque cuando hay muchas partículas constituyentes que se mueven al azar, podemos referirnos a su energía como energía *térmica*; véase §27.3). Para obtener la energía potencial total tenemos que saber algo de la naturaleza detallada de todas las fuerzas implicadas. Ni la energía cinética total ni la energía potencial total tienen que conservarse por separado, pero la energía total sí debe hacerlo. (El primer indicio de ello puede remontarse al estudio de Galileo del movimiento de los cuerpos bajo la gravedad. Cuando oscila la lenteja del péndulo, partiendo de una posición elevada, su energía potencial gravitatoria, medida por su altura sobre el suelo, se convierte en energía cinética, que después se convierte de nuevo en energía potencial, y luego de nuevo en energía cinética, etc.)

El *momento lineal* $\mathbf{p}$ de nuestra partícula es una magnitud vectorial, dada por la expresión

$$\mathbf{p} = m\mathbf{v},$$

donde $\mathbf{v}$ es el vector que describe su velocidad. Para obtener el momento lineal total tomamos la suma vectorial de todos los momentos individuales. Esta cantidad total también se conserva en el tiempo.[18.14]

Recordemos de §§17.2,3 que hay un principio de relatividad válido para la teoría newtoniana (relatividad de Galileo). ¿Cómo se las arreglan nuestras leyes de conservación para sobrevivir cuando ni la energía ni el momento lineal quedan invariables al pasar de un sistema inercial a otro? Si el segundo sistema se mueve uniformemente con respecto al primero, con una velocidad dada por el vector $\mathbf{u}$, entonces una partícula cuya velocidad es $\mathbf{v}$, en el primer sistema, tiene una velocidad descrita por $\mathbf{v} - \mathbf{u}$ en el segundo. Resulta que la conservación de la energía y el momento lineal en el primer sistema se corresponde con la conservación de la energía y el momento lineal en el segundo sistema, siempre que tengamos en cuenta que la masa también se conserva (y asimismo debemos utilizar la tercera ley de Newton; véanse la Fig. 17.4b y §17.3).[18.15]

[18.14] Utilice la conservación de la energía y del momento para demostrar que si una bola de billar en reposo es golpeada por otra de la misma masa, entonces salen a ángulos rectos (suponiendo una colisión elástica, de modo que no hay conversión de energía cinética en calor).

Habría que mencionar que en la mecánica newtoniana hay también otras magnitudes conservadas, la más importante de las cuales es el *momento angular* (o *momento* del momento lineal), tomado respecto a un punto origen O. Supongamos que el *vector de posición* de una partícula con respecto a O es

$$\mathbf{x} = (x^1, x^2, x^3),$$

siendo x^1, x^2, x^3 sus coordenadas cartesianas y $\mathbf{p}$ su momento lineal; entonces, el momento angular viene dado por

$$\mathbf{M} = 2\mathbf{x} \wedge \mathbf{p}$$

(véase §11.6 para el significado de $\wedge$).[10] Para obtener el momento angular total del sistema, simplemente sumamos las cantidades $\mathbf{M}$ para todas las partículas individuales.[18.16]

En la teoría newtoniana hay también otra magnitud que se conserva en el tiempo en ausencia de fuerzas externas que se suele discutir menos que el momento angular. Para una sola partícula, esta es

$$\mathbf{N} = t\mathbf{p} - m\mathbf{x},$$

donde t es el tiempo, y obtenemos el valor total de $\mathbf{N}$ sumando los valores individuales para cada partícula. Este total tiene la misma forma que el $\mathbf{N}$ dado antes, pero ahora $\mathbf{x}$ es el vector de posición del centro de masas y $\mathbf{p}$ el momento lineal total. La constancia de esta $\mathbf{N}$ total expresa el hecho de que el centro de masas se mueve uniformemente en línea recta; véase la Fig. 18.16.[18.17]

Tendremos que plantearnos la siguiente pregunta: ¿cómo queda afectado todo esto por las convulsiones de la relatividad especial? ¿Seguimos teniendo los conceptos de energía, momento, momento angular y movimiento del centro de masas conservados? ¿Qué hay de la

[18.15] Demuestre todo esto.
[18.16] ¿Por qué los patinadores que giran recogen sus brazos para incrementar su velocidad de rotación?
[18.17] Demuéstrelo. (N.B. El vector de posición del centro de masas es la suma de las cantidades $m\mathbf{x}$ dividida por la suma de las masas m.)

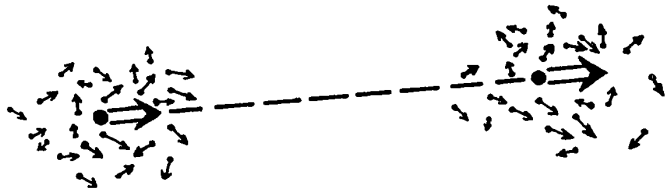


Fig. 18.16. Movimiento uniforme del centro de masas. La cantidad $\mathbf{N} = t\mathbf{p} - m\mathbf{x}$, donde t es el tiempo y $\mathbf{x}$ es el vector de posición del centro de masas, se conserva. Esto expresa el hecho de que el centro de masas se mueve uniformemente en línea recta, con velocidad $\mathbf{p}/m$.

conservación de la masa? La respuesta a las cuatro primeras preguntas es «sí», aunque tenemos que tener cuidado en definir estas magnitudes correctamente. Con respecto a la conservación de la *masa*, sucede algo muy curioso. Las dos leyes de conservación newtonianas independientes para energía y momento se subsumen en una. En un sentido claro, masa y energía se hacen completamente equivalentes entre sí, de acuerdo con la ecuación más famosa de Einstein

$$E = mc^2,$$

donde E es la energía total del sistema y m su masa total, siendo c la velocidad de la luz, como antes. En la última sección de este capítulo veremos cómo funciona todo esto.

18.7. Energía y momento (angular) relativistas

Recordemos de qué forma el espacio y el tiempo llegan a unirse en la teoría de la relatividad para convertirse en la entidad única «espacio-tiempo», de modo que la coordenada temporal t se añade al vector de posición en el 3-espacio $\mathbf{x} = (x^1, x^2, x^3)$ para dar el 4-vector x^a

$$(x^0, x^1, x^2, x^4) = (t, \mathbf{x}).$$

Encontraremos que el momento y la energía llegan a unirse de modo similar. Cualquier sistema finito en relatividad especial tendrá una energía total E y un 3-vector momento total $\mathbf{p}$. Estos se unen en

lo que se denomina el 4-vector *energía-momento*, cuyas componentes espaciales son

$$(p^1, p^2, p^3) = c^2\mathbf{p},$$

y cuya componente temporal p^0 mide no solo la energía total, sino también de forma equivalente la *masa* total m del sistema de acuerdo con

$$p^0 = E = mc^2,$$

que incorpora la famosa relación masa/energía de Einstein.

Con unidades más naturales con $c = 1$, energía y masa son simplemente iguales. Sin embargo, he escrito explícitamente la velocidad de la luz c (i.e., sin escoger unidades espacio/tiempo tales que $c = 1$) para facilitar la traducción a descripciones no relativistas. Los convenios que estoy utilizando consisten en tomar las componentes métricas g_{ab} como la matriz cuyas componentes no nulas son $(1, -c^{-2}, -c^{-2}, -c^{-2})$ por debajo de la diagonal principal, y cuya inversa g^{ab} tiene $(1, -c^{-2}, -c^{-2}, -c^{-2})$ por debajo de la diagonal principal.

Aunque inicialmente se puede considerar la energía-momento como un *vector* espaciotemporal en este sentido, resulta más conveniente (véanse §20.2 y §21.2) considerarlo como un *covector*, descrito por cantidad con subíndice p_a con componentes

$$(p_0, p_1, p_2, p_3) = (E, -\mathbf{p}).$$

Esto tiene un irritante signo menos (aunque la c ha desaparecido). Cualquiera que sea la versión utilizada (p_a o p^a), el 4-momento satisface una *ley de conservación*. Así pues, en una colisión entre dos o más partículas (o sistemas), o en la desintegración de una única partícula (o sistema) en dos o más, la suma de todos los 4-momentos antes de la colisión es igual a la suma de todos los 4-momentos después de la misma. Así pues, la ley de conservación de la energía, la de conservación del momento, y también la de conservación de la masa, están todas subsumidas en esta única ley. La razón para reunirlas de esta forma está en que, en un cambio de sistema de referencia, dichas cantidades se transforman entre sí de la forma correcta para la teoría de la relatividad, como requiere la notación de índices (véase §12.8).

Notamos que la masa total de un sistema no es una magnitud esca-

lar en la teoría de la relatividad, de modo que su valor depende del sistema de referencia con respecto al cual es medida. Por ejemplo, una partícula cuya masa es m, medida en su propio sistema de referencia, parece tener una masa mayor medida en un segundo sistema con respecto al cual se está moviendo. No obstante, para que el efecto sea significativo, la velocidad relativa entre los dos sistemas tendría que ser comparable a la velocidad de la luz.[18.18]

Sin embargo, estos comentarios se aplican solo al tipo de masa que se conserva en el sentido aditivo recién descrito (para un sistema sobre el que no actúan fuerzas externas). Existe otro concepto de masa en relatividad, a saber, la *masa en reposo* μ (≥ 0), que no depende del sistema de referencia. Esta es igual a la masa medida en el sistema de reposo, i.e., en el sistema para el que el momento es cero. La masa en reposo μ es c^{-2} multiplicado por la *energía en reposo* $(p_a p^a)^{1/2}$, de modo que

$$(c^2\mu)^2 = p_a p^a = E^2 - c^2\mathbf{p}^2;$$

y tenemos $\mu = c^{-2}(E^2 - c^2\mathbf{p}^2)^{1/2}$. Aquí estoy adoptando la notación vectorial 3-espacial en la que, para un 3-vector $\mathbf{a}$ arbitrario, definimos $\mathbf{a}^2 = \mathbf{a} \bullet \mathbf{a} = a_1^2 + a_2^2 + a_3^2$. El «punto» denota «producto escalar» (de modo análogo a la notación de §12.3):

$$\mathbf{a} \bullet \mathbf{b} = a_1 b_1 + a_2 b_2 + a_3 b_3,$$

con $\mathbf{a} = (a_1, a_2, a_3)$ y $\mathbf{b} = (b_1, b_2, b_3)$. (Esta notación nos será de utilidad más adelante.)

Para una única partícula que es *masiva* en el sentido de que $\mu > 0$, podemos tomar el 4-momento como la *4-velocidad* con una escala dada por la masa en reposo μ. La 4-velocidad v^a es un vector de género tiempo(-futuro) tangente a la línea de universo de la partícula, y cuya longitud (minkowskiana) es c (i.e., un vector *unidad* si $c = 1$):

$$p^a = \mu v^a, \quad \text{donde } v_a v^a = c^2;$$

véase la Fig. 18.17. Como se ha señalado antes, la masa en reposo de una partícula masiva es la masa (masa-energía) de dicha partícula medida en su propio sistema de reposo. Considerando que la 3-velocidad ordinaria de la partícula es $\mathbf{v}$, de modo que $\mathbf{v} = (dx^1/dt, dx^2/dt, dx^3/dt)$, donde $t = x^0$, obtenemos[18.19],[18.20]

$$\mathbf{p} = m\mathbf{v}, \quad m = \gamma\mu, \quad v^a = \gamma(c^2, \mathbf{v}),$$

donde

$$\gamma = (1 - \mathbf{v}^2/c^2)^{-1/2}.$$

Las partículas también pueden ser *sin masa* (i.e., con *masa en reposo nula*, $\mu = 0$), siendo el fotón el ejemplo primordial. Entonces el 4-mo-

[18.18] Demuestre que la fórmula para la masa aumentada es $m(1 - v^2/c^2)^{-1/2}$, donde v es la velocidad de la partícula en el segundo sistema; véase *infra*.

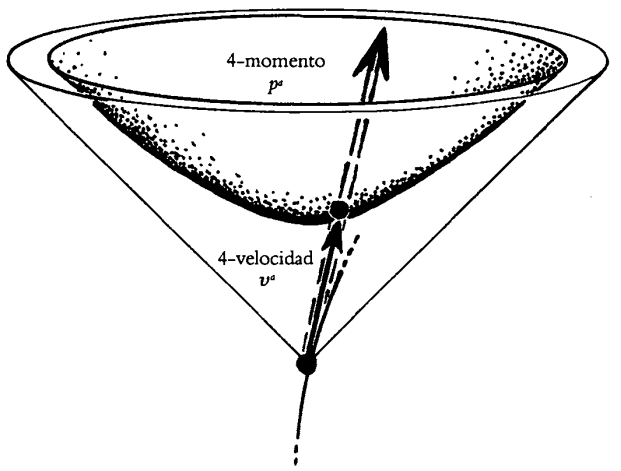


Fig. 18.17. Para una partícula con masa, el 4-momento p^a es la 4-velocidad v^a escalada por la masa en reposo μ (> 0), donde v^a es un 4-vector unidad (de género tiempo futuro) tangente a la línea de universo de la partícula (tomando $c = 1$).

[18.19] ¿Por qué?

[18.20] Utilice la serie de Taylor de §6.4 para obtener $(1 + x)^{1/2} = 1 + \frac{1}{2}x - \frac{1}{8}x^2 + \frac{1}{16}x^3 - \ldots$ A partir de aquí, obtenga un desarrollo en serie de potencias para la energía $E = [(c^2\mu)^2 + c^2\mathbf{p}^2]^{1/2}$ de una partícula de masa en reposo μ y 3-momento $\mathbf{p}$. Demuestre que el término dominante es precisamente el $E = mc^2$ de Einstein aplicado a la energía en reposo μ y que el término siguiente es la expresión newtoniana para la energía cinética. Escriba los dos términos siguientes para dar aproximaciones mejores a la energía relativista completa.

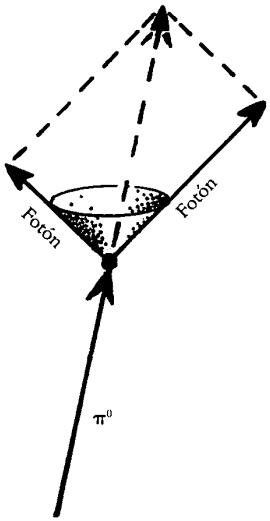


Fig. 18.18. La desintegración de un «pión neutro» masivo π^0 en 2 fotones sin masa. El 4-vector masa-energía se conserva aditivamente (aunque la masa en reposo no lo hace).

mento es un *vector nulo*. Puesto que la masa en reposo no se conserva, no hay nada que impida que una partícula masiva se desintegre en partículas sin masa, o que partículas sin masa se unan para producir partículas masivas. De hecho, una partícula conocida como «pión neutro» (denotado por π^0) se desintegrará normalmente en dos fotones en aproximadamente 10^{-16} segundos.

No obstante, en cualquier sistema de referencia particular la masa-energía total (*no* la masa en reposo) se conserva aditivamente, siendo no nula la masa-energía de cada fotón por separado. La forma en que se suman los 4-momentos se ilustra en la Fig. 18.18.

Veamos, finalmente, cómo hay que tratar el momento angular en relatividad especial. Se describe mediante una cantidad tensorial M^{ab}, antisimétrica en sus dos índices:

$$M^{ab} = -M^{ba}.$$

(Véase §22.12 para la relevancia de M^{ab} para la mecánica cuántica.) Para una simple partícula puntual sin estructura, tenemos[11]

$$M^{ab} = x^a p^b - x^b p^a,$$

donde x^a es el 4-vector de posición (de forma indexada) del punto en la línea de universo de la partícula en el instante en que se está considerando su momento angular. Si la partícula está en movimiento inercial, entonces M^{ab} es el mismo para *todos* los puntos en su línea de universo.[18.21] Para obtener el momento angular relativista total, simplemente sumamos los tensores momento angular para cada partícula por separado. Para una partícula individual (no giratoria), las tres componentes independientes puramente espaciales M^{23}, M^{31}, M^{12} son las componentes ($\times c^2$) del momento angular ordinario $\mathbf{M} = 2\mathbf{x} \wedge \mathbf{p}$ considerado antes en §18.6, y las restantes componentes independientes M^{01}, M^{02}, M^{03} constituyen la cantidad $\mathbf{N} = t\mathbf{p} - m\mathbf{x}$ ($\times c^2$). (La conservación del $\mathbf{N}$ total expresa el movimiento uniforme del centro de masas; véase la Fig. 18.16.)[18.22]

Recordemos de §18.2 que el grupo de Poincaré 10-dimensional de simetrías del espacio de Minkowski tiene 4 dimensiones que se refieren a las traslaciones espaciotemporales y las 6 restantes a rotaciones (Lorentz). Veremos en §20.6 que un principio importante de la mecánica clásica conocido como *teorema de Nöther* relaciona simetrías con leyes de conservación, y en §§21.1-5 y §22.8 que algo similar ocurre en la teoría cuántica. Esto proporciona una razón profunda para las leyes de conservación para el 4-momento p^a y el 6-momento angular M^{ab}, puesto que estas aparecen, respectivamente, a partir de las 4 simetrías traslacionales y las 6 simetrías rotacionales (Lorentz) del espacio de Minkowski. La conservación de p^a y M^{ab} tiene mucha importancia para el capítulo 21 y §§22.8,12,13.

Notas

Sección 18.1

18.1. Tom Banchoff, de la Brown University, lleva muchos años desarrollando sistemas interactivos de computación con el fin de desarrollar la intuición 4-dimensional, y en particular la visualización de funciones

[18.21] ¿Por qué?
[18.22] Explíquelo, en detalle, en el caso relativista.

complejas en términos de superficies de Riemann en $\mathbb{C}^2$. Véase Banchoff (1990, 1996).

18.2. Las cantidades «ds» en esta expresión deberían leerse simplemente como «cantidades infinitesimales» (como las ε de §13.6). Compárese con la nota 12.8.

Sección 18.2

18.3. Para una discusión particularmente detallada de los papeles de Lorentz, Poincaré y Einstein en el desarrollo de la relatividad especial, véase Stachel (1995), pp. 249-356. En mi opinión, ni siquiera Einstein tenía completa la relatividad especial en 1905, y se necesitó la perspectiva 4-dimensional de Minkowski de 1908 para completar la imagen; véase §17.8.

18.4. Existen también elementos *inversores del tiempo* en el grupo de Poincaré, que envían direcciones de género tiempo futuras a direcciones de género tiempo pasadas.

Sección 18.3

18.5. Debería resaltar, sobre todo para aquellos lectores ya familiarizados con la mecánica cuántica, que la noción compleja de «ortogonalidad» que estoy utilizando aquí es necesariamente la holomorfa (puesto que en esto consiste la «complexificación»), y no la noción hermítica de §13.9 que entra en la conjugación compleja, y que se utiliza en muchas otras áreas de la matemática y la física.

18.6. Véanse, por ejemplo, Rindler (1982, 2001), Synge (1956), Taylor y Wheeler (1963) y Hartle (2002). Para una aproximación geométrica axiomática, véase Schutz (1997).

Sección 18.5

18.7. Véanse, en particular, Newman y Penrose (1966) y Penrose y Rindler (1984, §§1.2-4, §4.15; 1986, §9.8).

18.8. Véase, por ejemplo, Rindler (1982, 2001).

18.9. Véanse, por ejemplo, Terrell (1959) y Penrose (1959).

Sección 18.6

18.10. Algunos lectores pueden estar confundidos por la presencia de un «2» en esta expresión, pero deberían reexaminar la definición de «$\wedge$» que

he dado en §11.6. Las componentes de $\mathbf{x} \wedge \mathbf{p}$ son $x^{[i}p^{j]} = \frac{1}{2}(x^ip^j - x^jp^i)$. Por lo tanto, las componentes de $\mathbf{M}$ son $x^ip^j - x^jp^i$.

Sección 18.7

18.11. Veremos en §22.8 que la mayoría de las partículas (cuánticas) también poseen un *espín intrínseco* que proporciona una contribución de «espín» constante a M^{ab} (véase §22.12) sumada al «M^{ab} orbital» que se da aquí.

19

Los campos clásicos de Maxwell y Einstein

19.1. Evolución fuera de la dinámica newtoniana

En el período comprendido entre la introducción del soberbio esquema dinámico de Newton, que podemos datar en la publicación de sus *Principia* en 1687, y la aparición de la teoría de la relatividad especial, que podría fecharse razonablemente en la primera publicación de Einstein sobre el tema (1905), se produjeron muchos desarrollos importantes para nuestras imágenes de la física fundamental. El mayor cambio ocurrido en ese período fue la comprensión, básicamente mediante los trabajos de Faraday y Maxwell en el siglo XIX, de que cierta noción de *campo físico*, que permea el espacio, debe coexistir con la previamente aceptada «realidad newtoniana» de las partículas individuales que interaccionan por medio de fuerzas instantáneas.[1] Más tarde, esta noción de «campo» se convirtió también en un ingrediente crucial de la teoría de la gravedad en un espaciotiempo curvo a la que llegó Einstein en 1915. Lo que ahora denominamos campos *clásicos* son el campo electromagnético de Maxwell y el campo gravitatorio de Einstein.

Pero hoy día sabemos que hay mucho más en la naturaleza del mundo físico que la sola física clásica. Ya en 1900, Max Planck había revelado los primeros indicios de la necesidad de una «teoría cuántica», aunque se necesitó un cuarto de siglo más antes de que pudiera ofrecerse una teoría bien formulada y global. También debería quedar claro que, además de todos estos profundos cambios que se han producido en los fundamentos «newtonianos» de la física, ha habido otros desarrollos importantes, tanto previos a dichos cambios como coexistentes con algunos de ellos en forma de poderosos avances matemáticos, dentro de

la propia teoría newtoniana. Estos avances matemáticos serán el tema del capítulo 20. Tienen interrelaciones importantes con la teoría de los campos clásicos y, lo que es incluso más significativo, constituyen un importante prerrequisito para la comprensión adecuada de la mecánica cuántica que se describirá en capítulos posteriores. Otra área importante de avance sobre la que habría que llamar la atención es la *termodinámica* (y su refinamiento conocido como *mecánica estadística*). Esta estudia el comportamiento de sistemas de un gran número de cuerpos, donde los detalles de los movimientos no se consideran importantes y el comportamiento del sistema se describe en términos de promedios de las magnitudes adecuadas. Esta fue una empresa iniciada entre mediados del siglo XIX y principios del XX, y los nombres de Carnot, Clausius, Maxwell, Boltzmann, Gibbs y Einstein son los protagonistas. Más adelante, en el capítulo 27, abordaré algunas de las cuestiones más fundamentales e intrigantes planteadas por la termodinámica.

En este capítulo describiré las teorías físicas de *campos* de Maxwell y Einstein: la «física clásica» del electromagnetismo y la gravitación. La teoría del electromagnetismo desempeña también un papel importante en la teoría cuántica, pues proporciona el «campo» arquetípico para el desarrollo posterior de la *teoría cuántica de campos* que veremos en el capítulo 26. Por el contrario, el enfoque cuántico apropiado del campo gravitatorio sigue siendo enigmático y controvertido. Abordar estas cuestiones constituirá una parte importante de los últimos capítulos de este libro (a partir del capítulo 28). No obstante, por lo que se refiere a la física que examinaremos inmediatamente, limitaremos nuestra comprensión a los campos físicos en su aspecto *clásico*.

Al inicio de este capítulo me he referido al hecho de que ya en el siglo XIX se había iniciado un cambio profundo en los fundamentos newtonianos, antes de las revoluciones de la relatividad y la teoría cuántica en el siglo XX. El primer indicio de que sería necesario un cambio semejante se produjo con los maravillosos descubrimientos experimentales de Michael Faraday hacia 1833, y de las representaciones de la realidad que encontró necesarias para acomodar dichos descubrimientos. Básicamente, el cambio fundamental consistió en considerar que las «partículas newtonianas» y las «fuerzas» que actúan entre ellas no son los únicos habitantes de nuestro universo. A partir de ahí había

que tomar en serio la idea de un «campo» con una existencia propia incorpórea. Fue el gran físico escocés James Clerk Maxwell quien en 1864 formuló las ecuaciones que debe satisfacer este «campo incorpóreo», y quien demostró que estos campos pueden transportar energía de un lugar a otro. Estas ecuaciones unificaban el comportamiento de los campos eléctricos, los campos magnéticos e incluso la luz, y hoy día son conocidas como las *ecuaciones de Maxwell*, las primeras entre las ecuaciones de campo relativistas.

Desde la perspectiva del siglo XX, cuando se han hecho profundos avances en las técnicas matemáticas (y aquí me refiero en particular al cálculo en variedades que hemos visto en los capítulos 12-15), las ecuaciones de Maxwell parecen tener una naturalidad y simplicidad convincentes que nos hacen preguntarnos cómo pudo considerarse alguna vez que el campo electromagnético pudiera obedecer a otras leyes. Pero semejante perspectiva ignora el hecho de que fueron las propias ecuaciones de Maxwell las que llevaron a muchos de estos mismos desarrollos matemáticos. Fue la forma de estas ecuaciones la que condujo a Lorentz, Poincaré y Einstein a las transformaciones espaciotemporales de la relatividad especial, que, a su vez, condujeron a la concepción de Minkowski del *espaciotiempo*. En el marco del espaciotiempo estas ecuaciones encontraron una forma que se desarrolló de manera natural en la teoría de Cartan de las formas diferenciales (§12.6); y las leyes de conservación de la teoría de Maxwell llevaron en definitiva al cuerpo de expresiones integrales que ahora se resumen de forma tan bella por esa maravillosa fórmula que se ha mencionado en §§12.5,6 como el *teorema fundamental del cálculo exterior*.

Quizá, pareciendo atribuir todos estos avances a la influencia de las ecuaciones de Maxwell, he adoptado una posición demasiado extrema en estos comentarios. En realidad, aunque no hay duda de que la influencia de las ecuaciones de Maxwell ha sido muy importante a este respecto, muchas precursoras de estas ecuaciones, tales como las ecuaciones de Laplace, D'Alembert, Gauss, Green, Ostrogradski, Coulomb, Ampère y otras, han tenido también influencias importantes. Pese a todo, seguía siendo la necesidad de entender los campos eléctrico y magnético la que básicamente proporcionó la fuerza impulsora que había tras estos desarrollos —estos y también el campo gravitatorio—. El resto de este capítulo está dedicado a la comprensión de los campos electromagnético y gravitatorio y de cómo encajan en el marco matemático moderno.

19.2. La teoría electromagnética de Maxwell

¿Cuáles son, entonces, las ecuaciones de Maxwell? Son ecuaciones en derivadas parciales (véase §10.2) que describen la evolución temporal de las tres componentes E_1, E_2, E_3 del campo eléctrico y las tres componentes B_1, B_2, B_3 del campo magnético, donde la densidad de carga eléctrica ρ y las tres componentes de la densidad de corriente eléctrica j_1, j_2, j_3 se consideran dadas. También pueden incluirse algunas otras cantidades que tienen que ver con el medio material dentro del que se considera que se propagan los campos. En las discusiones de la física *fundamental* es normal despreciar aquellos aspectos de las ecuaciones que se relacionan con dicho medio material, puesto que el propio medio consistiría, en realidad, en muchos constituyentes minúsculos, cada uno de los cuales podría tratarse en principio en el nivel más fundamental. También será conveniente escoger las que se denominan unidades «gaussianas», y utilizar *coordenadas de Minkowski estándar* (§18.1), a saber $x^0 = t, x^1 = x, x^2 = y, x^3 = z$ (signatura $+---$) con unidades espaciotemporales tales que la velocidad de la luz c se toma como unidad ($c = 1$).

El campo electromagnético y la densidad de corriente de carga se resumen, respectivamente (de acuerdo con una receta debida originalmente a Minkowski), en una 2-forma espaciotemporal $\boldsymbol{F}$, denominada *tensor de campo de Maxwell*, y un vector espaciotemporal $\boldsymbol{J}$, denominado *vector de carga-corriente*, con sus componentes mostradas en forma matricial como:

$$\begin{pmatrix} F_{00} & F_{01} & F_{02} & F_{03} \\ F_{10} & F_{11} & F_{12} & F_{13} \\ F_{20} & F_{21} & F_{22} & F_{23} \\ F_{30} & F_{31} & F_{32} & F_{33} \end{pmatrix} = \begin{pmatrix} 0 & E_1 & E_2 & E_3 \\ -E_1 & 0 & -B_3 & B_2 \\ -E_2 & B_3 & 0 & -B_1 \\ -E_3 & -B_2 & B_1 & 0 \end{pmatrix}$$

$$\begin{pmatrix} J^0 \\ J^1 \\ J^2 \\ J^3 \end{pmatrix} = \begin{pmatrix} \rho \\ j_1 \\ j_2 \\ j_3 \end{pmatrix}$$

Nótese que se satisface la antisimetría $F_{ba} = -F_{ab}$, como se exige para una 2-forma. También haré uso de lo que se conocen como los *duales de Hodge* de $\boldsymbol{F}$ y $\boldsymbol{J}$, que son, respectivamente, la 2-forma $^*\boldsymbol{F}$ y la 3-forma $^*\boldsymbol{J}$, definidas por

$$\begin{pmatrix} {}^*F_{00} & {}^*F_{01} & {}^*F_{02} & {}^*F_{03} \\ {}^*F_{10} & {}^*F_{11} & {}^*F_{12} & {}^*F_{13} \\ {}^*F_{20} & {}^*F_{21} & {}^*F_{22} & {}^*F_{23} \\ {}^*F_{30} & {}^*F_{31} & {}^*F_{32} & {}^*F_{33} \end{pmatrix} = \begin{pmatrix} 0 & -B_1 & -B_2 & -B_3 \\ B_1 & 0 & -E_3 & E_2 \\ B_2 & E_3 & 0 & -E_1 \\ B_3 & -E_2 & E_1 & 0 \end{pmatrix}$$

$$\begin{pmatrix} {}^*J_{123} \\ {}^*J_{023} \\ {}^*J_{013} \\ {}^*J_{012} \end{pmatrix} = \begin{pmatrix} -\rho \\ j_1 \\ -j_2 \\ j_3 \end{pmatrix}$$

donde se satisfacen las propiedades de antisimetría requeridas $^*F_{ab} = {}^*F_{[ab]}$, $^*J_{abc} = {}^*J_{[abc]}$. En términos del *tensor de Levi-Civita* $\boldsymbol{\varepsilon}$ (§12.7), con componentes totalmente antisimétricas $\boldsymbol{\varepsilon}_{abcd} (= \boldsymbol{\varepsilon}_{[abcd]})$ y normalizadas de modo que $\boldsymbol{\varepsilon}_{0123} = 1$, los duales pueden escribirse como

$$^*F_{ab} = \frac{1}{2}\boldsymbol{\varepsilon}_{abcd}F^{cd} \quad \text{y} \quad {}^*J_{abc} = \boldsymbol{\varepsilon}_{abcd}J^d,$$

donde la versión ascendida F^{ab} de F_{ab} es simplemente $g^{ac}g^{bd}F_{cd}$, de acuerdo con §14.8. Nótese que la versión «ascendida» $\boldsymbol{\varepsilon}^{abcd} = g^{ap}g^{bq}g^{cr}g^{ds}\boldsymbol{\varepsilon}_{pqrs}$ satisface $\boldsymbol{\varepsilon}^{0123} = -1$, de donde la $\boldsymbol{\epsilon}$ de §12.7 viene dada por[19.1] $\boldsymbol{\epsilon}^{abcd} = -\boldsymbol{\varepsilon}^{abcd}$. Véase la Fig. 19.1 para la forma diagramática de estas operaciones «dualizadoras» (y también de las propias ecuaciones de Maxwell). Veremos que la noción de un «dual» en este sentido (y otros relacionados) tendrá importancia más adelante, en varios contextos diferentes.

Habría que comentar algo sobre el significado geométrico del dual de Hodge. Recordemos de §12.7 que la operación de pasar de un *bivector* $\boldsymbol{H}$, como el descrito por la cantidad antisimétrica H^{ab}, a su 2-forma «dual» $\boldsymbol{H}^{\#}$, dada por $\frac{1}{2}\boldsymbol{\varepsilon}_{abcd}\boldsymbol{H}^{cd}$, no supone mucha diferencia para su interpretación geométrica. Si $\boldsymbol{H}$ fuera un bivector *simple*, por ejemplo, de modo que la 2-forma $\boldsymbol{H}^{\#}$ sería también simple (véase el final de §12.7), entonces el elemento 2-plano determinado por $\boldsymbol{H}^{\#}$ sería precisamente el mismo que el elemento 2-plano determinado por $\boldsymbol{H}$ (con la única diferencia de que, estrictamente, $\boldsymbol{H}^{\#}$ tiene el carácter de una densidad, como se ha indicado en §12.7). Por el contrario, la subida de índices que nos lleva de una 2-forma H_{ab} a un bivector H^{ab} $(= H_{cd}g^{ca}g^{db})$, tiene un efecto geométrico más significativo. En el caso de un bivector simple, el elemento 2-plano determinado por H_{ab} es el complemento ortogonal del elemento 2-plano determinado por H^{ab} (véase §18.3). El dual de Hodge, tal como se aplica a la 2-forma H_{ab}, que nos lleva a $\frac{1}{2}\boldsymbol{\varepsilon}_{abcd}H^{cd}$ (i.e., a $\boldsymbol{H}^{\#}$), emplea la subida de índices $H_{ab} \mapsto H^{ab}$, y por con-

[19.1] Compruebe ambas afirmaciones.

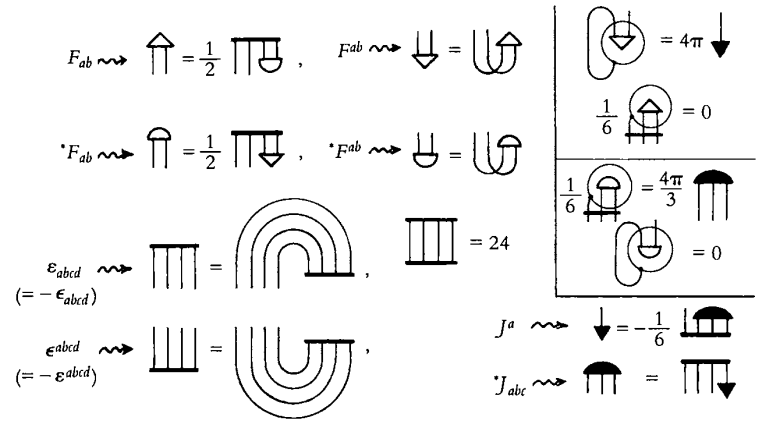


Fig. 19.1. Diagrama para los duales de Hodge y las ecuaciones de Maxwell. Las cantidades $\boldsymbol{\varepsilon}_{abcd} = (\boldsymbol{\varepsilon}_{[abcd]})$ y $\boldsymbol{\epsilon}^{abcd} = (\boldsymbol{\epsilon}^{[abcd]})$, normalizadas de modo que $\boldsymbol{\epsilon}_{0123} = \boldsymbol{\epsilon}^{0123} = 1$ en un marco de Minkowski estándar, están relacionadas con sus versiones de índices subidos/bajados (por g^{ab} y g_{ab}) por $\boldsymbol{\varepsilon}_{abcd} = -\boldsymbol{\epsilon}_{abcd}$ y $\boldsymbol{\epsilon}^{abcd} = -\boldsymbol{\varepsilon}^{abcd}$. En el diagrama (centro izquierda, dos líneas inferiores) este cambio de signo de absorbe por una inversión efectiva de índices. En la caja superior derecha están las ecuaciones de Maxwell, primero utilizando el tensor de campo $\boldsymbol{F}$ (con su forma elevada $F^{ab} = g^{ac}g^{bd}F_{cd}$; cf. Fig. 14.21), de modo que las ecuaciones son $\nabla_a F^{ab} = 4\pi J^b$, $\nabla_{[a}F_{bc]} = 0$, y por debajo de ello, utilizando correspondientemente el dual $^*\boldsymbol{F}$ (donde $^*F_{ab} = \frac{1}{2}\boldsymbol{\varepsilon}_{abcd}F^{cd}$, $^*\boldsymbol{J}_{abc} = \boldsymbol{\varepsilon}_{abcd}\boldsymbol{J}^d$), de modo que las ecuaciones son $\nabla {}^*_{[a}F_{bc]} = \frac{4\pi}{3}\,{}^*J_{abc}$, $\nabla {}^*_a F^{ab} = 0$.

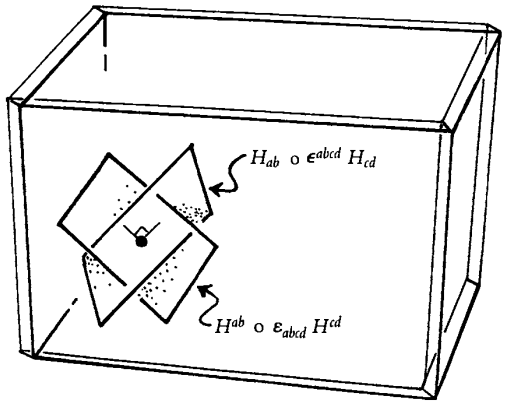


Fig. 19.2. En el 4-espacio un simple bivector $\boldsymbol{H}$ (H^{ab}) representa el mismo elemento 2-plano que su forma «dual» $H^{\#}(\frac{1}{2}\boldsymbol{\epsilon}_{abcd}H^{cd})$. Pero la versión de $\boldsymbol{H}$ de índice bajado, simple 2-forma H_{ab}, que es equivalente a su bivector «dual» $\frac{1}{2}\boldsymbol{\epsilon}_{abcd}H^{cd}$, representa el elemento 2-plano complemento ortogonal (véase la Fig. 18.4). De aquí que sea la subida/bajada de índices en el dual de Hodge la que lleva al paso al complemento ortogonal.

siguiente, implica pasar al complemento ortogonal. Véase la Fig. 19.2. En consecuencia, el dual de Hodge que nos lleva de $\boldsymbol{F}$ a $^*\boldsymbol{F}$ también implica un complemento ortogonal.

Una vez establecida esta notación, las *ecuaciones de Maxwell* pueden escribirse ahora de forma muy simple como[19.2]

$$d\boldsymbol{F} = 0, \quad d{}^*\boldsymbol{F} = 4\pi{}^*\boldsymbol{J}.$$

También podemos escribir las ecuaciones de Maxwell totalmente de forma indexada como[19.3]

$$\nabla_{[a}F_{bc]} = 0, \quad \nabla_a F^{ab} = 4\pi J^b.$$

[19.2] Escríbalas en detalle, en términos de las componentes de los campos eléctrico y magnético, mostrando que estas ecuaciones proporcionan una evolución temporal de los campos eléctrico y magnético, en términos del operador $\partial/\partial t$.
[19.3] Demuestre la equivalencia con el par de ecuaciones anterior.

Nótese que si aplicamos el operador derivada exterior d a ambos miembros de la segunda ecuación de Maxwell $d{}^*\boldsymbol{F} = 4\pi{}^*\boldsymbol{J}$, y utilizamos el hecho de que $d^2 = 0$ (§12.6), deducimos que el vector carga-corriente $\boldsymbol{J}$ satisface la ecuación de «divergencia nula»[19.4]

$$d{}^*\boldsymbol{J} = 0 \quad \text{o, lo que es equivalente,} \quad \nabla_a J^a = 0,$$

En este punto, y como pequeña digresión que más adelante tendrá una importancia considerable para nosotros (§32.2 y §§33.6,8,11; véanse §18.3 y el ejercicio [18.5](ii)), vale la pena señalar las partes *autodual* y *anti-autodual* del tensor de Maxwell, dadas respectivamente por

$${}^+\boldsymbol{F} = \frac{1}{2}(\boldsymbol{F} - i{}^*\boldsymbol{F}) \quad \text{y} \quad {}^-\boldsymbol{F} = \frac{1}{2}(\boldsymbol{F} + i{}^*\boldsymbol{F})$$

(que son complejas conjugadas una de otra). Resulta que en la teoría cuántica estas cantidades complejas describen, respectivamente, los fotones (cuantos del campo electromagnético) con espín *a derechas* o *a izquierdas*; véanse §§22.7,12 y la Fig. 22.7. Las propiedades autodual/anti-autodual se expresan en[19.5]

$${}^*({}^{\pm}\boldsymbol{F}) = \pm i{}^{\pm}\boldsymbol{F}.$$

Teniendo en cuenta que $^*\boldsymbol{J}$ es real, podemos combinar las dos ecuaciones de Maxwell (como partes imaginaria y real respectivamente) en la forma

$$d\,{}^+\boldsymbol{F} = -2\pi i\,{}^*\boldsymbol{J}.$$

Los fotones proporcionan la descripción corpuscular de la *luz*, y en el capítulo 21 veremos que la teoría cuántica permite que coexistan una descripción corpuscular y una descripción ondulatoria para la luz. Una de las grandes hazañas de Maxwell fue demostrar mediante sus ecuaciones que existen ondas electromagnéticas que viajan con la velocidad de la luz y tienen todas las propiedades de polariza-

[19.4] Demuestre que las dos versiones de esta divergencia nula son equivalentes.
[19.5] Demuéstrelo probando, primero, que dualizar dos veces produce *menos* la cantidad original. ¿Tiene relación este signo con la signatura lorentziana del espacio-tiempo? Explíquelo.

ción conocidas para la luz (y que examinaremos en §22.7). De acuerdo con estos hechos notables, Maxwell propuso que la luz es en realidad un fenómeno electromagnético. En 1888, casi un cuarto de siglo después de que Maxwell publicara sus ecuaciones, Heinrich Hertz confirmó experimentalmente la maravillosa predicción teórica de Maxwell.

En la descripción explícita anterior he supuesto que el espaciotiempo de fondo es el espacio de Minkowski plano $\mathbb{M}$, y las discusiones que siguen, en §§19.3,4, y en la primera parte de §19.5, también pueden considerarse sobre esta base. Sin embargo, esto no es realmente necesario, y todas las conclusiones se siguen aplicando en presencia de curvatura espaciotemporal. Para ello, las componentes dadas antes deben considerarse tomadas con respecto a cierto sistema de Minkowski local, y la notación de índices se ocupará del resto.[19.6]

19.3. Leyes de conservación y de flujo en la teoría de Maxwell

La divergencia nula del vector de carga-corriente nos proporciona la ecuación de *conservación de la carga eléctrica*. La razón de que se conozca como una «ecuación de conservación» se debe al hecho de que, por el teorema fundamental del cálculo exterior (véase §12.6), tenemos $\int_{\mathcal{R}} \mathrm{d}\,{}^*\boldsymbol{J} = \int_{\partial\mathcal{R}} {}^*\boldsymbol{J}$, de modo que

$$\int_{\mathcal{Q}} {}^*\boldsymbol{J} = 0,$$

integrado sobre cualquier 3-superficie cerrada $\mathcal{Q}$ en el espacio de Minkowski $\mathbb{M}$. (Cualquier 3-superficie cerrada en $\mathbb{M}$ es la frontera $\partial\mathcal{R}$ de cierta región 4-dimensional compacta $\mathcal{R}$ en $\mathbb{M}$.) Véase la Fig. 19.3. La magnitud ${}^*\boldsymbol{J}$ puede interpretarse como el «flujo de carga» a través de $\mathcal{Q} = \partial\mathcal{R}$. Así pues, lo que la ecuación anterior nos dice es que el flujo neto de carga eléctrica a través de esta frontera tiene que ser cero, i.e.,

[19.6] ¿Puede desarrollar esto? ¿Qué le sucede a las componentes de $\boldsymbol{F}$ y ${}^*\boldsymbol{F}$ en un sistema general de coordenadas curvilíneas? ¿Por qué no quedan afectadas las ecuaciones de Maxwell si se expresan correctamente?

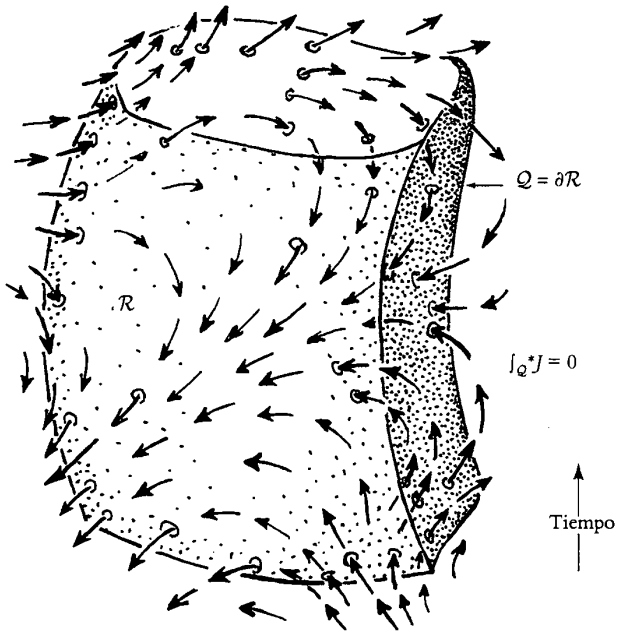


Fig. 19.3. Conservación de la carga eléctrica en el espaciotiempo. La 3-superficie cerrada $\mathcal{Q}$ es la frontera $\mathcal{Q} = \partial\mathcal{R}$ de un 4-volumen compacto $\mathcal{R}$, en el espaciotiempo de Minkowski $\mathbb{M}$, de modo que el teorema fundamental del cálculo exterior nos dice $\int_{\mathcal{Q}} {}^*\boldsymbol{J} = \int_{\mathcal{R}} \mathrm{d}\,{}^*\boldsymbol{J} = 0$, puesto que $\mathrm{d}\,{}^*\boldsymbol{J} = 0$. La cantidad ${}^*\boldsymbol{J}$ describe el «flujo» de carga a través de $\mathcal{Q}$, de modo que la carga total que atraviesa $\mathcal{Q}$ hacia dentro es igual a la que lo atraviesa hacia fuera, lo que expresa la conservación de carga.

la cantidad total que entra en $\mathcal{R}$ tiene que ser exactamente igual a la cantidad total que sale de $\mathcal{R}$: *la carga eléctrica se conserva*.[19.7]

También podemos utilizar la segunda ecuación de Maxwell $\mathrm{d}\,{}^*\boldsymbol{F} = 4\pi\,{}^*\boldsymbol{J}$ para obtener lo que se denomina una «ley de Gauss». Esta ley concreta se aplica en un instante dado $t = t_0$, de modo que ahora esta-

[19.7] Aunque correcto, este argumento ha sido dado de forma algo retórica. Desarrolle los detalles completamente, en el caso en que $\mathcal{R}$ es un «cilindro» espaciotemporal que consiste en una región espacial agotada que es constante en el tiempo, para un intervalo finito dado de la coordenada temporal t. Explique las diferentes nociones de «flujo de carga» implicadas, contrastando el flujo en la «base» y el «techo» de género espacio del cilindro con el flujo en los «lados» de género tiempo.

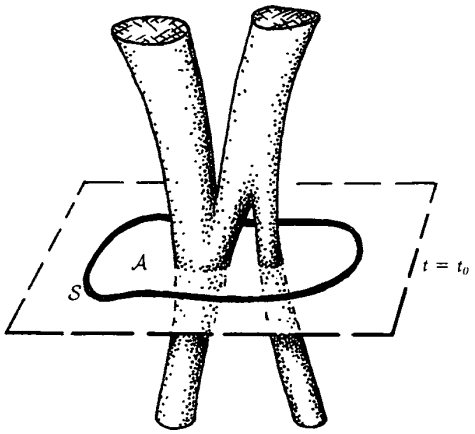


Fig. 19.4. Dentro de la 3-superficie de tiempo constante $t = t_0$, la $d{}^*\boldsymbol{F} = 4\pi{}^*\boldsymbol{J}$ de Maxwell nos da la ley de Gauss, por la que la integral del flujo eléctrico (integral de ${}^*\boldsymbol{F}$) sobre una 2-superficie espacial cerrada mide la carga total rodeada (por el teorema fundamental del cálculo exterior). De hecho, esto no está restringido a 2-superficies a tiempo constante, y la ley de Gauss se generaliza por ello.

mos utilizando la versión tridimensional del teorema fundamental del cálculo exterior. La ley nos dice cuál es el valor de la carga eléctrica total que hay dentro de una 2-superficie cerrada $\mathcal{S}$ en el instante t_0 (véase la Fig. 19.4), expresando esta carga como una integral sobre $\mathcal{S}$ del dual del tensor de Maxwell ${}^*\boldsymbol{F}$, lo que equivale a decir que podemos obtener la carga total encerrada por $\mathcal{S}$ si integramos el flujo total del campo eléctrico $\boldsymbol{E}$ a través de $\mathcal{S}$.[19.8]

De manera más general, esto se aplica incluso si $\mathcal{S}$ no se encuentra en un instante determinado $t = t_0$. Supongamos que $\mathcal{S}$ es la 2-frontera de género espacio de cierta región espacial compacta $\mathcal{A}$. Entonces, la carga total χ en la región $\mathcal{A}$, rodeada por $\mathcal{S}$ (o, en términos espaciotemporales, «ensartada en» $\mathcal{S}$; véase la Fig. 19.4), está dada por

$$\int_{\mathcal{S}} {}^*\boldsymbol{F} = 4\pi\chi, \quad \text{donde } \chi = \int_{\mathcal{A}} {}^*\boldsymbol{J}.$$

También podemos obtener un tipo relacionado de ley de conservación a partir de la primera ecuación de Maxwell $d\boldsymbol{F} = 0$. Esta tiene precisamente la misma forma que la segunda ecuación de Maxwell, excepto que $\boldsymbol{F}$ reemplaza a ${}^*\boldsymbol{F}$ y la «fuente» correspondiente a ${}^*\boldsymbol{J}$ es ahora cero. Por lo tanto, para cualquier 2-superficie cerrada en el espacio de Minkowski,[2] siempre tenemos la ley de flujo

[19.8] Explique en detalle por qué esto es precisamente el flujo eléctrico.

$$\int_{\mathcal{S}} \boldsymbol{F} = 0.$$

Nótese que al pasar de ${}^*\boldsymbol{F}$ a $\boldsymbol{F}$ (o de $\boldsymbol{F}$ a ${}^*\boldsymbol{F}$) simplemente intercambiamos los vectores campo eléctrico y campo magnético (con un cambio de signo para uno de ellos). La ausencia de una fuente para $\boldsymbol{F}$ es una expresión del hecho de que (hasta donde se sabe) *no existen monopolos magnéticos* en la naturaleza. Un monopolo magnético sería un polo norte magnético o un polo sur magnético aislado, en lugar de los polos norte y sur que siempre aparecen en pares, que es lo que sucede en un imán ordinario. (Estos últimos no son entidades físicas independientes, sino que aparecen a partir de la circulación de cargas *eléctricas*.) Parece que en la naturaleza no hay nunca una «carga magnética» neta («intensidad de polo» no nula) en un objeto físico. Desde el punto de vista de las ecuaciones de Maxwell únicamente, no parece haber ninguna buena razón para la ausencia de monopolos magnéticos, puesto que podríamos añadir un segundo miembro a la primera ecuación de Maxwell $d\boldsymbol{F} = 0$ sin ninguna pérdida de consistencia. De hecho, en ocasiones los físicos han contemplado la posibilidad de que realmente pudieran existir monopolos magnéticos y han intentado buscarlos. Su existencia tendría importantes implicaciones para la física de partículas (véase §28.2), pero de momento no hay ningún indicio de que existan tales monopolos en el universo real.

19.4. El campo de Maxwell como curvatura gauge

La *primera* ecuación de Maxwell $d\boldsymbol{F} = 0$ tiene también como consecuencia que

$$\boldsymbol{F} = 2d\boldsymbol{A},$$

para cierta 1-forma $\boldsymbol{A}$. (Esto se sigue del «lema de Poincaré», que establece que si la r-forma $\boldsymbol{\alpha}$ satisface $d\boldsymbol{\alpha} = 0$, entonces *localmente* existe siempre una $(r-1)$-forma $\boldsymbol{\beta}$ para la que $\boldsymbol{\alpha} = d\boldsymbol{\beta}$; véase §12.6.) Además, en una región con topología euclídea, este resultado local se amplía a uno global.[3] La magnitud $\boldsymbol{A}$ se denomina *potencial electromagnético*. No está unívocamente determinada por el campo $\boldsymbol{F}$, pero está

fijada salvo la adición de una cantidad $\mathrm{d}\boldsymbol{\Theta}$,[19.9] donde $\boldsymbol{\Theta}$ es un campo escalar real

$$\boldsymbol{A} \mapsto \boldsymbol{A} + \mathrm{d}\boldsymbol{\Theta}.$$

En forma de índices, estas relaciones son

$$F_{ab} = \nabla_a A_b - \nabla_b A_a$$

con libertad

$$A_a \mapsto A_a + \nabla_a \Theta.$$

Esta «libertad gauge» en el potencial electromagnético nos dice que $\boldsymbol{A}$ no es una magnitud localmente medible. No puede haber ningún experimento para medir «el valor de $\boldsymbol{A}$» en un punto porque $\boldsymbol{A} + \mathrm{d}\boldsymbol{\Theta}$ sirve exactamente para el mismo propósito físico que $\boldsymbol{A}$. Sin embargo, el potencial proporciona la clave matemática para la forma en que el campo de Maxwell interacciona con alguna otra entidad física $\boldsymbol{\Psi}$. ¿Cómo funciona esto? El papel específico de A_a consiste en que nos proporciona una *conexión gauge* (o conexión fibrada; véase §15.8)

$$\nabla_a = \partial/\partial x^a - \mathrm{i}eA_a,$$

donde e es un número real concreto que cuantifica la *carga eléctrica* de la entidad descrita por $\boldsymbol{\Psi}$. De hecho, esta «entidad» será en general una partícula cuántica cargada, tal como un electrón o un protón, y $\boldsymbol{\Psi}$ sería entonces su función de onda mecanocuántica. El significado completo de estos términos tendrá que esperar a la exposición del capítulo 21, cuando se explique la noción de una función de onda. Todo lo que necesitamos saber sobre ello ahora es que $\boldsymbol{\Psi}$ va a considerarse como una sección transversal de un fibrado (§15.3) que describe campos cargados, y es sobre este fibrado sobre el que actúa $\boldsymbol{\nabla}$ como una conexión.

Las magnitudes del campo electromagnético $\boldsymbol{F}$ y $\boldsymbol{A}$ no tienen carga ($e = 0$), de modo que ninguna de nuestras ecuaciones de Maxwell, etc., queda alterada por tener esta nueva definición para ∇_a; i.e., aún tenemos $\nabla_a = \partial/\partial x^a$ en dichas ecuaciones, en coordenadas de Minkowski planas, o la generalización adecuada (véase §14.3) si estamos considerando un espaciotiempo curvo. ¿Cuál es la naturaleza geométrica del fibrado sobre el que actúa esta conexión? Un punto de vista posible consiste en considerar que este fibrado tiene fibras que son círculos (S^1s), sobre el espaciotiempo $\mathbb{M}$, que describen una fase multiplicativa $\mathrm{e}^{\mathrm{i}\theta}$ para $\boldsymbol{\Psi}$. (Este es el tipo de cosas que suceden en la imagen «Kaluza-Klein» mencionada en §15.1 pero en ese caso el fibrado entero se considera «espaciotiempo».) Más adecuado es considerar el fibrado como el fibrado vectorial de los valores posibles de $\boldsymbol{\Psi}$ en cada punto, donde la libertad de multiplicación por una fase hace del fibrado un U(1)-fibrado sobre el espaciotiempo $\mathbb{M}$. (Una cuestión de este tipo ha sido considerada al final de §15.8.) Para que esto tenga sentido, $\boldsymbol{\Psi}$ debe ser un campo complejo cuya interpretación física es, en un sentido apropiado, insensible al reemplazamiento $\boldsymbol{\Psi} \mapsto \mathrm{e}^{\mathrm{i}\theta}\,\boldsymbol{\Psi}$ (donde θ es algún campo con valores reales en la variedad $\mathcal{M}$). Este reemplazamiento se conoce como una *transformación gauge* electromagnética, y el hecho de que la interpretación física sea insensible a este reemplazamiento se denomina *invariancia gauge*. La *curvatura* de nuestra conexión fibrada resulta ser entonces el tensor de campo de Maxwell F_{ab}.[19.10]

Antes de seguir con estas ideas es oportuno hacer unos breves comentarios históricos. Poco después de que Einstein introdujera su teoría de la relatividad general en 1915, Weyl sugirió en 1918 una generalización en la que la propia noción de *longitud* se hace dependiente del camino. (Hermann Weyl, 1885-1955, fue una importante figura matemática del siglo XX. En realidad, de entre todos los matemáticos cuya obra se realizó por completo en el siglo XX, él fue, en mi opinión, el más influyente, y destacó no solo como matemático puro, sino también como físico.) En la teoría de Weyl, los conos nulos retienen el papel fundamental que tienen en la teoría de Einstein (por ejemplo, definir las velocidades límite para partículas masivas y proporcionarnos el «grupo de Lorentz» local que va a actuar en la vecindad de cada punto), de modo que localmente se sigue requiriendo una métrica $\boldsymbol{g}$ lorentziana (digamos $+---$) con el objetivo de definir dichos conos. Sin embargo, en el esquema de Weyl no hay escala absoluta para medi-

[19.9] ¿Por qué se puede sumar una cantidad semejante?

[19.10] Demuéstrelo. *Sugerencia*: Eche una ojeada a §15.8.

das de tiempo o espacio, de modo que la métrica viene dada *salvo proporcionalidad*. Así pues, se permiten transformaciones de la forma

$$\boldsymbol{g} \mapsto \lambda \boldsymbol{g},$$

para una función escalar λ (digamos positiva) en el espaciotiempo $\mathcal{M}$, que no afectan a los conos nulos de $\mathcal{M}$. (Tales transformaciones se conocen como *reescalamientos conformes* de la métrica $\boldsymbol{g}$; en la teoría de Weyl, cada elección de $\boldsymbol{g}$ nos proporciona un *gauge* posible en términos del cual pueden medirse distancias y tiempos.) Aunque quizá Weyl estuviera pensando más en las separaciones espaciales, será oportuno que las consideremos en términos de medidas temporales (de acuerdo con el punto de vista del capítulo 17.) Así pues, en la geometría de Weyl no existen «relojes ideales» absolutos. El ritmo al que cualquier reloj mide el tiempo dependerá de su historia.

La situación es «peor» que la que se da en la «paradoja del reloj» estándar que he descrito en §18.3 (Fig. 18.6d). En la geometría de Weyl podemos imaginar a un viajero espacial que viaja a una estrella lejana y luego vuelve a la Tierra para encontrar no solo que los que están en la Tierra han envejecido mucho más, sino también que ¡los relojes de la Tierra corren ahora a un *ritmo* diferente del de los relojes del cohete! Véase la Fig. 19.5a. Utilizando esta sorprendente idea, Weyl fue capaz de incorporar las ecuaciones de la teoría electromagnética de Maxwell en la geometría espaciotemporal.

La forma en que lo hizo consistió esencialmente en codificar el potencial electromagnético en una conexión fibrada, igual que he hecho más arriba, pero sin la unidad imaginaria «i» en la expresión para ∇_a. Podemos considerar que el fibrado relevante sobre $\mathcal{M}$ está dado por las métricas lorentzianas $\boldsymbol{g}$ que comparten los mismos conos nulos. Así pues, la fibra sobre cierto punto x en $\mathcal{M}$ consiste en una familia de métricas *proporcionales* (en las que, si así lo deseamos, podemos escoger que los factores de proporcionalidad sean positivos). Estos factores son las posibles «λs» en la $\boldsymbol{g} \mapsto \lambda \boldsymbol{g}$ anterior. Para cualquier elección particular de métrica, tenemos un *gauge* por el que se definen distancias o tiempos a lo largo de curvas. Pero no hay ninguna elección absoluta de gauge, y por lo tanto ninguna elección preferida de métrica $\boldsymbol{g}$ de entre la clase de equivalencia de las métricas proporcionales. No obstante,

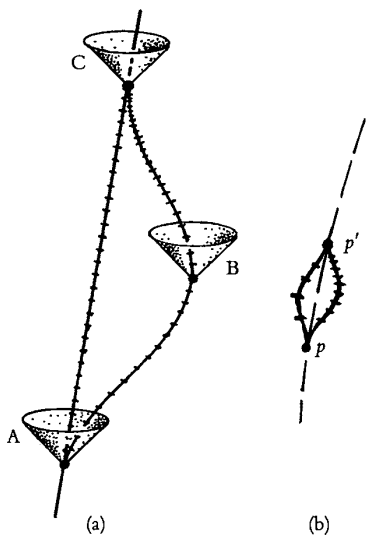


Fig. 19.5. En la teoría gauge del electromagnetismo original de Weyl, la noción de intervalo temporal (o de intervalo espacial) no es absoluta sino que depende del camino tomado. (a) Una comparación con la «paradoja del reloj» ilustrada en la Fig. 18.6; en la teoría de Weyl encontramos que el viajero espacial llega a casa (línea de universo ABC) para encontrar no solo lecturas de reloj diferentes entre las de la Tierra (ruta directa AB) y las de la nave espacial, ¡sino también diferentes *ritmos* de reloj! (b) La curvatura gauge de Weyl (que da el campo de Maxwell $\boldsymbol{F}$) resulta de este cambio de escala temporal (conforme) cuando rodeamos un lazo infinitesimal (diferencia entre dos rutas desde p a un punto vecino p').

existe cierta estructura adicional a esta estructura de conos nulos (i.e., a la estructura *conforme*), a saber, una conexión fibrada —o *conexión gauge*— que introdujo Weyl para obtener la $\boldsymbol{F}$ de Maxwell (i.e., F_{ab}) como su *curvatura*. Esta curvatura mide la discrepancia en los ritmos de los relojes como se ilustra en la Fig. 19.5a cuando las líneas de universo difieren solo en una parte infinitesimal; véase la Fig. 19.5b. (Esto puede compararse con el «fibrado tenso» $\mathcal{B}^{\mathbb{C}}$, sobre $\mathbb{C}$, considerado en §15.8, Figs. 15.16 y 15.19; el concepto fibrado básico es muy similar.)

Cuando Einstein supo de esta teoría, le dijo a Weyl que tenía una objeción fundamental que hacer desde el punto de vista *físico*, a pesar de la elegancia matemática de las ideas de este. Las frecuencias espectrales, por ejemplo, no parecen estar en absoluto afectadas por la historia de un átomo, mientras que la teoría de Weyl predecía lo contrario. Y lo que es más fundamental, aunque no todas las reglas mecanocuánticas relevantes habían sido completamente formuladas en esa época (y llegaremos a estas más adelante, en §21.4 y §§23.7,8), la teoría de Weyl está en conflicto con la identidad necesariamente exacta entre diferentes partículas del mismo tipo (véase §23.7). En particular, existe una relación directa entre ritmos de relojes y masas de partículas. Como ve-

remos más adelante, una partícula de masa en reposo m tiene una frecuencia natural mc^2h^{-1}, donde h es la constante de Planck y c la velocidad de la luz. Así pues, en la geometría de Weyl, no solo los ritmos de los relojes, sino también la *masa* de una partícula dependerán de su historia. En consecuencia, según la teoría de Weyl, dos protones que hubieran tenido historias diferentes tendrían casi con certeza masas diferentes, con lo que se viola el principio mecanocuántico de que las partículas del mismo tipo tienen que ser *exactamente* idénticas (véanse §§23.7,8).

Aunque esta observación suponía una condena para la versión original de la teoría de Weyl, más tarde se comprendió[4] que la misma idea funcionaría si el «gauge» de Weyl se refiriese no al escalamiento *real* (por λ), sino a un escalamiento por un *número complejo de módulo unidad* ($e^{i\theta}$). Esta puede parecer una idea extraña, pero como veremos en el capítulo 21 y siguientes (véanse muy en especial §§21.6,9), las reglas de la mecánica cuántica nos obligan al uso de números complejos en la descripción del estado de un sistema. Existe en particular un número complejo $e^{i\theta}$ de módulo unidad por el que puede multiplicarse este «estado cuántico» —el estado que se suele conocer como Ψ— sin consecuencias observables localmente. Este reemplazamiento «no observable» $\Psi \mapsto e^{i\theta}\Psi$ se conoce todavía hoy día como una «transformación gauge», incluso si ahora no hay implicado ningún cambio en la escala de longitud, pues el cambio es una rotación en el plano complejo (un plano complejo sin conexión directa con dimensiones espaciales ni temporales). De esta forma extrañamente retorcida, la idea de Weyl proporcionaba el escenario físico adecuado para una $U(1)$-conexión del tipo que he ilustrado al final del capítulo 15, y que ahora constituye la base de la imagen moderna de la forma de interacción real del campo electromagnético. El operador ∇ que se ha definido antes a partir del potencial electromagnético (i.e., $\nabla_a = \partial/\partial x^a - ieA_a$) proporciona una conexión $U(1)$-fibrada en el fibrado de funciones de ondas cuánticas cargadas ψ (véase §21.9).

Es interesante que la dependencia de camino de la conexión (que podemos comparar con la dependencia de camino ilustrada en la Fig. 19.5) se manifiesta de una forma sorprendente en ciertos tipos de situaciones experimentales, que ilustran lo que se conoce como el *efecto*

Aharonov-Bohm.[5] Puesto que nuestra conexión ∇ opera solamente en el nivel de los fenómenos cuánticos, no vemos esta dependencia del camino en los experimentos clásicos; en su lugar, el efecto Aharonov-Bohm depende de la *interferencia cuántica* (véanse §21.4 y la Fig. 21.4). En la versión más conocida se lanzan electrones desde una fuente para que atraviesen dos regiones que están libres de campo electromagnético ($\boldsymbol{F} = 0$), pero separadas una de otra por un largo solenoide cilíndrico (que contiene líneas de fuerza magnéticas), hasta llegar a una pantalla detectora que hay detrás (véase la Fig. 19.6a). En ningún momento encuentran los electrones un campo $\boldsymbol{F}$ no nulo. Sin embargo, la región relevante $\mathcal{R}$ libre de campo (que empieza en la fuente, se bifurca de modo que pasa a uno u otro lado del solenoide, y se vuelve a unir en la pantalla) no es simplemente conexa, y el campo $\boldsymbol{F}$ fuera de $\mathcal{R}$ es tal que no hay elección gauge para la que el potencial $\boldsymbol{A}$ desaparece en todos los lugares dentro de $\mathcal{R}$. La presencia de este potencial no nulo en la región no simplemente conexa $\mathcal{R}$ —o, más correctamente, la dependencia de camino de ∇ en $\mathcal{R}$— conduce a un desplazamiento en las franjas de interferencia en la pantalla.

De hecho, el efecto de desplazamiento de franjas no depende de ningún valor local especial que pudiera tomar $\boldsymbol{A}$ (lo que no puede hacer, porque $\boldsymbol{A}$ no es localmente observable, como se ha mencionado antes), sino de cierta integral no local de $\boldsymbol{A}$. Esta es la cantidad $\oint \boldsymbol{A}$, tomada a lo largo de un lazo topológicamente no trivial dentro de $\mathcal{R}$. Véase la Fig. 19.6b. Puesto que $d\boldsymbol{A}$ se anula dentro de $\mathcal{R}$ (porque $\boldsymbol{F} = 0$ en $\mathcal{R}$), la integral $\oint \boldsymbol{A}$ no se ve afectada si movemos de forma continua nuestro lazo cerrado dentro de $\mathcal{R}$.[19.11] De esto resulta evidente que la no anulación de $\oint \boldsymbol{A}$, dentro de una región libre de campo, y por lo tanto el propio efecto Aharonov-Bohm, depende de que esta región libre de campo sea topológicamente no trivial.

Debido a sus orígenes históricos en la notable idea de Weyl (donde originalmente sí desempeñó un papel como una «calibración» dependiente del camino), llamamos a esta conexión electromagnética ∇ una conexión *gauge*; este nombre se adopta también para las generalizaciones del electromagnetismo, conocidas como teorías de «Yang-

[19.11] Explíquelo.

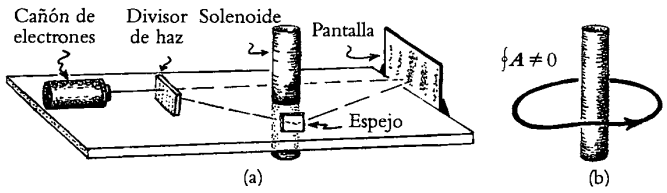


Fig. 19.6. Efecto Aharonov-Bohm. (a) Un haz de electrones se divide en dos caminos que pasan por uno u otro lado de una serie de líneas de flujo magnético (que se consiguen por medio de un largo solenoide). Los haces se reúnen en una pantalla, y la figura de interferencia cuántica resultante (compárese con la Fig. 21.4) depende de la intensidad de flujo magnético, pese al hecho de que los electrones solo encuentran una intensidad de campo nula ($\boldsymbol{F} = 0$). (b) El efecto depende del valor de $\oint \boldsymbol{A}$, que puede ser no nulo sobre el camino cerrado relevante topológicamente no trivial, pese a que $\boldsymbol{F}$ se anula sobre este camino. La cantidad $\oint \boldsymbol{A}$ queda inalterada por deformaciones continuas del camino dentro de la región libre de campo.

Mills», que se utilizan en la descripción de las interacciones débil y fuerte en la moderna física de partículas. Notemos que la idea de «conexión gauge» sí depende, estrictamente hablando, de la existencia de una simetría (que en el caso del electromagnetismo es la simetría $\Psi \mapsto e^{i\theta}\Psi$) que se supone exacta y no directamente observable. Recordemos que la objeción de Einstein a la idea gauge original de Weyl era, de hecho, que la masa de una partícula, y por consiguiente su frecuencia natural, es directamente medible, y por ello no puede utilizarse como un «campo gauge» en el sentido requerido. Más adelante veremos que esta cuestión resulta muy confusa en algunos usos modernos de la idea «gauge».

19.5. El tensor energía-momento

Como prerrequisito para dirigir nuestra atención a ese otro campo clásico fundamental con sus aspectos de «teoría gauge», a saber, el campo gravitatorio, será importante considerar primero la cuestión de la *densidad de energía* de un campo, pues esta densidad es la fuente de la gravedad. En efecto, la famosa ecuación de Einstein $E = mc^2$ nos dice que masa y energía son lo mismo (véase §18.6) y, como Newton ya nos ha-

bía informado, la masa es la fuente de la gravitación. Por ello necesitamos entender cómo se puede describir la densidad de energía de un campo, tal como el de Maxwell, y cómo esto puede actuar como fuente de gravedad. Lo que Einstein nos dice es que lo hace a través de una magnitud tensorial conocida como el *tensor energía-momento*. Este es un tensor simétrico $\left[{}^{0}_{2}\right]$-valente $\boldsymbol{T}$ (en forma de índices $T_{ab} = T_{ba}$) que satisface una «ecuación de conservación»

$$\nabla^a T_{ab} = 0.$$

(Para el resto del capítulo utilizamos el operador derivada covariante espaciotemporal ∇_a en lugar de $\partial/\partial x^a$. Puesto que todos nuestros campos carecen de carga, nuestras expresiones anteriores no se verán afectadas; véanse también el párrafo final de §18.2, la nota 19.2 y el ejercicio [19.6].) Podemos comparar esta expresión con la ecuación de conservación $\nabla^a J_a = 0$ para la carga eléctrica. La razón para el índice extra en T_{ab} es que lo que se conserva, a saber, la energía-momento, es una magnitud 4-(co)vectorial (el 4-(co)vector p_a energía-momento, considerado en §18.7) más que la carga eléctrica *escalar*. Para describir el contenido físico de T_{ab} con más detalle, es conveniente pasar a la magnitud equivalente $T^a{}_b = g^{ac}T_{cb}$, en la que se ha subido un índice mediante el uso del tensor métrico g^{ab}.[19.12] La cantidad $T^a{}_b$ reúne todas las diferentes densidades y flujos de energía y momento en los campos y partículas. Más concretamente, en un sistema de coordenadas de Minkowski estándar, el covector $T^0{}_b$ es la densidad de 4-momento y las tres cantidades $T^1{}_b$, $T^2{}_b$, $T^3{}_b$, proporcionan el flujo de 4-momento en las tres direcciones espaciales independientes. Esto es directamente análogo al caso de J^a, puesto que J^0 es la densidad de carga y las tres cantidades J^1, J^2, J^3 proporcionan el flujo de carga (i.e., la corriente) en las tres direcciones espaciales independientes. Es el índice extra b el que nos dice que nuestra ley de conservación se refiere ahora a una magnitud (co)vectorial. Resulta que la cantidad T_{00} mide la densidad de energía, y T_{11}, T_{22}, T_{33} miden la *presión* en la tres direcciones de los ejes de coordenadas espaciales.

[19.12] ¿Cómo se relacionan las componentes individuales $T^a{}_b$ con T_{ab}, en un sistema de Minkowski local, en el que las componentes g_{ab} tienen la forma diagonal $(1, -1, -1, -1)$?

Recordemos que, como Maxwell nos enseñó, los propios campos electromagnéticos transportan energía. En la notación de índices, el tensor energía-momento del campo electromagnético resulta ser[19.13]

$$\frac{1}{8\pi}(F_{ac}F^{c}{}_{b} + {}^{*}F_{ac}{}^{*}F^{c}{}_{b}).$$

Otros campos físicos tienen también sus tensores energía-momento, y varias de estas contribuciones diferentes tendrían que sumarse para dar el tensor energía-momento T completo, que satisface la ecuación de conservación $\nabla^{a}T_{ab} = 0$.

Sin embargo, algo muy diferente sucede con la energía-momento de la propia gravedad, como veremos enseguida. En ausencia de gravedad, el espaciotiempo es plano (i.e., espacio de Minkowski), y podemos utilizar coordenadas (minkowskianas) planas. Entonces cada uno de los cuatro vectores $T^{a}{}_{0}$, $T^{a}{}_{1}$, $T^{a}{}_{2}$ y $T^{a}{}_{3}$ satisface por separado exactamente la misma ecuación de conservación que J^{a} (a saber, $\nabla^{a}T^{a}{}_{0} = 0$, etc., análoga a $\nabla_{a}J^{a} = 0$), lo que tiene como consecuencia que existe una ley *integral* de conservación exactamente análoga a la de carga (i.e., análoga a $\int_{\mathcal{Q}}{}^{*}\boldsymbol{J} = 0$), para cada una de las 4 componentes de la energía-momento por separado. Así pues, la masa total se conserva, y también lo hacen las tres componentes del momento total. Pero recordemos la discusión dada en el capítulo 17 sobre el principio de equivalencia de Einstein, y de por qué esto nos lleva a un espaciotiempo curvo. Así pues, en presencia de gravedad debemos tener en cuenta el hecho de que «∇_{a}» ya no es simplemente «$\partial/\partial x^{a}$», sino que (de acuerdo con §14.3) existen términos extra $\Gamma^{b}{}_{ac}$ que confunden el propio significado de «$\nabla_{a}T^{a}{}_{0}$» y que ciertamente nos impiden obtener una ley integral de conservación para la energía y el momento a partir de nuestra «ecuación de conservación» $\nabla^{a}T_{ab} = 0$. El problema puede parafrasearse como el hecho de que el índice extra b en T_{ab} le impide ser el dual de una 3-forma, y no podemos escribir una formulación independiente de las coordenadas de una «ecuación de conservación» (como anulación de la derivada exterior de la 3-forma ${}^{*}\boldsymbol{J}$ en «d ${}^{*}\boldsymbol{J} = 0$»). ¡Parece que hemos perdido las leyes de conservación más fundamentales de la física, las leyes de conservación de la energía y el momento!

[19.13] Demuestre que esto satisface la ecuación de conservación $\nabla^{a}T_{ab} = 0$ si $\boldsymbol{J} = 0$. Obtenga la 00-componente de este tensor, y recupere la expresión original de Maxwell $(E^{2} + B^{2})/8\pi$ para la energía de un campo electromagnético en términos de (E_{1}, E_{2}, E_{3}) y (B_{1}, B_{2}, B_{3}).

De hecho, hay una perspectiva más satisfactoria sobre la conservación de energía-momento, que se refiere tanto a ciertos espaciotiempos curvos $\mathcal{M}$ como al espacio de Minkowski, y se aplica asimismo a la conservación del momento angular (véanse §18.5 y §§22.8,11). En esta perspectiva suponemos que tenemos un *vector de Killing* $\boldsymbol{\kappa}$ para $\mathcal{M}$ (que satisface $\nabla_{(a}\kappa_{b)} = 0$; véase §14.8), que describe una simetría continua de $\mathcal{M}$. En el espacio de Minkowski existen 10 de tales simetrías independientes, que se refieren a las 4 simetrías de *traslación* independientes (3 espaciales y 1 temporal) y 6 *rotaciones* espaciotemporales independientes (la parte no reflexiva del grupo de Lorentz O(1, 3)). Véase la Fig. 18.3b. Así pues, el espacio de Minkowski tiene 10 vectores de Killing independientes. Como veremos en el capítulo siguiente, el formalismo lagrangiano (teorema de Nöther) nos permite obtener una ley de conservación a partir de cada simetría continua que posean las leyes del sistema. La simetría de traslación temporal proporciona la conservación de la energía, mientras que la simetría de traslación espacial proporciona la conservación del 3-momento. La simetría rotacional da la conservación del momento angular. (Las rotaciones espaciales ordinarias nos dan las 3 componentes del momento angular ordinario, pero existen también 3 componentes que proceden de los «impulsos» lorentzianos, que nos llevan de una velocidad a otra. Estas nos dan la conservación del movimiento del centro de masas; véanse §§18.6,7 y la Fig. 18.16.) Para obtener la ley de conservación adecuada a partir de cualquier vector de Killing $\boldsymbol{\kappa}$, construimos la magnitud *flujo*

$$L_{a} = T_{ab}\kappa^{b},$$

que satisface la ley de conservación $\nabla_{a}L^{a} = 0$, siempre que el T_{ab} simétrico satisfaga $\nabla^{a}T_{ab} = 0$.[19.14] Por lo tanto, como en §19.3, hay una ley de conservación *integral* $\int_{\mathcal{Q}}{}^{*}\boldsymbol{L} = 0$.

[19.14] ¿Por qué? ¿Por qué este procedimiento se particulariza en la anterior $\nabla_{a}T^{a}{}_{0} = 0$, etc.? ¿Puede encontrar un análogo de la ley de conservación en campos continuos $\nabla^{a}(T_{ab}\kappa^{b}) = 0$, para un sistema discreto de partículas donde el 4-momento se conserva en las colisiones? *Sugerencia*: Encuentre una cantidad, dado el vector de Killing κ^{a}, que sea constante para cada partícula entre colisiones.

Estas leyes de conservación son válidas solamente en un espaciotiempo para el que existe la simetría adecuada, dada por el vector de Killing κ. Físicamente, la razón para ello es que los grados de libertad en la geometría espaciotemporal —i.e., la gravedad— están desacoplados de los campos. La geometría espaciotemporal sirve únicamente como fondo, de modo que no es perturbada por los campos que hay en su interior; además, los campos son incapaces de tomar la cantidad en cuestión del fondo (o perderla en el fondo) debido a la simetría. Estas consideraciones tendrán importancia más adelante, en especial en el capítulo 30 (§§30.6,7). De todas formas, no nos ayudan realmente a entender cuál será el destino de las leyes de conservación cuando la propia gravedad se convierta en un jugador activo. Todavía no hemos recuperado las leyes de conservación que nos faltan para la energía y el momento, cuando la gravedad entra en la imagen.

Desde los primeros días de la relatividad general, este hecho aparentemente complicado ha supuesto una de las objeciones más fuertes a dicha teoría, y ha dado razones para sentirse incómodos con ella, como han expresado numerosos físicos durante años.[6] Veremos más adelante, en §19.8, que la teoría de Einstein tiene en cuenta la conservación de la energía-momento de una forma muy sofisticada, al menos en aquellas circunstancias en las que más necesaria es una ley de conservación semejante. Por el momento, tomamos nota del hecho de que, en la teoría de Einstein, el tensor simétrico $\binom{0}{2}$-valente $\boldsymbol{T}$ que aparece en su ecuación de campo tiene que incluir la energía-momento de todos los campos (y partículas) *no* gravitatorios. Cualquier energía que haya en el propio campo gravitatorio debe ser excluida de tener una representación dentro de $\boldsymbol{T}$.

Este punto de vista se hace algo plausible si pensamos de nuevo en el principio de equivalencia. Imaginemos un observador en órbita libre, por ejemplo dentro de alguna nave espacial sin ventanas, de modo que parece, al menos en la primera aproximación, que no hay campo gravitatorio. El observador esperaría que la energía se conservara dentro de la nave espacial, y por consiguiente esperaría que la ecuación $\nabla^a T_{ab} = 0$ se mantuviera válida sin que haya ninguna contribución del campo gravitatorio. Sin embargo, esta «conservación» es solo una aproximación, y se espera que necesite corrección en cuanto empiecen a

desempeñar un papel los efectos (de marea) de la aceleración relativa debidos a la no uniformidad del campo gravitatorio (como se ha estudiado en §17.5; véanse las Figs. 17.8a, 17.8b y 17.9). Esta es una cuestión algo delicada, y resulta necesario examinar los «órdenes» en los que diferentes tipos de efectos empiezan a desempeñar un papel. El resultado final de todo esto es que la cantidad $\boldsymbol{T}$ y su ecuación $\nabla^a T_{ab} = 0$ deberían mantenerse inalteradas por la no uniformidad del campo gravitatorio —i.e., no ser afectadas por la curvatura $\boldsymbol{R}$ de la conexión espaciotemporal ∇— y que, de algún modo, las contribuciones de la gravedad a la conservación de la energía-momento deberían entrar *no localmente* como correcciones en el cálculo de la energía-momento total. (La única excepción real a este comentario podría ocurrir si hubiera que contemplar correcciones de la curvatura espaciotemporal para aquellas expresiones matemáticas que nos dicen cómo contribuyen los campos físicos a $\boldsymbol{T}$. Normalmente no hay tales correcciones, y no es una cuestión importante para nuestras consideraciones actuales.) Desde esta perspectiva, las contribuciones gravitatorias a la energía-momento «se filtran por las grietas» que separan la ecuación *local* $\nabla^a T_{ab} = 0$ de una ley de conservación integral para la energía-momento *total*.

19.6. La ecuación de campo de Einstein

Volveré a esta cuestión en §19.8, pero de momento necesitamos saber la forma real de la ecuación de campo de Einstein. Esta ecuación se expresa en términos del formalismo *tensorial* con el que ahora (¡así lo espero!) el lector no se sentirá demasiado incómodo. Parte de la razón de que se necesiten los tensores es que la curvatura espaciotemporal, en 4 dimensiones, es un asunto complicado. Recordemos a Albert, nuestro astronauta A de §17.5, orbitando libremente en el campo gravitatorio de la Tierra. En algunas direcciones que parten de A existen aceleraciones hacia dentro y en otras direcciones existen aceleraciones hacia fuera. Estas representan las fuerzas *de marea* experimentadas por A. Las fuerzas de marea son manifestaciones de la curvatura espaciotemporal. Para reunir estos efectos complicados, se utiliza una magnitud tensorial R_{abcd}, que tiene 10 componentes independientes en el espacio vacío, y

un total de 20 cuando hay también densidad de materia. De hecho, R_{abcd} es simplemente la forma indexada del tensor de Riemann-(Christoffel) $\boldsymbol{R}$ que ya hemos visto anteriormente en §14.7.

Pero existe otra razón, aparte de la de poner orden en la complicación, para que el tensor de Ricci desempeñe un papel fundamental en la teoría de Einstein. Esta se remonta al *principio de equivalencia* fundacional que puso en marcha toda la línea de pensamiento de Einstein. La gravitación no debe considerarse como una fuerza; para un observador que esté en caída libre (como nuestro astronauta A), no hay ninguna fuerza gravitatoria directa que sentir. En su lugar, la gravitación se manifiesta en forma de curvatura espaciotemporal. Ahora bien, para que funcione esta idea es importante que no haya «coordenadas privilegiadas» en la teoría.[7] En efecto, si cierta clase limitada de sistemas de coordenadas se considerasen elecciones privilegiadas de la naturaleza, entonces estas definirían «sistemas observadores naturales» con respecto a los cuales podría reintroducirse la noción de «fuerza gravitatoria», y se perdería el papel central del principio de equivalencia. El punto es bastante delicado, y ocasionalmente muchos físicos se han apartado de él. A mi modo de ver, es esencial para el espíritu de la teoría de Einstein que se mantenga esta noción de independencia de coordenadas. Esto es lo que se conoce como *principio de covariancia general*. Nos dice que no solo no debe haber coordenadas privilegiadas, sino que si tenemos dos espaciotiempos diferentes que representan dos campos gravitatorios físicamente distintos, entonces no tiene que haber una identificación punto a punto preferida entre ambos, de modo que no podemos decir qué punto espaciotemporal concreto de uno debe considerarse el *mismo* punto que cierto punto espaciotemporal del otro. Esta cuestión filosófica nos interesará más adelante (§30.11), en relación con la cuestión de cómo se relaciona la teoría de Einstein con los principios de la mecánica cuántica. Por el momento, la importancia del principio de covariancia general para nosotros es que nos obliga a una descripción libre de coordenadas de la física gravitatoria. Por esta razón en particular, el formalismo tensorial tiene una importancia tan capital en la teoría de Einstein.

Veamos ahora cuál es realmente la ecuación de Einstein. La forma de dicha ecuación está impulsada básicamente por los dos requisitos (i)

que la fuente (local) de la gravedad debería ser, en efecto, el tensor energía-momento $\boldsymbol{T}$, sujeto a $\nabla^a T_{ab} = 0$, y (ii) que, en el límite newtoniano adecuado (pequeñas velocidades, comparadas con la de la luz, y pequeños campos gravitatorios), debería recuperarse la teoría gravitatoria de Newton estándar. Debemos volver a la exposición de §17.5, en la que hemos visto que en la teoría newtoniana hay un efecto de reducción de volumen para geodésicas que están próximas, y son inicialmente paralelas, a una línea de universo geodésica γ de un observador. Estas geodésicas vecinas se aceleran con relación a γ de tal forma que el *3-volumen* (infinitesimal) δV de género espacio que encierran tiene una aceleración global que es igual a $-4\pi G\,\delta M$, donde δM es la masa gravitatoria activa del volumen (infinitesimal) encerrado por las geodésicas. El signo menos procede del hecho de que es una *reducción* de volumen lo que está implicado; véase la Fig. 17.8b. Esta es una expresión plena de la teoría de Newton, con respecto al efecto gravitatorio activo de una distribución de masa.

¿Cómo debemos traducir esto en una ecuación que relaciona la curvatura espaciotemporal $\boldsymbol{R}$ con el tensor energía-momento $\boldsymbol{T}$? El hecho geométrico clave es que la aceleración hacia dentro de volumen que se da en esta situación viene medida por un tensor simétrico $\binom{0}{2}$-valente, denominado el *tensor de Ricci*, definido por

$$R_{ab} = R_{acb}{}^{c},$$

donde R_{abcd} es la forma de índices del tensor de Riemann.[19.15] (Véase la Fig. 19.7 para la notación diagramática de esto.) Una vez más, existen innumerables convenios diferentes con respecto a signos, ordenación de índices, signaturas, etc., y, como antes, estoy imponiendo al lector mis propias preferencias; véase §14.4. Más concretamente, la aceleración de volumen (partiendo del reposo) viene dada por[19.16]

$$\mathbf{D}^2(\delta V) = R_{ab}t^a t^b \delta V.$$

[19.15] ¿Por qué es R_{ab} simétrico?

[19.16] Vea si puede demostrarlo utilizando la identidad de Ricci y las propiedades de la derivada de Lie.

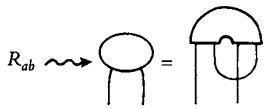


Fig. 19.7. Notación diagramática para la definición del tensor de Ricci $R_{ab} = R_{acb}{}^{c}$ (véase la Fig. 14.21).

Aquí, **D** representa el ritmo de cambio con respecto al *tiempo propio* del observador (véase §17.9), a lo largo de la línea de universo γ del observador, de modo que $\mathbf{D}^2$ denota en realidad la aceleración. Tenemos

$$\mathbf{D} = t^a \nabla_a = \nabla_{\!t},$$

donde t^a es el vector unidad de género tiempo futuro tangente a γ (de modo que $t^a t_a = 1$).

La *densidad de masa* (que es la misma que la densidad de *energía*, pues «$E = mc^2$» con $c = 1$; véase §18.6), medida por el observador, es la «00-componente» de T_{ab} en el sistema local del observador. Esta es simplemente la cantidad $T_{ab}t^a t^b$, de modo que la masa δM dentro del volumen δV encerrado por las geodésicas vecinas es

$$\delta M = T_{ab}t^a t^b \delta V.$$

Así pues, la «expectativa newtoniana» $-4\pi G\, \delta M$ (§17.5) para la aceleración de volumen debida a la densidad de materia es

$$-4\pi G T_{ab}t^a t^b \delta V.$$

Pero acabamos de ver que el efecto de aceleración de volumen debido a la curvatura espaciotemporal es $R_{ab}t^a t^b \delta V$, de modo que llegamos a la expectativa

$$R_{ab}t^a t^b \delta V = -4\pi G T_{ab}t^a t^b \delta V.$$

Dividiendo por δV y advirtiendo que esto se aplica a *todos* los observadores que pasan por el mismo suceso, de modo que podemos eliminar $t^a t^b$,[19.17] llegamos a la sugerida ecuación de campo

$$R_{ab} = -4\pi G T_{ab},$$

[19.17] Demuestre por completo por qué podemos «podar» todas las t, explicando el papel de la simetría de los tensores.

que, de hecho, era la propuesta inicial de Einstein. Sin embargo, esto no es satisfactorio porque la «ecuación de conservación» $\nabla^a T_{ab} = 0$ conduce entonces a $\nabla^a R_{ab} = 0$, ¡que, a su vez, lleva a dificultades!

¿Cuáles son estas dificultades? Recordemos, de §14.4, la ecuación *identidad de Bianchi* $\nabla_{[a}R_{bc]d}{}^{e} = 0$. Tomando una contracción de esta ecuación, obtenemos[19.18]

$$\nabla^a(R_{ab} - \tfrac{1}{2} R\, g_{ab}) = 0,$$

donde el *escalar de Ricci* (o *curvatura escalar*, aunque «$-R$» podría encajar mejor con la mayoría de los convenios matemáticos para el caso definido positivo) se define por

$$R = R_a{}^a$$

(donde «R» no debe confundirse con la «$\boldsymbol{R}$» que representa el tensor de curvatura *completo*). La «dificultad» con la ecuación propuesta más arriba, $R_{ab} = -4\pi G T_{ab}$, es que, cuando se combina con la identidad de Bianchi contraída, lleva a la conclusión de que la *traza* T del tensor energía-momento, definida por

$$T = T^a_a,$$

tiene que ser *constante* en todo el espaciotiempo.[19.19] Esto es absolutamente incompatible con la física (no gravitatoria) ordinaria. En consecuencia, Einstein concluyó finalmente, en 1915, que para que hubiera compatibilidad deberían *igualarse* (salvo un factor constante) los dos tensores que satisfacen la «ecuación de conservación» $\nabla^a(\ldots) = 0$, y llegó a lo que ahora se conoce como *ecuación de campo de Einstein*:[8,19.20]

$$R_{ab} - \tfrac{1}{2} R g_{ab} = -8\pi\, G T_{ab}.$$

En la situación concreta en que no hay materia presente (incluido el campo electromagnético), tenemos $T_{ab} = 0$. Esto se conoce como *vacío*. La ecuación de Einstein —la *ecuación de vacío*— se convierte en $R_{ab} - \frac{1}{2}Rg_{ab} = 0$, que puede reescribirse como[19.21]

[19.18] Demuéstrelo utilizando la notación diagramática, si lo desea.
[19.19] ¿Por qué?
[19.20] Explique el coeficiente $-8\pi G$, comparado con $-4\pi G$.
[19.21] ¿Por qué?

$$R_{ab} = 0.$$

Un espacio con tensor de Ricci nulo se conoce a veces como *Ricci-plano*.

19.7. Cuestiones adicionales: la constante cosmológica, el tensor de Weyl

En este punto deberíamos considerar el término adicional que sugirió Einstein en 1917, denominado *constante cosmológica*. Esta es una cantidad constante extraordinariamente minúscula Λ, cuya presencia real está sugerida por modernas observaciones cosmológicas, pero que no puede diferir de cero en más de la muy minúscula cantidad de aproximadamente 10^{-55} cm^{-2}. No tiene relevancia observacional directa hasta que se alcanzan escalas cosmológicas. La cantidad $R_{ab} - \frac{1}{2}Rg_{ab}$, en la expresión anterior, se reemplaza en consecuencia por $R_{ab} - \frac{1}{2}Rg_{ab} + \Lambda g_{ab}$. Esta sigue satisfaciendo la «ecuación de conservación», puesto que Λ es constante (y $\nabla \mathbf{g} = 0$). La ecuación de Einstein se lee ahora

$$R_{ab} - \tfrac{1}{2}Rg_{ab} - \Lambda g_{ab} = -8\pi G\, T_{ab}.$$

En un principio, Einstein introdujo este término extra para tener la posibilidad de un universo cerrado espacialmente y *estático* a escala cosmológica.[9] Pero cuando a partir de las observaciones de Edwin Hubble, en 1929, se supo que el universo se está expandiendo, y por consiguiente *no es* estático, Einstein retiró su apoyo a la constante cosmológica, asegurando que había sido «su mayor error» (¡quizá porque de no haberla introducido podría haber predicho la expansión del universo!). De todas formas, una vez propuestas las ideas, estas no desaparecen de manera tan sencilla. La constante cosmológica ha quedado latente en el fondo de la teoría cosmológica desde que Einstein la propuso por primera vez, causando preocupación a algunos y solaz a otros. Muy recientemente, observaciones de supernovas lejanas han llevado a algunos teóricos a reintroducir Λ, o algo similar conocido como «energía oscura», como una forma de hacer que estas observaciones sean compatibles con otros requisitos percibidos.[10] Volveré más adelante a la cuestión de la constante cosmológica (véase en especial §28.10). Por mi parte, en común con la mayoría de los teóricos de la relatividad, aun admitiendo la posibilidad de una Λ no nula en las ecuaciones, yo mismo hubiera sido bastante reacio a aceptar que la naturaleza hubiese utilizado este término. Sin embargo, como veremos en §28.10, evidencias cosmológicas muy recientes parecen apuntar en esta dirección.

También podemos escribir la ecuación de campo de Einstein (incluyendo la constante cosmológica) en sentido inverso:[19.22] $R_{ab} = -8\pi G(T_{ab} - \frac{1}{2}Tg_{ab}) + \Lambda g_{ab}$. Utilizando un sistema de coordenadas local con el eje temporal dado por t^a, de modo que al contraerlo con $t^a t^b$ da la 00-componente, observamos que la aceleración hacia dentro del volumen está dada por $8\pi G(T_{00} - \frac{1}{2}Tg_{00}) - \Lambda$, que es $4\pi G(\rho + P_1 + P_2 + P_3) - \Lambda$, donde P_1, P_2 y P_3 son los valores de la *presión* de la materia a lo largo de tres ejes espaciales (ortogonales). Por comparación con la «$4\pi G\,\delta M$» que da la teoría de Newton, encontramos que la densidad $\rho_{\mathcal{G}}$ de *masa gravitatoria activa*, en la relatividad general de Einstein es

$$\rho_{\mathcal{G}} = \rho + P_1 + P_2 + P_3 - \frac{\Lambda}{4\pi G},$$

antes que $\rho_{\mathcal{G}} = \rho$, siendo esto último lo que podríamos haber esperado simplemente a partir de «$E = mc^2$». (Se han elegido unidades de modo que $c = 1$.) La contribución de Λ es extraordinariamente minúscula y, de hecho, los términos de presión extra son también normalmente muy pequeños en comparación con la energía, porque, en términos generales, las pequeñas partículas que constituyen el material en cuestión se están moviendo con relativa lentitud comparadas con la velocidad de la luz. Sin embargo, las contribuciones de la presión a la masa gravitatoria activa desempeñan papeles importantes en ciertas condiciones extremas. Cuando una estrella muy masiva se acerca a una situación en la que está en peligro de colapsar bajo su propio tirón gravitacional hacia dentro, encontramos que una presión aumentada en la estrella, que

[19.22] ¿Por qué?

cabría esperar que ayudase a mantener la estrella, ¡*incrementa* en realidad la tendencia a colapsar debido a la masa gravitatoria activa extra que produce!

Tal como se ha señalado antes (§19.5), el tensor energía-momento T_{ab} es análogo, en la teoría de Einstein, al vector carga-corriente $\boldsymbol{J}_a$ de la teoría de Maxwell. Puede considerarse que la magnitud T_{ab} describe la *fuente* de la gravitación, de la misma forma que $\boldsymbol{J}_a$ es la fuente del electromagnetismo. ¿Podemos preguntar cuál sería el análogo aproximado del tensor de campo de Maxwell F_{ab} que describe los *grados de libertad gravitatorios*? La respuesta *no* es el tensor métrico $\mathbf{g}$, que es más parecido al potencial electromagnético $\mathbf{A}$. Algunos podrían considerar que el tensor de curvatura de Riemann completo R_{abcd} es el análogo de $\boldsymbol{F}$, pero resulta más adecuado escoger lo que se denomina el *tensor de Weyl* (o *tensor conforme*) C_{abcd}, que es parecido al tensor de Riemann completo pero del que se ha «eliminado» la parte del tensor de Ricci. Esto es razonable, porque el tensor de Ricci puede identificarse estrechamente con la fuente T_{ab}, de modo que tenemos que eliminar estos «grados de libertad fuente» si queremos identificar aquellos grados de libertad que describen directamente el campo gravitatorio. En el espacio libre, donde no hay materia (y donde tomamos nula la constante gravitatoria Λ), el tensor de Weyl es *igual* al tensor de curvatura de Riemann, si bien en general el tensor de Weyl se define por una fórmula de apariencia complicada que elimina la parte del tensor de Ricci de la curvatura total (donde he subido dos índices para hacer un uso completo de la notación de paréntesis cuadrados de §11.6):[19.23]

$$C_{ab}{}^{cd} = R_{ab}{}^{cd} - 2R_{[a}{}^{[c}g_{b]}{}^{d]} + \frac{1}{3}Rg_{[a}{}^{c}g_{b]}{}^{d}.$$

19.8. La energía del campo gravitatorio

Volvamos a la cuestión de la masa/energía en el propio campo gravitatorio. Aunque no hay lugar para algo semejante en el tensor energía-momento $\boldsymbol{T}$, es evidente que hay situaciones en que una energía

[19.23] Demuestre que todas las «trazas» de $\boldsymbol{C}$ se anulan (por ejemplo, $C_{abc}{}^{a} = 0$, etc.). Haga este cálculo de forma diagramática, si lo desea.

Fig. 19.8. No localidad de la energía potencial gravitatoria. Imaginemos dos planetas (que, por simplicidad, podemos suponer que están instantáneamente en reposo relativo). Si (a) están muy alejados, entonces la contribución negativa de la energía potencial (newtoniana) no es tan grande como (b) cuando están juntos. Así, la energía total (y por ello la masa total del sistema entero) es mayor en el caso (a) que en el caso (b), pese a que las densidades totales de energía, medidas por los tensores energía-momento, son prácticamente las mismas en los dos casos.

gravitatoria «incorpórea» está desempeñando realmente un papel físico. Imaginemos dos cuerpos masivos (por ejemplo, dos planetas). Si están próximos (y si podemos suponer que están instantáneamente en reposo uno con respecto al otro), entonces habrá una contribución de energía potencial gravitatoria (negativa) que hace que la energía total, y, por consiguiente, la masa total, sea menor que la que sería si estuviesen alejados (véase la Fig. 19.8). Ignorando efectos de energía muchísimo más pequeños, tales como distorsiones de la forma de cada cuerpo debidas al campo de marea gravitatorio del otro, vemos que las contribuciones totales al tensor de energía-momento real $\boldsymbol{T}$ serán las mismas, ya estén los dos cuerpos cerca o lejos. Pese a todo, la masa/energía total diferirá en los dos casos, y esta diferencia sería atribuida a la energía en el propio campo gravitatorio (de hecho, una contribución *negativa*, que es mayor cuando los cuerpos están próximos que cuando están alejados).

Consideremos ahora que los cuerpos están en movimiento, cada uno en órbita en torno al otro. Una consecuencia de la ecuación de campo de Einstein es que del sistema emanarán *ondas gravitatorias* —rizos en el propio tejido del espaciotiempo— que se llevarán energía (positiva). En circunstancias normales, esta pérdida de energía será muy pequeña. Por ejemplo, el mayor de estos efectos en nuestro propio sistema solar aparece en el sistema Júpiter-Sol, y la tasa de pérdida de ener-

Fig. 19.9. El sistema doble de estrellas de neutrones de Hulse-Taylor PSR 1913 + 16. Un miembro es un púlsar que emite señales electromagnéticas exactamente cronometradas que se reciben en la Tierra, lo que permite determinar las órbitas con extraordinaria precisión. Se observa que el sistema pierde energía en acuerdo exacto con la predicción de Einstein de ondas gravitatorias que transportan energía emitidas por dicho sistema. Estas ondas son rizos del vacío espaciotemporal, donde el tensor energía-momento se anula. (No está a escala.)

gía es tan solo aproximadamente igual a la que emite ¡una bombilla de 40 vatios! Pero en el caso de sistemas más masivos y violentos, tales como la coalescencia final de dos agujeros negros que han estado describiendo trayectorias espirales uno alrededor de otro, se espera que la pérdida de energía sería tan grande que detectores que se están construyendo en la actualidad aquí en la Tierra podrían registrar la presencia de tales ondas gravitatorias a una distancia de 15 megaparsecs, o aproximadamente $4{,}6 \times 10^{23}$ metros.

Entre estos dos extremos están las ondas gravitatorias emitidas por el extraordinario sistema doble de estrellas de neutrones conocido como PSR 1913 + 16, estudiado por el equipo formado por Joseph Taylor y Russell Hulse, que ganaron por ello un premio Nobel; véase la Fig. 19.9. (Una estrella de neutrones es una estrella extraordinariamente compacta, compuesta principalmente de neutrones, tan estrechamente apretados que la densidad global de la estrella es comparable a la de un núcleo atómico. ¡Una pelota de tenis llena de este material tendría una masa total comparable a la de Deimos, una de las lunas de Marte!) Este sistema ha sido observado durante unos veinticinco años, y su movimiento detallado ha sido registrado con gran precisión (lo que es posible debido a que una de las estrellas es un *púlsar* que emite «bips» electromagnéticos con una cadencia muy precisa unas 17 veces por segundo). La cadencia de estas señales es tan precisa, y el propio sis-

tema tan «limpio», que la comparación entre observación y predicción teórica proporciona una confirmación de la relatividad general de Einstein hasta aproximadamente una parte en 10^{14}, una precisión sin precedentes en la comparación científica entre teoría y observación. Esta cifra se refiere a la precisión del cronometraje global durante un período de unos 20 años.[11]

Con observaciones de esta naturaleza —y también con los impresionantes efectos de lente gravitatoria que serán abordados en §28.8—, la relatividad general observacional ha avanzado un largo camino desde los primeros días de la teoría. Sin embargo, durante el período 1915-1969 (donde 1969 marca el año en que las radioobservaciones de cuásares distantes iniciaron una nueva familia de tests de la relatividad general[12]) solo existían los famosos pero relativamente poco impresionantes «tres tests» para dar apoyo a esta teoría. El más significativo de estos era la explicación que dio Einstein del «avance anómalo del perihelio» de Mercurio. Esto constituía una desviación muy ligera de las predicciones de la teoría de la gravitación newtoniana (de solo 43 segundos de arco por siglo, o aproximadamente ¡una rotación orbital cada 3 millones de años!) que había sido observada durante más de medio siglo[13] (véanse también §30.1 y §34.9). Un segundo efecto observacional era la minúscula curvatura que la masa del Sol producía en la luz procedente de estrellas distantes que pasaba cerca de él, vista por la expedición de Arthur Eddington a la isla del Príncipe (en la costa del África occidental), durante el eclipse solar de 1919. Este es un efecto del mismo fenómeno de «lente gravitatoria» que se ha mencionado antes, y que ahora se utiliza de forma impresionante a distancias casi cosmológicas para obtener información importante sobre la distribución de masa en el universo (véase §28.8). Por último, está el efecto de frenado del ritmo de los relojes en un potencial gravitatorio, predicho por la teoría de Einstein. Este fue provisionalmente (y discutiblemente) observado por W.S. Adams en 1925, en el caso de una estrella enana blanca como la compañera de Sirio (una estrella algunos miles de veces más densa que el Sol). Pero más tarde se obtuvo una confirmación mucho más convincente, en un delicado experimento realizado por Pound y Rebka en 1960, para el propio campo gravitatorio de la Tierra. (Sin embargo, este efecto debe esperarse simplemente a partir de conside-

raciones cuánticas básicas y de energía en general, y es un test más bien débil de la teoría de Einstein.) Hay también un tipo diferente de efecto de «retardo temporal», para señales luminosas que llegan a la Tierra procedentes de objetos que están casi directamente detrás del Sol, que fue propuesto por primera vez (en 1964) por Erwin Shapiro, y más tarde confirmado por él en 1968-1971, en observaciones de Mercurio y Venus, y más exactamente (hasta un 0,1 por ciento, en 1971) por Reasenberg y Shapiro, utilizando transpondedores en la nave espacial Viking en órbita alrededor de Marte que se comparaban con uno colocado en el suelo de Marte.

Está claro que la teoría de Einstein está hoy día muy bien apoyada observacionalmente. La existencia de ondas gravitatorias parece estar claramente confirmada por las observaciones de Hulse-Taylor, incluso si esto no es una detección directa de tales ondas. En la actualidad hay varios proyectos para la detección directa de ondas gravitatorias que constituyen un esfuerzo mundial concertado para utilizar tales ondas con el fin de sondear actividad violenta (como colisiones de agujeros negros) en partes distantes del universo. En efecto, estos proyectos combinados[14] proporcionarán un telescopio de ondas gravitatorias, de modo que la teoría de Einstein tiene el potencial de darnos aún otra poderosa forma de explorar el universo distante.

Vemos que, a pesar de las preocupaciones de algunos sobre la conservación de la energía, la relatividad general tiene varias confirmaciones observacionales muy notables. Volvamos, por lo tanto, a la cuestión de la energía gravitatoria. Un punto esencial para la consistencia, tanto en la teoría como en la observación, es que los rizos del espacio vacío que constituyen las ondas gravitatorias emitidas por PSR 1913 + 16 y otros sistemas semejantes se lleven energía real del sistema. El tensor energía-momento en el espacio vacío es cero, de modo que la energía de la onda gravitatoria tiene que medirse de alguna otra manera que no sea localmente atribuible a una «densidad» de energía. La energía gravitatoria es una magnitud genuinamente no local. Sin embargo, esto no implica que no exista ninguna descripción matemática de la energía gravitatoria. Aunque creo que es justo decir que aún no tenemos una completa comprensión de la masa/energía gravitatoria, hay una serie significativa de situaciones en las que puede darse una respuesta muy

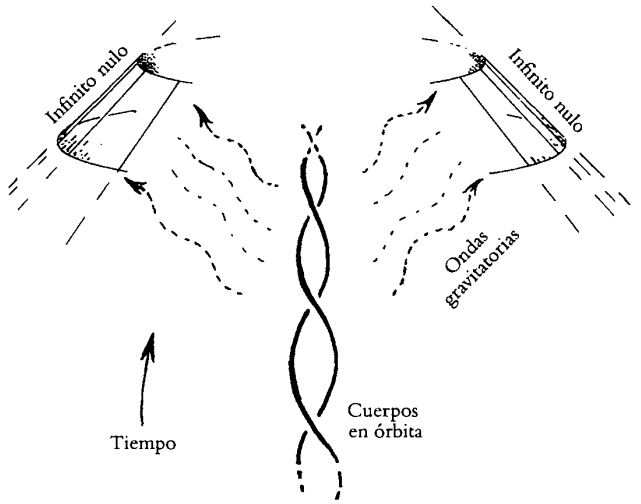


Fig. 19.10. Para un sistema aislado que emite ondas gravitatorias, donde puede suponerse que el espaciotiempo es asintóticamente plano, hay una medida precisa de la masa/energía-momento total y de su pérdida a través de la radiación gravitatoria, conocida como ley de conservación de masa/energía de Bondi-Sachs. Las cantidades matemáticas relevantes son no locales y están definidas en el «infinito nulo» (una noción geométrica que se expondrá en §27.12).

completa. Estas situaciones son las conocidas como *asintóticamente planas*, y se refieren a sistemas gravitantes que pueden considerarse aislados del resto del universo, debido esencialmente a su gran distancia a cualquier otro objeto. Podría ser, por ejemplo, un sistema de estrellas dobles, como el púlsar binario de Hulse-Taylor, en el que uno está interesado en la energía que se pierde por radiación gravitatoria. El trabajo de Hermann Bondi y sus colaboradores, generalizado por Rayner Sachs[15] (para eliminar la hipótesis simplificadora de Bondi de simetría axial), proporcionó una explicación matemática precisa de la masa/energía que sale de un sistema semejante en forma de ondas gravitatorias, y como consecuencia se consiguió una ley de conservación para la energía-momento;[16] véase la Fig. 19.10. Esta ley de conservación no tiene el carácter local de la que existe para campos no gravitatorios, tal como

se manifiesta en la «ecuación de conservación» $\nabla^a T_{ab} = 0$, y solo se aplica de una forma exacta en el límite en que el sistema llega a estar aislado espacialmente de cualquier otro objeto. Pese a todo, hay algo «milagroso» en cómo encajan todas estas cosas, entre ellos ciertos teoremas «de positividad» que fueron demostrados más tarde y que nos dicen que la *masa total* de un sistema (incluidas las «contribuciones de energía potencial gravitatoria negativa» que se han planteado antes) no puede ser negativa.[17]

Existen recetas generales para obtener leyes de conservación para sistemas de campos interactuantes, y proceden del enfoque lagrangiano que será introducido en el próximo capítulo. El enfoque lagrangiano es muy potente, general y bello, pese al hecho de que no parece darnos (al menos de forma directa) todo lo que necesitamos en el caso de la gravitación. Este, y el enfoque hamiltoniano íntimamente relacionado, constituyen partes centrales de la física moderna y es importante saber algo sobre ellos. Aventurémonos a continuación en este fabuloso territorio.

Notas

Sección 19.1

19.1. Parece dudoso que el propio Newton hubiera mantenido de forma tan dogmática una imagen semejante basada en partículas (véanse las *Queries* de Newton en su *Optiks* de 1730). No obstante, esta visión «newtoniana» fue defendida enérgicamente por R. G. Boscovich durante el siglo XVIII; véase Barbour (1989).

Sección 19.3

19.2 El resultado se aplicaría también en un espaciotiempo trivial topológicamente curvo, de modo que (más concretamente) una 2-superficie cerrada siempre llena un 3-volumen compacto.

Sección 19.4

19.3. Véase, por ejemplo, Flanders (1963).

19.4. Véase Weyl (1928), pp. 87-88 (trad. pp. 100-101). Esta observación fue

hecha también de forma independiente por W. Gordon, y por Pauli y Heisenberg; véase Pais (1986), p. 345.

19.5. Véase Aharonov y Bohm (1959). De hecho, este efecto ya había sido advertido diez años antes por Ehrenberg y Siday (1949). Fue verificado experimentalmente por Chambers, y luego establecido de forma más concluyente por Tonomura *et al.* (1982, 1986).

Sección 19.5

19.6. Véase Pais (1982).

Sección 19.6

19.7. El requisito, en el texto, de «no coordenadas privilegiadas» no solo es bastante vago, sino también algo que podría considerarse demasiado fuerte. En el espacio plano, por ejemplo, podría decirse razonablemente que las elecciones de «coordenadas cartesianas» (aquí las coordenadas de Minkowski (t, x, y, z) de §18.1, para las que la métrica toma la forma particularmente simple $ds^2 = dt^2 - dx^2 - dy^2 - dz^2$) están «privilegiadas» sobre todos los demás sistemas de coordenadas, y los modelos cosmológicos también tienen sistemas de coordenadas especiales en los que la forma métrica parece particularmente simple (véanse §27.11 y el ejercicio [27.18]). La clave está más bien en la cuestión más sutil de que tales coordenadas especiales no deberían tener ningún papel, y que las ecuaciones de la teoría deberían ser tales que su expresión más natural no dependa de ninguna elección de coordenadas particular.

19.8. Véase Stachel (1995), pp. 353-364. Entre los muchos textos excelentes sobre relatividad general, destacan Synge (1960), Weinberg (1972), Misner, Thorne y Wheeler (1973), Wald (1984), Ludvigsen (1999), Rindler (1977, 2001), Schutz (2003) y Hartle (2003).

Sección 19.7

19.9. El modelo de Einstein era el espacio ε, con topología $S^3 \times E^1$, que encontraremos en §31.16.

19.10. La introducción por parte de Einstein de un término cosmológico fue una de las varias modificaciones de la teoría de la relatividad general original que se han introducido a lo largo de los años. Además de la teoría de Weyl discutida en §19.4 y las ideas de Kaluza-Klein de dimensiones más altas mencionadas en §31.4 (ahora normalmente combinadas con supersimetría; véanse §§31.2,3), existe la modificación de

Brans-Dicke en la que hay un campo escalar adicional, y los numerosos intentos del propio Einstein por una «teoría de campo unificado» propuestos en el período comprendido entre 1925 y 1955. Véanse Einstein (1925, 1945), Einstein y Straus (1946), Einstein (1948), Einstein y Kaufman (1955) y Schrödinger (1950); para una referencia más reciente, véase Antoci (2001). La mayoría de estas propuestas pretendían incorporar el electromagnetismo, y quizá otros campos, en el armazón general de la relatividad general. También es digno de mención el esquema conocido como teoría Einstein-Cartan-Sciama-Kibble, en la que se introduce una torsión (§14.4) y se considera que describe un efecto gravitatorio directo de una densidad de espín (véase §22.8); véanse Kibble (1961), Sciama (1962) y las exposiciones de Trautman (1972, 1973).

Sección 19.8

19.11. Aquí se considera que la teoría de Einstein incluye la de Newton, y debería resaltarse que la cifra «10^{14}» no representa un aumento de precisión sobre el esquema de Newton. Además, habría que tener en cuenta que parte de la precisión del cronometraje se destina a determinar los parámetros desconocidos, tales como las masas, inclinación orbital, excentricidad, etc., que son necesarios para calcular los detalles del sistema. El «10^{14}» es en realidad una medida de la consistencia global de la imagen.

19.12. Los resultados de 1991 de D. S. Robinson y colaboradores, utilizando «interferometría de muy larga línea de base», confirman ahora los efectos de curvatura de la luz predichos por la relatividad general hasta una precisión de 10^{-4}.

19.13. Para una exposición detallada de la anomalía del perihelio de Mercurio, véase Roseveare (1982).

19.14. Estas búsquedas de ondas gravitatorias tienen acrónimos tan pintorescos como LIGO, LISA y GEO. Véanse Shawhan (2001), Abbott (2004), Grishchuk *et al.* (2001), Thorne (1995b), así como la muy útil página web de John Baez http://math.ucr.edu/home/baez/week143.html.

19.15. Bondi (1960), Bondi *et al.* (1962) y Sachs (1961, 1962a). Este trabajo fue anticipado en parte por Trautman (1958).

19.16. Véanse también Newman y Unti (1962), Penrose (1963, 1964), Sachs (1962b), Bannor y Rotenberg (1966) y Penrose y Rindler (1986), pp. 423-427.

19.17. Schoen y Yau (1979, 1982), Witten (1981), Nester (1981), Parker y Taubes (1982), Ludvigsen y Vickers (1982), Horowitz y Perry (1982) y Reula y Tod (1984); véanse también Penrose y Rindler (1986) y §32.3, en particular, la nota 32.11.

20

Lagrangianos y hamiltonianos

20.1. El mágico formalismo lagrangiano

En los siglos que siguieron a la introducción por parte de Newton de sus leyes dinámicas se construyó un impresionante cuerpo de trabajo teórico sobre estos fundamentos newtonianos. Euler, Laplace, Lagrange, Legendre, Gauss, Liouville, Ostrogradski, Poisson, Jacobi, Hamilton y otros aportaron ideas reformuladoras que condujeron a una profunda visión unificadora. En este capítulo haré una breve introducción a esta visión de conjunto dinámica, aunque me temo que mi explicación solo proporcione una impresión muy engañosa o inadecuada de la magnitud del logro. Debería señalarse también que la existencia de una imagen unificadora tan elegante matemáticamente parece decirnos algo profundo sobre los pilares matemáticos de nuestro universo físico, incluso en el nivel de las leyes que se revelaron en la mecánica newtoniana del siglo XVII. No muchas leyes sugeridas para un universo físico podían llevar a estructuras matemáticas de un esplendor tan imponentes.

¿Cuál es esta elegante imagen unificadora que resultó de la mecánica de Newton? Básicamente se presenta de dos formas diferentes pero íntimamente relacionadas, cada una de las cuales tiene sus virtudes peculiares. Llamaremos a la primera imagen *lagrangiana*, y a la segunda, imagen *hamiltoniana*. (La dificultad habitual con los nombres resulta inevitable. Al parecer, ambas imágenes eran conocidas para Lagrange, bastante anterior a Hamilton, y la imagen lagrangiana fue al menos parcialmente anticipada por Euler.) Consideremos un sistema newtoniano que consiste en un número (finito) de partículas individuales y

quizá algunos cuerpos rígidos, cada uno de ellos considerado como una entidad indivisible. Habrá un *espacio de configuración*, $\mathcal{C}$, de un número grande, N, de dimensiones, cada uno de cuyos puntos representa una única disposición espacial de todas estas partículas y cuerpos (véase §12.1). En el transcurso del tiempo, ese *único punto* de $\mathcal{C}$ que representa a todo el sistema se moverá dentro de $\mathcal{C}$ de acuerdo con cierta ley que engloba el comportamiento newtoniano del sistema; véase la Fig. 20.1. Resulta muy notable (y muy valioso computacionalmente) el hecho de que esta ley puede obtenerse por un procedimiento matemático directo a partir de una *única función*. En la imagen lagrangiana (al menos en su forma más simple y más habitual),[1] esta función —denominada la función *lagrangiana*— se define sobre la *fibra tangente* $\mathrm{T}(\mathcal{C})$ del espacio de configuración $\mathcal{C}$ (Fig. 20.2a); véase §15.5. En la imagen hamiltoniana, la función —denominada función *hamiltoniana*— se define sobre la *fibra cotangente* $\mathrm{T}^*(\mathcal{C})$ (véase §15.5), denominada *espacio de fases* (Fig. 20.2b). Notemos que tanto $\mathrm{T}(\mathcal{C})$ (cada uno de cuyos puntos representa un punto Q de $\mathcal{C}$, junto con un vector tangente en Q) como $\mathrm{T}^*(\mathcal{C})$ (cada uno de cuyos puntos representa un punto Q de $\mathcal{C}$, junto con un vector cotangente en Q) son variedades $2N$-dimensionales.

En esta sección investigaremos la imagen lagrangiana, dejando la hamiltoniana para la sección siguiente. Coordenadas para el $\mathrm{T}(\mathcal{C})$ de Lagrange servirían para determinar las posiciones de todos los cuerpos newtonianos (incluidos los ángulos apropiados para especificar las orien-

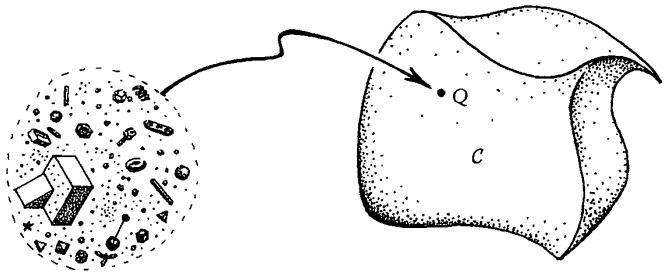


Fig. 20.1. Espacio de configuración. Cada punto Q de la variedad N-dimensional $\mathcal{C}$ representa una configuración posible entera de (digamos) una familia newtoniana de partículas puntuales y cuerpos rígidos. A medida que el sistema evoluciona en el tiempo, Q describe una curva en $\mathcal{C}$.

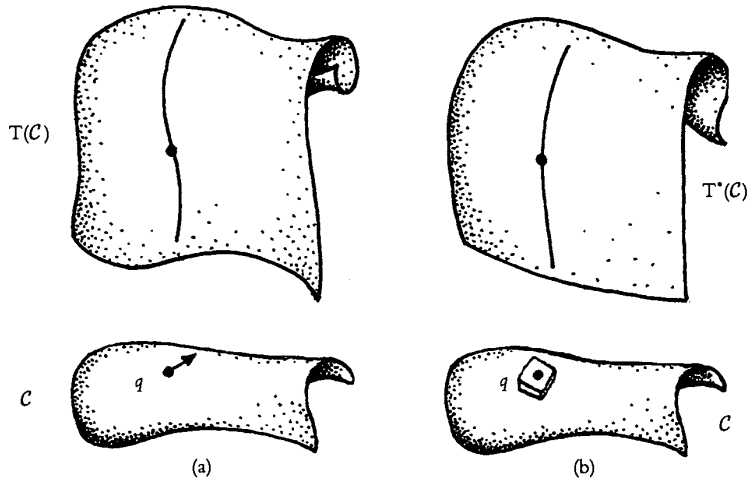


Fig. 20.2. (a) En la imagen lagrangiana estándar, el lagrangiano $\mathcal{L}$ es una función suave sobre el fibrado tangente $\mathrm{T}(\mathcal{C})$ del espacio de configuración $\mathcal{C}$. (b) En la imagen hamiltoniana, el hamiltoniano $\mathcal{H}$ es una función suave del fibrado cotangente $\mathrm{T}^*(\mathcal{C})$ llamado espacio de fases.

taciones espaciales de los cuerpos rígidos, etc.) y también sus *velocidades* (incluidas las correspondientes velocidades angulares de los cuerpos rígidos, etc.). Las coordenadas de posición $q^1, \ldots, q^N$, normalmente denominadas «coordenadas generalizadas», etiquetan los diferentes puntos q del espacio de configuración $\mathcal{C}$ (quizá dadas solo «al modo de atlas»; véase §12.2). Cualquier sistema de coordenadas (adecuado) servirá. No hace falta que sean «cartesianas» ni de ningún otro tipo estándar. Esta es la belleza del enfoque lagrangiano (y también del hamiltoniano). La elección de coordenadas está gobernada simplemente por la conveniencia. Este es precisamente el mismo papel de las coordenadas utilizadas en los capítulos 8, 10, 12, 14 y 15, etc., cuando considerábamos variedades generales de diversos tipos. En correspondencia con el conjunto escogido de coordenadas generalizadas están las «velocidades generalizadas» $\dot{q}^1, \ldots, \dot{q}^N$, donde el «punto» significa el ritmo de cambio con respecto al tiempo «d/dt»:

$$\dot{q}^1 = \frac{\mathrm{d}q^1}{\mathrm{d}t}, \ldots, \dot{q}^N = \frac{\mathrm{d}q^N}{\mathrm{d}t}.$$

El lagrangiano $\mathcal{L}$ se escribirá como una función de *todas* ellas:[2]

$$\mathcal{L} = \mathcal{L}(q^1, \ldots, q^N; \dot{q}^1, \ldots, \dot{q}^N).$$

Cada $\dot{q}^r$ tiene que ser tratada como una variable *independiente* (independiente de q^r, en particular). Esta es una de las características inicialmente desconcertantes de los lagrangianos, ¡pero funciona![3]

La interpretación física normal del valor real de la función $\mathcal{L}$ sería la diferencia $\mathcal{L} = K - V$ entre la energía cinética K del sistema y la energía potencial debida a las fuerzas externas, expresadas en dichas coordenadas (véase §18.6). Las ecuaciones de movimiento del sistema —que codifican *todo* su comportamiento newtoniano— vienen dadas por lo que se denominan las ecuaciones de *Euler-Lagrange*, que son sorprendentes por su extraordinario alcance y esencial simplicidad:

$$\frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial \mathcal{L}}{\partial \dot{q}^r} = \frac{\partial \mathcal{L}}{\partial q^r} \quad (r = 1, \ldots, N).$$

Recordemos que cada $\dot{q}^r$ debe tratarse como una variable independiente, de modo que la expresión «$\partial\mathcal{L}/\partial\dot{q}^r$» (que significa «derivar *formalmente* $\mathcal{L}$ con respecto a $\dot{q}^r$, manteniendo fijas todas las demás variables) ¡tiene sentido realmente!

Estas ecuaciones expresan un hecho notable, a veces conocido como *principio de Hamilton* o *principio de acción estacionaria*. Su significado se hace quizá más claro si pensamos en términos del movimiento del punto Q, en $\mathcal{C}$, donde recordamos que $\mathcal{C}$ representa el espacio de configuraciones espaciales posibles de todo el sistema (i.e., todas las posiciones de todas sus partes). El punto Q, cuya posición en cualquier instante está etiquetada por las q^r, se mueve a lo largo de cierta curva en $\mathcal{C}$ a una cierta velocidad que, junto con la dirección tangente a la curva, está determinada por los valores de las $\dot{q}^r$. Las ecuaciones de Euler-Lagrange nos dicen básicamente que el movimiento de Q en $\mathcal{C}$ es tal que minimiza la *acción*, siendo esta «acción» la integral de $\mathcal{L}$ a lo largo de la curva, tomada entre dos puntos extremos fijos, a, b, en el espacio de configuración $\mathcal{C}$; véase la Fig. 20.3.

Dicho de forma más correcta, esta acción puede no ser realmente un «mínimo», pero el término «estacionario» sería siempre ade-

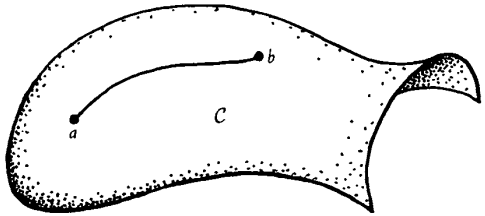


Fig. 20.3. Principio de Hamilton. Las ecuaciones de Euler-Lagrange nos dicen que el movimiento de Q a través de $\mathcal{C}$ es tal que hace la acción —la integral de $\mathcal{L}$ a lo largo de una curva, tomada entre dos puntos fijos, a, b, en $\mathcal{C}$— estacionaria bajo variaciones de la curva.

cuado. La situación es básicamente similar a lo que sucede en el cálculo infinitesimal ordinario (véase §6.2), donde el *mínimo* de una función suave de valor real $f(x)$ ocurre donde $\mathrm{d}f/\mathrm{d}x = 0$, pero a veces $\mathrm{d}f/\mathrm{d}x = 0$ ocurre cuando la función f no alcanza un mínimo: podría ser un máximo o un punto de inflexión o, en dimensiones más altas, lo que se denomina un *punto de ensilladura* (Fig. 20.4b). Todos los lugares donde $\mathrm{d}f(x)/\mathrm{d}x = 0$ se denominan *estacionarios*. Véanse las Figs. 6.4 y 20.4. Recordemos la caracterización básicamente similar, dada en §14.8, §17.9 y §18.3 de una *geodésica* en el espacio (pseudo)riemanniano como un «camino de longitud mínima» en el caso (localmente) definido positivo, y a veces como un «camino de género tiempo de longitud máxima» en el caso lorentziano, aunque simplemente de «longitud estacionaria» en el caso general. Así pues, la trayectoria de Q puede considerarse como un tipo de «geodésica» en el espacio $\mathcal{C}$.

Resulta útil considerar un sencillo ejemplo de un lagrangiano, como es el de una única partícula newtoniana de masa m que se mueve en cierto campo externo fijo dado por un potencial V que depende de la posición: $V = V(x, y, z; t)$. El significado de V es que define la *energía potencial* de la partícula debida a dicho campo externo. En el caso del campo gravitatorio de la Tierra (cerca de la superficie de la Tierra), considerado como una atracción constante hacia abajo, podemos tomar $V = mgz$, donde z es la altura sobre el suelo y g es la aceleración hacia abajo de la gravedad. Las tres componentes de la velocidad son $\dot{x}$,

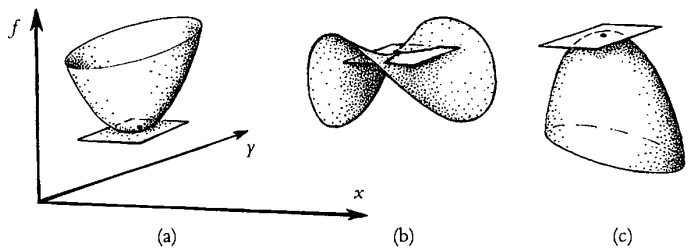


Fig. 20.4. Los valores estacionarios de una función suave de valor real f de varias variables. Se ilustra el caso de una función $f(x, y)$ de dos variables. Esta es estacionaria, donde su gráfica (una superficie 2-dimensional) es horizontal ($\partial f/\partial x = 0 = \partial f/\partial y$). Esto ocurre (a) donde f tiene un mínimo, pero también en otras situaciones tales como (b) en un punto de ensilladura y (c) en un máximo. En el caso del principio de Hamilton (Fig. 20.3) —o una geodésica que conecta dos puntos a, b—, el lagrangiano $\mathcal{L}$ toma el valor de f, pero la especificación de un camino requiere infinitos parámetros, en lugar de solo x e y. De nuevo, $\mathcal{L}$ puede no ser un mínimo, aunque es un punto estacionario de algún tipo.

$\dot{y}, \dot{z}$, de modo que utilizando la expresión $\frac{1}{2}mv^2$ para la energía cinética (véase §18.6), encontramos el lagrangiano

$$\mathcal{L} = \frac{1}{2}m(\dot{x}^2 + \dot{y}^2 + \dot{z}^2) - mgz.$$

La ecuación de Euler-Lagrange para z nos da ahora $\mathrm{d}(m\dot{z})/\mathrm{d}t = -mg$, de la que se sigue la constancia de aceleración de Galileo, en la dirección de la Tierra.[20.1]

20.2. La más simétrica imagen hamiltoniana

En la imagen hamiltoniana seguimos utilizando coordenadas generalizadas, pero ahora las coordenadas de posición generalizadas $q^1, \ldots, q^N$ se toman junto con las que se denominan sus coordenadas de *momento generalizado* $p_1, \ldots, p_N$ (en lugar de las velocidades). Para una única partícula libre, el momento es simplemente la velocidad de la partícula multi-

[20.1] Llene todos los detalles, completando el razonamiento para obtener el movimiento parabólico de Galileo para la caída libre bajo gravedad.

plicada por su masa. Pero en general, la expresión para el momento generalizado no tiene por qué ser esta. Sin embargo, siempre podemos obtenerla a partir del lagrangiano mediante la fórmula definitoria

$$p_r = \frac{\partial \mathcal{L}}{\partial \dot{q}^r}.$$

En cualquier caso, estos parámetros p_r sirven para proporcionar coordenadas para los *espacios cotangentes* a $\mathcal{C}$, de modo que un covector puede escribirse como

$$p_a \mathrm{d}q^a$$

(donde recordamos el convenio de suma de §12.7, que adoptamos aquí, aunque es legítimo leer esto también como una expresión de índices abstractos, como en §12.8). Esta es, por supuesto, una 1-forma, y su derivada exterior (§12.6)

$$\boldsymbol{S} = \mathrm{d}p_a \wedge \mathrm{d}q^a$$

es una 2-forma (que satisface $\mathrm{d}\boldsymbol{S} = 0$)[20.2] que asigna una *estructura simpléctica* natural al espacio de fases $\mathrm{T}^*\mathcal{C}$) (véase §14.8). Gran parte de la fuerza de la imagen hamiltoniana reside en el hecho de que los espacios de fases son *variedades simplécticas*, y esta estructura simpléctica es independiente del hamiltoniano concreto que se elija para dar la dinámica. La física clásica está así íntimamente relacionada con la bella y sorprendente geometría de las variedades simplécticas a la que llegaremos en §20.4.

Como paso preliminar para comprender el papel que desempeña esta geometría, veamos la forma de las ecuaciones dinámicas de Hamilton. Estas describen la evolución temporal de un sistema como una trayectoria, dentro del espacio de fases $\mathrm{T}^*(\mathcal{C})$, de un punto P que representa todo el sistema newtoniano. La evolución está gobernada por completo por la *función hamiltoniana*

$$\mathcal{H} = \mathcal{H}(q^1, \ldots, q^N; p^1, \ldots, p^N),$$

que (en el caso de lagrangianos y hamiltonianos independientes del tiempo que nos interesan aquí) describe la *energía total* del sistema, en

[20.2] ¿Por qué?

términos de las posiciones y los momentos (generalizados). En realidad, podemos obtenerla a partir del lagrangiano por medio de la expresión (convenio de suma de índices abstractos)

$$\mathcal{H} = \dot{q}^r \frac{\partial \mathcal{L}}{\partial \dot{q}^r} - \mathcal{L},$$

que luego tiene que reescribirse eliminando todas las velocidades generalizadas en favor de los momentos generalizados (¡lo que en general no es una tarea fácil!). En términos de dichas coordenadas de posición y momento, las ecuaciones de evolución de Hamilton tienen una bella simetría

$$\frac{\mathrm{d}p_r}{\mathrm{d}t} = -\frac{\partial \mathcal{H}}{\partial q^r}, \quad \frac{\mathrm{d}q^r}{\mathrm{d}t} = \frac{\partial \mathcal{H}}{\partial p_r}.$$

Estas ecuaciones describen la velocidad de un punto P en $\mathrm{T}^*(\mathcal{C})$. Esta velocidad está definida para todo P, de modo que tenemos un campo vectorial en $\mathrm{T}^*(\mathcal{C})$ definido por el hamiltoniano $\mathcal{H}$. En términos de la notación con el «operador de derivada parcial» para un campo vectorial dado en §12.3, esto es[20.3]

$$\frac{\partial \mathcal{H}}{\partial p_r}\frac{\partial}{\partial q^r} - \frac{\partial \mathcal{H}}{\partial q^r}\frac{\partial}{\partial p_r},$$

escrito como $\{\mathcal{H}, \ \}$, en §20.4. Esto proporciona un «flujo» en $\mathrm{T}^*(\mathcal{C})$ que describe el comportamiento newtoniano del sistema (Fig. 20.5).

En el ejemplo concreto de una partícula que cae en un campo gravitatorio constante, como el dado antes (§20.1) en forma lagrangiana, el hamiltoniano es

$$\mathcal{H} = \frac{p_x^2 + p_y^2 + p_z^2}{2m} + mgz$$

$$= \frac{p^2}{2m} + mgz,$$

donde p_x, p_y y p_z son las componentes espaciales ordinarias del momento en las direcciones de los ejes cartesianos x, y y z, respectivamente. Esto puede calcularse directamente si se conoce cuál debería ser la energía total de la partícula cuando se expresa en términos de las componentes de posición y momento, o puede obtenerse a partir del lagrangiano, como se da en los procedimientos anteriores.[20.4]

[20.3] Explíquelo.
[20.4] Hágalo explícitamente. Utilice las ecuaciones newtonianas del movimiento para una partícula que cae en un campo gravitatorio constante.

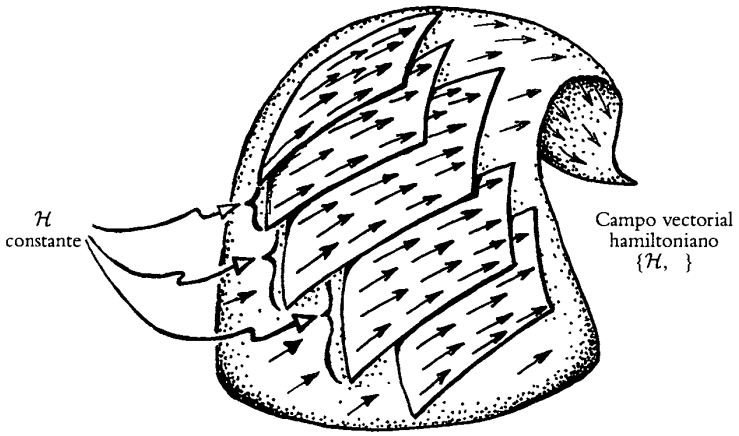


Fig. 20.5. El flujo hamiltoniano $\{\mathcal{H}, \ \}$, que representa la evolución temporal newtoniana del sistema (véase §20.4), es un campo vectorial sobre un espacio de fases $T^*(\mathcal{C})$. Para las hipersuperficies de valores de $\mathcal{H}$-fijos (energía fija, tomando $\mathcal{H}$ independiente del tiempo), las trayectorias permanecen dentro de la hipersuperficie de $\mathcal{H}$ fijo, de acuerdo con la conservación de la energía (véase la nota 27.36 para el término «hipersuperficie»).

En este punto debería confesar una dificultad notacional que no sé cómo evitar, de modo que ¡mejor lo hubiera dejado estar! En §18.7 hemos visto que las componentes espaciales del momento p_1, p_2, p_3 en coordenadas estándar de Minkowski para el espaciotiempo plano, con mi signatura preferida $(+---)$, son las *negativas* de las componentes habituales del momento. Así, en el ejemplo anterior, tenemos $p_x = -p_1$, $p_y = -p_2$, y $p_z = -p_3$. En la discusión *general* de los hamiltonianos es natural utilizar las versiones «abajo» de los momentos, a saber p_a, pese a que esto es inconsistente con las «p_a» (i.e., p_1, p_2, p_3), que son naturales en relatividad con signatura $(+---)$. La forma en que estoy tratando este problema notacional en este libro consiste simplemente en dar el formalismo general utilizando q^a y p_a con los convenios de signo habituales que conectan las p y las q, aunque sin ser específico acerca de la interpretación particular que pudiera tener cada una de las q o las p (¡de modo que el lector puede hacer su propia elección de signos!). Cuando, por el contrario, utilizo x^a y p_a, entonces *entiendo* en realidad la notación consistente con la de §18.7, de modo que $-p_1, -p_2, -p_3$ son las

componentes ordinarias (iguales a p^1, p^2, p^3 en un sistema de Minkowski estándar) del momento espacial ordinario. La consecuencia de esto es que, cuando se escriben en términos de las x en lugar de las q, mis ecuaciones hamiltonianas aparecen con el signo opuesto

$$\frac{\mathrm{d}p_r}{\mathrm{d}t} = -\frac{\partial \mathcal{H}}{\partial x^r}, \quad \frac{\mathrm{d}x^r}{\mathrm{d}t} = \frac{\partial \mathcal{H}}{\partial q_r}.$$

Recomiendo al lector no demasiado interesado en todos los detalles del formalismo que voy a presentar que simplemente ignore esta cuestión por completo. (La mayoría de los expertos harían lo mismo, ¡hasta que llega el momento en que tienen que escribir artículos o libros sobre el tema!)

20.3. Pequeñas oscilaciones

Antes de pasar en la próxima sección a la extraordinaria geometría a la que nos lleva la descripción hamiltoniana, será de gran ayuda considerar el importante tema de las vibraciones de un sistema físico en torno a un estado de equilibrio. El tema tiene una relevancia considerable en diversas áreas, y más adelante tendrá una relevancia especial para nosotros en el contexto de la mecánica cuántica (§22.13). La teoría de vibraciones puede describirse convenientemente en el formalismo lagrangiano o en el hamiltoniano, cada uno de los cuales está perfectamente adaptado para su tratamiento. Aquí daré explícitamente mis descripciones en el formalismo hamiltoniano porque este nos lleva de forma más directa a la versión mecanocuántica de las vibraciones, a la que echaremos una ojeada en §22.13. La teoría lagrangiana de las vibraciones, que es muy similar a la hamiltoniana, queda para el lector (véase el ejercicio [20.10]).

Un sencillo ejemplo de sistema vibrante es un péndulo ordinario que oscila bajo la acción de la gravedad. Cuando las oscilaciones son pequeñas, el movimiento de la lenteja de un lado a otro describe una onda sinusoidal como función del tiempo (véase la Fig. 20.6). (Este es el tipo de comportamiento que se encuentra en las «componentes de Fourier» individuales estudiadas en §9.1.) Para dichas pequeñas oscilaciones, el período de vibración es independiente de la amplitud de la

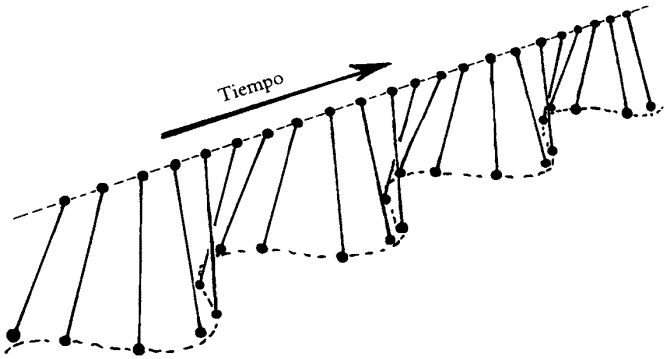


Fig. 20.6. Un péndulo que oscila sometido a la gravedad. Para pequeñas oscilaciones, el movimiento de la lenteja se aproxima a un movimiento armónico simple, y el desplazamiento de la lenteja (representado en función del tiempo) da una «onda seno».

oscilación (i.e., de la distancia que recorre la lenteja en su oscilación), una famosa observación temprana de Galileo, en 1583. Este tipo de movimiento se conoce como *movimiento armónico simple.*

En esta sección veremos cuán ubicuo es este movimiento. Una estructura física general (suponiendo que puedan descartarse los efectos de fricción) puede «oscilar» en torno a su estado de equilibrio solo de maneras muy específicas. Encontraremos que toda oscilación a pequeña escala puede descomponerse en modos particulares de vibración —llamados *modos normales*— en los que toda la estructura participa de un movimiento armónico simple con una frecuencia muy específica, llamada *frecuencia normal.*

Veamos primero cómo se describe analíticamente el movimiento armónico simple. Denotemos por q la distancia horizontal desde la lenteja del péndulo al punto más bajo, o si no el desplazamiento respecto del equilibrio de cualquier otra cantidad vibrante que pudiéramos estar considerando. Entonces, la ecuación de movimiento, para pequeños desplazamientos q, es

$$\frac{\mathrm{d}^2q}{\mathrm{d}t^2} = -\omega^2 q,$$

donde la cantidad positiva $\omega/2\pi$ es la *frecuencia* de la oscilación. Esto nos dice que la aceleración hacia dentro $\mathrm{d}^2q/\mathrm{d}t^2$ es proporcional (con

factor de proporcionalidad ω^2) al desplazamiento hacia fuera. Vemos de §6.5 que tanto $q = \cos \omega t$ como $q = \operatorname{sen} \omega t$ satisfacen esta ecuación, y por lo tanto también lo hace la combinación lineal general

$$q = a \cos \omega t + b \operatorname{sen} \omega t,$$

donde a y b son constantes.[20.5] Para un péndulo de longitud h que oscila (en un plano) bajo la acción de la gravedad, encontramos una ecuación de movimiento que se aproxima estrechamente a la antes dada cuando q es pequeña, con $\omega^2 = g/h$; pero para valores mayores de q, aparecen desviaciones de dicha ecuación.[20.6]

Supongamos que tenemos un sistema hamiltoniano general que está en equilibrio cuando las qs toman ciertos valores especiales $q^a = q^a_0$. Será conveniente escoger el *origen* de nuestras coordenadas generalizadas, de modo que represente nuestro estado de equilibrio, i.e., escogemos $q^a_0 = 0$. «Equilibrio» se refiere a una configuración en la que si inicialmente no hay movimiento, entonces el sistema *permanecerá* estacionario. Podemos estar interesados en si el equilibrio es *estable*, por lo que se entiende la situación en la que, si se perturba ligeramente el sistema en la configuración de equilibrio, entonces el sistema no se alejará mucho del equilibrio, sino que oscilará en torno a él. En nuestro estudio de las vibraciones estamos interesados en oscilaciones en torno a una configuración de equilibrio estable. Por lo tanto, nos interesan solo *pequeños* valores de las coordenadas generalizadas q^a. Además, puesto que nuestras oscilaciones son el resultado de pequeñas perturbaciones con pequeñas velocidades, nos interesarán también pequeños valores del momento p_a.

Supongamos que nuestro hamiltoniano es una expresión analítica en las q y las p —véase §6.4 para el significado de «analítica»—, de

[20.5] Confirme esto, explicando por qué $\omega/2\pi$ es la frecuencia. Explique por qué el gráfico de esta función sigue pareciéndose a una curva sinoidal. ¿Por qué es esta la solución *general*?

[20.6] Demuéstrelo, encontrando la ecuación *completa*, (a) utilizando el método lagrangiano, (b) utilizando el método hamiltoniano, (c) directamente de las leyes de Newton. *Sugerencia*: Demuestre que $\mathcal{L} = \frac{1}{2}mh^2\dot{q}^2(h^2 - q^2)^{-1} + mg(h^2 - q^2)^{1/2}$. (Nótese que los métodos lagrangiano y hamiltoniano no nos hacen ganar nada en este caso sencillo; su potencia reside en el tratamiento de situaciones más generales.)

modo que podamos desarrollarla en serie de potencias en las q y las p. Para una configuración de equilibrio estable, $q^a = 0$ debe representar un *mínimo* (local) de energía potencial.[20.7] Además, la introducción del movimiento solo puede aumentar la energía (la energía cinética); la energía cinética es mínima cuando $p_a = 0$. La energía *total* —que es el valor del hamiltoniano $\mathcal{H}$— tiene así un mínimo local en $q^a = 0 = p_a$. Se sigue de ello que nuestro desarrollo en serie de potencias debe empezar (puesto que están ausentes todos los términos *lineales* en las q, las p o ambas) como

$$\mathcal{H} = \text{constante} + \frac{1}{2}Q_{ab}q^a q^b + \frac{1}{2}P^{ab}p_a p_b$$

$$+ \text{términos de orden 3 o mayor en } q \text{ y } p\text{s},$$

donde Q_{ab} y P^{ab} son las componentes de matrices simétricas constantes y *definidas positivas* (de modo que $Q_{ab}q^a q^b > 0$, si $q^a \neq 0$ y $P^{ab}p_a p_b > 0$, si $p_a \neq 0$; véase §13.8). Los factores 1/2 se introducen por conveniencia.[20.8]

Ignoremos los términos de orden superior para encontrar así la naturaleza de las pequeñas oscilaciones. Las ecuaciones de Hamilton dan entonces

$$\frac{dq^a}{dt} = \frac{\partial \mathcal{H}}{\partial p_a} = P^{ab}p_b,$$

de donde, derivando una vez más con respecto a t,

$$\frac{d^2q^a}{dt^2} = \frac{d}{dt}P^{ab}p_b = P^{ab}\frac{dp_b}{dt}$$

$$= -P^{ab}\frac{\partial \mathcal{H}}{\partial q^b} = -P^{ab}Q_{bc}q^c = -W^a{}_c q^c,$$

donde $W^a{}_c = P^{ab}Q_{bc}$ es el producto matricial de Q_{ab} por P^{ab} (véase §13.3), que podemos escribir de la forma

$$\mathbf{W} = \mathbf{PQ},$$

de modo que la conclusión de nuestra ecuación anterior puede reescribirse ahora

[20.7] ¿Por qué?
[20.8] ¿Puede explicar todo esto con más detalle? ¿Podemos tener los términos lineales si el equilibrio es inestable? Explíquelo.

$$\frac{d^2\mathbf{q}}{dt^2} = -\mathbf{Wq}.$$

Estamos interesados en los *autovectores* de la matriz $\mathbf{W}$ (véase §13.5), que son los vectores $\mathbf{q}$ que satisfacen

$$\mathbf{Wq} = \omega^2\mathbf{q},$$

donde ω^2 es el *autovalor* de $\mathbf{W}$ correspondiente a $\mathbf{q}$. De hecho, este autovalor debe ser positivo, porque las matrices $\mathbf{P}$ y $\mathbf{Q}$ son ambas definidas positivas.[20.9] De modo que podemos escribirlo como el cuadrado de la cantidad positiva ω. Vemos que cualquiera de estos autovectores $\mathbf{q}$ debe satisfacer la ecuación

$$\frac{d^2\mathbf{q}}{dt^2} = -\omega^2\mathbf{q},$$

que representa un movimiento armónico simple con frecuencia $\omega/2\pi$.[20.10]

Cada uno de estos autovectores $\mathbf{q}$ es una cierta combinación de coordenadas generalizadas q^a, de modo que la oscilación correspondiente a $\mathbf{q}$ requeriría que todas estas coordenadas vibren juntas, todas a la misma frecuencia. Esto se conoce como un *modo normal* de oscilación, y la $\omega/2\pi$ correspondiente se denomina la *frecuencia normal* correspondiente a dicho modo. En el caso más general, todas estas frecuencias son distintas, pero en casos especiales «degenerados», algunas de estas frecuencias normales pueden coincidir.[20.11] Los autovalores degenerados deben contar con sus *multiplicidades* adecuadas, por lo que el número total de modos normales sigue siendo igual al número N de coordenadas generalizadas $q_1, q_2, \ldots, q_N$. Puede señalarse que dos modos normales $\mathbf{q}$ y $\mathbf{r}$ cualesquiera, correspondientes a frecuencias diferentes, son «ortogonales» entre sí con respecto a la «métrica» definida por $\mathbf{Q}$, en el sentido $\mathbf{r}^T\mathbf{Qq} = 0$.[20.12]

¿Qué hemos aprendido de todo esto? Hemos llegado a una con-

[20.9] Trate de demostrar esta deducción. *Sugerencia*: Demuestre que la inversa de una matriz definida positiva es definida positiva.
[20.10] Trate de llevar a cabo el análisis anterior de forma lagrangiana en lugar de hamiltoniana.
[20.11] Describa el sistema de autovectores en tales casos degenerados.
[20.12] Demuéstrelo. (Recuerde de §13.7 que «T» significa «traspuesto».)

clusión muy general y notable acerca de cómo puede vibrar un sistema clásico, con N grados de libertad, en torno a una configuración de equilibrio estable. Cualquier vibración semejante está compuesta de modos normales —que pueden tratarse como si fueran independientes uno de otro—, cada uno de los cuales tiene su propia frecuencia característica, y de los que hay N en total. En esta descripción ignoramos los efectos de la *disipación*, por la cual, en la práctica, una vibración en un sistema macroscópico se amortiguaría hasta detenerse, transfiriéndose su energía a los movimientos aleatorios de las partículas constituyentes. Cuando se tienen en cuenta todas las partículas constituyentes (como sucede con una molécula, por ejemplo), entonces no hay disipación.

Hasta ahora he considerado la situación ordinaria en la que el número de grados de libertad N en el sistema es finito, pero la teoría anterior también se aplica a sistemas que son —al menos idealmente— de dimensión infinita. Estamos familiarizados con esta idea cuando nos interesamos por los sonidos que puede emitir un instrumento musical. Un tambor, por ejemplo, o un triángulo musical, oscilarán de acuerdo con diversas frecuencias cuando son percutidos, frecuencias que determinan su timbre especial. Lo mismo sucede con el sonido de un instrumento de viento, aunque en este caso la oscilación procede de la columna de aire en su interior. Similar también es la vibración de una cuerda de un instrumento de cuerda, etc.

El análisis de Fourier que hemos estudiado en el capítulo 9 nos permite expresar las vibraciones de una cuerda de longitud finita. Podríamos considerar que esta está fija en sus extremos, o quizá curvada en círculo. El análisis de Fourier expresa las vibraciones generales como una combinación lineal de modos, siendo estos los tonos puros de onda seno o coseno, y son infinitos en número. En este caso, todas las frecuencias son múltiplos enteros de la del modo fundamental. ¡este es el tipo de cosas en las que uno se esfuerza al construir un instrumento musical sonoro! Pero en general (como sucede con un tambor o una campana), las frecuencias normales no están relacionadas de forma tan simple.

En tales situaciones, los formalismos hamiltoniano o lagrangiano pueden ampliarse inmediatamente para cubrir el caso $N = \infty$, pero es necesario tener cierto cuidado. En cierto sentido, nos vemos llevados

de forma natural a la teoría lagrangiana (o hamiltoniana) de *campos*, a la que echaremos una ojeada en §20.5. Esta tiene muchas aplicaciones en la física moderna. En particular, la aproximación a una teoría fundamental de la naturaleza —conocida como *teoría de cuerdas*— en la que las partículas puntuales son reemplazadas por pequeños lazos (o si no, «cuerdas» con extremos abiertos) requiere este formalismo. Aquí se considera que los diversos campos o partículas de la naturaleza aparecen a partir de modos normales de vibración de las «cuerdas» (véanse §§31.5,7,14).

Habría que hacer una última observación. La exposición de esta sección se ha centrado solo en oscilaciones en torno a un equilibrio estable, pero se aplica también a movimientos a partir de un *equilibrio inestable*. La diferencia básica es que nuestra matriz simétrica real $\mathbf{Q}$ no es ahora definida positiva (ni siquiera definida no negativa), de modo que $\mathbf{W} = \mathbf{PQ}$ puede tener autovalores *negativos*. Las correspondientes pequeñas perturbaciones divergen entonces exponencialmente a partir del equilibrio.[20.13]

20.4. La dinámica hamiltoniana como geometría simpléctica

Retrocedamos para ver cómo se ligan las ecuaciones de Hamilton, en un número finito de dimensiones, con la geometría simpléctica. Como se ha descrito en §14.8, cualquier variedad simpléctica posee una operación, que puede realizarse sobre pares de campos escalares Φ, Ψ en la variedad para dar otro campo escalar Θ, denominado su *paréntesis de Poisson*.[20.14]

$$\Theta = \{\Phi, \Psi\} = \frac{\partial \Phi}{\partial p_a}\frac{\partial \Psi}{\partial q^a} - \frac{\partial \Phi}{\partial q^a}\frac{\partial \Psi}{\partial p_a}.$$

Si dejamos en blanco el «espacio de la Ψ», obtenemos un operador diferencial $\{\Phi, \ \}$, un *campo vectorial* (véase §12.3) cuya acción sobre Ψ da $\{\Phi, \Psi\}$. Sustituyamos $\mathcal{H}$ en lugar de Φ. Observamos que el campo vectorial $\{\mathcal{H}, \ \}$ «apunta a lo largo de» las trayectorias en $\mathrm{T}^*(\mathcal{C})$ que re-

[20.13] Describa este comportamiento.
[20.14] Confirme que esta expresión para $\{\Phi, \Psi\}$ coincide con la de §14.9.

presentan *evolución temporal*; de hecho $\{\mathcal{H}, \ \}$, es precisamente esta evolución de acuerdo con las ecuaciones de Hamilton (véase §20.2). Una de las características notables de la geometría simpléctica es que la evolución dinámica de un sistema puede así ser englobada geométricamente en una única función escalar (a saber, el hamiltoniano).

La geometría simpléctica hace muchas otras cosas por nosotros. Por ejemplo, hay un famoso resultado debido a Liouville que afirma que el volumen del espacio de fases es siempre conservado por la dinámica; véase la Fig. 20.7. Como elemento de *volumen* en el espacio de fases se toma la $2N$-forma

$$\boldsymbol{\Sigma} = \boldsymbol{S} \wedge \boldsymbol{S} \wedge \ldots \wedge \boldsymbol{S},$$

con N de los $\boldsymbol{S}$s en producto cuña; aquí, como recordamos, la 2-forma simpléctica $\boldsymbol{S}$ viene dada por $\boldsymbol{S} = \mathrm{d}p_a \wedge \mathrm{d}q^a$. Ahora bien, no es difícil comprobar que el propio $\boldsymbol{S}$ se conserva en la evolución hamiltoniana (i.e., que se anula la derivada de Lie de $\boldsymbol{S}$ con respecto al campo vectorial $\{\mathcal{H}, \ldots\}$.[20.15] Se sigue inmediatamente que la forma de volumen completa $\boldsymbol{\Sigma}$ se conserva también en dicha evolución. Este es el teorema de Liouville.

Puesto que $\{\mathcal{H}, \mathcal{H}\} = 0$,[20.16] se sigue que el propio hamiltoniano se conserva, i.e., es constante a lo largo de las trayectorias, lo que es un reflejo del hecho de que *la energía total de un sistema cerrado es constante*. Así pues, cada trayectoria reside en una superficie $(2N-1)$-dimensional dada por $\mathcal{H}$ = constante; véase la Fig. 20.5. Ahora podemos considerar la historia entera del sistema representada por su trayectoria en $\mathrm{T}^*(\mathcal{C})$. El espacio de dichas trayectorias, para un valor dado de $\mathcal{H}$, es $(2N-2)$-dimensional, véase la Fig. 20.8. (Perdemos una dimensión debido a que mantenemos $\mathcal{H}$ fijo, y perdemos otra porque «dividimos» por las trayectorias 1-dimensionales.) Es un hecho sorprendente e importante que la $(2N-2)$-variedad resultante es de nuevo simpléctica. Este procedimiento (no solamente cuando se escoge que Φ sea $\mathcal{H}$) tiene muchas aplicaciones elegantes en la mecánica clásica y la geometría simpléctica.

[20.15] Demuéstrelo.
[20.16] ¿Por qué?

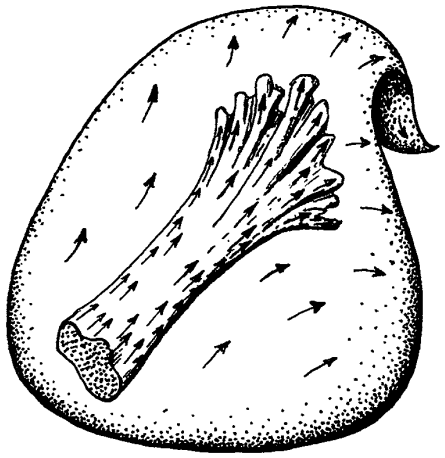


Fig. 20.7. Teorema de Liouville. El flujo hamiltoniano conserva el volumen de la región del espacio de fases inicial (que representa un rango de estados iniciales posibles), incluso si la forma de esta región puede distorsionarse fuertemente en la evolución temporal.

Hay una belleza indudable en esta imagen maravillosamente global de la dinámica newtoniana. De todas formas, como veremos también en relación con teorías físicas posteriores, es importante que no nos dejemos arrastrar por la belleza y aparente finalidad de tales esquemas matemáticos aparentemente bien entretejidos. En el pasado, la naturaleza solía tentarnos primero a una complacencia eufórica por la potencia y elegancia de las estructuras matemáticas que parecía obligarnos a aceptar como guía para su mundo, pero en ocasiones nos sacaba precipitadamente de nuestro torpor conceptual mostrándonos que nuestra imagen ¡no podía haber sido correcta a fin de cuentas! No obstante, el cambio ha sido siempre un cambio sutil que deja al edificio anterior orgulloso y en pie, pese al hecho de que los cimientos sobre los que se había sostenido han sido ahora completamente reemplazados.

La visión hamiltoniana proporciona un ejemplo maravilloso. Aunque la mecánica clásica que encarna está contradicha por algunos hechos brutos del mundo cuántico, la estructura hamiltoniana nos ofrece un camino importante hacia la teoría real de la mecánica cuántica. Más aún: las versiones cuánticas de los hamiltonianos proporcionan ingredientes esenciales para el formalismo cuántico estándar. Debería decir que esto es válido en el caso de la teoría cuántica no relativista estándar, en la que no hay un intento serio de combinar tiempo y espacio de acuerdo con los principios de la relatividad. Sin embargo, en el caso

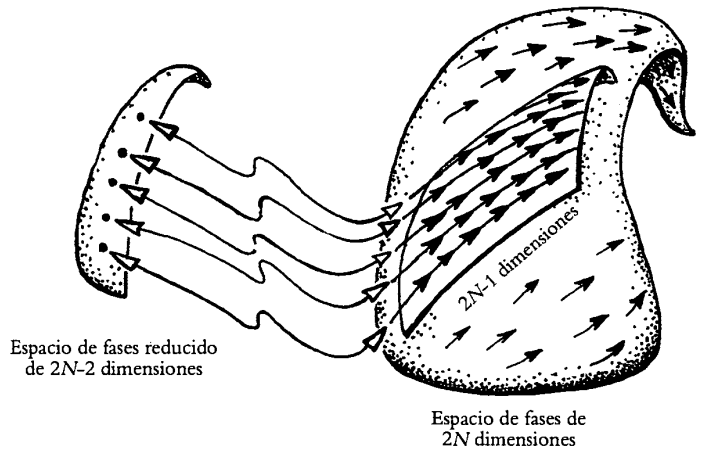


Fig. 20.8. El espacio de fases $T^*(\mathcal{C})$ es una variedad simpléctica $2N$-dimensional, para un $\mathcal{C}$ N-dimensional. Para un valor dado de la energía (H constante, como en la Fig. 20.5), tenemos una región $(2N-1)$-dimensional que contiene una familia $(2N-2)$-dimensional de las trayectorias del flujo hamiltoniano. El espacio de fases reducido, cuyos puntos representan dichas trayectorias, es una variedad simpléctica $(2N-1)$-dimensional.

de la teoría cuántica relativista es el marco lagrangiano el que generalmente se ha encontrado que ofrece el punto de partida más natural. Pero ¿hacia dónde vamos? ¡Es la necesidad de una combinación adecuada de los principios de la relatividad especial con los de la mecánica cuántica la que nos tienta a sumergirnos en el profundo barrizal de la teoría cuántica de campos!

Más adelante, en los capítulos 21-23 y 26, llegaremos a los procedimientos de la teoría cuántica y la teoría cuántica de campos. Pero antes de que podamos hacerlo será necesario preparar mejor el terreno. El propio término «teoría cuántica de campos» implica que son los campos, y no solo las partículas, los que hay que introducir en el marco de las reglas mecanocuánticas. Así pues, tendremos que ver cómo deben tratarse los campos utilizando los métodos lagrangianos (o hamiltonianos).

20.5. Tratamiento lagrangiano de los campos

En las discusiones anteriores sobre los lagrangianos (y los hamiltonianos) se han considerado sistemas newtonianos consistentes en un número finito de partículas y cuerpos rígidos. En estos sistemas hay solo un número finito de grados de libertad, de modo que la variedad $\mathcal{M}$ del espacio de configuración, y su fibra tangente $T(\mathcal{M})$ (así como su fibra cotangente $T^*(\mathcal{M})$) son variedades normales ordinarias de dimensión finita. Sin embargo, el formalismo lagrangiano (y el hamiltoniano) es más general que esto, y puede aplicarse también a campos físicos. Un campo varía de forma continua de un lugar a otro y no puede ser especificado por un número finito de parámetros. El espacio de configuración de los campos de Maxwell libres en cierta región, por ejemplo, será de dimensión infinita.

Sigue siendo posible utilizar el formalismo lagrangiano (o el hamiltoniano) en el caso de un espacio de configuración de dimensión infinita; de hecho, este es el procedimiento adoptado de forma estándar tanto en la teoría de campos clásica como en la cuántica. La principal novedad en los procedimientos matemáticos formales necesarios es el concepto de *diferenciación funcional*. Se considera que, en lugar de ser una función de un número finito de coordenadas generalizadas $q^1, q^2, \ldots, q^N$ y un número finito de velocidades generalizadas $\dot{q}^1, \dot{q}^2, \ldots, \dot{q}^N$, el lagrangiano es una función de cierto número de *campos* $\boldsymbol{\Phi}, \ldots, \boldsymbol{\Psi}$ (cada uno de los cuales es en sí mismo una función en el espaciotiempo, y quizá poseen índices para indicar su carácter tensorial o espinorial) y las *derivadas* de dichos campos $\nabla_a\boldsymbol{\Phi}, \ldots, \nabla_a\boldsymbol{\Psi}$ (donde normalmente aparecen solo derivadas primeras, aunque también se permiten derivadas de orden superior). Nótese que ahora no hay ningún papel especial para las *derivadas temporales* (como se indicaba mediante el «punto» en los argumentos para nuestros lagrangianos originales), y estamos siendo más imparciales utilizando el operador ∇_a en su lugar. En consecuencia, el formalismo entra ahora en línea con los requisitos de la relatividad.

Especialmente en este tipo de situaciones, los lagrangianos se suelen denominar *funcionales*, porque estamos interesados en su forma funcional y no simplemente en los valores que toman para valores concretos de sus argumentos. Las ecuaciones de Euler-Lagrange incluyen ahora

«derivadas con respecto a los campos» y «con respecto a los gradientes de los campos». La realización formal de dichas operaciones refleja con fidelidad las operaciones del cálculo infinitesimal ordinario, como se han descrito en el capítulo 6. A menudo hay sutilezas matemáticas implicadas, si se quiere estar seguro de que los resultados sean rigurosamente ciertos, pero es normal que los físicos no se preocupen demasiado por estas, y el interés principal está en seguir correctamente las reglas formales.

No es mi intención aquí entrar en detalle en dichas cuestiones, pero vale la pena escribir las ecuaciones de Euler-Lagrange para el caso de «derivada funcional» (donde la derivada funcional se denota utilizando «δ» en lugar de «∂»):

$$\nabla_a \frac{\delta \mathcal{L}}{\delta \nabla_a \Phi} = \frac{\delta \mathcal{L}}{\delta \Phi}, \ldots, \nabla_a \frac{\delta \mathcal{L}}{\delta \nabla_a \Psi} = \frac{\delta \mathcal{L}}{\delta \Psi}.$$

Como se ha mencionado más arriba, los campos $\Phi, \ldots, \Psi$ también pueden poseer índices. Realizar una derivada funcional en la práctica es esencialmente aplicar las mismas reglas que en el cálculo ordinario, y utilizar la cantidad justa de «sentido común matemático» (por ejemplo, si $\mathcal{L} = \Phi^c \Phi^b \nabla_a \Psi_b$, entonces $\delta \mathcal{L}/\delta \Phi^c = \Phi^b \nabla_c \Psi_b + \Phi^a \nabla_a \Psi_c$, $\delta \mathcal{L}/\delta \nabla_c \Phi^d = 0$, $\delta \mathcal{L}/\delta \Psi_c = 0$, $\delta \mathcal{L}/\delta \nabla_c \Psi_d = \Phi^c \Phi^d$).

Existe un análogo del principio de Hamilton para tales lagrangianos. Recordemos que este principio expresa las ecuaciones de Euler-Lagrange como la estacionariedad de la *acción*, siendo la acción la integral del lagrangiano a lo largo de una curva que une dos puntos fijos a y b del espacio de configuración (recordemos la Fig. 20.3). En las situaciones más generales que consideramos aquí, los puntos extremos fijos a y b en $\mathcal{C}$ están reemplazados por configuraciones del campo en una región o regiones 3-dimensionales del espaciotiempo. A menudo estas se toman como dos regiones $\mathcal{A}$, $\mathcal{B}$ 3-espaciales, en el espaciotiempo, que llenan el mismo 2-espacio $\mathcal{S}$ (donde quizá $\mathcal{S}$ está tomando en el infinito) —véase la Fig. 20.9—, y esta imagen es también importante en la formulación de integrales de camino de la teoría cuántica de campos a la que llegaremos más adelante (§26.6). Si queremos, podemos tomar $\mathcal{A}$ y $\mathcal{B}$ juntas (invirtiendo la orientación de una de ellas) para constituir la *frontera* $\partial \mathcal{D}$ de un 4-volumen $\mathcal{D}$ espaciotemporal (posiblemente compacto; véase §12.6). Véase la Fig. 20.10. En cual-

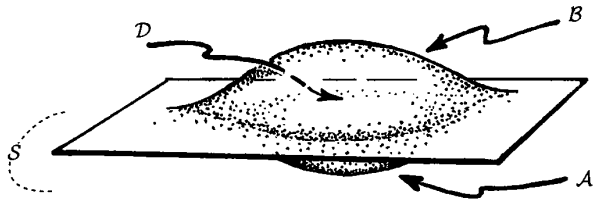


Fig. 20.9. Principio de Hamilton para lagrangianos de *campo*. Los dos puntos extremos fijos a, b, en $\mathcal{C}$, de la Fig. 20.3, representan configuraciones de campo en dos regiones espaciotemporales 3-dimensionales $\mathcal{A}$, $\mathcal{B}$, respectivamente, que forman una «ampolla», que encierra la 4-región $\mathcal{D}$. Podemos tomar $\mathcal{A}$ y $\mathcal{B}$ juntos, terminando en una 2-superficie finita $\mathcal{S}$ (no dibujada en la figura), o podemos considerar «$\mathcal{S}$» en el infinito, quizá a lo largo de una hipersuperficie de género espacio a lo largo de la cual $\mathcal{A}$ y $\mathcal{B}$ coinciden más allá de la región $\mathcal{D}$ (el caso ilustrado).

quier caso, el principio de Hamilton expresa la *estacionariedad* de la integral espaciotemporal del lagrangiano sobre la región $\mathcal{D}$. Así pues, el lagrangiano $\mathcal{L}$ debe considerarse como una *densidad* espaciotemporal, lo que, estrictamente hablando, significa que la entidad invariante es la 4-forma $\mathcal{L}\boldsymbol{\varepsilon}$, donde la 4-forma natural $\boldsymbol{\varepsilon}$ es la cantidad[4] que normalmente se expresa como $\boldsymbol{\varepsilon} = dx^0 \wedge dx^1 \wedge dx^2 \wedge dx^3 \sqrt{(-\det \boldsymbol{g})}$. La integral de *acción* es entonces

$$S = \int_{\mathcal{D}} \mathcal{L}\boldsymbol{\varepsilon}.$$

Las ecuaciones de campo surgen de la afirmación de que la cantidad S es estacionaria con respecto a variaciones de todas las variables (de modo que da el análogo de una geodésica; véase la Fig. 20.3), lo que significa que la derivada variacional de $\mathcal{L}$ con respecto a todos los campos contituyentes y sus derivadas tiene que anularse. Esta condición se escribe

$$\delta S = 0.$$

La magnitud S es fundamental para el enfoque de integrales de camino de la teoría cuántica de campos, a la que llegaremos en §26.6.

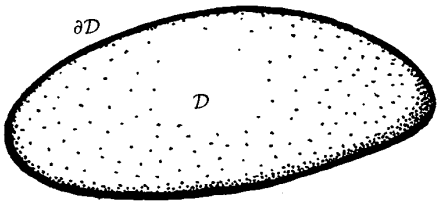


Fig. 20.10. Si lo deseamos, podemos considerar que las $\mathcal{A}$ y $\mathcal{B}$ de la Fig. 20.9 se unen —pero tomadas con orientaciones opuestas (véase la Fig. 12.16)—, de modo que constituyen la frontera $\partial\mathcal{D}$ de un 4-volumen espaciotemporal (compacto) $\mathcal{D}$. El principio de Hamilton (Fig. 20.3) expresa la estacionariedad de $\int_{\mathcal{D}} \mathcal{L}\boldsymbol{\varepsilon}$ para una configuración de campo dada en la frontera $\partial\mathcal{D}$.

20.6. Cómo impulsan los lagrangianos la teoría moderna

La teoría lagrangiana (así como la hamiltoniana) tiene un papel muy influyente en la física moderna, pues se le pueden dar muchos usos notables. Por ejemplo, existe un teorema importante, conocido como *teorema de Nöther*, que nos dice que si un lagrangiano ordinario posee una simetría continua (suave), entonces habrá una ley de conservación asociada a dicha simetría. Por ejemplo, si existe invariancia del lagrangiano bajo *traslación temporal* (i.e., independiente del tiempo), entonces hay una *energía* conservada; si hay invariancia bajo una traslación *espacial*, entonces se conserva un *momento*. Además, si existe invariancia bajo rotación angular en torno a un eje, entonces hay un *momento angular* conservado en torno al mismo eje. Para un sistema aislado en el espaciotiempo plano, cabe esperar estas simetrías. Si escogemos las coordenadas de modo que nuestra simetría dada del lagrangiano $\mathcal{L}$ se expresa en el hecho de que $\mathcal{L}$ es independiente de cierta coordenada de «posición» generalizada q_r, entonces la magnitud conservada será el «momento conjugado» p_r de dicha coordenada q_r, dado por la receta anterior (en §20.2): $p_r = \partial\mathcal{L}/\partial\dot{q}^r$. Se sigue inmediatamente de las ecuaciones de Euler-Lagrange que este p_r es constante en el tiempo.[20.17]

Este procedimiento puede generalizarse a los funcionales lagrangianos de campos. Por ejemplo, si existe una «invariancia gauge», entonces esperamos encontrar una correspondiente «carga conservada» (por ejemplo, la carga eléctrica en el caso de la invariancia gauge electromagnética bajo $\Psi \mapsto e^{i\theta}\Psi$). No obstante, en tales situaciones empiezan a aparecer cuestiones complicadas. Por ejemplo, no es en absoluto una cuestión obvia cómo se pueden aplicar estas ideas para obtener conservación de energía-momento en la relatividad general y, estrictamente hablando, el método no funciona realmente en este caso. El análogo gravitatorio evidente de la simetría gauge $\Psi \mapsto e^{i\theta}\Psi$ es «invariancia bajo transformaciones de coordenadas generales» (de las que la relatividad general se ocupa al tener sus ecuaciones escritas en términos de operaciones tensoriales), pero el teorema de Nöther no funciona en esta situación, pues da algo del tipo «0 = 0». Parece que se necesitan algunas ideas completamente diferentes para la relatividad general, pese a la poderosa comprensión que puede obtenerse utilizando el teorema de Nöther en otras circunstancias, parte de la cual se traslada de forma impresionante notable a la teoría cuántica, como se indicará en §21.1 Como ejemplo de las limitaciones del teorema de Nöther en el caso de la teoría gravitatoria, debería señalarse que un interrogante significativo se cierne aún sobre la cuestión del momento angular en la relatividad general, incluso en el caso de espaciotiempos asintóticamente planos.[5]

La teoría de Einstein puede derivarse ciertamente de un enfoque lagrangiano, como fue demostrado por primera vez por el profundo y versátil matemático David Hilbert (1915). El lagrangiano gravitatorio de Hilbert es básicamente la curvatura escalar R, dividida por la constante $-16\pi G$, pero hace falta convertir esta en una densidad (o una 4-forma) multiplicándola por la 4-forma natural $\boldsymbol{\varepsilon}$ de §20.5. Esta tiene que sumarse al lagrangiano de materia $\mathcal{L}$, y obtenemos para la acción total

$$S = \int_{\mathcal{D}} \left(\mathcal{L} - \frac{1}{16\pi G} R\right)\boldsymbol{\varepsilon}.$$

Cuando Hilbert sugirió esta acción, estaba obsesionado con una teoría de la materia que era muy popular en su época, a saber, la *teoría de Mie*, y estableció su principio de acción gravitatoria solo para el caso en que el lagrangiano de materia fuera el apropiado para la teoría de

[20.17] Explique por qué.

Mie. Al parecer, creía que su lagrangiano total nos daba lo que ahora llamaríamos una «teoría de todo». Eso era en 1915. ¿Quién se acuerda hoy de la teoría de Mie?

Aunque la teoría de Mie implicaba apartarse de la teoría de Maxwell, hacía muchos años que se conocía un lagrangiano apropiado para la teoría de Maxwell *estándar* del campo electromagnético,[6] a saber,

$$\mathcal{L}_{\text{EM}} = -\frac{1}{4} F_{ab}F^{ab}.$$

Pero para hacer que esto funcione necesitamos asegurar que este lagrangiano esté expresado en términos del *potencial electromagnético* A_a. Cuando hay también campos cargados, se necesitan términos adicionales que expresen esta interacción, y estos también incluyen a A_a. Como punto importante, hay que comprobar la invariancia gauge del conjunto. Cuando se incorpora también la gravedad, entonces se necesita que exista la «invariancia gauge» apropiada para la gravedad, a saber, la invariancia de coordenadas. Esto se trata normalmente escribiendo las cosas de manera apropiada de forma tensorial (o alguna otra de acuerdo con otras recetas invariantes que utilizan marcos o formalismos espinoriales apropiados).

Cuando en los intentos modernos en física fundamental se propone alguna nueva teoría, esta viene dada casi invariablemente en forma de cierto funcional lagrangiano. Esto tiene muchas ventajas, tal como el hecho de que existe una mayor probabilidad (pero no una certeza absoluta) de que la teoría resultante tenga las propiedades de consistencia e invariancia requeridas, y que hay implícita alguna forma de la «tercera ley de Newton» (en el sentido de que si dos campos interaccionan, entonces la interacción es mutua: si uno actúa sobre el otro, entonces el segundo retroactúa también sobre el primero). Además, los lagrangianos tienen la agradable propiedad de que si se introduce un nuevo campo, sus contribuciones pueden ser simplemente añadidas al lagrangiano que se tenía antes. Más importante quizá es que existe una ruta directa hacia la formulación de una teoría cuántica mediante el enfoque de integrales de camino al que he aludido antes, y a la que llegaremos con más extensión en §26.6.

No obstante, debo confesar que siento cierta incomodidad al considerar esto como un enfoque *fundamental*. Tengo dificultades para formular mi incomodidad, pero tiene algo que ver con la propia generalidad del enfoque lagrangiano, generalidad que puede ofrecer poca guía para encontrar las teorías correctas. También suele haber una falta de unicidad al escoger un lagrangiano, y quizá haya cierta artificiosidad en muchos casos, por no hablar de complicación no disimulada. Tiende a haber un alejamiento de la comprensión física real y «práctica», especialmente en el caso de lagrangianos para campos. Ni siquiera el lagrangiano para la teoría de Maxwell libre $\frac{1}{4}F_{ab}F^{ab}$ tiene un obvio significado físico (pues esta cantidad es 1/8 de la diferencia entre las longitudes al cuadrado de los vectores de los campos eléctrico y magnético, en términos 3-dimensionales.)[20.18] Además, el «lagrangiano de Maxwell» no funciona como un lagrangiano, a menos que se exprese en términos de un potencial, aunque el valor real del potencial A_a no es una cantidad directamente observable. En el caso de la gravedad (a diferencia del caso del electromagnetismo), el lagrangiano para la teoría de Einstein libre *se anula idénticamente* cuando se satisface la ecuación de campo (puesto que $R_{ab} - \frac{1}{2}Rg_{ab} = 0$ implica $R = 0$). Una vez más, R no funciona como un lagrangiano, a menos que se exprese en términos de magnitudes (normalmente las componentes métricas en cierto sistema de coordenadas) que una vez más no son directamente observables. En la mayoría de las situaciones, la densidad lagrangiana no parece tener por sí misma un significado físico claro; además, suele haber muchos lagrangianos diferentes que llevan a las mismas ecuaciones de campo.

Los lagrangianos para campos son sin duda extraordinariamente útiles como artificios matemáticos, y nos permiten escribir muchísimas sugerencias para teorías físicas. Pero sigo sintiéndome incómodo al poner demasiada confianza en ellos para nuestra búsqueda de mejores teorías físicas fundamentales. Esta incomodidad tiene también relevancia para las cuestiones de la teoría cuántica de campos a la que llegaremos en §26.6, pero esto es suficiente por ahora.

[20.18] Demuéstrelo.

Notas

Sección 20.1

20.1. Tipos más generales de lagrangianos (para sistemas no newtonianos) pueden incluir derivadas superiores y están definidos en lo que se denominan «fibras en chorro» de $\mathcal{C}$, pero ahora no tenemos que preocuparnos por eso.

20.2. En mi exposición simplifico la discusión general de los lagrangianos al suponer que nuestro sistema es lo que se denomina *holónomo*. Con un sistema no holónomo no hay suficientes coordenadas de velocidad disponibles, en relación con las posiciones generalizadas. Un buen ejemplo de sistema no holónomo es un aro que rueda sobre un plano horizontal con la restricción de que no puede deslizar, de modo que su punto de contacto solo puede moverse en la dirección de la tangente al aro por un movimiento de rodadura. Se necesitan dos coordenadas para la localización de este punto de contacto, pero solo hay una disponible para su velocidad.

Se puede considerar que para sistemas tratados en el nivel *fundamental* de descripción no se da tal no holonomicidad. En el caso de nuestro aro, la restricción de rodadura es una idealización en la que se niega la posibilidad de deslizamiento. Una vez que se permite una pequeña cantidad de deslizamiento, el sistema se convierte en holónomo.

20.3. Aquí considero el caso de un «lagrangiano independiente del tiempo» para simplificar la descripción. Pero es fácil introducir una dependencia temporal de las fuerzas externas, incluyendo simplemente otra «coordenada generalizada» $q^0 = t$ y una cantidad formal $\dot{q}^0$ que finalmente toma el valor 1.

Sección 20.5

20.4. Otra forma de especificar $\boldsymbol{\varepsilon}$ es decir que la componente $\boldsymbol{\varepsilon}_{0123}$ de $\boldsymbol{\varepsilon}$, en un sistema de referencia ortonormal dextrógiro local, satisface $\boldsymbol{\varepsilon}_{0123} = 1$ (§19.2). El $\left[{}^{0}_{4}\right]$-tensor $\boldsymbol{\varepsilon}$ está determinado, salvo signo, por la métrica por $\boldsymbol{\varepsilon}_{abcd}\boldsymbol{\varepsilon}_{pqrs}g^{ap}g^{bq}g^{cr}g^{ds} = -24$, donde la elección de signo *proporciona* la orientación del volumen espaciotemporal.[20.19]

[20.19] Demuestre que esta prescripción es equivalente a la dada en el texto central.

Sección 20.6

20.5. Véanse Penrose (1982), Penrose y Rindler (1986), Winicour (1980) y Rizzi (1998).

20.6. Véase Pais (1986), p. 342, y las referencias 46, 47, 48 en p. 357.

21

La partícula cuántica

21.1. Variables no conmutativas

Es probable que la mayoría de los físicos consideren que los cambios que ha aportado la mecánica cuántica para nuestra imagen del mundo son mucho más revolucionarios incluso que el extraordinario espaciotiempo curvo de la relatividad general de Einstein. De hecho, como veremos en este capítulo y en los dos siguientes, lo que realmente nos dice la teoría cuántica que debemos creer acerca de la «realidad» en los niveles submicroscópicos de átomos o de partículas fundamentales está tan enormemente alejado de nuestras imágenes macroscópicas ordinarias que simplemente sería mejor abandonar cualquier imagen a nivel cuántico. Tanto es así que muchos físicos parecen incluso dudar de la propia existencia de una «realidad» genuina en las escalas cuánticas, y en su lugar confían simplemente en el formalismo matemático mecanocuántico para obtener respuestas. (En el capítulo 29 volveré con más detalle a la controvertida cuestión de la «realidad cuántica».)

Pese a todo, es muy notable cuánto de los procedimientos lagrangianos/hamiltonianos del capítulo 20 —ese esquema global aunque enteramente clásico que se desarrolló a partir de la mecánica newtoniana del siglo XVII— proporciona el fondo esencial para la teoría mecanocuántica. Por supuesto, hubo que hacer cambios en el formalismo matemático; si no fuera así, la nueva teoría sería simplemente otra copia de la antigua. Pero es como si el formalismo que se desarrolló a partir del esquema de Newton estuviera ya esperando la llegada de la mecánica cuántica, con piezas de la forma y el tamaño correctos en su

maquinaria, para que los nuevos ingredientes cuánticos pudieran ser colocados simplemente en su lugar.

La propiedad matemática clave que permite que suceda esto es una aparente «curiosidad» que ya había sido advertida en el siglo XIX por el muy original ingeniero electrónico y físico matemático Oliver Heaviside (1850-1925), de quien ya he hablado en §6.1. La observación de Heaviside consistía en que los operadores diferenciales pueden ser tratados a menudo exactamente de la misma forma que los números ordinarios, un hecho que suele ser extraordinariamente útil en la resolución de ciertos tipos de ecuaciones diferenciales. Veamos un ejemplo. Consideremos la ecuación diferencial[1]

$$y + \frac{d^2y}{dx^2} = x^5,$$

(véase §6.3 para el significado de los símbolos). Supongamos que queremos encontrar una función particular $y = y(x)$ que satisfaga esta relación. El método de Heaviside consiste en tratar «d/dx» como si este operador fuera un número ordinario. Para hacer que esto se vea más «plausible», denotemos el operador por la simple letra D:

$$D = \frac{d}{dx}.$$

La entidad representada por «D^2» es entonces la derivación repetida $d^2/dx^2 = (d/dx)^2$, que es un operador de derivada *segunda*; el representado por «D^3» es la derivada tercera d^3/dx^3, etc. Entonces, nuestra ecuación se convierte en $y + D^2y = x^5$, que puede expresarse como

$$(1 + D^2)y = x^5.$$

Podemos «resolverla» formalmente «dividiendo por $(1 + D^2)$», y escribiendo la respuesta como $y = (1 + D^2)^{-1}x^5$. Desarrollando $(1 + D^2)^{-1}$ como una «serie de potencias en D», encontramos

$$y = (1 - D^2 + D^4 - D^6 + \ldots)x^5.$$

(Recordemos que ya hemos considerado esta serie en §4.3, con x en lugar de D.) Notando (§6.5) que $Dx^5 = 5x^4$, $D^2x^5 = 20x^3$, $D^3x^5 =$

$= 60x^2$, $D^4x^5 = 120x$, $D^5x^5 = 120$, $D^6x^5 = 0$, etc., encontramos la solución particular (¡correcta!)[21.1],[21.2],[21.3]

$$y = x^5 - 20x^3 + 120x.$$

Con cuidadosa atención a las reglas apropiadas, este tipo de procedimiento formal puede hacerse perfectamente riguroso, ¡a pesar de que Heaviside encontró al principio mucha oposición a su uso!

Aunque la cantidad $D(= \mathrm{d}/\mathrm{d}x)$ puede tratarse algebraicamente (aunque con el debido cuidado) como un número ordinario, debemos ser prudentes cuando tenemos Ds y xs mezcladas, porque no conmutan. Tenemos que pensar que «x» y «D» están actuando sobre alguna función invisible, digamos $\Psi(x)$, que hay a la derecha. El operador x simplemente multiplica lo que está a su derecha por x, mientras que D deriva con respecto a x lo que está a su derecha. Vemos entonces que tenemos la relación de *conmutación*

$$Dx - xD = 1.$$

¿Por qué es así? Recordemos la propiedad de la «ley de Leibniz» de §6.5, que nos dice que $D(x\psi) = (D(x))\psi + xD(\psi)$, i.e., $D(x\psi) - xD(\psi) = (D(x))\psi$. Esta es simplemente la relación $(Dx - xD)\psi = 1\psi$, donde tenemos en cuenta que $D(x) = 1$ (i.e., D aplicada directamente a x es 1), que es la relación antes mostrada aplicada a una $\psi = \psi(x)$ a la derecha.

Extendamos esto ahora a muchas variables $x^1, \ldots, x^N$, y a los operadores correspondientes $D_1 = \partial/\partial x^1, \ldots, D_N = \partial/\partial x^N$ (ahora operadores de derivada parcial, y recordemos que x^N es simplemente la coordenada N-ésima, y no x multiplicada por sí misma N veces), donde la función «invisible» del segundo miembro es ahora una función de todas

[21.1] Demuestre que $(1 + D^2)\cos x = 0$ y $(1 + D^2)\mathrm{sen}\, x = 0$ (remitiendo a las fórmulas en §6.5, si las necesita).

[21.2] Tomando nota del ejercicio [21.1], encuentre la solución general de $(1 + D^2)y = x^5$, dando una demostración de que su solución es, de hecho, la más general.

[21.3] Trate de explicar por qué el procedimiento dado en el texto pasa por alto muchas de las soluciones dadas en el ejercicio [21.2] ¿Puede sugerir un procedimiento general modificado que las encuentre todas? *Sugerencia*: ¿Hasta qué punto «$1 - D^2 + D^4 - D^6 + \ldots$» satisface realmente los requisitos para un inverso de $1 + D^2$? Trate de actuar sobre $(1 + D^2)\cos x$ con esta expresión infinita.

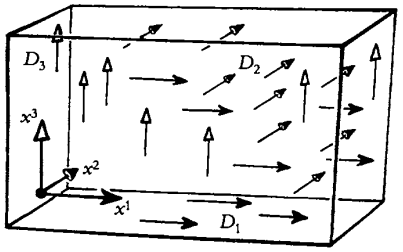


Fig. 21.1. En el espacio N-euclídeo (afín) $\mathbb{E}^N$, existen N simetrías de traslación independientes generadas por los operadores (campos vectoriales) $D_1 = \partial/\partial x^1, D_2 = \partial/\partial x^2, \ldots, D_N = \partial/\partial x^N$, que satisfacen relaciones de conmutación $D_b x^a - x^a D_b = \delta^a_b$ con las coordenadas cartesianas respectivas $x^1, x^2, \ldots, x^N$. (Se ilustra el caso $N = 3$).

estas variables, $\psi = \psi(x^1, \ldots, x^N)$. Obtenemos las relaciones de conmutación

$$D_b x^a - x^a D_b = \delta^a_b.$$

(Recordemos la delta de Kronecker δ^a_b de §13.3; la expresión anterior contiene tanto el conmutador previo, cuando $a = b$, como el hecho de que la x y la D conmutan[21.4] cuando $a \neq b$.) Podríamos suponer que las coordenadas x^a son coordenadas espaciales ordinarias o coordenadas espaciotemporales en el espacio plano, pero también podrían ser algo más generales, tales como las coordenadas generalizadas q^a de los formalismos lagrangiano o hamiltoniano. No obstante, hay algunas dificultades importantes para llevar demasiado lejos esta generalidad. Por ello, al menos por el momento, será mejor imaginar que estamos trabajando con cierto N-espacio *plano* $\mathbb{E}^N$ (no necesariamente de solo 3 o 4 dimensiones). Los operadores $D_1, \ldots, D_N$ describen entonces traslaciones infinitesimales de $\mathbb{E}^N$ en las direcciones de cada uno de los ejes (Fig. 21.1), cada una de las cuales expresa una simetría independiente del espacio afín $\mathbb{E}^N$.

Recordemos, del teorema de Nöther (en §20.6), que existe una íntima asociación entre tales simetrías del espacio y la *conservación del momento lineal*: si un lagrangiano es invariante bajo traslación en cierta dirección, entonces el momento en dicha dirección se conserva. Este es un hecho elegante e importante, y es perfectamente comprensible desde el punto de vista matemático. La mecánica cuántica hace algo que se parece un poco a esto, aunque no es tan matemáticamente compren-

[21.4] ¿Por qué?

sible. De hecho, ¡parece algo sin sentido matemáticamente hablando! Pero la mecánica cuántica va aún más lejos. Pese a todo, existe una indudable elegancia matemática en este extraño procedimiento mecanocuántico; pues en mecánica cuántica no solo hay un momento conservado asociado con tal simetría, sino que ¡el propio momento se *identifica* realmente con el operador diferencial que genera esa simetría concreta!

21.2. Hamiltonianos cuánticos

¿Cómo puede identificarse realmente un momento con un operador diferencial? ¡Parece una locura! Para ser más correctos, hay que introducir un factor $\hbar$ (la versión de Dirac de la constante de Planck, a saber, $h/2\pi$, donde h es la constante de Planck original; véase *infra*) y también la unidad imaginaria i. Así pues, hacemos la definición de apariencia absurda $p_a = i\hbar D_a$, es decir,

$$p_a = i\hbar \frac{\partial}{\partial x^a},$$

para el momento asociado con x^a. Esto nos conduce a una ley de conmutación llamada *regla de conmutación canónica* que relaciona posición y momento lineal

$$p_b x^a - x^a p_b = i\hbar\, \delta^a_b.$$

¿Qué vamos a hacer con este operador/momento de apariencia absurda? El papel de este «momento mecanocuántico», $i\hbar\partial/\partial x^a$, consiste en que debe ser introducido en la función hamiltoniana clásica $\mathcal{H}(p_1, \ldots, p_N; x^a, \ldots, x^N)$, precisamente allí donde solía estar el antiguo momento clásico p_a. Esta es la clave del procedimiento conocido como *cuantización* (canónica). No nos preocupamos de momento por la relatividad, así que los momentos considerados más arriba son los momentos espaciales,[2] y no la energía. Es probable que nuestro espacio $\mathbb{E}^N$ sea mucho mayor que solo 3-dimensional, porque podría haber muchas partículas u otras estructuras implicadas, y todas las diferentes componentes de las posiciones y los momentos tienen que estar en la lista. De acuerdo con la exposición general del capítulo 20, no estoy

admitiendo la posibilidad de una dependencia temporal explícita del hamiltoniano.[3]

La interpretación normal de estas coordenadas x^a será que proporcionan las posiciones de diversas partículas (o quizá otros parámetros apropiados). En este capítulo solo me interesaré en detalle en la mecánica cuántica de una única partícula, pero será bueno tener listo el formalismo general para cuando consideremos sistemas de muchas partículas en el capítulo 23. En el caso especial de una única partícula, existe una simetría relativista evidente entre su componente temporal x^0 y sus tres componentes espaciales x^1, x^2, x^3. Enseguida veremos que esto desempeña un papel importante al definir la evolución temporal real de la mecánica cuántica. De todas formas, tal como están los procedimientos de «cuantización» (canónica), en particular cuando hay muchas partículas involucradas, estos proporcionan un procedimiento no relativista en el que los aspectos espaciales y temporales de la física se tratan de forma muy diferente.

Echemos una ojeada a un sencillo ejemplo de hamiltoniano cuántico para ver cómo funciona esta idea absurda. Podemos considerar el caso de una única partícula newtoniana de masa m que se mueve en un campo externo dado por una función de energía potencial V que puede depender de la posición: $V = V(x, y, z)$. Ya hemos visto el hamiltoniano *clásico*, en §20.2, y recordemos que este es $\mathcal{H} = (p_x^2 + p_y^2 + p_z^2)/2m + V(x, y, z)$, donde p_x, p_y, p_z son los momentos espaciales en las direcciones de los ejes cartesianos x, y y z. El hamiltoniano cuántico (canónicamente cuantizado) es, por lo tanto,

$$\mathcal{H} = \frac{p_x^2 + p_y^2 + p_z^2}{2m} + V(x, y, z) = -\frac{\hbar^2}{2m}\nabla^2 + V(x, y, z),$$

donde $\nabla^2 = (\partial/\partial x)^2 + (\partial/\partial y)^2 + (\partial/\partial z)^2$ (lo que significa $\partial^2/\partial x^2 + \partial^2/\partial y^2 + \partial^2/\partial z^2$) es el laplaciano (considerado antes en §10.5, pero ahora en el caso tridimensional).

En este ejemplo, todo ha procedido de forma suave (¡aunque tendremos que esperar a la próxima sección para ver *hacia dónde*!). No obstante, en general el reemplazamiento de momentos clásicos por momentos cuánticos en el hamiltoniano puede ser un procedimiento no exento de ambigüedades debido principalmente a la no conmutación entre el p mecanocuántico y su correspondiente x. Por ejemplo, si en

el hamiltoniano clásico aparece un término *producto* de la forma px, no está claro si en el correspondiente hamiltoniano cuántico debería aparecer como px, o como xp, o quizá como $\frac{1}{2}(px + xp)$, o como cualquiera entre un número infinito de otras posibilidades. Este tipo de ambigüedad se conoce como el *problema del orden de factores*. En muchas circunstancias prácticas, esta ambigüedad quizá no sea muy grave, pues con frecuencia resulta haber alguna elección «obvia». La elección puede estar gobernada por algún principio guía dominante, como un requisito de simetría o invariancia, o quizá por alguna convincente demanda física, matemática o estética. O a veces puede suceder que diferentes alternativas den lugar a teorías cuánticas equivalentes. Pese a todo, el hecho de que tales ambigüedades existan significa que el proceso de «cuantizar» una teoría clásica dada podría implicar serias cuestiones de elección.

Hay una cuestión relacionada que concierne a la «generalidad» de la elección de coordenadas $x^1, \ldots, x^N$. Recordemos que en §§20.1,2 admitíamos una completa libertad en nuestra elección de coordenadas generalizadas $q^1, \ldots, q^N$ en el espacio de configuración $\mathcal{C}$. Podemos preguntar: ¿sigue estando permitida esta completa libertad cuando pasamos a la teoría cuántica? De hecho, la respuesta es «no», si estamos esperando que el momento conjugado clásico p_a, de cada q^a, tenga que ser «cuantizado» simplemente como $i\hbar\partial/\partial q^a$. La cuestión es muy delicada, y nos lleva a un área fascinante conocida como *cuantización geométrica*.[4] Tiene particular importancia en relación con la relatividad general, ya nos propongamos «cuantizar el campo gravitatorio», o meramente discutir campos cuánticos en un fondo espaciotemporal curvo. (Volveré a la cuestión de la teoría cuántica en fondos curvos en §30.4.) Sin embargo, existen muchas situaciones estándar en las que podemos arreglárnoslas con coordenadas que son más generales que las planas, siempre que seamos cuidadosos. En particular, es útil usar coordenadas angulares, y los momentos conjugados son entonces momentos angulares. Más adelante nos centraremos en el momento angular (§22.8, y §22.12 para el caso relativista).

21.3. La ecuación de Schrödinger

Ignoremos estas cuestiones de orden de factores y coordenadas generalizadas, etc., al menos por el momento, y supongamos que tenemos un hamiltoniano mecanocuántico con el que estamos satisfechos. ¿Qué uso vamos a darle? La respuesta es que desempeña un papel crucial en esa ecuación, fundamental para nuestra comprensión de cómo evoluciona un sistema cuántico con el tiempo, conocida como la *ecuación de Schrödinger*. De hecho, la forma de esta ecuación ya está efectivamente determinada por las reglas establecidas antes. ¿Cómo funciona esto? En primer lugar, debemos hacer visible la función «invisible» ψ que estaba oculta en el extremo derecho de todas nuestras relaciones de conmutación. Ahora el hamiltoniano es un operador después de todo, debido a todas esas $\partial/\partial x$, y necesita algo (al menos potencialmente) a su derecha sobre lo que operar. Puesto que la ecuación de Schrödinger, al ser una ecuación de evolución temporal, va a hacer que ψ varíe con el tiempo, tenemos que escribir ψ como una función de t, además de todas nuestras x^a espaciales:

$$\psi = \psi(x^1, \ldots, x^N; t).$$

Pero esta función no puede depender de los p_a, porque estas cantidades *no* son ahora «variables independientes», sino que deben interpretarse como diferenciaciones con respecto a las x^a. Semejante función ψ se denomina una *función de onda*. Proporciona el *estado cuántico* del sistema. A su debido tiempo consideraremos la interpretación física de las funciones de onda.

¿Cómo encaja en esto la derivada con respecto a t? Aquí es donde interviene la notable evolución temporal de Schrödinger. Recordemos de §20.2 que el hamiltoniano (clásico independiente del tiempo) representa la energía total del sistema. Tomemos también nota del hecho —que se ha sugerido antes (en §21.2)— de que si nuestra teoría cuántica tiene que satisfacer los requisitos de la relatividad, la regla cuántica $p_a = i\hbar\partial/x^a$ (para una única partícula) debería extenderse a la componente $a = 0$ tanto como a las tres componentes espaciales (véase §18.7). En consecuencia, en el procedimiento de «cuantización» la energía debería quedar reemplazada por la derivada con respecto al

tiempo ($E = i\hbar\partial/\partial t$). Lo que expresa la *ecuación de Schrödinger* es precisamente este «papel cuántico» de la interpretación del hamiltoniano como energía total:

$$i\hbar\frac{\partial\psi}{\partial t} = \mathcal{H}\psi,$$

donde

$$\mathcal{H} = \mathcal{H}\left(i\hbar\frac{\partial}{\partial x^1}, \ldots, i\hbar\frac{\partial}{\partial x^N}; x^1, \ldots, x^N\right).$$

A modo de ejemplo sencillo, utilizando el caso particular de un hamiltoniano cuántico dado antes en §21.2, podemos escribir la ecuación de Schrödinger para una única partícula de masa m que se mueve en un campo eléctrico cuya contribución a la energía es $V = V(x, y, z)$:[21.5],[21.6]

$$i\hbar\frac{\partial\psi}{\partial t} = -\frac{\hbar^2}{2m}\nabla^2\psi + V\psi.$$

Por supuesto, todo este reemplazamiento de momento y energía por operadores diferenciales se parece mucho a un truco matemático, y podríamos preguntarnos qué tienen que ver estas diversiones con el momento impartido por el puño de un boxeador o por el swing de un jugador de golf. ¡Es una buena pregunta! Pero según la mecánica cuántica, tiene todo que ver. La clave está en que el momento se conserva, y el efecto de un golpe en quien lo recibe es simplemente un resultado de la inevitabilidad de dicha conservación. El momento debe ir a alguna parte; no puede desaparecer simplemente, porque se conserva. Lo mismo se aplica a la energía.

Pero, por supuesto, ya tenemos conservación del momento y conservación de la energía en la teoría hamiltoniana clásica, podría decir el lector escéptico, así que ¿por qué la necesidad de esta curiosa identifi-

[21.5] Resuelva esta ecuación de Schrödinger explícitamente en el caso de una partícula de masa m en un campo gravitatorio newtoniano constante: $V = mgz$. (Aquí z es la altura sobre la superficie de la Tierra y g es la aceleración de la gravedad.)

[21.6] Transformando al sistema en caída libre con coordenadsa $X = x$, $Y = y$, $Z = z - \frac{1}{2}t^2g$, $T = t$, demuestre que la ecuación de Schrödinger del ejercicio [21.5] transforma a uno sin un campo gravitatorio, con función de onda $\Psi = e^{i(1/6m\hbar^2g^2 + mtzg)}\psi$. ¿Qué nos dice esto sobre el principio de equivalencia de Einstein (véase §17.4), aplicado a sistemas cuánticos? (Tome nota de §21.9.)

cación entre una magnitud física y un operador diferencial incorpóreo, por mucho bien que haga?[21.7] Para intentar responder a esto, y para hacer más plausible la propuesta, hay que apelar al experimento. (De otra forma, ¡una propuesta de aspecto tan ridículo difícilmente tendría mucho peso!) Ahora no entraré en cuestiones experimentales detalladas, pero el punto esencial que emerge de esa enorme cantidad de evidencia experimental es que existe una asociación directa entre frecuencia y energía, y una asociación correspondiente entre número de onda (el recíproco de la longitud de onda) y momento; más aún, estas asociaciones parecen ser universales en todos los fenómenos. En §21.5 veremos la relevancia que tiene esto para «$p_a = i\hbar\partial/\partial x^a$». Mientras tanto, echemos una ojeada a algunas de las razones experimentales para creer que energía y momento tienen este tipo de asociación «ondulatoria».

21.4. La base experimental de la teoría cuántica

Quizá una de las manifestaciones más directas de este tipo de asociaciones se da en los materiales cristalinos. En tales estructuras tenemos una periodicidad espacial en las disposiciones atómicas del cristal. Como fue demostrado por primera vez en un famoso experimento por C. J. Davisson y L. H. Germer (1927), si se lanzan contra un material semejante electrones con una elección apropiada del 3-momento inicial, entonces los electrones son desviados (o reflejados) por el material en ciertos ángulos muy especiales. Una consecuencia manifiesta de estos resultados experimentales es que existe una clara relación de proporcionalidad inversa entre los 3-momentos de los electrones y una distancia de desplazamiento periódico; véase la Fig. 21.2. Lo mismo sucede con otros tipos de partículas. El resultado es que una partícula de momento p parece ser un objeto periódico, como una onda, para el

[21.7] Demuestre que, si el hamiltoniano cuántico $\mathcal{H}$ tiene invariancia traslacional, por ejemplo por ser independiente de la variable de posición x^3, entonces el momento correspondiente p_3 se conserva en el sentido de que el operador p_3 conmuta con la evolución temporal $\partial/\partial t$. A la luz de las interpretaciones que se dan más adelante, explique por qué esta conmutación implica conservación.

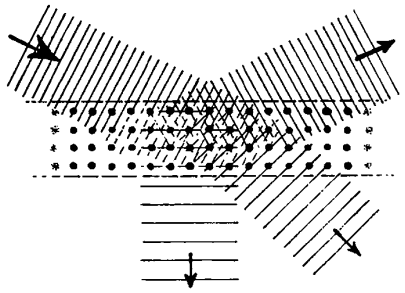


Fig. 21.2. El experimento de Davisson-Germer. Un haz de electrones de 3-momento **p** incide en un material de estructura cristalina periódica. Hay dispersión o reflexión cuando la pauta de los átomos encaja con la de los electrones, siendo estos considerados como ondas con una longitud de onda λ relacionada con la magnitud del momento p de acuerdo con $\lambda = h/p$, donde h es la constante de Planck.

que existe una relación universal entre la longitud de onda λ y la magnitud de su momento p, de acuerdo con la fórmula recíproca (que incluye la constante de Planck $h = 2\pi\hbar$)

$$\lambda = hp^{-1} = \frac{2\pi\hbar}{p}.$$

La longitud de onda λ, asociada al momento p de una partícula, se denomina su longitud de onda de *De Broglie*, por el aristócrata y muy intuitivo físico francés príncipe Louis de Broglie, que sugirió por primera vez (1923), que las partículas materiales tienen una naturaleza ondulatoria con una longitud de onda dada por la fórmula anterior. Además, de acuerdo con los requisitos de la relatividad (véase §18.7), la partícula también debería tener una frecuencia ν, dada en términos de su energía E por la fórmula de Planck

$$E = h\nu = 2\pi\hbar\nu,$$

a la que llegaremos enseguida.[21.8] En su propio sistema de reposo la energía de la partícula es la $E = \mu c^2$ de Einstein, donde μ es su masa en reposo, de modo que está asociada de forma natural con la frecuencia $\mu c^2/2\pi\hbar$, es decir, $\mu c^2/h$.

[21.8] Intente ver por qué los requisitos de la relatividad especial permiten que la fórmula $E = h\nu$ de Planck se deduzca de la $p = h\lambda^{-1}$ de la de De Broglie. (*Sugerencia*: Puede suponer que los hiperplanos en $\mathbb{M}$ en los que la onda toma un valor constante son Lorentz-ortogonales a la 4-velocidad de la partícula.)

Consideraciones como esta llevaron a la conclusión de que una partícula ordinaria manifiesta un comportamiento ondulatorio, que guarda con la masa en reposo de la partícula una relación universal determinada por las fórmulas de Planck y De Broglie. Pero en las dos décadas anteriores ya se había establecido una relación recíproca a esta, que mostraba que las entidades que previamente se consideraban puramente ondulatorias —básicamente los campos eléctrico y magnético oscilantes de Maxwell como constituyentes de la luz (véae §19.2)— también tenían que verse como poseedoras de una naturaleza de partícula, compatible una vez más con las fórmulas de Planck y De Broglie. La prueba más convincente de esto se encontraba en el efecto fotoeléctrico, observado por primera vez por Heinrich Hertz en 1887 y cuyos aspectos más enigmáticos —como demostró Philipp Lenard en 1902— fueron magníficamente explicados por Einstein en 1905 utilizando una imagen de la luz como partícula. (Esto es lo que le valió a Einstein el premio Nobel de Física en 1921, y no la teoría de la relatividad.) El efecto fotoeléctrico ocurre cuando luz de frecuencia ν adecuadamente alta incide sobre un material metálico apropiado, haciendo que se emitan electrones. El enigma surge del hecho de que la energía de los electrones emitidos no depende en absoluto de la intensidad de la luz (si su frecuencia ν se toma constante). En una imagen de onda se esperaría que cuanto mayor fuera la intensidad, más energéticos serían los electrones expulsados. No sucede así (aunque salen *más* electrones cuando la intensidad es mayor). Einstein lo explicó sobre la base de considerar la luz como partículas incidentes —ahora llamadas *fotones*— cuya energía individual está dada por la $E = h\nu$ de Planck, y cada electrón expulsado es el resultado del impacto de uno de estos fotones sobre un átomo individual. Einstein utilizó la fórmula de Planck con gran efecto, haciendo varias predicciones que más tarde fueron confirmadas, en particular por el inicialmente escéptico físico estadounidense Robert Millikan en los años anteriores a 1916.

De hecho, esta naturaleza mecanocuántica corpuscular de la luz ya había comenzado a manifestarse algo antes. Fue en 1900, cuando Max Planck lanzó la revolución cuántica. Lo hizo al realizar un notable análisis de la *radiación de cuerpo negro*, que es la radiación electromagnética en equilibrio con su entorno material «negro»,[5] todo ello mantenido a

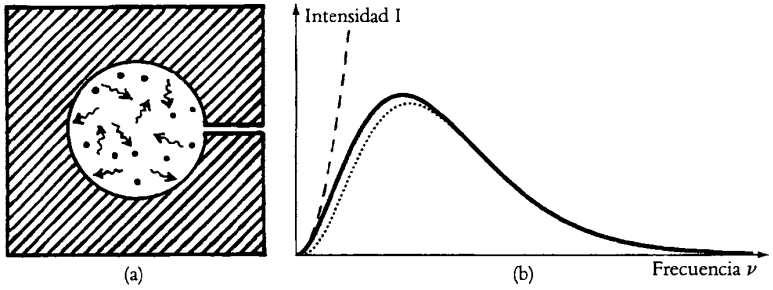


Fig. 21.3. Radiación de cuerpo negro. (a) La cavidad «negra» asegura que la radiación contenida está en equilibrio técnico, a temperatura T, con su entorno caliente. (b) Para una T dada, se encuentra que la intensidad I para cada frecuencia ν es una función específica de ν. La curva continua es la observada, dada por la famosa fórmula de Planck $I = 2h\nu^3/(e^{h\nu/kT} - 1)$ (donde h y k son las constantes de Planck y de Boltzmann, respectivamente). La curva de trazos es la de Rayleigh-Jeans, $I = 2kT\nu^2$, en la que la radiación se trata como ondas clásicas, y se aproxima a la fórmula de Planck para ν pequeña, pero diverge para ν grande. La curva de puntos muestra la ley de Wien $I = 2h\nu^3 e^{-h\nu/kT}$, en la que la radiación se trata como partículas clásicas.

una temperatura específica T (véase la Fig. 21.3a). Planck obtuvo la fórmula (correcta) representada en la Fig. 21.3b

$$\frac{2h\nu^3}{(e^{h\nu/kT} - 1)}$$

para la *densidad de energía* I, en función de la frecuencia ν, donde k es la constante de Boltzmann (relacionada con las unidades en que se mide la temperatura; cf. §27.3).

Resulta que la fórmula de Planck ajusta perfectamente las observaciones. Antes de Planck, la naturaleza del espectro de cuerpo negro había sido un misterio. La imagen enteramente ondulatoria de la radiación electromagnética había llevado a la paradójica fórmula de Rayleigh-Jeans $I = 2kT\nu^2$, aproximada para ν pequeña, pero en la que la intensidad divergiría a infinito para ν grande. Una mejora aparente fue la propuesta de Wien $I = 2h\nu^3 e^{-h\nu/kT}$, aproximada para ν grande, a la que podía darse una justificación sobre la base de tratar la radiación como si fuera un baño de partículas clásicas. La cantidad h, tal como aparecía en la propia fórmula de Planck, fue postulada por él como una *nueva*

constante fundamental de la naturaleza (ahora denominada *constante de Planck*), cuyo valor extraordinariamente minúsculo resulta ser aproximadamente $6{,}62 \times 10^{-34}$ julios/segundo. Para obtener su fórmula, el propio Planck se vio forzado a la idea de que las oscilaciones electromagnéticas podían ser absorbidas o emitidas solo en paquetes de energía específica E, directamente relacionada con la frecuencia ν de la oscilación de acuerdo con la relación anterior

$$E = h\nu,$$

y donde también adoptó un recuento estadístico «extraño» que, sorprendentemente, prefiguraba la estadística de Bose-Einstein (mecanocuánticamente correcta) a la que llegaremos en §23.7.

Aquí el enigma físico era el opuesto al que planteaban los electrones que inciden sobre un cristal, puesto que entonces se pensaba que los efectos electromagnéticos eran simplemente ondas, ¡pero ahora parecían tener propiedades de partículas! Utilizando la forma de Dirac de la constante de Planck, encontramos $E = 2\pi\hbar\nu$, de modo que el período temporal de la oscilación ν^{-1} satisface la fórmula correspondiente a nuestra fórmula anterior ($\lambda = 2\pi\hbar/p$) que relaciona longitud de onda y momento, a saber, $\nu^{-1} = 2\pi\hbar/E$. En la actualidad (siguiendo las ideas posteriores de Einstein, Bose y otros) entendemos que esta relación de Planck se refiere no solo a «oscilaciones del campo electromagnético», sino a «partículas» *reales* —los quanta del electromagnetismo de Maxwell que llamamos fotones—, aunque se necesitaron muchos años para que las ideas que originalmente defendía Einstein en solitario fueran aceptadas. Entre las confirmaciones posteriores, tras estos éxitos con el efecto fotoeléctrico, estaba el experimento crucial de Arthur Compton (1923), que demostró que los fotones, en sus choques con partículas cargadas, se comportaban como partículas sin masa, de acuerdo con la dinámica relativista de §18.7 (véanse §25.4 y la Fig. 25.10). En consecuencia, tanto la energía como el momento están recíprocamente relacionados con un período (de tiempo para la energía y de espacio para el momento), período que está escalado en términos de $2\pi\hbar$.

Una de las razones más convincentes (y mejor conocidas) para que tengamos que enfrentarnos al hecho de que las partículas pueden com-

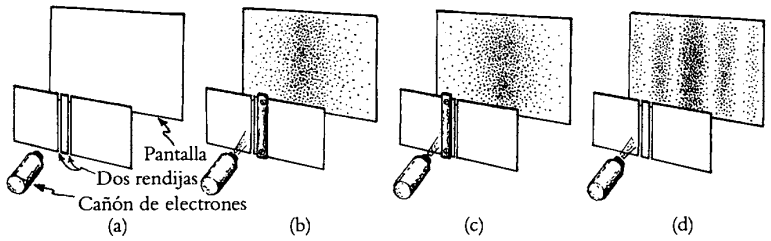


Fig. 21.4. (a) Disposición para el experimento de la doble rendija. Se emite un electrón cada vez que se dirige a la pantalla a través del par de rendijas. (b) Figura en la pantalla cuando la rendija derecha está tapada. (c) Lo mismo cuando la rendija izquierda está tapada. (d) La interferencia ocurre cuando ambas rendijas están abiertas. Algunas regiones en la pantalla no pueden ser alcanzadas ahora, pese al hecho de que pueden serlo cuando solo una o la otra rendija está abierta.

portarse como ondas y las ondas como partículas es el *experimento de la doble rendija*.[6] Aquí tenemos una fuente de partículas y una pantalla detectora, y situada entre la fuente y la pantalla hay una barrera con un par de rendijas paralelas ligeramente separadas. Véase la Fig. 21.4a. Supongamos que la fuente emite partículas de una en una, dirigidas a la pantalla. Si empezamos con una rendija abierta y la otra cerrada, entonces aparecerá en la pantalla una figura aleatoria de puntos que se va formando de uno en uno a medida que las partículas individuales que proceden de la fuente inciden sobre ella. La intensidad de la figura (en el sentido de la máxima densidad de puntos) es más extrema en una franja central, próxima al plano que conecta la fuente con la rendija, como cabe esperar, y decrece uniformemente en ambas direcciones a partir de esta franja central (Fig. 21.4b). Esta figura es efectivamente la misma si el experimento se repite cuando es la otra rendija la que está abierta (Fig. 21.4c). No hay aquí ningún misterio. Pero si el experimento se realiza otra vez cuando ambas rendijas están abiertas, entonces sucede algo extraordinario; véase la Fig. 21.4d. Las partículas siguen produciendo puntos en la pantalla de uno en uno, pero ahora hay una figura de *interferencia* ondulatoria formada por bandas paralelas de intensidad alta y baja alternativamente, y observamos incluso que existen regiones de la pantalla a las que no llega ninguna partícula de la fuente, ¡pese al hecho de que cuando solo una u otra de las rendijas estaban

abiertas, las partículas podían llegar perfectamente a dichas regiones! Aunque los puntos aparecen en la pantalla de uno en uno en posiciones localizadas, y aunque cada incidencia de una partícula que llega a la pantalla puede identificarse con un suceso concreto de emisión de una partícula por la fuente, el comportamiento de la partícula *entre* fuente y pantalla, incluyendo su ambiguo encuentro con las dos rendijas en la barrera, es similar al de una onda, en cuanto que la onda-partícula siente ambas rendijas durante este encuentro. Además —y esta es la cuestión de relevancia particular para nuestros propósitos más inmediatos—, el espaciado entre las bandas en la pantalla nos dice cuál debe ser la longitud de onda de nuestra onda/partícula, y esta longitud de onda λ viene dada realmente en términos del momento p de las partículas precisamente por la misma fórmula que antes, a saber $\lambda = 2\pi\hbar/p$.

21.5. Comprendiendo la dualidad onda-partícula

Quizá sea así, podría decir el escéptico pragmático, ¡pero esto aún no nos obliga a hacer esta identificación aparentemente absurda entre energía-momento y un operador! Realmente no, ¡pero no deberíamos rechazar un *milagro* cuando se nos presenta! ¿Qué milagro es este? El milagro es el hecho de que estos grandes absurdos aparentes en los hechos experimentales —que las ondas son partículas y las partículas son ondas— pueden ser acomodados dentro de un bello formalismo matemático, un formalismo en el que el momento es realmente identificado con la «derivación con respecto a la posición», y la energía, con la «derivación con respecto al tiempo».

¿Cómo nos ayuda este formalismo a entender esta misteriosa dualidad onda-partícula? Para describir nuestra onda-partícula, requeriremos una entidad matemática que pueda proporcionarnos un 4-momento P_a de la partícula claramente definido, y que al mismo tiempo posea una periodicidad espacial y temporal, de tipo ondulatorio, de la cantidad prescrita antes. (Ahora utilizo una *P mayúscula* porque en este caso particular me refiero al valor «clásico» concreto del 4-momento que nuestra partícula pudiera tener. Aún seguimos considerando que el «4-momento cuántico» está descrito por un operador diferencial.) Una

entidad matemática natural sería una función de onda con una dependencia espaciotemporal concreta de la forma (véase §5.3)

$$\psi(x^a) = e^{-iP_a x^a/\hbar}$$

(una onda plana). Esta magnitud se transforma en sí misma si incrementamos $P_a x^a$ en $2\pi\hbar$ (puesto que suma $-2\pi i$ al exponente, de modo que la expresión se multiplica por $e^{-2\pi i} = 1$). Por lo tanto, tiene una periodicidad temporal de $2\pi\hbar/P_0$ y una periodicidad espacial de $2\pi\hbar/P_1$ en la x_1-dirección, y análogamente para las otras direcciones espaciales. Esto concuerda exactamente con los requisitos establecidos antes.

Ahora bien, ¿qué hay de especial en esta entidad particular? Es lo que se denomina una *autofunción* o *función propia* de nuestro operador momento cuántico

$$p_a = i\hbar\frac{\partial}{\partial x^a}.$$

Esto significa que si aplicamos este operador a la $\psi(x^a)$ anterior volvemos a obtener simplemente un múltiplo constante de $\psi(x^a)$ (§6.5):

$$i\hbar\frac{\partial}{\partial x^a}\psi(x^a) = i\hbar\frac{\partial}{\partial x^a}e^{-iP_a x^a/\hbar} = P_a e^{-iP_a x^a/\hbar} = P^a\psi(x^a).$$

Nótese que este multiplicador constante es precisamente el 4-momento P_a (clásico) que exigimos que tenga nuestra entidad. Así pues, cuando $\psi(x^a)$ tiene la forma correcta, a saber, la dada anteriormente, nuestro misterioso momento cuántico $p_a = i\hbar\partial/\partial x^a$ se convierte en un simple momento clásico P_a cuando se aplica a esta ψ:

$$p_a\psi = P_a\psi,$$

pero no lo hace para otro tipo de estados. Decimos que la función ψ anterior tiene un valor definido para el 4-momento, y la llamamos un *estado de momento*. Tenemos que pensar en una partícula libre, con un momento definido P_a clásicamente identificable, que está matemáticamente descrita por esta función de onda particular ψ, que es autofunción del operador cuántico p_a con *autovalor* P_a. Las únicas funciones de onda que tienen un valor definido del momento clásico son aquellas que son autofunciones del operador momento cuántico.

Recordemos que en §13.5 se ha introducido la noción de *autovector* (o *vector propio*) de un operador lineal $\boldsymbol{T}$ que es un vector $\boldsymbol{v}$ para el que $\boldsymbol{T}\boldsymbol{v} = \lambda\boldsymbol{v}$ para una cantidad escalar λ, llamada su *autovalor*. Esta es exactamente la situación que tenemos aquí, donde $i\hbar\partial/\partial x^a$ ocupa el lugar del operador $\boldsymbol{T}$, y donde P_a ocupa el lugar de λ (tomando sucesivamente cada valor concreto de a), excepto que en §13.5 nos referíamos básicamente a espacios vectoriales de *dimensión finita* y sus transformaciones lineales. Aquí nos interesa el espacio vectorial $\mathbf{W}$ de las posibles $\psi(x^a)$, y ese será un espacio vectorial de dimensión *infinita*. (Es un espacio vectorial porque podemos sumar funciones de x^a, y podemos multiplicar funciones de x^a por números, y en ambos casos obtenemos simplemente nuevas funciones de x^a. Es de dimensión infinita porque, por ejemplo, todas las funciones del tipo particular que acabamos de considerar para diferentes valores de P_a, son linealmente independientes.)[21.9]

Las autofunciones (o *autoestados*, como se les llama con frecuencia en mecánica cuántica) desempeñan un papel clave en el formalismo cuántico. En el lenguaje de la mecánica cuántica, los diversos operadores (como $p_a = i\hbar\partial/\partial x^a$ que acabamos de considerar, aunque más tarde consideraremos otros tales como la posición y el momento angular) se denominan *variables dinámicas*. Nuestra función de onda ψ, que inicialmente desempeñaba solo el papel de la «función invisible» que imaginábamos en las sombras, a la derecha de todos nuestros operadores, empieza ahora a desempeñar decididamente un papel activo. La consideramos el *estado* del sistema físico, como se ha mencionado antes. A veces se denomina un *vector de estado* (aunque este es en realidad un término más general, para el que no necesitan aplicarse las descripciones particulares en términos de coordenadas espaciales y temporal que he utilizado para ψ). Como sucede con el caso del 4-momento, que se ha considerado antes, los *autoestados* de una variable dinámica son aquellos estados para los que dicha variable dinámica concreta tiene lo que se denomina un «valor definido», y el autovalor es el «valor» real que tiene la variable dinámica para dicho estado.

Debería hacer una puntualización sobre el hecho de que, hasta este

[21.9] ¿Por qué? Aquí la dependencia lineal puede incluir sumas *continuas*, a saber, integrales.

momento, he estado tratando a nuestro estado de momento de una forma absolutamente 4-dimensional espaciotemporal, compatible con los requisitos de la relatividad especial. Esto es económico, en cuanto que la expresión[21.10]

$$e^{-iP_a x^a/\hbar} = e^{-iEt/\hbar} e^{i\mathbf{P}\cdot\mathbf{x}/\hbar}$$

(con $P_a = (E, -\mathbf{P})$ y $x^a = (t, \mathbf{x})$, como en §18.7) contiene tanto la dependencia *espacial*, que hace de él un autoestado del momento espacial ordinario 3-dimensional

$$\mathbf{p} = (-p_1, -p_2, -p_3) = -i\hbar\left(\frac{\partial}{\partial x^1}, \frac{\partial}{\partial x^2}, \frac{\partial}{\partial x^3}\right)$$

con autovalor $\mathbf{P}$, como la dependencia *temporal*, que hace de él una solución de la ecuación de Schrödinger con autovalor E para la energía. Sin embargo, el formalismo de Schrödinger en conjunto no es un esquema relativista, en cuanto que trata el tiempo de forma diferente a las variables espaciales, de modo que será mejor que en las discusiones que siguen en este capítulo vuelva a descripciones no relativistas.

21.6. ¿Qué es la «realidad» cuántica?

Retrocedamos por un momento y preguntémonos qué está tratando esto de decirnos acerca de la «realidad». ¿Son las variables dinámicas «objetos reales»? ¿Son «reales» los estados? ¿O deberíamos decir que hemos alcanzado la realidad solo cuando hayamos llegado a las magnitudes aparentemente «clásicas» que surgen como autovalores de las variables dinámicas (o de otros operadores)? De hecho, los físicos cuánticos no suelen ser muy claros sobre esta cuestión. La mayoría de ellos se sienten característicamente incómodos al abordar la cuestión de la «realidad». Pueden decir que adoptan lo que llaman una postura «positivista», y rechazan abordar la cuestión de lo que se supone que quiere decir «realidad», considerando que una pregunta semejante es «acientífica». Todo lo que deberíamos pedir a nuestro formalismo, podrían decir, es que dé respuestas a preguntas apropiadas que podamos plantear

[21.10] ¿Por qué puedo separarla de esta manera?

a un sistema, y que dichas respuestas estén de acuerdo con los hechos observados.

Si hay que creer que alguna cosa en el formalismo cuántico es «real», para un sistema cuántico, entonces creo que tiene que ser la función de onda (o vector de estado) que describe la realidad cuántica. (En el capítulo 29 plantearé otras posibilidades; véase también el final de §22.4.) Mi propio punto de vista es que la cuestión de la «realidad» *debe* ser abordada en mecánica cuántica —especialmente si se adopta el punto de vista (como muchos físicos parecen hacer) de que el formalismo cuántico debe aplicarse universalmente a la totalidad de la física—, pues si no hay realidad cuántica, no puede haber realidad en absoluto a ningún nivel (ya que según esta idea todos los niveles son niveles cuánticos). En mi opinión, no tiene sentido negar la realidad en general. Necesitamos una noción de realidad física, incluso aunque solo sea provisional o aproximada, pues sin ella nuestro universo objetivo, y con ello la totalidad de la ciencia, ¡se evapora ante nuestra mirada contemplativa!

Todo esto está muy bien, pero, entonces, ¿qué pasa con el vector de estado? ¿Cuál es la dificultad de tomarlo como representación de la realidad? ¿A qué se debe que los físicos se muestren a menudo muy reacios a adoptar esta posición filosófica? Para entender las dificultades, debemos examinar más cuidadosamente la naturaleza de las funciones de onda y sus interpretaciones físicas.

Para empezar, examinemos más de cerca nuestro estado de momento $\psi = e^{i\mathbf{P}\cdot\mathbf{x}/\hbar}$ (que, por simplicidad, he tomado en el instante $t = 0$). Notemos que de ningún modo está localizado como una partícula ordinaria. Se extiende de manera uniforme sobre la totalidad del universo. Su «magnitud», medida por su módulo $|e^{i\mathbf{P}\cdot\mathbf{x}/\hbar}|$, tiene el mismo valor 1 en todo punto del espacio (§5.1). El lector será perdonado si piensa que esta es una imagen extraña para mantener, en el caso de una única partícula con un momento perfectamente definido en cierta dirección espacial. ¿Qué ha sucedido con nuestra imagen ordinaria de una partícula como algo localizado (al menos aproximadamente) en un solo punto? Bien, podríamos decir que un estado de momento es tan solo una idealización. Aún podemos salir del paso y tener un momento muy bien definido (si no perfectamente definido) si pasamos a estados algo

parecidos conocidos como «paquetes de ondas». Estos están dados por funciones de onda cuyo valor tiene un pico abrupto en cierta posición y son «casi» autofunciones del momento, en un sentido apropiado. En una dimensión tales paquetes de onda pueden presentarse de forma explícita, tomando el producto de un autoestado de momento por una gaussiana e^{-x^2} o, lo que es más apropiado, por una gaussiana general

$$Ae^{-B^2(x-C)^2},$$

(donde A, B y C son constantes reales). Esta es la bien conocida curva «en forma de campana» de la estadística (véase la Fig. 27.5 en §27.4 para una ilustración de la misma), donde, en la expresión anterior, su «pico» está centrado en el punto $x = C$. Es de cierto interés (y una ventaja para el cálculo) que el paquete de ondas obtenido al tomar dicho producto puede expresarse sucintamente permitiendo que en la expresión anterior C sea un *número complejo*.[21.11] De la misma forma pueden construirse paquetes de onda en las tres dimensiones del espacio utilizando una gaussiana $Ae^{-B^2(x^2+y^2+z^2)}$, con su pico desplazado en una dirección compleja. En cada caso B^{-1} proporciona una medida de su dispersión. De hecho, existe un teorema, subyacente en lo que se denomina «principio de incertidumbre de Heisenberg», que nos dice que existe un límite absoluto para cuán pequeña puede ser su dispersión en relación con cuán próximo está a un «casi» estado de momento. Veremos esto de forma más explícita en §21.11.

Ahora intentemos obtener una imagen mejor de cómo son realmente los estados de momento y los paquetes de ondas. Debemos tener en cuenta que una función de onda es una función con valor complejo, y su carácter «ondulatorio» no debe verse necesariamente como una oscilación en su magnitud (o intensidad). En el caso de un estado de momento, es el *argumento* de la función de onda (§5.1), a saber $-P_a x^a/\hbar$, medido alrededor de un círculo —i.e., $e^{-iP_a x^a/\hbar}$ tomado sobre el círculo unidad en el plano complejo—, el que tiene un carácter ondulatorio. En teoría cuántica solemos referirnos al argumento del valor de la función de onda como su *fase*. Encontramos que la fase no es tanto

[21.11] Reemplazando el número real C en la expresión que se ha mostrado antes por el número complejo $C + iD$ (donde C y D son reales), encuentre la frecuencia del paquete de ondas y la localización de su pico.

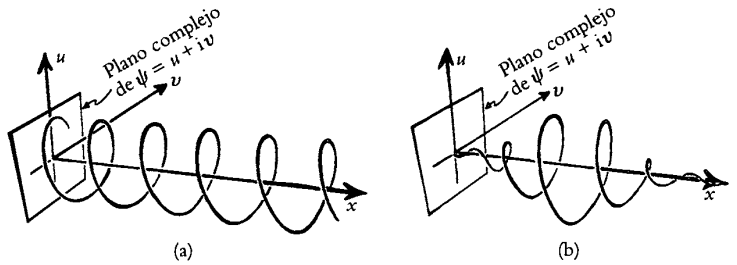


Fig. 21.5. Una función de onda de una partícula: ψ es una función compleja de la posición x. (a) Estado de momento $e^{-iPx/\hbar}$, representado como un sacacorchos (autofunción de momento p). (b) Un paquete de ondas $e^{-A^2x^2}e^{-iPx/\hbar}$.

«ondulatoria», sino que «da una y otra vuelta». En la Fig. 21.5a he tratado de mostrar este comportamiento de la función de onda en una dirección concreta, representando dicha dirección en un ángulo inclinado hacia la derecha (eje x de la figura), y tomando un plano perpendicular a dicha dirección (los restantes ejes u e v en la figura) para representar el plano complejo de los valores que puede tomar la función ψ (de modo que la figura representa $\psi = u + iv$ en dicho plano). Así pues, la dirección x de mi figura corresponde a cierta dirección real en el espacio ordinario, pero las direcciones u y v no son direcciones espaciales ordinarias, sino que representan el plano complejo de valores posibles de la función de onda. Notemos que para nuestro estado de momento la función de onda es una especie de sacacorchos (que es dextrógiro para momento positivo en la dirección espacial representada por la dirección x de nuestra figura). En la Fig. 21.5b he dibujado la imagen correspondiente para un paquete de ondas. Se parece a un sacacorchos en un tramo (de modo que tiene un momento moderadamente bien definido), pero luego su «sacacorcheidad» se desvanece en ambas direcciones, y la función de onda se hace muy pequeña fuera de cierto intervalo.

Por supuesto, para obtener la imagen *completa* de estas ondas tendríamos que tratar de imaginar que esto ocurre en las tres direcciones del espacio a la vez, lo que es algo difícil porque ¡necesitaríamos dos dimensiones extra (cinco en total) para encajar el plano complejo, además de las direcciones espaciales! Pero las cosas no están tan mal, en el

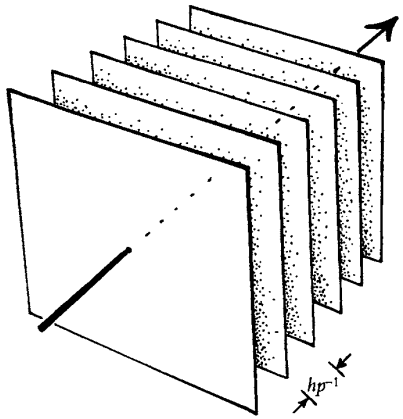


Fig. 21.6. Planos de fase dada, para un autoestado de momento con espaciado hp^{-1}, donde p es la magnitud del 3-momento espacial. (Compárese con la Fig. 21.2.)

caso del estado de momento, si pensamos solo en *planos de fase constante*. Estos son planos paralelos perpendiculares a la dirección del momento, con un espaciado entre cada plano y el siguiente de $2\pi\hbar/p$, donde p es la magnitud del 3-momento (espacial). Véase la Fig. 21.6. Este tipo de descripción es útil para considerar cosas tales como la imagen de la función de onda de un fotón que incide sobre un cristal, como se ha hecho en la Fig. 21.2. También se puede utilizar esta descripción en el caso del experimento de la doble rendija si consideramos que las rendijas están a una gran distancia de la pantalla; entonces es posible pensar que la función de onda de cada partícula, a medida que se aproxima a cierta región localizada de la pantalla, está compuesta de la suma de dos partes, cada una de las cuales está muy próxima a un estado de momento (siendo esencialmente una onda plana de una única frecuencia, debido a la gran distancia de las rendijas a la pantalla), pero la dirección de cada una de las dos partes componentes es ligeramente diferente. En algunos lugares de la pantalla las dos ondas se reforzarán mutuamente, mientras que en otros lugares se cancelarán, dando lugar a las bandas de mayor y menor intensidad que he descrito antes (en la Fig. 21.4d). Podemos ver esta geometría en la Fig. 21.7, donde los planos representan regiones del espacio en las que cada onda componente tiene un valor constante para su fase. En la función de onda completa estas dos partes componentes deben sumarse de modo que, suponiendo que ambas partes por separado tienen la misma in-

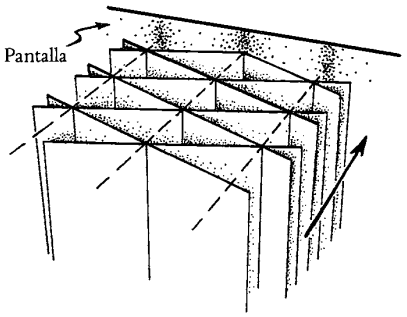


Fig. 21.7. La función de onda de un electrón que se acerca a la pantalla de la Fig. 21.4 en el experimento de la doble rendija puede considerarse como una superposición de dos de las ondas planas de la Fig. 21.6, que forman un pequeño ángulo entre sí. Donde las fases coinciden (a lo largo de las líneas de trazos), las dos se refuerzan mutuamente, dando lugar a la máxima probabilidad de llegada en la pantalla. A medio camino entre estos máximos, las fases son opuestas y las ondas se cancelan, dando una probabilidad cero de que el electrón llegue a la pantalla.

tensidad, se cancelarán en lugares donde estén en oposición de fase y se reforzarán cuando estén en fase. Esto nos da las bandas de intensidad alterna que se observan realmente en el experimento de la doble rendija.

Sí, sí, podrá interrumpir el lector impaciente, pero así es precisamente como se comportan las *ondas*. ¡No me he enfrentado al hecho de que nuestras ondas-partículas son ondas-*partículas*! Aparte del embellecimiento aparentemente mucho menor que supone el que mis ondas son ondas complejas, todo lo que he hecho es describir un tipo de interferencia que podría darse con las ondas de agua en el mar, o las ondas sonoras, o las ondas de Maxwell construidas a partir del campo electromagnético clásico (ondas de radio, luz visible, rayos X, etc.). Pero la cuestión general del experimento de la doble rendija —y en eso se supone que yo estaba insistiendo— es que el experimento muestra un conflicto entre la imagen ondulatoria y la imagen corpuscular. Así es; la manifestación más obvia de la naturaleza corpuscular, en este experimento, ¡ocurre cuando estos pequeños amigos dejan sus marcas minúsculas en la pantalla: de uno en uno...!

21.7. La naturaleza «holística» de una función de onda

Aquí habría que resaltar algo. Uno podría imaginar que ocasionalmente se produce un punto pequeño en la pantalla, cuando la intensidad local de la onda alcanza cierto valor crítico o, más bien, que existe cierta probabilidad de que aparezca un pequeño punto en la pantalla, probabilidad que aumenta a medida que aumenta la intensidad de la onda. ¡Buen intento! Pero, tal como he formulado antes, el experimento de la doble rendija (en su forma idealizada), sencillamente no funciona. En efecto, si se tratara tan solo de probabilidades individuales en lugares individuales, cabría esperar que a veces aparecerían dos puntos en la pantalla, en posiciones ampliamente separadas donde la intensidad es apreciable, aunque solo hay una función de onda para describir la emisión de una única partícula en la fuente. La dificultad se hace más manifiesta si imaginamos que nuestras partículas son partículas cargadas, tales como los electrones; pues si la emisión de un único electrón en la fuente pudiera dar como resultado un par de electrones que llegan a la pantalla, incluso si solo se produjera ocasionalmente, entonces tendríamos una violación de la ley de la conservación de la carga. Lo mismo se aplicaría a cualquier otro número cuántico conservado para una partícula, tal como la conservación del número bariónico (§25.6), por ejemplo, si utilizáramos neutrones.[21.12] Semejante comportamiento sin conservación estaría en flagrante contradicción con una enorme masa de evidencia experimental. Pese a todo, ¡tanto los electrones como los neutrones muestran el tipo de autointerferencia en un experimento de doble rendija tal como acabo de describir!

¡Así que hemos llegado exactamente a *ninguna parte* en la comprensión de ondas-partículas!, objetará seguramente algún lector airado, con impaciencia cada vez más justificada. Pero conténganse, por favor, no hemos acabado con la interpretación de nuestras funciones de onda. Tenemos que pensar que la onda entera describe (o «es») solo

[21.12] Demuestre que la probabilidad de semejantes apariciones de puntos dobles, de acuerdo con tal imagen, debe ser muy apreciable, cualquiera que pudiera ser la ley de probabilidad para la aparición de puntos en términos de la intensidad de la función de onda. *Sugerencia*: Divida la pantalla en dos partes, con la misma probabilidad de aparición de puntos en cada una de ellas.

una única partícula. Aunque en un sentido definido determina la probabilidad de que se produzca un punto en los diversos lugares de la pantalla, esta probabilidad se refiere solo a una partícula. Esta interpretación no funcionará si consideramos la función de onda de un modo local, como algo que proporciona independientemente una probabilidad de formación de un punto en cada lugar separado de la pantalla. Debemos considerar la función de onda como un objeto entero. Si hace que un punto aparezca en un lugar, entonces ha realizado su trabajo, y este aparente acto de creación le prohíbe hacer que un punto aparezca también en algún otro lugar. Las funciones de onda son completamente diferentes de las ondas clásicas en este aspecto importante. No hay que pensar en las diferentes partes de la onda como perturbaciones locales, portadoras cada una de ellas de lo que va a suceder en una región remota. Las funciones de onda tienen un fuerte carácter no local; en este sentido, son entidades holísticas por completo.

Este punto puede establecerse de forma más clara en una situación experimental algo diferente. Esto tiene la ventaja adicional de aclararnos que la imagen de paquete de ondas de una onda-partícula es, por sí misma, completamente inadecuada para explicar el comportamiento cuántico tipo partícula. Imaginemos que hay una fuente de partículas, igual que antes, y supongamos que solo emite una única partícula. En lugar de utilizar una barrera con un par de rendijas, supongamos que hay lo que se denomina un *divisor de haz* en el camino de la partícula. Nos ayudará considerar que la partícula es un fotón, y podemos imaginar que el divisor de haz es un tipo de «espejo semiplateado».[7] Imaginemos que nuestro «experimento» se lleva a cabo en el espacio interestelar (y el lector debería ser advertido de que no estoy proponiendo nada remotamente factible; nuestro ejemplo solo servirá para mostrar algunas predicciones muy básicas de la mecánica cuántica en circunstancias tan extremas). Si así lo decidimos, podemos imaginar que la función de onda del fotón parte de la fuente en la forma de un pequeño paquete de ondas pero, después de incidir en el divisor de haz, se dividirá en dos, con una parte del paquete de ondas reflejada y la otra parte del paquete de ondas transmitida por el divisor de haz, por ejemplo en direcciones perpendiculares (Fig. 21.8). La función de onda

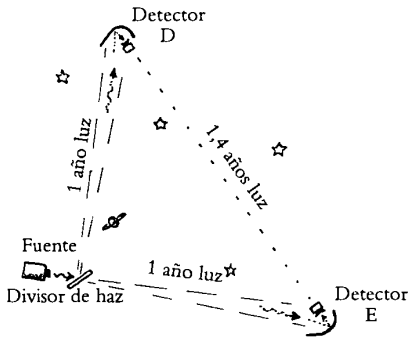


Fig. 21.8. Un experimento espacial hipotético que ilustra la naturaleza no local de una función de onda bajo medida. La función de onda del fotón parte de la fuente como un pequeño paquete de ondas, pero se divide en dos después de incidir en el divisor de haz, para llegar, al cabo de un año, a los detectores D y E a una distancia de un año luz. Pero solo uno de D o E puede registrar el fotón.

completa es la suma de estas dos partes. Si quisiéramos, podríamos esperar un año antes de decidir interceptar la función de onda del fotón con una placa fotográfica u otro tipo de detector. Las dos partes estarán entonces a una gran distancia la una de la otra, pero imaginemos que tengo dos colegas (en dos diferentes laboratorios espaciales), separados por una distancia de más de 1,4 años luz. Cada uno de ellos tiene un detector independiente, y aunque cada una de las dos partes del paquete de ondas puede estar ahora considerablemente dispersa, cada colega tiene un gran espejo reflectante parabólico que recoge el paquete de ondas disperso y lo concentra en el detector de ese colega concreto. ¿Qué dice la mecánica cuántica que sucederá? Dice que uno u otro de mis colegas detectará realmente el fotón, pero no pueden hacerlo *ambos*. Este no es el tipo de cosas que hace una onda clásica. Recordemos que mis dos colegas están a más de 1,4 años luz de distancia. La relatividad afirma que ninguna señal puede transmitirse entre ellos en menos de 1,4 años luz; pese a todo, el hecho es que una parte del paquete de onda que da un fotón impide que lo haga la otra, a 1,4 años luz de distancia, y viceversa. Al cabo de solo un año, sé por cada uno de ellos lo que ha sucedido, y veo que solo uno de ellos ha recibido un fotón. ¡Parece que la parte de la función de onda a la que tiene acceso cada uno de los colegas «sabe» lo que va a hacer la otra parte! Cada vez que realizo este experimento, veo que uno u otro colega recibe el fotón, pero no ambos. Ningún efecto ondulatorio de tipo clásico conseguiría esta aparente «comunicación instantánea» entre las dos partes de

la función de onda. Las funciones de onda cuánticas son simplemente *diferentes* de las ondas clásicas.

Pese a todo, quizá el lector escéptico no esté aún convencido, pues tal comunicación no sería necesaria si el fotón hace simplemente su elección acerca de seguir un camino u otro *en el divisor de haz*. Absolutamente cierto. Lo que el anterior montaje experimental está mostrando es el aspecto corpuscular de un fotón. Si el fotón permaneciera localizado y fuera corpuscular, entonces su decisión acerca del camino a seguir tendría que tomarse en el divisor de haz. (¡Una partícula localizada no puede extenderse sobre intervalos de años luz!) Si todos los experimentos a los que pueden someterse los fotones fueran como este, entonces las funciones de onda no serían necesarias. Pero se pueden realizar otros experimentos sobre el fotón una vez que sale del divisor de haz. ¿Cómo puede saber nuestro pobre fotón, cuando está a punto de salir, que mis colegas no tienen planeado para él un destino diferente? Supongamos que, en lugar de tratar cada uno de ellos de detectar individualmente el fotón, han maquinado el siguiente plan: reflejarán por separado sus partes de la función de onda hacia un cuarto lugar, donde las dos partes reflejadas incidirían al cabo de un año, pongamos por caso, sobre un segundo divisor de haz (Fig. 21.9). Allí, cada parte del paquete de onda que llega se dividiría en dos, de modo que una mitad emerge de este divisor de haz en una dirección para incidir en un detector A, y la otra mitad emerge en otra dirección para ir a parar a otro detector B. (Esto se aplica separadamente a cada una de las dos partes del paquete de ondas, procedentes de las vecindades separadas de cada uno de mis dos colegas.) Si se fijan adecuadamente las longitudes de todas las trayectorias (por ejemplo, todas iguales), entonces observamos que el fotón emergente solo puede activar uno de los detectores, digamos el A, y no el B, debido a la interferencia constructiva entre las dos partes de la función de onda en A y la interferencia destructiva en B.

Ninguna imagen puramente corpuscular de un fotón puede conseguir esto. La función de onda es ahora decididamente necesaria para explicar el aspecto onda de la dualidad onda-partícula. Si el fotón hubiera hecho ya su elección respecto de hacia cuál de mis colegas va a viajar tras dejar el primer divisor de haz, entonces la otra ruta sería

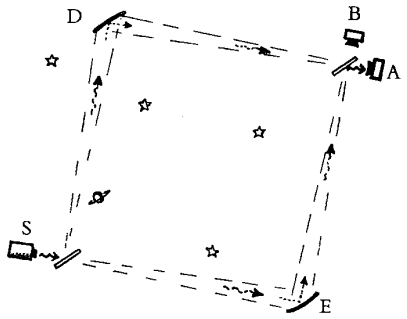


Fig. 21.9. Interferómetro de Mach-Zehnder a una escala interestelar. ¿Cómo puede saber el fotón, tras salir del primer divisor de haz, que, en lugar de la configuración de la Fig. 21.8, los espejos en D y E reflejan las porciones de la función de onda a un segundo divisor de haz? Tras este encuentro, solo el detector A puede recibir el fotón.

irrelevante. En ese caso, cuando el fotón llega finalmente al segundo divisor de haz procede de una sola dirección, y podría seguir cualquiera de los dos caminos para llegar *o bien a* A, *o bien a* B. No hay ahora ninguna posibilidad para la necesaria interferencia destructiva que le impida alcanzar el detector B. Puesto que A es siempre el detector que registra, debe darse el caso de que el fotón no haya tomado su decisión cuando deja el primer divisor de haz. Es necesario que el fotón experimente simultáneamente las dos rutas alternativas que podría tomar en su paso del primero al segundo divisor de haz.[8]

Por supuesto, he sido demasiado fantasioso en las situaciones concretas de escala astronómica que acabo de describir. ¡Es evidente que todavía no se ha llevado a cabo ningún experimento cuántico con una línea de base parecida! Por el contrario, sí se han realizado a menudo versiones basadas en tierra de este tipo de experimentos (el último de los cuales se denomina un interferómetro de Mach-Zehnder), con longitudes de brazo de quizá metros en lugar de años luz, y nunca han quedado contradichas las predicciones de la mecánica cuántica. El enigma clave es que parece que un fotón (u otra partícula cuántica) tiene que «saber» de algún modo qué tipo de experimento se va a realizar con él mucho antes de que se produzca la realización de dicho experimento. ¿Cómo puede saber por adelantado si debe colocarse en «modo partícula» o en «modo onda» cuando deja el (primer) divisor de haz?

La teoría cuántica trabaja de forma que no da a la partícula ninguna «capacidad de previsión» semejante, sino que acepta simplemente el carácter holístico no local de una función de onda. En los dos experimentos anteriores consideramos que la función de onda se divide en dos partes en el divisor de haz inicial, y el aspecto corpuscular de la onda-partícula solo se manifiesta en el detector, cuando se realiza finalmente la medida. La medida pone de manifiesto el carácter holístico de la función de onda, en el sentido de que la partícula siempre se manifiesta en un solo lugar, de modo que su aparición en una posición prohíbe su aparición simultánea en cualquier otra.

21.8. Los misteriosos «saltos cuánticos»

Pero ahora surge otra pregunta: ¿cómo sabemos qué circunstancia física constituye una medida? ¿Por qué, una vez que hemos utilizado felizmente esta descripción de una partícula como una onda que se dispersa, deberíamos volver repentinamente a una descripción de la misma como una partícula localizada en cuanto se lleva a cabo la detección? Esta misma imagen extraña de una partícula cuántica parece también apropiada para la detección en la pantalla en nuestro experimento de la doble rendija, como sucedía con los «detectores» (no especificados) utilizados por mis lejanos colegas. En las descripciones que he dado hasta ahora parece ciertamente que hay que mantener los aspectos tipo onda hasta que decidimos «realizar una medida» para detectar la partícula, pero luego volvemos repentinamente a una descripción tipo partícula, en la que se produce un complicado cambio discontinuo (y no local) del estado —un *salto cuántico*— cuando pasamos de la imagen de función de onda a la «realidad» presentada por la medida. ¿Por qué? ¿Qué hay en el proceso de detección que exige que con ocasión de una «medida» debería adoptarse un procedimiento matemático diferente (y altamente no local) del procedimiento estándar de evolución cuántica que es la evolución de Schrödinger?

Más adelante, en los capítulos 23, 29 y 30, trataré de abordar esta cuestión enigmática con más profundidad. Pero incluso si aceptamos que, al menos en el nivel de descripción matemática formal, debemos adoptar este curioso procedimiento de «saltos», queda la cuestión de lo que nos dice ello acerca de la «realidad» de la función de onda. Este «salto» del estado cuántico —un proceso que no parece explicable por

ninguna evolución continua de acuerdo con la ecuación de Schrödinger— es lo que lleva a muchos físicos a dudar de que la evolución del vector de estado pueda considerarse seriamente como una descripción adecuada de la realidad física. El propio Schrödinger se sentía extraordinariamente incómodo con los «saltos cuánticos», y en cierta ocasión comentó en una conversación con Niels Bohr:[9]

> Si todos estos condenados saltos cuánticos fueran a quedarse realmente, entonces lamentaría haber tenido algo que ver con la mecánica cuántica.

Por el momento aceptemos, al menos como un modelo matemático del mundo cuántico, esta curiosa descripción según la cual el estado cuántico evoluciona durante un tiempo en forma de una función de onda que se extiende normalmente por el espacio (pero que quizá se concentra de nuevo en una región más localizada); pero luego, cuando se realiza una medida, el estado colapsa a algo localizado y específico. Esta localización instantánea sucede con independencia de cuán dispersa pudiera estar la función de onda antes de la medida, y a partir de entonces el estado evoluciona de nuevo como una onda guiada por Schrödinger, partiendo de esta configuración localizada específica y dispersándose de nuevo normalmente hasta que se realiza la siguiente medida. A partir de las situaciones experimentales (y «experimentos mentales») anteriores, se podría tener la impresión de que los aspectos corpusculares de una onda-partícula son los que se manifiestan en una medida, mientras que son los aspectos tipo onda los que se manifiestan entre medida y medida.

Esto no está tan lejos de la verdad de lo que nos dice la mecánica cuántica, pero los dos aspectos onda-partícula no están en absoluto tan delimitados. Aunque algunos físicos han aceptado la idea de que todas las medidas son en definitiva medidas de posición,[10] considero esta perspectiva demasiado estrecha de miras. De hecho, la forma en que se presenta normalmente el formalismo cuántico no requiere que las medidas sean medidas de posición. Por ejemplo, la medida del momento de una partícula (o, digamos, su momento angular respecto a cierto eje) constituiría una medida tan buena como lo sería una medida de posición. En §21.11 expondré la relación entre medidas de posición y de momento, pero dejaré para el próximo capítulo la cuestión general del tratamiento de las medidas que hace el formalismo cuántico. Observaremos que la descripción matemática de la medida física de un sistema cuántico es algo muy diferente de una evolución cuántica (de Schrödinger). Las cuestiones controvertidas que surgen de este curioso hecho se analizarán más adelante, especialmente en el capítulo 29.

21.9. Distribución de probabilidad en una función de onda

Abordemos primero la cuestión más limitada de lo que se supone que nos está diciendo la función de onda ψ acerca de la posición de la partícula. Las reglas de la teoría cuántica nos dicen que el *módulo al cuadrado* de ψ, es decir, $|\psi|^2 (= \bar{\psi}\psi$; véase §10.1), debe interpretarse como la distribución de probabilidad de que una medida de posición encuentre la partícula en las diferentes localizaciones espaciales posibles. Así pues, allí donde la función de onda es máxima en valor absoluto existe la máxima probabilidad de encontrar la partícula; y allí donde la función es nula, la partícula no será encontrada. Ahora bien, la probabilidad total de encontrar la partícula en cualquier parte del espacio tiene que ser 1 y, por consiguiente, la integral de $|\psi|^2$ extendida a todo el espacio,[11] i.e.,

$$\|\psi\| = \int_{E^3} |\psi(\mathbf{x})|^2 \mathrm{d}x^1 \wedge \mathrm{d}x^2 \wedge \mathrm{d}x^3,$$

es igual a 1:

$$\|\psi\| = 1.$$

Decimos que la función de onda ψ está *normalizada* si se satisface esta condición.

Este requisito de normalización tiene la irritante consecuencia de que descarta las funciones de onda «estados de momento» $\psi = \mathrm{e}^{\mathrm{i}\mathbf{P}\cdot\mathbf{x}/\hbar}$ con las que hemos empezado, puesto que para ellas $|\psi|^2 = 1$ en toda la infinitud del espacio, de modo que la integral anterior (que es igual al volumen total del espacio) *diverge*. Así pues, tenemos que considerar los

estados de momento como idealizaciones irrealizables. Por otra parte, podemos hacer la vida un poco más fácil para los estados de momento si adoptamos una actitud algo más relajada con respecto a las funciones de onda. Podemos seguir llamando a ψ «función de onda» incluso si no satisface esta condición de normalización, pero la llamamos una *función de onda normalizada* si lo hace.

Una función de onda ψ será *normalizable* si la integral que define a $\|\psi\|$ converge. En este caso podemos dividir ψ por la raíz cuadrada de $\|\psi\|$ para obtener la función de onda normalizada: $\psi\|\psi\|^{-1/2}$. Solo las funciones de onda normalizables tienen posibilidad de ser realizadas físicamente. Las otras (tales como los estados de momento) representan idealizaciones físicas. El espacio vectorial complejo de las funciones de onda (no necesariamente normalizadas) es nuestro espacio de estados $\mathbf{W}$. Tendré que admitir también que algunas de nuestras funciones de onda podrían ser en realidad hiperfuncionales (§9.7), y la razón para esto se hará evidente enseguida.

Con respecto a la interpretación física (admitiendo esta actitud más relajada), consideramos que si se multiplica ψ por un número complejo constante, distinto de cero, sigue representando la misma situación física que antes. En cualquier caso, es normal en la teoría cuántica considerar que ψ y $e^{i\theta}\psi$, donde θ es una constante real, son físicamente equivalentes. En otras palabras, multiplicar la función de onda por una fase constante no supone ninguna diferencia para el estado físico. (Evidentemente, esto no afecta al valor de $|\psi(\boldsymbol{x})|^2$.) Resulta razonable llevar esto más lejos y permitir la multiplicación por una constante compleja no nula κ y seguir considerando equivalentes las funciones de onda:

$$\psi \equiv \kappa\psi.$$

(Evidentemente, la ecuación de Schrödinger tampoco se ve alterada por este reemplazamiento.) Esto equivale a pasar del espacio vectorial complejo $\mathbf{W}$ de las funciones de onda a su espacio proyectivo $\mathbb{P}\mathbf{W}$ de «estados físicos» idealizados. (Véase §15.6 para la noción de espacio proyectivo.)[12] Por supuesto, el reescalado general constante $\psi \mapsto \kappa\psi$ no conserva $|\psi|^2$, así que tenemos que reinterpretar la densidad de probabilidad para la localización de la partícula, de modo que sea aplicable

cuando ψ no esté normalizado. Hacemos esto ofreciendo la regla revisada según la cual la densidad de probabilidad se obtiene tomando $|\psi|^2$ dividido por la integral de $|\psi|^2$ extendida a todo el espacio:

$$\frac{|\psi(\boldsymbol{x})|^2}{\|\psi\|}.$$

Para algunos estados, tales como los estados del momento, $\|\psi\|$ diverge, de modo que no obtenemos así una distribución de probabilidad que sea práctica (ya que la *densidad* de probabilidad es cero en todas partes, lo que resulta razonable para una única partícula en un universo infinito).

De acuerdo con esta interpretación probabilista, no es inusual que a la función de onda se la denomine «onda de probabilidad». Sin embargo, creo que esta es una descripción muy insatisfactoria. En primer lugar, la propia $\psi(\boldsymbol{x})$ es compleja, de modo que no puede ser una probabilidad. Además, la fase de ψ (salvo un factor multiplicactivo general constante) es un ingrediente esencial de la evolución de Schrödinger. Ni siquiera me parece muy razonable considerar $|\psi|^2$ (o $|\psi|^2/\|\psi\|$) como una «onda de probabilidad». Recordemos que para un estado de momento, el módulo $|\psi|$ de ψ es realmente *constante* en todo el espaciotiempo. ¡No hay información en $|\psi|$ que nos diga siquiera la dirección de movimiento de la onda! Es la fase, sola, la que da a esta onda su carácter «tipo onda».

Además, las probabilidades no son nunca negativas, y mucho menos complejas. Si la función de onda fuera solo una onda de probabilidades, entonces nunca tendríamos las cancelaciones de la interferencia destructiva. Esta cancelación es una característica de la mecánica cuántica, ¡tan vívidamente retratada (Fig. 21.4d) en el experimento de la doble rendija!

En este punto es oportuno ampliar ligeramente la discusión y tomar contacto con las consideraciones que se han hecho en §19.4 acerca del campo electromagnético y la conexión gauge ∇ asociada con él. Si nuestra función de onda describe una partícula cargada, entonces podemos hacer transformaciones *gauge* de la forma $\psi \mapsto e^{i\theta}\psi$, donde θ $(= \theta(\mathbf{x}))$ es una función arbitraria con valor real de la posición, que proporciona la necesaria «simetría gauge» que permite que el electro-

magnetismo actúe como una conexión gauge. Pero ¿no acabo de afirmar que la evolución temporal de Schrödinger depende *esencialmente* del conocimiento de cómo varían de un lugar a otro las fases de la función de onda? La aplicación de una transformación gauge $\psi \mapsto e^{i\theta}\psi$ nos permitiría cambiar la forma en que varían las fases por algo que nos gustase. ¿No contradice esto lo que acabo de afirmar sobre la importancia física crucial de la forma en que varían las fases?

En absoluto: aunque estos cambios de fase no constantes *sí están* permitidos, esto sucede solo si van acompañados de un cambio compensador en los operadores $\partial/\partial x^a$ (i.e., en el momento). Este cambio ($\partial/\partial x^a \mapsto \partial/\partial x^a - ieA_a$, donde $A_a = \nabla_a\theta$ y e = 1) es precisamente tal que deja inalterado el operador ∇. La «información de la fase» sigue estando allí, pero ahora está mezclada con la definición de ∇. Uno no puede limitarse a aplicar solamente $\psi \mapsto e^{i\theta}\psi$, con una θ que varía arbitrariamente, y confiar en que quede inalterada la situación física. Los detalles de la variación espacial de θ (en relación con ∇) son esenciales para la evolución dinámica del estado, y argumentaré que ψ es claramente mucho más que una onda de probabilidad. En cualquier caso, si ψ describe una onda-partícula sin carga ($e = 0$), entonces la situación es exactamente la misma que antes.

21.10. Estados de posición

En mi opinión, está claro que la función de onda debe ser algo mucho más «real» que meramente una «onda de probabilidad». La ecuación de Schrödinger nos proporciona una evolución precisa en el tiempo para esta entidad (esté cargada o no), una evolución que depende críticamente de cómo varíe la fase de un lugar a otro. Pero si preguntamos a una función de onda «¿dónde está la partícula?», realizando sobre ella una medida de posición, debemos estar preparados para perder esta información de la distribución de fases. De hecho, después de la medida tenemos que empezar todo de nuevo con una nueva función de onda. Si el resultado de la medida afirma «la partícula está aquí», entonces nuestra nueva función de onda tiene que tener un máximo pronunciado en la posición «aquí», pero luego se dispersa rápidamente otra vez, de

acuerdo con la evolución de Schrödinger. Si nuestra medida de posición fuera *absolutamente* precisa, entonces el nuevo estado tendría un «pico infinito» en dicha posición; de hecho, tendría que describirse mediante una función delta de Dirac, una cantidad con la que nos hemos encontrado brevemente en §6.6 y en §9.7 disfrazada de hiperfunción.

Veamos cómo maneja esto el formalismo de la mecánica cuántica. Por simplicidad, consideraremos la medida de una sola componente de la posición de una partícula, digamos la coordenada x^1. El *resultado* de nuestra medida debería ser un estado con «un valor definido para x^1»; de modo que, de acuerdo con lo que se ha dicho en el caso del momento, exigimos que ψ sea un autoestado del operador x^1 (i.e., de multiplicación por x^1), siendo el autovalor el valor concreto X^1 de la coordenada x^1 que se encuentra para la partícula. Para que la actuación de x^1, a saber,

$$\psi \mapsto x^1\psi,$$

tenga el valor definido X^1 (un número real) para la coordenada x^1, exigimos la ecuación de autovalores

$$x^1\psi = X^1\psi$$

(donde recordamos que x^1 es un operador lineal y X^1 es un número). Esta es satisfecha por

$$\psi = \delta(x^1 - X^1),$$

donde $\delta(x)$ es la «función delta» de Dirac, que se ha definido (como hiperfunción) en §9.7. En efecto, tiene la propiedad[21.13] de que $x\delta(x) = 0$, de donde $(x^1 - X^1)\delta(x^1 - X^1) = 0$, i.e., $x^1\delta(x^1 - X^1) = X^1\delta(x^1 - X^1)$, como se requería. Esta «función de onda» no es una función en el sentido ordinario, sino una función idealizada (una hiperfunción o distribución), que tiene un pico infinito en el autovalor $x^1 = X^1$, como se ha mencionado antes.

Esta medida particular no dice nada sobre las restantes coordenadas espaciales, y la función de onda aún puede tener una variación arbitraria en estas coordenadas, lo que nos proporciona un escalamiento para

[21.13] Compruébelo a partir de la definición hiperfuncional dada en §9.7.

la función delta en la forma de una función arbitraria de las restantes coordenadas x^2, x^3, de modo que obtenemos

$$\psi = \phi(x^2, x^3)\ \delta(x^1 - X^1)$$

para el autoestado general del operador x^1. Podemos ir más lejos y buscar un estado que sea simultáneamente un estado propio de las tres coordenadas espaciales. (Esta es una petición legítima porque x^1, x^2 y x^3 conmutan. Existe, de hecho, una propiedad general de los observables mecanocuánticos por la que, si tenemos un conjunto de ellos, todos los cuales conmutan entre sí, entonces existen autoestados comunes para todos ellos; véase §22.13.)[13] La respuesta es que, para el valor resultante (*autovalor*) $\mathbf{X} = (X^1, X^2, X^3)$ para la (triple) medida espacial, exigimos (salvo un factor de escala global)

$$\psi = \delta(x^1 - X^1)\ \delta(x^2 - X^2)\ \delta(x^3 - X^3) = \\ = \delta(\mathbf{x} - \mathbf{X}),$$

estando la segunda línea *definida* por la que está sobre ella.[14] Esta es la forma de un *estado de posición*.

Tales «estados de posición» son funciones de onda idealizadas en sentido contrario al de los estados de momento. Mientras que los estados de momento están infinitamente dispersos, los estados de posición están infinitamente concentrados. Ninguno de ellos es normalizable (en el caso de $\psi = \delta(\boldsymbol{x} - \boldsymbol{X})$, la dificultad reside en que las funciones delta no pueden elevarse al cuadrado; cf. §9.7). Terminaré este capítulo señalando que hay una dualidad importante entre posición y momento que dilucida esta cuestión.

21.11. Descripción en el espacio de momentos

Hasta ahora he representado los estados cuánticos enteramente como funciones de la posición: funciones de onda. Esto significa, de hecho, que cada *estado* —cada elemento de **W**— se considera una combinación lineal de estados propios del operador posición $\mathbf{x}$, i.e., de *estados de posición* (estados $\delta(\mathbf{x} - \mathbf{X})$). Expresar una función de onda ψ como una función de la posición significa, en efecto, que se considera una combinación lineal de tales funciones delta. Conseguimos esto mediante la fórmula $\psi(\mathbf{x}) = \int \psi(\mathbf{X})\delta(\mathbf{x} - \mathbf{X})\mathrm{d}^3\mathbf{X}$, que expresa $\psi(\mathbf{x})$ como una combinación *continua* de ellas, donde $\mathrm{d}^3\mathbf{X} = \mathrm{d}X^1 \wedge \mathrm{d}X^2 \wedge \mathrm{d}X^3$. En esta fórmula, los «coeficientes» en tal combinación lineal son los números complejos $\psi(\mathbf{X})$.

Pero hay muchas otras maneras de representar un estado cuántico ψ. Alternativamente, podemos representarlo como una combinación lineal de estados de *momento* $\mathrm{e}^{\mathrm{i}\mathbf{P}\cdot\mathbf{x}/\hbar}$. Ahora los «coeficientes» son números complejos diferentes, que tomamos como los productos de $(2\pi)^{-3/2}$ por las cantidades $\tilde{\psi}(\mathbf{P})$, de modo que llegamos a la fórmula:

$$\psi(\mathbf{x}) = (2\pi)^{-3/2} \int_{\mathbb{P}^3} \tilde{\psi}(\mathbf{P})\ \mathrm{e}^{\mathrm{i}\mathbf{P}\cdot\mathbf{x}/\hbar}\ \mathrm{d}^3\mathbf{P}.$$

(La razón para el factor «$(2\pi)^{-3/2}$» se explicará enseguida.) Esta fórmula expresa $\psi(\mathbf{x})$ como una transformada de Fourier de cierta función $\tilde{\psi}(\mathbf{P})$ igual que se ha hecho en §9.4, salvo que aquí tenemos una transformada de Fourier 3-dimensional, que equivale a aplicar la fórmula de §9.4 tres veces.

Esto sugiere que $\tilde{\psi}$ (como función de $\mathbf{P}$, pero ahora podemos escribirla como función de $\mathbf{p}$) proporciona una representación tan buena de los estados cuánticos de la partícula como lo hace la función original $\psi(\mathbf{x})$. En realidad, hay una simetría muy precisa entre las variables de posición y momento. Ahora podemos considerar las variables de *momento* $\mathbf{p}$ como primarias y representar las variables de *posición* $\mathbf{x}$ como «derivación con respecto a $\mathbf{p}$», de modo que podemos hacer la interpretación *inversa* (advirtiendo el cambio de signo)

$$x^a = -\mathrm{i}\hbar\frac{\partial}{\partial p_a}$$

(al menos para las variables espaciales x^1, x^2, x^3).[21.14] Se satisfacen relaciones de conmutación que son *idénticas* a las que teníamos antes:

$$p_b x^a - x^a p_b = \mathrm{i}\hbar\ \delta^a_b.$$

[21.14] Demuestre que reemplazar ψ por $x^1\psi$ o por $\mathrm{i}\hbar\partial\psi/\partial x^1$ corresponde, respectivamente, a sustituir $\tilde{\psi}$ por $-\mathrm{i}\hbar\partial\tilde{\psi}/\partial p_1$ o por $p_1\tilde{\psi}$. Demuestre que reemplazar $\psi(x^a)$ por $\psi(x^a + C^a)$ corresponde a reemplazar $\tilde{\psi}$ por $\mathrm{e}^{-\mathrm{i}C_P/\hbar}\tilde{\psi}$ (donde a recorre los valores 1, 2, 3).

La función «invisible» a la derecha es ahora una función del momento p_a y no de la posición x^a. Son los estados de momento los que están ahora representados por funciones delta $\delta(\mathbf{p} - \mathbf{P})$ y los estados de posición están representados como ondas planas $e^{-i\mathbf{p}\cdot\mathbf{X}/\hbar}$. La representación de «funciones de onda» del momento en términos de estados de posición $e^{-i\mathbf{p}\cdot\mathbf{X}/\hbar}$ viene dada por la transformada de Fourier (inversa) prácticamente idéntica

$$\tilde{\psi}(\boldsymbol{p}) = (2\pi)^{-3/2}\int_{\mathbb{E}^3}\psi(\mathbf{X})e^{i\mathbf{p}\cdot\mathbf{X}/\hbar}\,d^3\mathbf{X},$$

con tan solo un pequeño cambio de signo en el exponente. (Ahora vemos la razón para el «$(2\pi)^{-3/2}$»; es para equilibrar las cosas, de modo que la transformada de Fourier inversa sea prácticamente la misma que la original.)

Los paquetes de onda pueden describirse tanto en la representación del espacio de momentos como en la representación de posición.[21.15] Podemos introducir una noción precisa de la «dispersión» (o falta de localización) de un paquete de ondas, ya sea en la descripción de posición o en la descripción de momento. Denotamos estas medidas de la dispersión por Δx y Δp, respectivamente. Las *relaciones de incertidumbre de Heisenberg* nos dicen que el producto de estas dispersiones no puede ser menor que algo del orden de la constante de Planck, y tenemos[15]

$$\Delta p\,\Delta x \gtrsim \tfrac{1}{2}\hbar.$$

En la Fig. 21.10 se ilustran estados de posición, estados de momento y paquetes de onda en las dos representaciones. Notemos que en el caso extremo de un estado de momento puro, la dispersión en el momento es cero, de modo que $\Delta p = 0$ (i.e., una función delta en el espacio de momentos). Por la relación de Heisenberg, Δx es ahora infinita, de acuerdo con la imagen antes descrita (en §21.6), donde la función de onda se dispersa de manera uniforme sobre todo el espacio de posiciones. La situación es exactamente la contraria con un estado de posición, donde ahora $\Delta x = 0$ y la posición está definida con completa

[21.15] Utilice los resultados de los ejercicios [21.10], [21.13] y [21.14] para demostrar que la transformada de Fourier del paquete de ondas $\psi = Ae^{-B^2(x-C)^2}e^{i\omega x}$ es $\tilde{\psi} = (Ae^{i\omega C}/B\sqrt{2})e^{-(p-u)^2/4B^2}e^{-iCp}$ (haciendo $\hbar = 1$ por conveniencia).

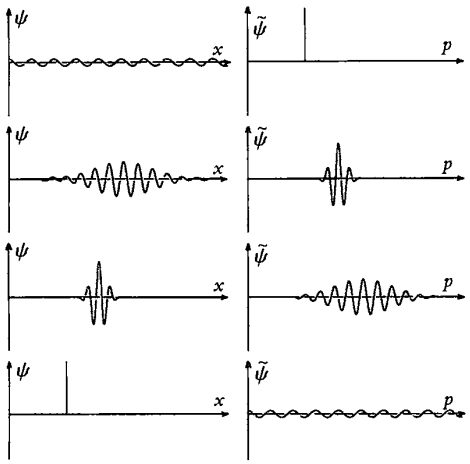


Fig. 21.10. Las representaciones de las funciones de onda ψ en el espacio de posición están a la izquierda, y las correspondientes imágenes de ψ en el espacio de momento están a la derecha. El par superior representa un estado de momento, y el par inferior, un estado de posición. Las dos representaciones entre ellos muestran paquetes de onda. La relación de incertidumbre de Heisenberg se ilustra por el hecho de que la mayor dispersión en posición va acompañada de una menor dispersión de momento, y viceversa.

precisión, pero la dispersión Δp en el momento se hace ahora infinita.

Resulta interesante ver que aquí tenemos ejemplos que ilustran claramente la incompatibilidad de medidas no conmutativas en mecánica cuántica (que es un fenómeno general que encontraremos frecuentemente en nuestras consideraciones posteriores). Una medida del momento de una partícula la colocaría en un estado de momento, correspondiente a cierto valor clásico P, y cualquier medida posterior del momento en dicho estado daría el mismo resultado P. Sin embargo, si el estado fuera sometido a una posterior medida de *posición*, el resultado sería del todo incierto y cualquier resultado para la posición sería tan probable como cualquier otro. Esta medida hace del estado una función delta en la posición. En el espacio de momentos, este estado es una onda plana, dispersa uniformemente sobre todos los valores posibles para el momento. Una medida posterior del *momento* sería entonces completamente incierta. Así pues, el propio acto de una medida intermedia de la posición ha arruinado por completo la pureza del estado de momento original.

Debería mencionarse también que, consistentemente con la relati-

vidad (§18.7), existe una relación de incertidumbre similar entre energía y tiempo

$$\Delta E\, \Delta t \gtrsim \frac{1}{2}\hbar.$$

Normalmente se considera que esta relación tiene un estatus físico algo diferente de la más familiar relación de incertidumbre posición-momento, puesto que en la mecánica cuántica estándar el tiempo se trata como un parámetro externo, en lugar de como una variable dinámica. La interpretación habitual de la incertidumbre energía-tiempo es que si se determina la energía de un sistema cuántico mediante una medida que se realiza en un intervalo de tiempo Δt, entonces hay una incertidumbre ΔE en esta medida de energía que debe satisfacer la relación anterior.

Esto tiene particular relevancia para, por ejemplo, los núcleos inestables. El hecho de que un núcleo semejante (por ejemplo, de uranio) sea inestable significa que hay un límite para el tiempo —a saber, la vida media de la partícula— durante el cual puede determinarse la energía de la partícula. En consecuencia, la relación de Heisenberg nos da una incertidumbre fundamental en la energía, para una partícula o núcleo inestable, que está relacionada de forma inversa con su vida media. Debido a la $E = mc^2$ de Einstein (véase §18.7), esto nos da una incertidumbre fundamental en su *masa*. Por ejemplo, la vida media de un núcleo de uranio U_{238} es de unos 10^9 años, de modo que en este caso hay una incertidumbre en la energía de unos 10^{-51} julios, con lo que la correspondiente incertidumbre en la masa es absolutamente minúscula, a saber unos 10^{-68} kg. (La función de onda de una partícula inestable se desvía de la forma estacionaria e^{-iEt} para un valor real[16] definido E de la energía, y contiene también un factor de decaimiento exponencial. Al no ser un autoestado de la energía, hay una dispersión resultante en la energía medida, lo que da la incertidumbre en la energía.) ¡La relación de incertidumbre energía-tiempo desempeñará un papel especial en §30.11, en relación con una aproximación concreta a la solución del enigma de la medida cuántica!

Notas

Sección 21.1

21.1. Este es un ejemplo de lo que se denomina una ecuación diferencial *ordinaria*, o EDO, porque es una ecuación que solo incluye operadores diferenciales ordinarios tales como d/dx, d/dy, etc., o sus potencias, por ejemplo, d^3/dx^3. Una ecuación en derivadas parciales, o EDP, sería una ecuación que incluye los operadores en derivadas parciales $\partial/\partial x$, $\partial^2/\partial x^2$, $\partial^2/\partial x \partial y$, etc., como sucede con las ecuaciones de Maxwell o las ecuaciones de Einstein del capítulo 19.

Sección 21.2

21.2. Sin embargo, por razones de consistencia (véase también §21.3), me estoy ateniendo a la notación apropiada para la relatividad (§18.7), de modo que las componentes espaciales de los momentos en «p_a» son las *negativas* de las componentes usuales del momento (que son las componentes espaciales p^a divididas por c^2). Esta elección es compatible con mis comentarios de §20.2 porque ahora estoy utilizando «x» (en lugar de las «q» del formalismo general de Lagrange/Hamilton).

21.3. Esta independencia temporal asegura que puede mantenerse la interpretación de $\mathcal{H}$ como una *energía total* conservada. El lector podría estar preocupado por el hecho de que, puesto que se ha permitido una dependencia de las coordenadas espaciales, los requisitos de una invariancia relativista en un nivel fundamental podrían exigir que admitamos también una dependencia temporal (véase la nota 20.3). Pero en un nivel *fundamental*, un requisito ordinario sería la independencia espacial tanto como temporal.

21.4. Véase Woodhouse (1991).

Sección 21.4

21.5. El término «negro» se refiere aquí a la naturaleza completamente absorbente (tanto como sea posible) del cuerpo que encierra la radiación. En estos primeros experimentos se utilizaba una cavidad oscura casi completamente esférica como contenedor de la radiación, con una abertura muy estrecha que conectaba el volumen interior con el exterior. Sin embargo, el cuerpo podía estar brillando perfectamente, debido a su temperatura, de modo que en realidad no parecería negro.

21.6. En mis descripciones de este experimento estoy idealizando la situación, dejando fuera todas las dificultades prácticas para ir al punto esencial.

Sección 21.7

21.7. En experimentos precisos, sería poco probable que un objeto semejante incluyera un plateado real, sino que utilizaríamos efectos de interferencia entre las ondas reflejadas en las dos caras de una lámina de material transparente.

21.8. En la descripción de la mecánica cuántica conocida como teoría de De Broglie-Bohm (Bohm y Hiley, 1994), tanto los aspectos de onda como los de partícula se retienen simultáneamente. Aquí la partícula hace realmente su elección en el divisor de haz, pero la onda continúa, explorando ambos caminos simultáneamente. Cuando se alcanza el divisor de haz final, es la onda la que instruye a la partícula para alcanzar el detector en A, prohibiéndole que llegue a B. Trataré de evaluar este punto de vista interesante pero «poco convencional» (y también no local) en §§29.2,9.

Sección 21.8

21.9. Tal como informa Heisenberg (1971), p. 73.

21.10. Véanse Goldstein (1987) y Bell (1987).

Sección 21.9

21.11. Muchos autores podrían definir la «norma» como la *raíz cuadrada* de lo que entiendo aquí por $\|\psi\|$, es decir, su $\|\psi\|^2$ es mi $\|\psi\|$.

21.12. Varios autores han desarrollado el formalismo mecanocuántico de una manera elegante dentro del marco proyectivo. Véanse en particular Brody y Hughston (2001), Hughston (1995), Ashtekar y Schilling (1998).

Sección 21.10

21.13. Esta propiedad de observables que conmutan se expone en cualquier texto de mecánica cuántica; véase, por ejemplo, Shankar (1994).

21.14. Es legítimo multiplicar funciones delta si se refieren a variables *diferentes*. Véase Arfken y Weber (2000) para las propiedades de las funciones delta.

Sección 21.11

21.15. Véanse Shankar (1994) y Hannabuss (1997).

21.16. Es una práctica común, en física de partículas, utilizar la misma dependencia temporal $e^{-iEt/\hbar}$, aunque con una E compleja, cuya parte real es el valor medio de la energía y cuya parte imaginaria es $-\frac{1}{2}\hbar \log 2$ multiplicado por el recíproco de la semivida (véase, por ejemplo, Das y Ferbel, 2004).[21.16]

[21.16] ¿Puede ver la forma de justificar el factor $-\frac{1}{2}\hbar \log 2$? (La semivida es el tiempo en que la probabilidad de desintegración alcanza el valor un medio.)

22

Álgebra, geometría y espín cuánticos

22.1. Los procedimientos cuánticos **U** y **R**

La naturaleza no intuitiva de la mecánica cuántica —o más bien, de la propia naturaleza en el nivel de actividad mecanocuántica— lleva a muchos a desesperar de encontrar imágenes fidedignas de los fenómenos a nivel cuántico. Pese a todo, además de su elegante estructura algebraica, hay mucha belleza geométrica asociada con la mecánica cuántica, y sería una pena pensar que necesariamente debemos confiar en un formalismo sin imágenes e invisualizable para seguir avanzando en la descripción de las acciones cuánticas. Aunque hemos visto que incluso una sola «partícula puntual» indiferenciada parece ser, en el formalismo cuántico, un misterioso objeto ondulatorio y difuso, se trata de un «objeto» que puede representarse y que tiene una fascinante estructura matemática en la que empiezan a manifestarse muchos aspectos de la magia de los números complejos.

Esta imagen nos permite empezar a entender la descripción cuántica de una única partícula puntual. Cabría suponer que, una vez que hayamos entendido a qué se parece una única partícula cuántica, seremos capaces de sentarnos y relajarnos un poco, puesto que esto nos ofrece seguramente una comprensión, en teoría, de sistemas complicados que incluyen muchos tipos de partículas diferentes. Por desgracia, esta expectativa es prematura, y necesitaremos una perspectiva más amplia si queremos llegar a una imagen cuántica global del mundo. En el capítulo 23 veremos cuánto más confusa se hace nuestra imagen del mundo cuando hay que considerar varias partículas juntas en un sistema. En lugar de que cada partícula tenga individualmente su «vector de

estado» independiente, encontramos que el sistema cuántico entero requiere un *único* vector de estado muy complicado.

Pero incluso las «partículas puntuales» individuales suelen tener más estructura que la recogida por las descripciones que he ofrecido hasta ahora, pues con frecuencia poseen lo que se denomina *espín*, lo que lleva a una complicación extra. Afortunadamente, como veremos más adelante en este capítulo, el espín es en sí mismo un fenómeno con una descripción matemática de especial riqueza y elegancia, donde pasan a primer plano otros aspectos de la magia de los números complejos.

Revisemos las descripciones del capítulo anterior, donde hemos tenido que acostumbrarnos a que la partícula cuántica (no relativista) es algo descrito por lo que hemos denominado un *vector de estado* (o función de onda) cuya evolución viene dada, de una forma muy exacta, por la ecuación de Schrödinger, hasta que se realiza una medida sobre el sistema. Como veremos más explícitamente en el capítulo 23, lo mismo se aplicará a los vectores de estado que describen complicados sistemas cuánticos completos. La medida propiamente dicha se describe matemáticamente de una forma completamente diferente de la evolución de Schrödinger. Hemos visto indicios de esto en §§21.4,7,8. En §§21.10,11 hemos considerado medidas de posición, en el curso de las cuales el estado de una partícula *saltaría* a un estado (generalmente diferente), ahora localizado en una localización concreta, i.e., a un estado que es un autovector del operador de posición **x** (que en coordenadas de posición es una función delta). También consideramos el resultado de medidas del momento lineal (en §§21.5,6,11), por las que el estado de una partícula se ha hecho saltar a un autoestado del operador momento **p**, de modo que el estado de la partícula está ahora disperso en una forma ondulatoria (en principio sobre todo el espacio). De manera más general, una medida correspondería a un operador **Q** de cierto tipo (normalmente, un operador hermítico; véase §22.5), y el efecto de la medida sobre el estado sería hacerlo saltar a algún autoestado de **Q**. ¿A *qué* autoestado de **Q** va a saltar? Según la teoría cuántica, esta es una cuestión de puro azar, pero hay reglas precisas para calcular las probabilidades (véase §22.5).

El salto del estado cuántico[1] a uno de los autoestados de **Q** es el

proceso conocido como *reducción del vector de estado* o *colapso de la función de onda*. Es uno de los aspectos más enigmáticos de la teoría cuántica, y volveremos a esta cuestión muchas veces en este libro. Creo que la mayoría de los físicos cuánticos no consideran que la reducción del estado sea una acción *real* del mundo físico, sino que refleja el hecho de que no deberíamos considerar que el vector de estado describe una «realidad» física en el nivel cuántico. Llegaremos a tan controvertida cuestión con más detalle en el capítulo 29. De todas formas, con independencia de cuál sea la actitud que se pueda adoptar sobre la realidad física del fenómeno, la utilización que se hace de la mecánica cuántica en la práctica consiste en considerar que el estado salta de esta curiosa manera cuando quiera que se estima que tiene lugar una medida. Inmediatamente después de la medida, la evolución de Schrödinger domina de nuevo, hasta que se realiza otra medida sobre el sistema, y así sucesivamente.

Denoto la evolución de Schrödinger por **U** y la reducción de estado por **R**. ¡Esta alternancia entre estos dos procesos de aspecto completamente diferente parece una forma decididamente singular para el comportamiento de un universo! Véase la Fig. 22.1. De hecho, cabría imaginar que en realidad esta es una aproximación a alguna otra cosa todavía desconocida. ¿Es posible que exista una ecuación matemática más general, o un principio de evolución de un tipo matemático coherente, del que **U** y **R** son aproximaciones límite? Mi opinión personal es que es muy probable que un cambio de este tipo para la teoría cuántica sea correcto —quizá como parte de una nueva física del siglo XX—, y en el capítulo 30 haré algunas sugerencias concretas orientadas a esta posibilidad. Sin embargo, la mayoría de los físicos *no* parecen creer que sea fructífero seguir este camino.

Su razón para preferir no alterar el marco básico de la mecánica cuántica es (además de la gran elegancia matemática de su formalismo **U**) el acuerdo extraordinariamente impresionante y preciso que hay entre la teoría cuántica y los hechos experimentales, en los que no se conoce nada que contradiga la teoría (en su forma híbrida actual) y muchos resultados diversos la confirman con gran precisión. En consecuencia, la mayoría de los físicos cuánticos adoptarán un punto de vista filosófico (o más bien uno entre varios y diferentes puntos de vista

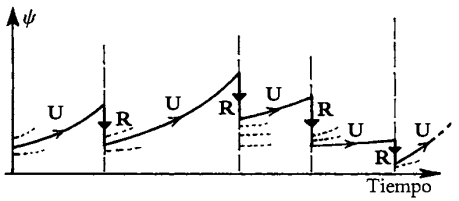


Fig. 22.1. La evolución temporal del estado ψ para un sistema físico, según principios aceptados de la mecánica cuántica, alterna entre dos procedimientos completamente diferentes: evolución unitaria **U** (continua, determinista) y reducción de estado **R** (discontinua, probabilista).

filosóficos alternativos que se describen en §29.1) que trate de entender la aparente contradicción entre los procedimientos **U** y **R**, aunque sin intentar cambiar de ninguna forma significativa el formalismo cuántico actual. Uno de mis objetivos en este capítulo y en el siguiente es abordar estas cuestiones, pero sin desviarnos de lo que es ahora convencional en la teoría cuántica. Más adelante volveré a la cuestión **U**/**R**, particularmente en §§29.1,2,7-9, y también en §§30.10-13, donde expondré mis propias opiniones sobre la cuestión de forma más completa.

Creo que sería justo decir que un hilo común a muchas de las que podrían llamarse actitudes «convencionales» respecto a la mecánica cuántica es que el proceso **U** debe tomarse como una «verdad subyacente», mientras que hay que tratar de entender **R**, de una forma u otra, como un tipo de aproximación, ilusión o conveniencia, y hay muchas exposiciones en la literatura científica que siguen este enfoque.[2] Incluso aquellos (yo incluido) que son de la opinión de que es necesario algún cambio en alguna etapa del formalismo cuántico, argumentarán que el esquema actual es al menos una aproximación maravillosa, de modo que es necesario entenderlo por completo para que haya alguna esperanza de ir más allá del mismo. En consecuencia, debemos tratar de ver más profundamente cómo opera **U** y, más aún, cómo puede encajar de forma tan bella con **R**, ¡aun siendo en cualquier caso incompatible con este!

También debería explicar el uso de la letra **U**. Significa *evolución unitaria*. Tendremos que ver en qué sentido es «unitaria» (véase §13.9) la ecuación de Schrödinger; llegaremos a ello de inmediato, en §22.4.

Asimismo, hay otras formas (equivalentes) de expresar esta «evolución unitaria», muy en especial la que se conoce como *imagen de Heisenberg*, a la que también llegaremos en §22.4. De todas formas, la imagen que ofrece la ecuación de Schrödinger es la más conveniente para nuestras descripciones.

22.2. La linealidad de **U** y sus problemas para **R**

Antes de abordar la cuestión de la unitariedad, examinemos la cuestión más básica de la *linealidad* de **U**. Veremos que este aspecto de **U** por sí solo presenta una seria incompatibilidad con **R**. Por consiguiente, examinemos de nuevo la ecuación de Schrödinger $i\hbar\partial\psi/\partial t = \mathcal{H}\psi$. Imaginaremos que el hamiltoniano $\mathcal{H}$ es una cosa conocida (estando especificado por la naturaleza de las partículas que describe y las fuerzas entre ellas, y por cualesquiera fuerzas externas conservativas —i.e., que conservan la energía— que pudieran influir en el sistema). Hay ciertas consecuencias que son inmediatas a partir de la forma general de la ecuación, y son así completamente independientes de la naturaleza detallada del hamiltoniano.

Advertimos que es una ecuación *determinista* (pues la evolución temporal está completamente determinada una vez que se conoce el estado en un instante cualquiera). Esto puede resultar una sorpresa para algunas personas que quizá hayan oído hablar de la «incertidumbre cuántica» y de que los sistemas cuánticos se comportan de forma no determinista. La falta de determinismo viene únicamente con la aplicación del proceso **R**. No se encuentra en la evolución temporal (**U**) del estado cuántico, tal como la describe la ecuación de Schrödinger. Otra cosa que vemos inmediatamente en la ecuación de Schrödinger es que es una ecuación *compleja*, debido a la aparición manifiesta de «i» en el primer miembro (y existen muchas más posibilidades para las «i» en el hamiltoniano).

Finalmente, vemos que la ecuación de Schrödinger es lineal, en el sentido de que si ψ y Φ son soluciones (con el mismo $\mathcal{H}$) de

$$i\hbar\frac{\partial\psi}{\partial t} = \mathcal{H}\psi, \quad i\hbar\frac{\partial\phi}{\partial t} = \mathcal{H}\phi,$$

entonces también lo es cualquier *combinación lineal* $w\psi + z\phi$, donde w y z son constantes complejas. En efecto, multiplicando la primera de las ecuaciones anteriores por w, la segunda por z y sumándolas obtenemos (§6.5):

$$i\hbar\frac{\partial}{\partial t}(w\psi + z\phi) = \mathcal{H}(w\psi + z\phi).$$

A partir de esto, vemos que la ecuación de Schrödinger conserva la estructura de *espacio vectorial complejo* del espacio de estados **W** (que normalmente es un espacio de dimensión infinita).

El hamiltoniano $\mathcal{H}$ define la transformación lineal infinitesimal de **W** que describe el cambio que ha sufrido el estado cuando ha evolucionado durante un tiempo infinitesimal. Esta acción hamiltoniana viene descrita por un *campo vectorial* en **W** (véase la Fig. 22.2). Al cabo de un tiempo finito, los estados habrían cambiado de acuerdo con una transformación lineal finita, obtenida mediante lo que se denomina «exponenciación» de la acción hamiltoniana infinitesimal. Esta es muy similar a la «exponenciación» que hemos encontrado antes (§14.6) al describir el proceso por el que se obtiene un elemento de un grupo de Lie a partir de la exponenciación de un elemento del álgebra de Lie correspondiente. Sin embargo, la exponenciación en la evolución hamiltoniana puede ser mucho más difícil de llevar a cabo. (Además, aparecen dificultades adicionales debido a la dimensión infinita de **W**.)

Pero difícil o no, el punto esencial aquí es que al cabo de cualquier tiempo finito T la transformación del espacio **W** de estados cuánticos será siempre lineal. Esto equivale a la afirmación siguiente (donde utilizaré el símbolo $\rightsquigarrow$ para indicar cómo habrá evolucionado un estado al cabo del período de tiempo especificado T):

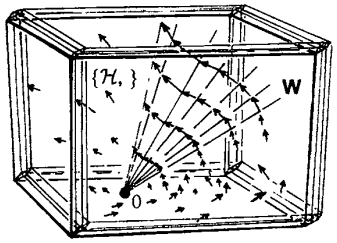


Fig. 22.2. El flujo hamiltoniano $\{\mathcal{H}, \}$ (un campo vectorial) define una transformación lineal infinitesimal del espacio de estados **W**, que da el cambio en el estado tras un tiempo infinitesimal. Para obtener el cambio (unitario) tras un tiempo finito, debemos «exponenciar» esta acción hamiltoniana infinitesimal.

Si

$$\psi \rightsquigarrow \psi' \text{ y } \phi \rightsquigarrow \phi',$$

entonces

$$w\psi + z\phi \rightsquigarrow w\psi' + z\phi'.$$

Aquí, ψ y ϕ son dos estados (funciones de onda) escogidos arbitrariamente, y w y z son constantes complejas arbitrarias.[22.1]

Esto tiene algunas implicaciones muy curiosas si tratamos de adoptar el punto de vista de que **U** es toda la historia y el proceso de medida es tan solo algún tipo de «conveniencia» que se invoca para manejar situaciones donde el estado cuántico se hace intratablemente complicado, por incluir quizá un número tremendo de partículas «entrelazadas» en el sistema y su aparato de medida. (Llegaremos a la noción mecanocuántica de «entrelazamiento» más específicamente en el capítulo 23. Entonces veremos que los estados cuánticos son entidades «holísticas» de una forma más seria que en §21.7, entidades en las que diferentes partes del sistema no tienen estados cuánticos independientes, sino que son partes de un «todo» entrelazado.) De acuerdo con tal punto de vista de «conveniencia» para **R**, uno imagina que **R** emergería como cierto tipo de aproximación a una «genuina» evolución **U** subyacente. Pero este punto de vista conduce a serias paradojas.

Recordemos, por ejemplo, el experimento mental de §21.7, en el que mis dos colegas espaciales tenían detectores individuales, y tratemos de imaginar que la respuesta de cada detector es simplemente el resultado de una evolución de Schrödinger que parte de su interacción con la parte del paquete de ondas que recibe. El estado cuántico antes de la detección es en realidad una suma de las dos partes individuales del paquete de ondas, una que llega a un detector y la otra que llega al otro detector; por lo tanto, debido a la linealidad, la consiguiente respuesta de cada detector a partir de la evolución de Schrödinger debe coexistir en superposición con una respuesta en el otro. La evolución de Schrödinger conduce a la respuesta de un detector *más* la respuesta

[22.1] Deje claro por qué la acción de cualquier evolución de Schrödinger es lineal, pese al hecho de que $\mathcal{H}$ puede ser una función altamente no lineal de las p y las x.

de otro detector («más» en el sentido de superposición cuántica de las dos respuestas de los detectores), y no a la respuesta de un detector *o* la respuesta del otro detector (y es este «o» lo que realmente sucede siempre). Me parece que no es sostenible tratar de mantener que **U** describe la historia completa (y ciertamente no es esto lo que trata de hacer la mecánica cuántica «convencional» en la «interpretación de Copenhague» de Niels Bohr, pues trata a los propios detectores como «entidades clásicas»).

Por lo que puedo ver, la única manera de mantener **U** para todos los procesos que incluyen medidas sería adoptar un punto de vista del tipo «muchos universos» (véase §29.1), en el que las dos respuestas de los detectores coexisten realmente en «universos diferentes».[3] Pero ni siquiera así **U** puede ser «la historia completa», porque necesitaríamos una teoría para explicar por qué solo percibimos siempre respuestas individuales de los detectores, ¡mientras que nunca se perciben conscientemente superposiciones de respuestas con no respuestas! (Volveremos a estas cuestiones en §§29.1,8). Debería decir, en este momento, que yo mismo no creo que los «muchos universos» sean el camino correcto a seguir; simplemente estoy argumentando que parece que uno se ve llevado a ello si insiste en mantener el punto de vista «**U** en todos los niveles».

Volveremos a estas cuestiones más adelante, en los capítulos 29 y 30, donde se abordará la cuestión de si **U** y **R** deben o no tratarse como aproximaciones a alguna teoría futura más amplia. Por el momento, sigamos las recetas del formalismo convencional. Si es necesaria una teoría mejorada, en cualquier caso tendrá que estar de acuerdo hasta un grado muy alto de precisión con las recetas de la teoría convencional. Cualquier lector con aspiraciones de encontrar una nueva teoría (¡y espero que haya algunos!) haría bien en entender perfectamente lo que tiene que decir la teoría convencional.

22.3. Estructura unitaria, espacio de Hilbert, notación de Dirac

No he abordado todavía adecuadamente el aspecto «unitario» de la evolución de Schrödinger. Este tiene que ver con la propiedad de «nor-

malización» de las funciones de onda mencionada en el capítulo anterior. Recordemos que para la función de onda ψ de una única partícula (sin espín), la «norma» se refiere a la cantidad $\|\psi\|$, definida como la integral de $|\psi(\boldsymbol{x})|^2$ sobre todo el espacio. La condición de normalización sobre ψ es que $\|\psi\| = 1$ (y cuando se impone esta condición, entonces $|\psi(\boldsymbol{x})|^2$ es la densidad de probabilidad de que una medida de posición encuentre la partícula en el punto $\boldsymbol{x}$). En una situación mecanocuántica general, donde quizá haya muchas partículas con espín interactuantes (o quizá entidades de un tipo más general, tales como cuerdas, etc.), siempre exigimos algo de esta naturaleza para definir una correspondiente noción de *norma* $\|\psi\|$, que debe ser un número real positivo[4] para cualquier estado cuántico ψ aceptable. Aunque esta norma es algo pertinente a la parte **U** del formalismo cuántico, también desempeña un papel crucial en la parte **R**, pues determina, en efecto, todas las *probabilidades* que aparecen.

Podemos pensar que la norma, matemáticamente hablando, proporciona una noción de *longitud al cuadrado*, que debería ser finita para vectores «aceptables» que pertenecen al espacio de estados **W**. El adjetivo «unitaria», aplicado a la evolución temporal, nos dice que esta norma se conserva a lo largo de la evolución. Veremos muy pronto (en §22.4) por qué esto se aplica realmente en el caso de la evolución de Schrödinger.

En primer lugar, será útil establecer cierta notación e investigar algunas propiedades de los estados cuánticos normalizables. Será provechoso pensar en esta norma como un caso particular de un producto escalar hermítico entre estados (§13.9). Para estados ψ y ϕ, este se escribe normalmente $\langle\phi|\psi\rangle$ en la literatura mecanocuántica, y la norma de ψ es el caso especial cuando $\phi = \psi$:

$$\|\psi\| = \langle\psi|\psi\rangle.$$

En el caso de una única partícula (sin espín), el producto escalar es:

$$\langle\phi|\psi\rangle = \int_{\mathbb{E}^3} \bar{\phi}\psi \mathrm{d}x^1 \wedge \mathrm{d}x^2 \wedge \mathrm{d}x^3,$$

que generaliza la expresión particular para $\|\psi\|$ que se ha dado en §21.9. Esto nos da un producto escalar hermítico definido positivo entre dos funciones de onda ψ y ϕ normalizables de 1-partícula.[22.2]

De hecho, las funciones de onda normalizables constituyen un espacio vectorial complejo **H** (un subespacio de **W**), que es de un tipo particular conocido como un *espacio de Hilbert*.[22.3] Un espacio de Hilbert se define como un espacio vectorial complejo que posee una operación de producto escalar $\langle\,|\,\rangle$, cuyo valor es un número complejo, que satisface las propiedades algebraicas

$$\begin{gathered}\langle\phi|(\psi+\chi)\rangle = \langle\phi|\psi\rangle + \langle\phi|\chi\rangle,\\ \langle\phi|(a\psi)\rangle = a\langle\phi|\psi\rangle,\\ \langle\phi|\psi\rangle = \overline{\langle\psi|\phi\rangle},\\ \psi \neq 0 \text{ implica } \langle\psi|\psi\rangle > 0\end{gathered}$$

(todas las cuales son inmediatas en el caso de la integral para 1-partícula dada antes).[22.4] Estas ecuaciones implican también $\langle(\phi+\chi)|\psi\rangle = \langle\phi|\psi\rangle + \langle\chi|\phi\rangle$, $\langle(a\phi)|\psi\rangle = \bar{a}\langle\phi|\psi\rangle$.[22.5] Más aún, una vez que la norma es conocida, el producto escalar puede definirse en términos de ella,[22.6] de modo que las transformaciones lineales que conservan la norma deben conservar también el producto escalar. Además, un espacio de Hilbert debería satisfacer ciertas propiedades básicas de continuidad.[5]

La notación anterior forma parte de la valiosa y ampliamente utilizada herramienta notacional para la mecánica cuántica introducida

[22.2] Trate de explicar por qué la integral $\langle\phi|\psi\rangle$ converge siempre que $\langle\phi|\phi\rangle$ y $\langle\psi|\psi\rangle$ convergen. *Sugerencia*: Considere lo que está implícito en el hecho de que la integral de $|\phi - \lambda\psi|^2$ sea no negativa sobre cualquier región *finita* de $\mathbb{E}^3$, obteniendo una desigualdad que relacione el módulo al cuadrado de la integral de ψ con el producto de la integral de $\phi\phi$ por la integral de $\psi\psi$. Como paso intermedio, encuentre las condiciones mínimas sobre los números complejos a, b, c, d que implican $a + \lambda b + \lambda c + \lambda\lambda d \geq 0$ para todo λ.

[22.3] Siguiendo el ejercicio [22.2], demuestre que las funciones de onda normalizables constituyen un espacio vectorial.

[22.4] Verifíquelo estableciendo cuidadosamente qué propiedades de la integración se están utilizando.

[22.5] Demuestre por qué.

[22.6] Demuestre cómo puede definirse $\langle\phi|\psi\rangle$ a partir de la norma. *Sugerencia*: Calcule las normas de $\phi + \psi$ y $\phi + \mathrm{i}\psi$.

por el gran físico del siglo XX Paul Dirac. Como ingrediente de este esquema general, se muestra útil considerar que expresiones como

$$|\psi\rangle,\ |\uparrow\rangle,\ |\rightarrow\rangle,\ |\leftrightarrow\rangle,\ |0\rangle,\ |7\rangle,\ |+\rangle,\ |\boldsymbol{X}\rangle,\ |\,\text{MUERTO}\rangle,\ \text{u}\ |\,\text{OFF}\rangle$$

representan diversos vectores de estado pertenecientes al espacio de Hilbert $\mathbf{H}$, donde el símbolo dentro del $|\ldots\rangle$ es alguna etiqueta apropiada (y quizá fácil de recordar) que indica el estado en cuestión. Se les suele denominar vectores «ket». Para cada uno de estos kets, habrá un miembro concreto del espacio dual $\mathbf{H}^*$, denominado el correspondiente vector «bra», que es el *conjugado hermítico* de dicho estado (en el sentido de §13.9), escritos respectivamente

$$\langle\psi|,\ \langle\uparrow|,\ \langle\rightarrow|,\ \langle\leftrightarrow|,\ \langle 0|,\ \langle 7|,\ \langle +|,\ \langle\boldsymbol{X}|,\ \langle\text{MUERTO}|\ \text{u}\ \langle\text{OFF}|\,.$$

Puesto que los vectores bra son duales de los vectores ket, tienen un producto escalar en el mismo sentido que el producto «•» de §12.3. Este producto escalar —o «bracket» (paréntesis)— de un vector bra $\langle\psi|$ por un vector ket $|\phi\rangle$ es precisamente el producto escalar hermítico escrito antes como $\langle\psi|\phi\rangle$. Esto es consistente con que el número complejo $\langle\psi|\phi\rangle$ sea el complejo conjugado de $\langle\phi|\psi\rangle$. Se dice que los dos estados $|\phi\rangle$ y $|\psi\rangle$ son *ortogonales* si $\langle\phi|\psi\rangle = 0$, i.e., si $\langle\psi|\phi\rangle = 0$.

La acción de cierto operador lineal $\boldsymbol{L}$ sobre $|\psi\rangle$ se escribe $\boldsymbol{L}|\psi\rangle$, y el producto escalar del ket $|\phi\rangle$ por $\boldsymbol{L}|\psi\rangle$ se escribe

$$\langle\phi|\boldsymbol{L}|\psi\rangle.$$

Este es también el producto escalar de cierto bra «$\langle\phi|\boldsymbol{L}$» por $|\psi\rangle$. ¿Qué es el bra «$\langle\phi|\boldsymbol{L}$»? Es el complejo conjugado de cierto ket $\boldsymbol{L}^*|\phi\rangle$, donde $\boldsymbol{L}^*$ es el *adjunto*[6] de $\boldsymbol{L}$. Esta operación «adjunta», aplicada a un operador lineal $\boldsymbol{L}$, es precisamente la operación de «conjugación hermítica» * que se ha considerado en §13.9, en el caso de la dimensión finita. El conjugado complejo del número complejo $\langle\phi|\boldsymbol{L}|\psi\rangle$ es el número complejo $\langle\psi|\boldsymbol{L}^*|\phi\rangle$.

22.4. Evolución unitaria: Schrödinger y Heisenberg

Ahora estamos en una buena posición para considerar la naturaleza «unitaria» de la evolución de Schrödinger. Ya hemos visto en §22.3 que esta evolución es lineal, de modo que todo lo que tenemos que establecer es que conserva el producto escalar $\langle\phi|\psi\rangle$ entre dos elementos $|\phi\rangle$ y $|\psi\rangle$ de $\mathbf{H}$. Es decir, $\langle\phi|\psi\rangle$ es *constante* en el tiempo: $\mathrm{d}\langle\phi|\psi\rangle/\mathrm{d}t = 0$. (A partir de lo que se ha dicho antes, la conservación de la norma y la conservación del producto escalar son requisitos equivalentes.) Básicamente, lo que necesitamos de nuestro hamiltoniano cuántico $\mathcal{H}$ es (i) que nos mantenga en el espacio de Hilbert, y (ii) que sea hermítico. Estos son requisitos muy mínimos, y serán satisfechos por cualquier sugerencia razonable para un hamiltoniano. Su naturaleza hermítica, por ejemplo, es un requisito natural para garantizar que sus autovalores —los valores posibles para la energía del sistema— sean números reales. También es habitual exigir que $\mathcal{H}$ sea *definido positivo*, lo que significa que $\langle\psi|\mathcal{H}|\psi\rangle > 0$ para todo $|\psi\rangle$ no nulo, de donde se sigue que todos los autovalores de $\mathcal{H}$ (valores de la energía) son positivos —aunque esto no es necesario para la naturaleza unitaria de la evolución—. Obtenemos rápidamente (utilizando la propiedad de Leibniz para la derivada de un producto; véase §6.5 y las propiedades anteriores)

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\langle\phi|\psi\rangle &= \left\langle\frac{\mathrm{d}}{\mathrm{d}t}\phi\middle|\psi\right\rangle + \left\langle\phi\middle|\frac{\mathrm{d}}{\mathrm{d}t}\psi\right\rangle\\ &= \langle -\mathrm{i}\hbar^{-1}\mathcal{H}\phi|\psi\rangle + \langle\phi| -\mathrm{i}\hbar^{-1}\mathcal{H}\psi\rangle\\ &= \mathrm{i}\hbar^{-1}\langle\phi|\mathcal{H}|\psi\rangle - \mathrm{i}\hbar^{-1}\langle\phi|\mathcal{H}|\psi\rangle = 0,\end{aligned}$$

lo que demuestra que los productos escalares se conservan, i.e., la evolución de Schrödinger es unitaria.[22.7] El mismo argumento se aplica a otros operadores hermíticos, tales como los generadores de traslaciones o rotaciones espaciales, lo que demuestra que estas operaciones corresponden también a transformaciones unitarias de $\mathbf{H}$.

La ecuación anterior demuestra que la *tasa* de cambio de un producto escalar $\langle\phi|\psi\rangle$ es cero. De esto se sigue que $\langle\phi|\psi\rangle$ permanece in-

[22.7] Desarrolle este argumento con más detalle. ¿Puede explicar por qué cabría esperar que la propiedad de Leibniz sea válida para un producto escalar en el espacio de Hilbert?

variable para *todo* instante, mientras cada uno de los $|\phi\rangle$ y $|\psi\rangle$ experimentan individualmente la evolución de Schrödinger de acuerdo con el mismo $\mathcal{H}$. Supongamos que tenemos estados cuánticos $|\phi\rangle$ y $|\psi\rangle$ en el instante $t = 0$, y consideremos que evolucionan mediante la receta de Schrödinger hasta un tiempo posterior, T, en el que los estados se convierten respectivamente en $|\phi_T\rangle$ y $|\psi_T\rangle$:

$$|\phi\rangle \rightsquigarrow |\phi_T\rangle \text{ y } |\psi\rangle \rightsquigarrow |\psi_T\rangle$$

(utilizando la notación de §22.2). Entonces,

$$\langle\phi|\,\psi\rangle = \langle\phi_T|\,\psi_T\rangle.$$

Esto nos dice que la acción lineal de la evolución de Schrödinger, en el espacio de Hilbert **H**, tomada desde $t = 0$ hasta cierto instante definido $t = T$, es *unitaria*, en el sentido de que existe un operador $\boldsymbol{U}_T$ que efectúa esta transformación

$$|\phi_T\rangle = \boldsymbol{U}_T|\,\phi\rangle,\ |\psi_T\rangle = \boldsymbol{U}_T|\,\psi\rangle,\ \text{etc.}$$

y este operador $\boldsymbol{U}_T$ es *unitario* en el sentido de §13.9, a saber, que su inverso es igual a su adjunto:

$$\boldsymbol{U}_T^{-1} = \boldsymbol{U}_T^*,\ \text{i.e.,}\ \boldsymbol{U}_T\boldsymbol{U}_T^* = \boldsymbol{U}_T^*\boldsymbol{U}_T = \boldsymbol{I}.$$

Aquí $\boldsymbol{I}$ es el operador identidad en **H**. (Véase §13.9 para la demostración de esta propiedad de $\boldsymbol{U}_T$.)

Como se ha mencionado en §22.1, hay otras maneras de representar la evolución de un sistema cuántico, y lo que se denomina la *imagen de Heisenberg* es la alternativa más familiar. En la imagen de Heisenberg, el «estado» del sistema se considera *constante* en el tiempo, y son las variables dinámicas las que asumen la evolución temporal en su lugar. El lector podría preguntarse cómo puede considerarse «invariable» el estado cuántico ¡aun cuando puede estar produciéndose algún cambio físico real en el sistema cuántico! Cierto; pero pasar de la imagen de Schrödinger a la imagen de Heisenberg es en realidad solo cuestión de redefinir nuestros símbolos.

Consideremos, en primer lugar, la imagen de Schrödinger ordinaria que hemos adoptado hasta ahora. Tenemos cierto estado cuántico $|\psi\rangle$ en el instante $t = 0$, que consideramos que evoluciona según la receta

de Schrödinger definida por un hamiltoniano cuántico dado $\mathcal{H}$, de modo que en algún instante posterior T el estado es $|\psi_T\rangle$:

$$|\psi\rangle \rightsquigarrow |\psi_T\rangle = \boldsymbol{U}_T|\psi\rangle.$$

Recordemos que la acción de $\boldsymbol{U}_T$ se aplica linealmente al espacio de Hilbert *entero* **H**, de modo que cualquier otro estado $|\phi\rangle$ experimentaría una evolución correspondiente $|\phi\rangle \rightsquigarrow |\phi_T\rangle = \boldsymbol{U}_T|\phi\rangle$, con el mismo $\boldsymbol{U}_T$ que se ha utilizado para $|\psi\rangle$. En la imagen de Heisenberg, simplemente *redefinimos* el estado, en el instante T, como

$$|\psi\rangle_{\text{H}} = \boldsymbol{U}_T^{-1}|\psi\rangle = \boldsymbol{U}_T^*|\psi\rangle.$$

Es evidente que este «estado de Heisenberg» $|\psi\rangle_{\text{H}}$ no cambia (¡básicamente por definición!) con el paso del tiempo. Por otra parte, para que todos los procedimientos algebraicos continúen como antes, de modo que los autovalores (parámetros físicos medidos) sean los mismos que en la imagen de Schrödinger, exigimos que las variables dinámicas evolucionen para compensar. Así pues, cualquier operador lineal $\mathbf{Q}$ (en **H**) debe ser reemplazado por su versión de Heisenberg

$$\mathbf{Q}_{\text{H}} = \boldsymbol{U}_T^{-1}\mathbf{Q}\boldsymbol{U}_T = \boldsymbol{U}_T^*\mathbf{Q}\boldsymbol{U}_T.$$

Se sigue directamente que la versión de Heisenberg de cualquier autovalor o de cualquier producto escalar es la misma que la versión de Schrödinger.[22.8] La evolución de Heisenberg se aplica ahora a los operadores $\mathbf{Q}$ (que se suponían constantes en la imagen de Schrödinger) y en particular a las variables dinámicas. Encontramos que[22.9]

$$\mathrm{i}\hbar\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{Q}_{\text{H}} = (\mathcal{H}\mathbf{Q}_{\text{H}} - \mathbf{Q}_{\text{H}}\mathcal{H}),$$

que son las *ecuaciones de movimiento de Heisenberg*. (Nótese que una consecuencia obvia de esto es la conservación de la energía, dada cuando $\mathbf{Q}_{\text{H}} = \mathcal{H}$.)

El lector puede preguntarse qué hemos ganado reexpresando las cosas de este modo. En algunos contextos existen ventajas técnicas en la imagen de Heisenberg, pero la imagen de Heisenberg no mejora las cosas con respecto a los enigmas interpretativos de la mecánica cuánti-

[22.8] Explique todo esto en detalle.
[22.9] Trate de confirmarlo.

ca. El problema de los «saltos cuánticos» no ha desaparecido, ¡pero ahora podemos elegir entre echar la culpa al estado, permitiendo que $|\psi\rangle_H$ «salte» a algún otro cuando opera **R**, o, en su lugar, considerar que son las variables dinámicas de Heisenberg las que dan el «salto»! Por mi parte, encuentro que estas cuestiones del «salto» se hacen más oscuras en la imagen de Heisenberg, sin que resuelvan nada.

En la imagen de Schrödinger tenemos al menos un vector de estado que evoluciona ¡y que tiene una posibilidad de ofrecernos un atisbo del aspecto que podría tener la «realidad cuántica»! La imagen de Heisenberg no parece tener muchas posibilidades de hacer esto, puesto que su vector de estado se queda inmóvil incluso si se está produciendo una acción física. Además, la evolución de las variables dinámicas no puede representar el cambio en un sistema físico específico, porque no describe sistemas específicos en absoluto, sino más bien las preguntas que pueden hacerse a un sistema, tales como «¿cuál es tu posición?», etc.

La razón de tener estas dos imágenes diferentes es en buena medida histórica. Heisenberg fue el primero en presentar su esquema, en julio de 1925, y Schrödinger presentó su propuesta medio año más tarde, en enero de 1926, advirtiendo la equivalencia entre los dos esquemas inmediatamente después. Fue Max Born el primero en reconocer la interpretación probabilista para el módulo al cuadrado $|\psi|^2$ de la función de onda de Schrödinger (§21.9) en junio de 1926. El propio Schrödinger había tratado de mantener una imagen de ψ más similar a un «campo clásico». El armazón general de operadores de la mecánica cuántica salió del trabajo de Heisenberg, Born y Pascual Jordan, y fue completamente formulado por Dirac, y descrito en detalle en su muy influyente libro *Los principios de la mecánica cuántica*, publicado por primera vez en 1930.[7]

Por supuesto, puede darse el caso de que cuando se introduce un cambio en la teoría cuántica haya buenas razones para preferir un formalismo a otro y se rompa la equivalencia entre ambos. Este es aproximadamente el caso incluso con la teoría cuántica de campos (véase el capítulo 26), que trata de unir la teoría cuántica y la teoría de la relatividad (especial) en un esquema consistente. Dirac ha dado algunos argumentos para preferir[8] la imagen de Heisenberg en este caso. Sin embargo, ni la imagen de Heisenberg ni la de Schrödinger son relativísticamente invariantes, y en este contexto se prefiere a veces una «imagen de interacción» híbrida.[9]

22.5. «Observables» cuánticos

Consideremos ahora cómo debe representarse en el formalismo una medida de un sistema cuántico. Como se ha señalado en §22.1, los ejemplos de medidas de posición y momento dados en el capítulo 21 son ilustrativos de lo que sucede en el caso general de una medida cuántica. Una cualidad «medible» de un sistema cuántico vendrá representada por cierto tipo de operador **Q**, llamado un *observable*, y este operador podrá aplicarse al estado cuántico. Las variables dinámicas (la posición o el momento, pongamos por caso) serán ejemplos de observables.[10] La teoría exige que un observable **Q** esté representado por un operador lineal (como sucede con los ejemplos de los operadores de posición y momento), de modo que su acción sobre el espacio **H** consistirá en efectuar una transformación lineal de **H** —aunque posiblemente una transformación singular (§13.3). Decimos que el estado ψ tiene un *valor definido* para el observable **Q** si ψ es un *autoestado* de **Q**, y el correspondiente autovalor q será dicho valor definido.[11] Esta es precisamente la terminología que ya hemos encontrado en §§21.5,10,11 para posición y momento.

En mecánica cuántica convencional se exige normalmente que todos los autovalores tengan que ser números reales. Se puede garantizar esto (suponiendo que los autovectores son normalizables) exigiendo que **Q** sea hermítico en el sentido de que **Q** es igual a su adjunto $\mathbf{Q}^*$:[22.10]

$$\mathbf{Q}^* = \mathbf{Q},$$

En mi opinión, este requisito de hermiticidad sobre un observable **Q** es un requisito irrazonablemente fuerte, puesto que los números complejos se utilizan a menudo en física clásica, como sucede en la re-

[22.10] Demuestre que cualquier autovalor de un operador hermítico **Q** es un número *real*.

presentación como esfera de Riemann de la esfera celeste (§18.5), y en muchas discusiones estándar del oscilador armónico (§22.13), etc.[12] Un requisito esencial para un observable es que sus autovectores correspondientes a autovalores distintos sean mutuamente ortogonales. Esta es una propiedad característica de los operadores que se conocen como «normales». Un operador *normal* $\mathbf{Q}$ es un operador que conmuta con su adjunto:

$$\mathbf{Q}^*\mathbf{Q} = \mathbf{Q}\mathbf{Q}^*,$$

y cualquier par de autovectores (normalizables) de un operador normal $\mathbf{Q}$, correspondientes a autovalores diferentes, deben ser, de hecho, ortogonales.[22.11] Puesto que acepto con gusto que los resultados de las medidas (autovalores) sean números complejos, aunque insistiendo en el requisito estándar de ortogonalidad entre los estados alternativos que pueden resultar de una medida, exigiré que mis «observables» cuánticos sean operadores lineales normales, en lugar del requisito convencional de que sean hermíticos.

Debería hacer aquí un comentario sobre un requisito adicional para los observables cuánticos, y es que sus autovectores abarcan el espacio de Hilbert $\mathbf{H}$ (de modo que cualquier elemento de $\mathbf{H}$ puede expresarse linealmente en términos de estos autovectores). En el caso de la dimensión finita, esta propiedad es una consecuencia matemática de la naturaleza hermítica (o normal) de $\mathbf{Q}$. Pero para un $\mathbf{H}$ de dimensión infinita la necesitamos como una hipótesis independiente para cualquier $\mathbf{Q}$ que vaya a tener un papel como observable cuántico. Un $\mathbf{Q}$ hermítico con esta propiedad se denomina *autoadjunto*.

El requisito de ortogonalidad para un observable cuántico es importante para el proceso de medida cuántica. De acuerdo con las reglas de la mecánica cuántica, el resultado de una medida, correspondiente a cierto operador $\mathbf{Q}$, será siempre uno de sus autoestados: este es el «salto» del estado cuántico que ocurre con el proceso $\mathbf{R}$ (véase §22.1). De acuerdo con $\mathbf{R}$, cualquiera que sea el estado del sistema antes de la medida, salta a uno de los autoestados de $\mathbf{Q}$ en cuanto es medido. Después de la medida, el estado adquiere un valor definido para el observable $\mathbf{Q}$, a saber, el correspondiente autovalor q. Así pues, para cada uno de los diferentes resultados posibles de la medida del observable $\mathbf{Q}$ —es decir, para cada uno de los diferentes autovalores $q_1, q_2, q_3, \ldots$— obtenemos uno de un conjunto de estados resultantes alternativos, todos los cuales son mutuamente ortogonales.

¿Por qué es esto importante? Enseguida veremos cuáles son las reglas cuánticas para calcular las probabilidades de cada uno de estos resultados alternativos. Una consecuencia de estas reglas será que la probabilidad de que, como resultado de la medida, un estado salte a un estado ortogonal es siempre cero. En consecuencia, si se repite la medida definida por un observable $\mathbf{Q}$, entonces la segunda medida dará el mismo autovalor —i.e., el mismo resultado de la medida— que dio la primera medida. Dar un resultado diferente implicaría dar un salto desde un estado a otro ortogonal, lo que las reglas de la probabilidad no permiten. Pero esta feliz conclusión depende de la ortogonalidad de los autoestados de $\mathbf{Q}$ para diferentes autovalores, y por esto es por lo que exigimos que $\mathbf{Q}$ sea un operador normal.

Vayamos ahora a la asignación de probabilidades a los diferentes autoestados alternativos del observable $\mathbf{Q}$ cuando se le presenta el estado ψ que está siendo «observado». Una característica notable del proceso $\mathbf{R}$ cuántico es que la probabilidad mecanocuántica depende solo de los estados antes y después de la medida, y no de cualquier otro aspecto del observable $\mathbf{Q}$ (tal como el autovalor medido, por ejemplo). La regla es que la probabilidad de que el estado salte de $|\psi\rangle$ al autoestado $|\phi\rangle$ de $\mathbf{Q}$, viene dada por

$$|\langle\psi|\phi\rangle|^2,$$

suponiendo que $|\psi\rangle$ y $|\phi\rangle$ están normalizados ($\||\psi\rangle\| = 1 = \||\phi\rangle\|$). Si no es así, tenemos que dividir lo anterior por $\||\psi\rangle\|$ y por $\||\phi\rangle\|$ antes de obtener la probabilidad. Podemos escribir esta probabilidad, para estados no normalizados, de la elegante forma

$$\frac{\langle\phi|\psi\rangle\,\langle\psi|\phi\rangle}{\langle\psi|\psi\rangle\,\langle\phi|\phi\rangle}.$$

[22.11] Trate de demostrarlo. *Sugerencia*: Considerando la siguiente expresión $\langle\psi|(\mathbf{Q}^* - \bar{\lambda}\mathbf{I})(\mathbf{Q} - \lambda\mathbf{I})|\psi\rangle$, demuestre primero que si $\mathbf{Q}|\psi\rangle = \lambda|\psi\rangle$, entonces $\mathbf{Q}^*|\psi\rangle = \bar{\lambda}|\psi\rangle$.

Esto es siempre un número real comprendido entre 0 y 1, que solo toma el valor 1 si los estados son proporcionales.[22.12] Recordemos de la exposición anterior que la evolución de Schrödinger conserva los productos escalares $\langle\phi|\psi\rangle$. Esta es una importante relación de compatibilidad entre los procesos **U** y **R** y expresa el hecho de que, a pesar de ser mutuamente incompatibles, **U** y **R** «encajan» limpiamente uno con otro. Vemos que un estado nunca salta directamente, en una medida, a un estado ortogonal, porque $\langle\phi|\psi\rangle = 0$ implica que la probabilidad para ello sería nula.

En una superposición cuántica entre estados ortogonales y normalizados ψ y ϕ, por ejemplo $w\psi + z\phi$, los factores de peso complejos, w y z, se suelen denominar *amplitudes*, o «amplitudes de probabilidad». En este caso, un experimento preparado para distinguir ψ de ϕ en el estado $w\psi + z\phi$ daría ψ con probabilidad $\bar{w}w = |w|^2$ y daría ϕ con probabilidad $\bar{z}z = |z|^2$, i.e., tomamos los módulos al cuadrado de las amplitudes para obtener las probabilidades. Un comentario similar se aplica a superposiciones de más de dos estados.

Una propiedad útil de un operador normal **Q** (suponiendo que sus autovectores abarcan la totalidad de **H**) es que siempre posee una familia de autoestados que constituyen una base ortonormal para el espacio de Hilbert. Una *base ortonormal* (compárese con §13.9) es un conjunto de elementos $\boldsymbol{e}_1, \boldsymbol{e}_2, \boldsymbol{e}_3, \ldots$ de **H** tales que

$$\langle \mathbf{e}_i | \mathbf{e}_j \rangle = \delta_{ij}$$

(siendo δ_{ij} la delta de Kronecker) y donde cada elemento ψ de **H** puede expresarse como

$$\psi = z_1\boldsymbol{e}_1 + z_2\boldsymbol{e}_2 + z_3\boldsymbol{e}_3 + \ldots,$$

donde $z_1, z_2, z_3, \ldots$ son «coordenadas cartesianas» complejas para ψ). Esto es similar a la expresión de una función de onda general, para una simple partícula sin estructura, como combinación lineal continua de estados de momento (como se consigue utilizando una transformada de Fourier) o de estados de posición (utilizando $\psi(\boldsymbol{x}) = \int \psi(\boldsymbol{X})\delta(\boldsymbol{x} - \boldsymbol{X})\mathrm{d}^3\boldsymbol{X}$),

[22.12] Demuéstrelo a partir de las propiedades algebraicas de $\langle|\rangle$, mediante los métodos utilizados en el ejercicio [22.2].

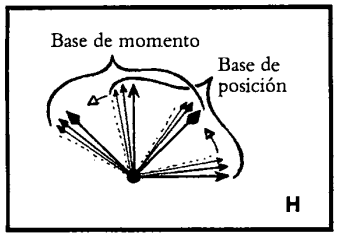


Fig. 22.3. Pasar de una representación de posición a una representación de momento es simplemente cambiar la base en el espacio de Hilbert **H** (aunque técnicamente ni los estados de posición ni los de momento, al ser no normalizables, pertenecen realmente a **H**).

puesto que los estados de momento y posición son los autoestados de los operadores momento y posición **p** y **x**, respectivamente. Pasar de la representación de posiciones a la representación de momentos equivale a un cambio de base en el espacio de Hilbert **H** (véase la Fig. 22.3). Sin embargo, ni los estados de momento ni los estados de posición forman realmente una base en sentido técnico, porque no son normalizables ¡y desde luego no pertenecen realmente a **H**! (La mecánica cuántica está llena de cuestiones irritantes de este tipo. Tal como están las cosas, se puede o bien pasar decididamente por encima de tales sutilezas matemáticas e incluso pretender que los estados de posición y estados de momento son realmente estados, o por el contrario, uno puede pasarse todo el tiempo insistiendo en obtener las matemáticas correctas, en cuyo caso existe el peligro opuesto de quedarse atrapado en un *rigor mortis*. Estoy haciendo lo posible por mantenerme en un término medio, ¡pero no estoy en absoluto seguro de cuál es la respuesta correcta para hacer progresos en el tema!)

22.6. Medidas sí/no; proyectores

En el caso de operadores tales como los operadores de posición o de momento, cuyos autoestados no son normalizables, obtenemos una probabilidad cero de encontrar una partícula en uno de tales estados. Esta es realmente la respuesta «correcta», ya que la probabilidad de que la posición o el momento tengan un valor concreto con precisión exacta sería realmente cero (al ser la posición y el momento parámetros continuos). Esto no nos ayuda mucho, de modo que sería preferible utilizar otros tipos de observables, tales como el que plantea la si-

guiente pregunta: «¿Está la posición dentro de cierto intervalo de valores?», y una pregunta similar podría plantearse para el momento (o para cualquier otro observable continuo). Este tipo de preguntas SÍ/NO pueden ser incorporadas en el formalismo cuántico asignando, por ejemplo, el autovalor 1 a la respuesta sí y el autovalor 0 a NO. Un observable de este tipo se describe mediante lo que se denomina un *proyector*.

Un proyector $\boldsymbol{E}$ tiene la propiedad de que es autoadjunto y su cuadrado es igual a sí mismo[22.13]

$$\boldsymbol{E}^2 = \boldsymbol{E} = \boldsymbol{E}^*.$$

Tales objetos proporcionan el tipo más primitivo de medida, y para muchos propósitos las cuestiones que plantea la «medida» en mecánica cuántica se discuten mejor en términos de tales operadores. Hay, sin embargo, una cuestión concreta que resulta especialmente importante cuando se lleva a cabo una de estas medidas SÍ/NO, porque (en más de 2 dimensiones) tales operadores son (completamente) *degenerados*.

Decimos que $\mathbf{Q}$ es degenerado con respecto a algún valor propio q si el espacio de autovectores correspondiente a q es más que 1-dimensional, i.e., si existen varios autovectores de $\mathbf{Q}$, no proporcionales, correspondientes al mismo valor propio q. El subespacio lineal de $\mathbf{H}$ que consiste en todos los autovectores correspondientes al mismo autovalor q se conoce como el *espacio propio* de $\mathbf{Q}$ correspondiente a q (§13.5). El subespacio lineal entero de $\mathbf{H}$ consistente en todos los autovectores correspondientes al mismo autovalor q se conoce como el autoespacio de $\mathbf{Q}$ correspondiente a q. En tales casos, la obtención del «resultado» de la medida (i.e., la determinación del autovalor) no nos dice por sí solo a qué estado se supone que «salta» el vector de estado. La cuestión se resuelve mediante el denominado *postulado de proyección* que afirma que el estado $|\psi\rangle$ que se está sometiendo a la medida se proyecta ortogonalmente en el autoespacio[13] de $\mathbf{Q}$ correspondiente a q. De hecho, el término «postulado de proyección» es con frecuencia utilizado simplemente para el procedimiento mecanocuántico estándar de §22.1

[22.13] (Demuestre que si un observable $\mathbf{Q}$ satisface una ecuación polinómica, entonces cada uno de sus autovalores satisface la misma ecuación.

(como hizo explícito Von Neumann),[14] por el que, como resultado de la medida de un observable $\mathbf{Q}$, el estado salta a un autoestado de $\mathbf{Q}$, correspondiente al autovalor que da la medida. En esta sección y en la siguiente subrayo la importancia del aspecto de proyección de este postulado en el caso de autovalores degenerados.[15]

Una de las mejores formas de expresar esta proyección es mediante el uso de un proyector adecuado $\boldsymbol{E}$, a saber, aquel cuyo autoespacio correspondiente a su autovalor sí 1 es idéntico al autoespacio de $\mathbf{Q}$ correspondiente a q. (Esto puede hacerse siempre; $\boldsymbol{E}$ está planteando simplemente una pregunta más básica que la que plantea $\mathbf{Q}$, a saber: «¿Es q el resultado de la medida $\mathbf{Q}$?».) Entonces, lo que el postulado de proyección afirma es que el resultado de la medida (de o bien $\mathbf{Q}$ con el resultado q, o bien $\boldsymbol{E}$ con el resultado 1) es que

$$|\psi\rangle \text{ salta a } \boldsymbol{E}|\psi\rangle.$$

En esto no me he preocupado de las normalizaciones (y no tenemos necesidad de preocuparnos si no queremos hacerlo). Si pedimos que el estado resultante sea normalizado, podemos considerar que $|\psi\rangle$ salta a $\boldsymbol{E}|\psi\rangle\langle\psi|\boldsymbol{E}|\psi\rangle^{-1/2}$. No obstante, en mis descripciones encontraré más conveniente no tener que normalizar mis estados. Esto hace que muchas de las fórmulas parezcan más simples que si lo hiciera.

En la Fig. 22.4 he indicado la naturaleza geométrica, dentro del espacio de Hilbert $\mathbf{H}$, del postulado de proyección. Nótese que si reemplazamos el proyector $\boldsymbol{E}$ por $\boldsymbol{I} - \boldsymbol{E}$ (también un proyector), encontramos que los espacios propios sí y NO se intercambian. (Aquí $\boldsymbol{I}$ es el operador identidad en $\mathbf{H}$.) Así pues, si la medida obtiene 0 para la medida $\boldsymbol{E}$, entonces $|\psi\rangle$ salta a $(\boldsymbol{I} - \boldsymbol{E})|\psi\rangle (= |\psi\rangle - \boldsymbol{E}|\psi\rangle)$. Nótese que $|\psi\rangle$ es la suma de los dos estados $\boldsymbol{E}|\psi\rangle$ e $(\boldsymbol{I} - \boldsymbol{E})|\psi\rangle$, que son mutuamente ortogonales,[22.14] y la medida $\boldsymbol{E}$ decide entre los dos, sí para el primero y NO para el segundo:

$$|\psi\rangle = \boldsymbol{E}|\psi\rangle + (\boldsymbol{I} - \boldsymbol{E})|\psi\rangle.$$

Hay una forma geométrica directa de expresar las probabilidades de estas dos alternativas, que está dada por el factor en que se reduce la

[22.14] Demuéstrelo.

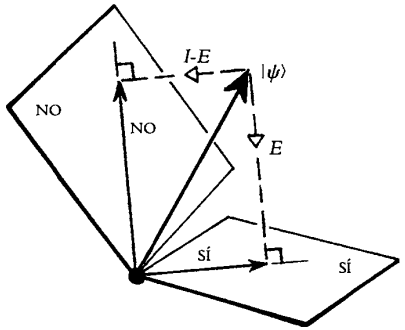


Fig. 22.4. La naturaleza geométrica, dentro de **H**, del postulado de proyección. Se indican los autoespacios del proyector $\boldsymbol{E}$, donde el plano horizontal representa el autovalor 1 (SÍ) y el plano vertical, el autovalor 0 (NO). La imagen ilustra la descomposición $|\psi\rangle = \boldsymbol{E}|\psi\rangle + (\boldsymbol{I}-\boldsymbol{E})|\psi\rangle$ de $|\psi\rangle$ en dos partes ortogonales, donde $E|\psi\rangle$ es la proyección de $|\psi\rangle$ dentro del espacio SÍ (el resultado de una medida que da SÍ) e $(\boldsymbol{I}-\boldsymbol{E})|\psi\rangle$, la proyección dentro del espacio NO (a partir del resultado NO). La probabilidad en cada caso viene dada por el factor de proporcionalidad exacto por el que la longitud (hermítica) al cuadrado $|\psi|^2$ se reduce en la proyección (vectores de estado no normalizados).

«norma» (longitud al cuadrado) del estado en cada proyección.[22.15] ¡Este hecho geométrico simple queda oscurecido si insistimos en normalizar nuestros estados!

22.7. Medidas nulas, helicidad

Algunos físicos han expresado dudas sobre el postulado de proyección (o que es «innecesario» o «inobservable»). La dificultad reside en que quizá no tenemos medios de determinar qué ha sido realmente del estado después de la medida, quizá porque el propio proceso de medida ha hecho que la entidad observada se entrelace con el aparato de medida, de modo que el estado de la entidad que está siendo observada ya no puede considerarse independiente. En realidad, eso podría ser a veces una cuestión que complica las cosas, pero hay en verdad circunstancias en las que el postulado de proyección describe manifiestamente una medida (degenerada, si es necesario). El caso más claro de esto ocurre con lo que se denomina una medida *nula* (o *libre de interacción*). Este fascinante tipo de situación interesa por sí misma, e ilustra uno de los aspectos más extraños del comportamiento mecanocuántico. En consecuencia, vale la pena que nos fijemos en uno o dos ejemplos.

[22.15] ¿Por qué?

Consideremos una situación del tipo que se ha examinado en §21.7, en la que un único fotón es dirigido hacia un divisor de haz donde su estado es parcialmente reflejado y parcialmente transmitido. Después del choque, el estado es una suma de estas dos partes ortogonales, la parte transmitida $|\tau\rangle$ y la parte reflejada $|\rho\rangle$ (donde, para hacer de ello una bonita suma directa, absorbemos cualquier factor de fase relativa en las definiciones de $|\tau\rangle$ y $|\rho\rangle$ y no insistimos en la normalización):

$$|\psi\rangle = |\tau\rangle + |\rho\rangle$$

(véase la Fig. 22.5). Supongamos que en el camino del haz transmitido se coloca un detector que, a efectos del argumento, suponemos que tiene una eficiencia de un ciento por ciento. Además, la fuente de fotones debe ser tal que cada suceso de emisión de fotón es registrado (en la fuente) con un ciento por ciento de eficiencia. (Estas son evidentemente idealizaciones; en un experimento real sería difícil acercarse a tales eficiencias. De todas formas, son idealizaciones razonables para ilustrar cómo funciona la mecánica cuántica). Si observamos que en algunas ocasiones la fuente ha emitido un fotón pero el detector no lo ha recibido, entonces podemos estar seguros de que en dichas ocasiones el fotón ha «seguido la otra vía» y su estado es por consiguiente el estado reflejado $|\rho\rangle$. El hecho notable es que la medida de *no* detección del fotón ha provocado que el estado del fotón sufra un salto cuántico (desde la superposición $|\psi\rangle$ al estado reflejado $|\rho\rangle$), ¡pese al hecho de que el fotón no ha interaccionado en absoluto con el aparato de medida! Este es un ejemplo de una medida *nula*.

Un uso impresionante de este tipo de cosas ha sido sugerido por Avshalom Elitzur y Lev Vaidman.[16] Supongamos que nuestro divisor de haz es parte de un interferómetro de Mach-Zehnder (recordemos la parte final de mi experimento mental astronómico descrito en §21.7; véase la Fig. 21.9), aunque no sabemos si un detector C ha sido

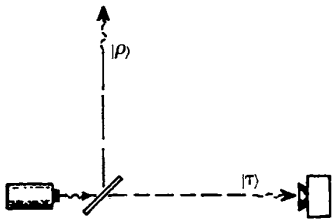


Fig. 22.5. Medida nula que requiere el postulado de proyección. Un único fotón se dirige hacia un divisor de haz. El estado resultante $|\psi\rangle$, al ser parcialmente reflejado y parcialmente transmitido, es la suma $|\psi\rangle = |\tau\rangle + |\rho\rangle$ de la parte transmitida $|\tau\rangle$ y la parte reflejada $|\rho\rangle$ (absorbiendo cualquier factor de fase relativa en las definiciones y sin insistir en la normalización). Si se encuentra que la fuente ha emitido un fotón pero el detector no lo ha recibido, entonces sabemos que el fotón está en el estado $|\rho\rangle$ incluso si no ha interaccionado en absoluto con el detector.

colocado o no en el haz transmitido en el primer divisor de haz. Supongamos que el detector C detona una bomba, de modo que la bomba explotaría si C recibiera el fotón. Hay dos detectores finales, A y B, y sabemos (de §21.7) que solo A, y no B, puede registrar la recepción del fotón si C está ausente (véase la Fig. 22.6). Queremos determinar la presencia de C (y de la bomba) en una situación sin perderla realmente en una explosión. Esto se consigue en aquellas circunstancias en las que el detector B registra realmente el fotón; ¡pues ello solo puede ocurrir si el detector C hace la medida de que *no* recibe el fotón! En efecto, en este caso el fotón ha tomado en realidad la otra ruta, de modo que ahora cada uno de los detectores A y B tiene una probabilidad 1/2 de recibir el fotón (porque ahora no hay interferencia entre los dos haces), mientras que en ausencia de C, solo A puede recibir el fotón.[17]

En los ejemplos que acabamos de ver no hay degeneración, de modo que no se presenta realmente la cuestión que se ha abordado antes acerca de que el mero resultado de la medida es insuficiente para determinar el estado al que «salta» el sistema. Recordemos de §22.6 que necesitamos el uso adecuado del postulado de proyección. En consecuencia, introduzcamos otro grado de libertad, y es conveniente hacerlo teniendo en cuenta el fenómeno de polarización del fotón. Este es un ejemplo de la cualidad física mencionada antes, del espín mecanocuántico. Volveré a las ideas del espín de forma más detallada en §§22.8-11. Por el momento necesitamos solo unas pocas propie-

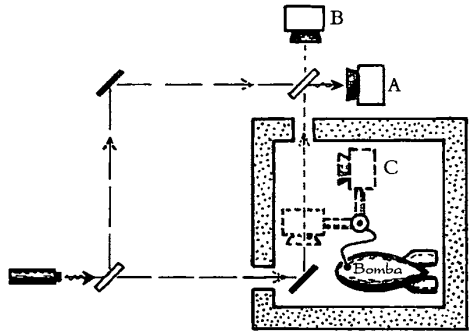


Fig. 22.6. La prueba de la bomba de Elitzur-Vaidman. Un detector C, unido a una bomba, puede o no estar insertado en un interferómetro de tipo Mach-Zehnder (véase la Fig. 21.9). (Los rectángulos blancos delgados representan divisores de haz; los negros, espejos.) Las longitudes de los brazos dentro del interferómetro son iguales, de modo que un fotón emitido por la fuente debe llegar al detector A cuando quiera que C no esté insertado. En el caso de que el detector B reciba el fotón (sin que explote la bomba), sabemos que C está en su lugar en el haz, incluso si no ha encontrado el fotón.

dades básicas del espín, pero únicamente en el caso de una partícula sin masa. Los fotones son, de hecho, partículas que poseen espín, pero, al no tener masa, su espín se comporta de una forma ligeramente diferente del espín más habitual de una partícula masiva (por ejemplo, un electrón o un protón) al que llegaremos en §§22.8-10. Debemos considerar que un fotón (u otra partícula sin masa) está girando necesariamente alrededor de su dirección de movimiento; véase la Fig. 22.7.

La cantidad $|s|$ de este espín es siempre la misma para un tipo dado de partícula sin masa, pero el espín puede ser dextrógiro ($s > 0$) o levógiro ($s < 0$) en torno a la dirección de movimiento. Además, de acuerdo con los principios generales de la mecánica cuántica, el estado del espín puede ser cualquier combinación lineal (cuántica) de los dos. La magnitud s se denomina *helicidad* de la partícula sin masa (§22.12), y su valor tiene que ser siempre entero o semientero (o, introduciendo las unidades apropiadas, deberíamos decir que la helicidad es un múltiplo entero de $\frac{1}{2}\hbar$). Se dice que una partícula sin masa tiene *espín*

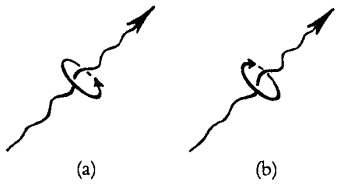


Fig. 22.7. Una partícula sin masa, tal como un fotón, solo puede girar alrededor de su dirección de movimiento. La magnitud $|s|$ de este giro es siempre la misma para un tipo dado de partícula sin masa, pero si la helicidad s es no nula (como en el caso de un fotón), entonces el giro puede ser o bien (a) dextrógiro ($s > 0$, helicidad positiva), o bien (b) levógiro ($s < 0$, helicidad negativa). Para un fotón tenemos $|s| = 1$ (en unidades de $\hbar$), que da los dos casos $s = 1$, para la polarización circular dextrógira, y $s = -1$, para la polarización circular levógira. Por el principio de superposición cuántico, podemos formar combinaciones lineales complejas de estos, que dan los otros estados posibles de polarización del fotón, como se muestra en las Figs. 21.12 y 21.13.

j si $|s| = j$ (o, con unidades, $|s| = j\hbar$). Un fotón tiene espín 1 (de modo que su helicidad es ± 1); un gravitón tiene espín 2 (helicidad ± 2). Los neutrinos tienen espín $1/2$, y si existen neutrinos sin masa,[18] tales neutrinos tendrán helicidad $-1/2$, y sus correspondientes antineutrinos tendrán helicidad $1/2$.

En el caso de un fotón, los estados de helicidad (estados de helicidad definida) son los estados de *polarización circular*, dextrógira para $s = 1$ y levógira para $s = -1$, respectivamente. Existen otros estados posibles para la polarización de un fotón, tales como la polarización plana, pero estos son simplemente combinaciones lineales de los estados dextrógiro y levógiro. Al final de §22.9 llegaré a la geometría de todo esto, pero por el momento no será necesario. Todo lo que necesitamos por ahora es un hecho particular acerca de cómo se comporta la polarización circular bajo reflexión. Estoy suponiendo que un fotón en un estado circularmente polarizado incide *perpendicularmente* sobre el divisor de haz (o sobre cualquier otro tipo de espejo que pudiéramos utilizar), de modo que el haz reflejado vuelve directamente en la dirección de la que procedía el fotón. El hecho que necesitamos es que el estado de polarización del fotón reflejado es entonces *opuesto* al del fotón emitido en la fuente, mientras que la parte transmitida tiene una

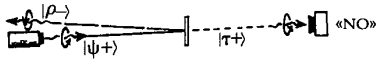


Fig. 22.8. Una vuelta al experimento de la Fig. 22.5, pero ahora el fotón incide casi perpendicularmente. La fuente emite un fotón dextrógiro. Una vez que el fotón ha encontrado el divisor de haz, su estado es $|\psi+\rangle = |\tau+\rangle + |\rho-\rangle$, donde los «+» y «–» dentro del ket se refieren al signo de la helicidad. Si el detector (insensible a la polarización) no registra la recepción del fotón, concluimos que el estado ha saltado (por la ausencia de detección) al estado levógiro reflejado $|\rho-\rangle$. Esto exige el postulado de proyección completo (para el punto de Lüders, véase la Fig. 22.9), porque hay una degeneración tanto en el caso NO (el 2-espacio generado por $|\rho+\rangle$ y $|\rho-\rangle$) como en el caso SÍ (el generado por $|\tau+\rangle$ y $|\tau-\rangle$). El estado de partida real $|\tau+\rangle + |\rho-\rangle$ es necesario para determinar dónde salta el estado por la medida (aquí una no detección).

polarización que es la *misma* que la del fotón emitido.[22.16] Si lo deseamos, podemos suponer que hay una minúscula inclinación en la dirección del haz inicial, de modo que el haz de fotones reflejados no vuelve a entrar en la fuente. Esto no afectará de forma importante a nuestras consideraciones.

Volvamos a nuestro experimento original de «medida nula» de la Fig. 22.5, pero ahora con el fotón incidiendo perpendicularmente, como en la Fig. 22.8. Supongamos que nuestra fuente puede sintonizarse de modo que emita sus fotones en un estado polarizado dextrógiro o levógiro. En una ocasión concreta emite un fotón dextrógiro (y toma nota de este hecho). Una vez que el fotón ha incidido en el divisor de haz, el estado del fotón es una combinación lineal (una *suma* con los convenios apropiados sobre factores de fase, como antes):

$$|\psi+\rangle = |\tau+\rangle + |\rho-\rangle,$$

donde el «+» o el «–» dentro del ket se refieren al signo de la helicidad. Coloquemos nuestro detector en el haz transmitido, como antes (y supongamos que es insensible a la polarización). Entonces si, como antes, la fuente registra que ha emitido el fotón dextrógiro pero el detector no lo registra, debe concluirse que el estado ha saltado (por la «no detección» por la fuente) al estado levógiro reflejado $|\rho-\rangle$. El punto que

[22.16] ¿Puede sugerir una sencilla razón para esto?

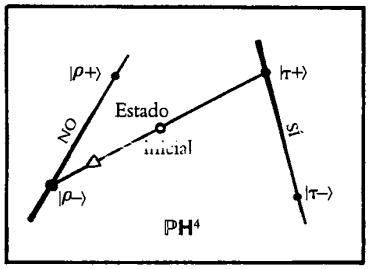


Fig. 22.9. Una descripción en el espacio de Hilbert proyectivo $\mathbb{P}\mathbf{H}^4$ (véase la Fig. 15.15) del postulado de proyección de la Fig. 22.4, para los estados de polarización del fotón en la Fig. 22.8. El estado inicial es $|\tau+\rangle + |\rho-\rangle$ indicado dentro de $\mathbb{P}\mathbf{H}^2$, y $|\tau+\rangle$, $|\tau-\rangle$, $|\rho+\rangle$ y $|\rho-\rangle$ llenan el espacio completo. La flecha triangular blanca muestra la proyección en $|\rho-\rangle$ (punto de Lüders), y está a lo largo de la línea que es la única transversal a las líneas sí y NO desde el punto inicial ($|\tau+\rangle + |\rho-\rangle$). La (no) detección nos diría meramente que el estado resultante yace en la línea NO, pero la elección del estado inicial rompe esta degeneración, según el postulado de proyección completo.

estoy señalando es que se requiere el postulado de proyección para descubrir la naturaleza de este estado resultante; véase la Fig. 22.9. La medida es de un carácter puramente sí/NO, porque el resultado es o bien «no detección» (NO), o bien «detección» (sí). Hay una degeneración para estas dos alternativas, porque el autoespacio de la respuesta NO es el 2-espacio generado por $|\rho+\rangle$ y $|\rho-\rangle$, y el autoespacio de la respuesta sí es el generado por $|\tau+\rangle$ y $|\tau-\rangle$. Puesto que el estado inicial es en este caso, $|\tau+\rangle + |\rho-\rangle$, el postulado de proyección[19] nos lleva correctamente a $|\rho-\rangle$, en el caso de NO, en lugar de a $|\rho+\rangle$ o a $|\rho+\rangle + |\rho-\rangle$ (o a cualquier otra combinación lineal de $|\rho+\rangle$ y $|\rho-\rangle$) en caso de un resultado de no detección.[20,[22.17]]

22.8. Espín y espinores

Este no es un experimento muy apasionante, pero ilustra un punto importante.Veremos algunas cosas mucho más notables en el capítulo 23.

[22.17] Explique con más detalle por qué la respuesta correcta está dada por «protección».

Pero, como preparación para ello, será conveniente decir algo más sobre el espín. Este se refiere, en el caso de una partícula masiva, al momento angular respecto a su centro de masas.[21] En §§21.1-5 hemos visto la importancia de la conservación de la masa-energía y la conservación del momento lineal como características, respectivamente, de la simetría de traslación temporal y la simetría de traslación espacial de nuestras leyes cuánticas. De modo análogo, la simetría rotacional da lugar a la conservación del momento *angular* (véanse también §18.7 y §20.6). En el caso de una partícula masiva, podemos imaginar que estamos en el sistema de reposo de la partícula, y entonces las rotaciones relevantes son aquellas que constituyen el grupo de rotación O(3) tomado respecto a la posición de la partícula en dicho sistema.

Del mismo modo que en mecánica cuántica una componente del momento lineal se representa como $-i\hbar$ multiplicado por el operador que genera traslaciones infinitesimales en la dirección de la coordenada de posición correspondiente (§§21.1,2), también hay una componente del momento angular que se representa por $i\hbar$ multiplicado por el generador de las rotaciones infinitesimales en torno al eje (espacial cartesiano) correspondiente. Las componentes del momento angular, en mecánica cuántica, remiten así al *álgebra* de las rotaciones infinitesimales (§§13.6-10), i.e., el álgebra de Lie del grupo de rotación O(3) o equivalentemente SO(3), ya que el álgebra de Lie no distingue entre los dos.

Puesto que SO(3) es no abeliano, no todos los elementos del álgebra de Lie conmutan; de hecho, los generadores de esta álgebra ℓ_1, ℓ_2, ℓ_3, las rotaciones infinitesimales alrededor de los tres ejes espaciales cartesianos, satisfacen[22.18]

$$\ell_1\ell_2 - \ell_2\ell_1 = \ell_3, \quad \ell_2\ell_3 - \ell_3\ell_2 = \ell_1, \quad \ell_3\ell_1 - \ell_1\ell_3 = \ell_2.$$

Estas están relacionadas, según las reglas de la mecánica cuántica, con las componentes $\boldsymbol{L}_1, \boldsymbol{L}_2, \boldsymbol{L}_3$ del momento angular respecto a los tres ejes, de acuerdo con:

$$\boldsymbol{L}_1 = i\hbar\ell_1, \quad \boldsymbol{L}_2 = i\hbar\ell_2, \quad \boldsymbol{L}_3 = i\hbar\ell_3.$$

[22.18] Utilice cuaterniones para comprobarlo.

Así, nuestras reglas de conmutación para el momento angular son[22]

$$\boldsymbol{L}_1\boldsymbol{L}_2 - \boldsymbol{L}_2\boldsymbol{L}_1 = \mathrm{i}\hbar\boldsymbol{L}_3, \quad \boldsymbol{L}_2\boldsymbol{L}_3 - \boldsymbol{L}_3\boldsymbol{L}_2 = \mathrm{i}\hbar\boldsymbol{L}_1, \quad \boldsymbol{L}_3\boldsymbol{L}_1 - \boldsymbol{L}_1\boldsymbol{L}_3 = \mathrm{i}\hbar\boldsymbol{L}_2.$$

Igual que sucede con prácticamente cualquier otra cosa en mecánica cuántica, las componentes del momento angular $\boldsymbol{L}_1$ $\boldsymbol{L}_2$ $\boldsymbol{L}_3$ deben actuar como operadores lineales en el espacio de Hilbert $\mathbf{H}$. Así pues, los sistemas cuánticos que poseen momento angular proporcionan una representación del álgebra de Lie de SO(3) en términos de transformaciones lineales de $\mathbf{H}$ (véanse §§13.6,10, §14.6).

Esto conduce a uno de los aspectos más elegantes y reveladores de la mecánica cuántica, y es un tema que recompensa con creces un estudio detallado. Sin embargo, este no es lugar para tales detalles, de modo que solo señalaré algunos puntos de particular importancia. En primer lugar, tomemos nota del hecho de que las matrices explícitas

$$\boldsymbol{L}_1 = \frac{\hbar}{2}\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \boldsymbol{L}_2 = \frac{\hbar}{2}\begin{pmatrix} 0 & -\mathrm{i} \\ \mathrm{i} & 0 \end{pmatrix}, \boldsymbol{L}_3 = \frac{\hbar}{2}\begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix},$$

llamadas (sin los $\hbar/2$) *matrices de Pauli*, satisfacen las reglas de conmutación requeridas.[22.19] Proporcionan la representación más sencilla (no trivial) del momento angular, e imaginamos que estas matrices 2×2 actúan sobre una función de onda con dos componentes $\{\psi_0(\boldsymbol{x}), \psi_1(\boldsymbol{x})\}$. Cuando empezamos a rotar este estado, las componentes $\psi_0(\boldsymbol{x})$ y $\psi_1(\boldsymbol{x})$ se mezclan de acuerdo con las reglas de multiplicación de matrices que generan las matrices de Pauli.

Podemos llamar ψ_A a esta función de onda de 2-componentes, utilizando un subíndice A (que toma los valores 0 y 1, o cualquier otra cosa que podamos considerar como un índice abstracto de acuerdo con la «notación de índices abstractos» mencionada en §12.8). La magnitud descrita por ψ_A se denomina un *espinor*, y su índice A se conoce como un índice *2-espinorial*. Resulta que ψ_A es un objeto espinorial en el sentido descrito en §11.3 (una rotación continua de 2π lo convierte en su negativo). De hecho, si «exponenciamos» de forma continua (véase §14.6) una de las matrices de Pauli hasta que obtengamos una rotación completa de 2π, obtenemos el operador $-\boldsymbol{I}$, que envía ψ_A a $-\psi_A$.[22.20]

[22.19] Compruébelo. Explique cómo se relacionan estas reglas de multiplicación con las de los cuaterniones.

Esta notación es parte de un potente formalismo que puede desarrollarse para complementar (o incluso reemplazar)[23] el formalismo del cálculo tensorial, utilizando cantidades «tipo tensor» construidas a partir de cosas como «ψ_A». Aunque aquí no es completamente necesario, su poder real se manifiesta cuando nos aprovechamos de la versión relativista de este formalismo. Para esto necesitamos también índices «primados» A', B', C', ..., además de los «no primados» A, B, C, ..., que son, en un sentido apropiado, *complejos conjugados* unos de otros; véase §13.9. La notación tiene gran valor en la teoría cuántica de campos (un hecho quizá menos apreciado de lo que debiera;[24] véanse §25.2 y §34.3) y en la relatividad general[25] (y desempeña un papel básico en la teoría de twistores; véase §33.6). No es oportuno que entre ahora en detalles, pero será útil tomar prestado un poco de este formalismo 2-espinorial. Todo lo que necesitaremos aquí y en §§22.9-11 es representar estados de espín general de una forma clara. No necesitaremos los índices primados (hasta §§25.2,3 y §§33.6,8), puesto que aquí solo estamos haciendo física no relativista.

Antes de entrar en esto, quiero hacer una simplificación notacional. Para el resto de esta sección y hasta el final de §§22.11, supondré por conveniencia que se han escogido las unidades de modo que $\hbar = 1$. De hecho, esto es siempre posible, y veremos en §27.10 (y §31.1) que podríamos ir mucho más lejos y describir las cosas en términos de lo que se denominan «unidades de Planck», en las que la velocidad de la luz y la constante gravitatoria se hacen también iguales a la unidad. No hay necesidad de ir tan lejos aquí, y en cualquier caso no es difícil restaurar $\hbar$, si hace falta, por simples consideraciones de dimensiones físicas. (Por ejemplo, para restaurar $\hbar$ en cualquier fórmula en la que $\hbar$ se ha hecho igual a la unidad, reemplazamos cualquier magnitud que escale como la q-ésima potencia de la masa, ignorando la longitud y el tiempo, por $\hbar^{-q}$ multiplicado por dicha magnitud. En particular, masa, energía, momento y momento angular quedarían simplemente divididos por $\hbar$.)

[22.20] Hágalo explícitamente.

Volviendo al formalismo 2-espinorial, recordemos que una magnitud espinorial univalente ψ_A puede utilizarse para describir una partícula de espín 1/2. El mismo tipo de notación puede adoptarse para valores más altos del espín, correspondientes a otras representaciones del álgebra de Lie de SO(3). El valor del espín es siempre un múltiplo entero no negativo de 1/2:

$$0, \frac{1}{2}, 1, \frac{3}{2}, 2, \frac{5}{2}, \ldots$$

(o, restaurando $\hbar$, diríamos que el espín/$\hbar$ toma estos valores) y la función de onda puede describirse mediante un objeto $\psi_{AB\ldots F}$ (un «tensor-espín) que es completamente simétrico en sus n índices en el caso del espín $n/2$

$$\psi_{AB\ldots F} = \psi_{(AB\ldots F)}$$

(donde los paréntesis redondos denotan simetrización sobre todos los n índices; véase §12.7). De hecho, todas las representaciones de SO(3) —donde incluimos las espinoriales 2-valuadas— pueden construirse como *sumas directas* de estas representaciones concretas, las representaciones *irreducibles* (véase §13.7). Esto equivale a decir que la representación general puede expresarse como una colección (posiblemente infinita) de funciones de onda

$$\{\psi_{AB\ldots F},\ \phi_{GH\ldots K},\ \chi_{LM\ldots R}, \ldots\},$$

cada una de las cuales es totalmente simétrica en sus índices espinoriales.

Para una partícula individual habría solo un campo simétrico semejante, por ejemplo, $\psi_{AB\ldots F}$, para su función de onda. (Sería un error comprensible pensar que para dos partículas habría dos de ellos; para tres partículas, tres de ellos, etc. Veremos cómo se describen realmente los sistemas de más de una partícula en el próximo capítulo. Es algo bastante más sutil que esto.) Para una partícula de espín 0, denominada una partícula *escalar* (tal como un mesón π), la función de onda tiene 0 índices, y esta era la situación que se ha tratado en el capítulo 21. Las partículas más familiares, electrones, muones, neutrinos, protones, neutrones, y también sus quarks constituyentes, tienen todas espín 1/2 (solo 1 índice). El deuterón (núcleo del hidrógeno pesado) y los bosones W (véase §25.4) tienen espín 1 (2 índices espinoriales simétricos). Muchos núcleos más pesados, e incluso átomos enteros, pueden ser tratados como una única partícula con espín mucho mayor. Para espín $\frac{1}{2}n$, el objeto de n índices $\psi_{AB\ldots F}$ tiene $n + 1$ componentes complejas independientes.[26,[22.21]] Aunque el tensor-espín $\psi_{AB\ldots F}$ se conoce frecuentemente como un *espinor* de n-índices, solo es un objeto espinorial (§11.3) cuando n es impar, casos estos en que el espín es un número semiimpar, y no entero. Debería señalarse también que el propio valor de espín, $j = \frac{1}{2}n(\geq 0)$, determina (y es determinado por) el autovalor $j(j + 1)$ del operador «espín total»[27]

$$\mathbf{J}^2 = \boldsymbol{L}_1^2 + \boldsymbol{L}_2^2 + \boldsymbol{L}_3^2,$$

siendo esta la «longitud al cuadrado» del operador 3-vectorial $\mathbf{J} = (\boldsymbol{L}_1, \boldsymbol{L}_2, \boldsymbol{L}_3)$.

El espín total $\mathbf{J}^2$ *conmuta*[22.22],[22.23] con cada componente $\boldsymbol{L}_1, \boldsymbol{L}_2, \boldsymbol{L}_3$ del momento angular (pese al hecho de que tales componentes no conmutan entre sí). Esta propiedad caracteriza a $\mathbf{J}^2$ como un *operador de Casimir* para SO(3); véase §22.12. Para delinear por completo los estados cuánticos, formamos un conjunto completo de operadores conmutativos (§22.12), y buscamos estados que sean simultáneamente autoestados de todos los operadores del conjunto. En el caso del momento angular lo que se hace normalmente es tomar, además de $\mathbf{J}^2$, el operador $\boldsymbol{L}_3$ del momento angular en torno a la dirección hacia arriba («z»). Los dos «números cuánticos» j y m etiquetan el estado, siendo $j(j + 1)$ el autovalor de $\mathbf{J}^2$ y m el autovalor de $\boldsymbol{L}_3$. Tomamos $j \geq 0$ y $-j \leq m \leq j$,

[22.21] Trate de calcularlo a partir de la información dada.

[22.22] Compruebe esta conmutación directamente a partir de las reglas de conmutación del momento angular.

[22.23] Considere los operadores $\boldsymbol{L}^+ = \boldsymbol{L}_1 + \mathrm{i}\boldsymbol{L}_2$ y $\boldsymbol{L}^- = \boldsymbol{L}_1 - \mathrm{i}\boldsymbol{L}_2$ y calcule sus conmutadores con $\boldsymbol{L}_3$. Escriba $\mathbf{J}^2$ en términos de $\boldsymbol{L}^\pm$ y $\boldsymbol{L}_3$. Demuestre que si $|\psi\rangle$ es un autoestado de $\boldsymbol{L}_3$, entonces también lo son $\boldsymbol{L}^\pm|\psi\rangle$, siempre que sean no nulos, y encuentre sus autovalores en términos del autovalor de $|\psi\rangle$. Demuestre que si $|\psi\rangle$ pertenece a un espacio de representación irreducible de dimensión finita generado por tales autoestados, entonces la dimensión es un entero $2j$, donde $j(j + 1)$ es el autoestado de $\mathbf{J}^2$ para todos los estados en el espacio.

donde j y m son ambos o bien semienteros impares (caso espinorial), o bien enteros. Los $2j + 1$ $(= n + 1)$ diferentes valores posibles de m corresponden a las diferentes componentes de $\psi_{AB\ldots F}$.

La elección de la dirección hacia arriba es, por supuesto, arbitraria (y corresponde a escoger una base arriba/abajo (la $|\uparrow\rangle$, $|\downarrow\rangle$ de §22.9) para las componentes espinoriales. Igualmente, podría escoger cualquier otra dirección espacial en lugar de «arriba». En consecuencia, en ocasiones mencionaré el «valor m» en alguna otra dirección dada (como sucede con la descripción de Majorana de §22.10).

22.9. La esfera de Riemann de los sistemas de dos estados

Consideremos la extraordinariamente concisa —e incluso mágica— geometría cuántica de los estados de espín individuales para el espín 1/2 (por ejemplo, electrón, protón, neutrón, quark). Esto es también ilustrativo para la comprensión de los sistemas cuánticos de 2-estados en general. Un sistema semejante está descrito por un espacio de Hilbert 2-dimensional $\mathbf{H}^2$, y el caso del espín 1/2 representa muy bien su geometría.

En el caso de nuestra partícula de espín 1/2 nos interesaremos solo en los *grados de libertad de espín*, en el sistema de reposo de la partícula. Para hacer esto explícito, imaginemos que la partícula está «en reposo» en el sentido de que está en el autoestado de momento lineal cero, de modo que su estado tiene que ser constante[22.24] en las variables espaciales $\mathbf{x}$. Entonces, ψ_0 y ψ_1 son simplemente números complejos, digamos $\psi_0 = w$ y $\psi_1 = z$, y escribimos el estado como $\{w, z\}$. Podemos disponer las cosas de modo que «espín arriba» $|\uparrow\rangle$ (dextrógiro respecto a la vertical hacia arriba) sea el estado de espín $\{1, 0\}$; en correspondencia, «espín abajo» $|\downarrow\rangle$ (dextrógiro respecto a la vertical hacia abajo) será $\{0, 1\}$. Estos dos estados de base son ortogonales:

$$\langle\uparrow|\downarrow\rangle = 0.$$

[22.24] ¿Por qué?

También normalizamos:

$$\langle\uparrow|\uparrow\rangle = 1 = \langle\downarrow|\downarrow\rangle.$$

El estado de espín 1/2 general $\psi_A = \{w, z\}$ (elemento general de $\mathbf{H}^2$) es la combinación lineal

$$\{w, z\} = w|\uparrow\rangle + z|\downarrow\rangle$$

de estos dos estados. El producto escalar de otro estado general $\{a, b\}$ (i.e., $a|\uparrow\rangle + b|\downarrow\rangle$) por $\{w, z\}$ está dado por[22.25]

$$\langle\{a, b\}|\{w, z\}\rangle = \bar{a}w + \bar{b}z.$$

Resulta ahora que todo estado de espín 1/2 debe ser realmente un estado puro de espín que es dextrógiro respecto a *alguna* dirección en el espacio, de modo que podemos escribir (digamos)

$$w|\uparrow\rangle + z|\downarrow\rangle = |\nearrow\rangle,$$

donde «$\nearrow$» es una dirección real en el espacio.[22.26] Esto nos da una notable identificación entre el espacio proyectivo $\mathbb{P}\mathbf{H}^2$ (§15.6) y la geometría de direcciones en el espacio, consideradas estas como direcciones de espín. Los estados de espín 1/2 físicamente distintos son proporcionados por este espacio proyectivo (véase §21.9), y los diferentes puntos de $\mathbb{P}\mathbf{H}^2$ están etiquetados por las distintas *razones*

$$z : w.$$

En otras palabras, $\mathbb{P}\mathbf{H}^2$ es tan solo una copia de nuestra vieja amiga, la *esfera de Riemann*, con la que nos hemos familiarizado por primera vez en §8.3. Cada punto de esta esfera de Riemann etiqueta un estado de espín 1/2 distinto, que es el autoestado «$m = +1/2$» de la medida de espín concreta que se toma en la dirección que va desde el centro de la esfera a dicho punto (Fig. 22.10).

[22.25] Obtenga esta expresión a partir de lo que se ha dicho antes.

[22.26] Trate de obtener esto de dos formas diferentes: (i) encontrando la dirección explícitamente en algún sistema cartesiano apropiado, donde el estado $\{a, b\}$ define b/a como un punto en el plano complejo de la Fig. 8.7a; (ii) sin cálculo directo, utilizando el hecho de que, puesto que $\mathbf{H}^2$ es un espacio de representación de SO(3), todas las direcciones de espín están incluidas, pese a que $\mathbb{P}\mathbf{H}^2$ no es «suficientemente grande» para contener más estados que esto.

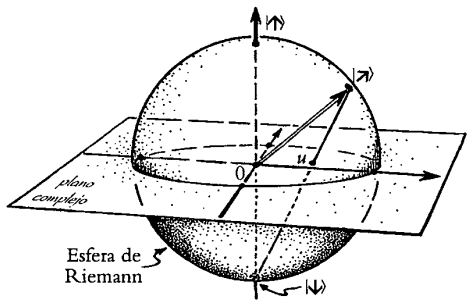


Fig. 22.10. El espacio proyectivo $\mathbb{P}\mathbf{H}^2$ para un sistema de dos estados es una esfera de Riemann (véase la Fig. 8.7). Para los estados de espín de una partícula masiva de espín 1/2, podemos utilizar el polo norte para representar el estado de espín $|\uparrow\rangle$ (espín «arriba») y el polo sur, el estado $|\downarrow\rangle$ (espín «abajo»). Un estado de espín general $|\nearrow\rangle$ está representado (con fases apropiadas para $|\uparrow\rangle$ y $|\downarrow\rangle$) por el punto en la esfera cuya dirección a partir del centro es la de $|\nearrow\rangle$ (i.e., que da el resultado «sí» *con certeza* para una medida de espín $\boldsymbol{E}_{\nearrow}$ en dicha dirección), como se ilustra por la flecha de doble trazo. Podemos expresar el estado $|\nearrow\rangle$ como una combinación lineal $|\nearrow\rangle = w\,|\uparrow\rangle + z|\downarrow\rangle$ (donde podemos considerar los números complejos z, w, como las componentes $w = \psi_0$, $z = \psi_1$ de un 2-espinor ψ_A). Los puntos en la esfera corresponden a las distintas razones $z : w$. Cada una de estas puede representarse por un número complejo $u = z/w$ (admitiendo ∞) en el plano complejo, tomando este plano como el plano ecuatorial de la esfera. El punto u proyecta estereográficamente desde el polo sur al punto en la esfera que representa $|\nearrow\rangle$.

Vemos esta relación geométrica de forma más explícita si utilizamos la proyección estereográfica de la esfera desde su polo sur en su plano ecuatorial descrita en §8.3 (Fig. 8.7a). Hay que considerar este plano como el plano complejo de la *razón* z/w entre las amplitudes mecanocuánticas z y w (y no de la «z» de §8.3). Esto relaciona directamente el punto concreto de la esfera, correspondiente a la dirección espacial $\nearrow$, con la razón z/w.

Utilicemos el proyector $\boldsymbol{E}_{\nearrow}$ para denotar la medida que plantea la pregunta «¿Está el espín en la dirección $\nearrow$?», de modo que el valor propio es 1 (sí), si se encuentra que el estado de espín es (o se proyecta en) $|\nearrow\rangle$ y es 0 (NO), si el espín se proyecta en el estado de espín ortogonal $|\swarrow\rangle$ en la dirección espacial *opuesta* (correspondiente al punto antipodal en la esfera de Riemann). (Nótese que en este ejemplo «ortogonal»

en el espacio de Hilbert no corresponde a «formando ángulos rectos» en el espacio, sino a «en dirección opuesta».) Si partimos del estado $|\uparrow\rangle$, entonces la probabilidad de sí para la medida $\boldsymbol{E}_{\nearrow}$ es $|w|^2/(|w|^2 + |z|^2)$. Si el espín está inicialmente en cierto estado $|\nwarrow\rangle$, y se realiza una medida sobre él para establecer si el estado está en alguna otra dirección $|\nearrow\rangle$, donde el ángulo 3-espacial euclídeo ordinario entre $\nwarrow$ y $\nearrow$ es θ, entonces la probabilidad de encontrar el resultado sí es[22.27]

$$\frac{1}{2}(1 + \cos\theta).$$

Podemos también expresar esta probabilidad directamente en términos de la geometría de la esfera, donde $\nwarrow$ y $\nearrow$ vienen dados por dos puntos A y B, respectivamente, sobre la esfera y proyectamos ortogonalmente B en un punto C sobre el diámetro que pasa por A (Fig. 22.11). Si A′ es el punto antipodal de A, entonces la probabilidad de sí es la longitud A′C dividida por el diámetro AA′ de la esfera.[22.28]

Nótese que la «esfera de Riemann» utilizada aquí tiene más estructura que la de §8.3 y la esfera celeste de §18.5, pues ahora la noción de «punto antipodal» es parte de la estructura de la esfera (para que podamos decir qué estados son «ortogonales» en el sentido de espacio de Hilbert). La esfera es ahora una «esfera métrica» antes que una «esfera conforme», de modo que sus simetrías vienen dadas por rotaciones en el sentido ordinario, y perdemos los movimientos conformes que se manifestaban en efectos de aberración en la esfera celeste. De todas formas, nuestro uso actual de la esfera de Riemann muestra claramente una conexión explícita entre las razones de números complejos que aparecen en mecánica cuántica y las direcciones ordinarias del espacio. Vemos, una vez más, que los números complejos que aparecen en el formalismo de estados cuánticos no son cosas completamente abstractas; están íntimamente relacionados con el comportamiento geométrico y dinámico. (Recordemos también el papel de las fases complejas para determinar la dinámica de un estado del momento, como se describe en §21.6.)

Habría que señalar que la geometría de la Fig. 22.11, que expresa

[22.27] Demuéstrelo.
[22.28] Confírmelo.

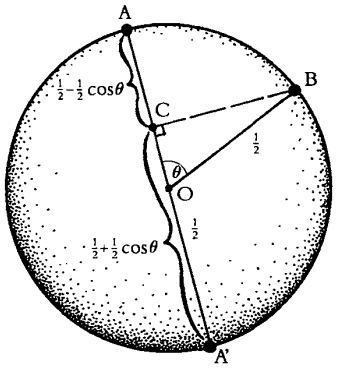


Fig. 22.11. Supongamos que el estado inicial de un sistema de dos estados (como el de la Fig. 22.10) está representado por el punto B en la esfera de Riemann y queremos ejecutar una medida SÍ/NO correspondiente a algún otro punto en la esfera, donde sí encontraría el estado en A y no lo encontraría en el punto A′, antipodal de A. Suponiendo que la esfera tiene radio $\frac{1}{2}$, y proyectando A ortogonalmente a C en el eje A′A, encontramos que la probabilidad de sí es la longitud A′C, que es $\frac{1}{2}(1 + \cos\theta)$, y la probabilidad de NO es la longitud CA, que es $\frac{1}{2}(1 - \cos\theta)$, donde θ es el ángulo entre OB y OA, siendo O el centro de la esfera.

las probabilidades que aparecen en una medida cuántica en relación con $\mathbb{P}\mathbf{H}^2$, no está restringida al caso del espín, sino que es completamente general para un sistema de 2-estados. Lo que resulta especial para el caso de espín 1/2 es la asociación inmediata entre *direcciones espaciales* y los puntos de la esfera de Riemann $\mathbb{P}\mathbf{H}^2$. La esfera de Riemann está siempre allí, en un sistema de 2-estados, proporcionando la «difusión cuántica» de un par de alternativas clásicas. Sin embargo, en muchas situaciones físicas el papel geométrico de esta esfera, y de los números complejos mecanocuánticos subyacentes, no es muy directo y existe una tendencia entre los físicos a considerarlos como cantidades puramente «formales». Esta actitud surge en parte del hecho de que la fase global del vector de estado para un sistema físico completo se considera inobservable, de modo que se suele ignorar la riqueza geométrica potencial de los coeficientes complejos internos. Las fases *relativas* entre una parte y otra desempeñan ciertamente un papel observable. Una manera de expresarlo está en el hecho de que la geometría compleja del espacio de Hilbert proyectivo $\mathbb{P}\mathbf{H}$ para un sistema es físicamente significativa. Aunque la fase global queda fuera en la definición de $\mathbb{P}\mathbf{H}$, todas las fases relativas intervienen en su geometría. De hecho,

hay enfoques elegantes de la mecánica cuántica que explotan la geometría proyectiva compleja de $\mathbb{P}\mathbf{H}$.[28]

Hay también otras situaciones en las que la geometría de la esfera de Riemann relaciona directamente los números complejos de la mecánica cuántica con las propiedades espaciales del espín. Lo que resulta más significativo es que esto se aplica a los estados de espín generales de una partícula masiva, como se describirá enseguida (en §22.11). Pero, para terminar esta sección, volvamos a la polarización del fotón que hemos visto brevemente en §22.7. Recordemos que el estado general de polarización de un fotón es una combinación lineal compleja de los estados de helicidad positiva $|+\rangle$ y helicidad negativa $|-\rangle$:

$$|\phi\rangle = w|+\rangle + z|-\rangle.$$

La interpretación física de un estado semejante se hace en términos de lo que se denomina *polarización elíptica*, que generaliza los casos particulares de polarización plana y polarización circular. No es mi intención describir esto aquí con todo detalle, pero se obtiene una imagen suficientemente buena si pensamos en términos de una onda plana electromagnética clásica. Los «planos» son los frentes de onda, que son perpendiculares a la dirección de movimiento. En cada punto del espacio habrá un vector eléctrico $\mathbf{E}$ y un vector magnético $\mathbf{B}$, y en el caso de una onda plana estos son siempre perpendiculares y yacen en los frentes de onda. Si imaginamos que fijamos el punto en el espacio y dejamos que la onda pase por él, el vector eléctrico gira a medida que pasa la onda, de modo que la punta de dicho vector describe una elipse en el plano del frente de onda. El vector magnético le sigue, describiendo una elipse idéntica pero rotada un ángulo recto.Véase la Fig. 22.12. En ciertos casos particulares, la elipse se aplasta hasta convertirse en un segmento de línea: son los casos de polarización *plana*. La polarización *circular* ocurre cuando la elipse se convierte en un círculo. Si orientamos las cosas de modo que la onda se dirija directamente hacia nosotros, entonces los vectores giran en sentido contrario a las agujas del reloj para la helicidad positiva, y en el sentido de las agujas del reloj para la helicidad negativa.

Veamos cómo encaja la esfera de Riemann en todo esto. Consideremos que el polo norte representa el estado de helicidad positiva $|+\rangle$,

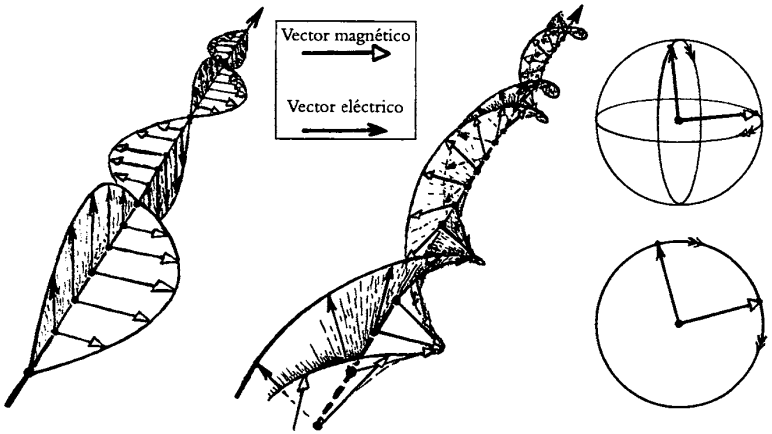


Fig. 22.12. Polarización del fotón (véase la Fig. 21.7) como una característica de las ondas planas electromagnéticas. (a) Una onda plano-polarizada que se aleja del observador. Los vectores eléctricos (flechas de punta negra) y los vectores magnéticos (flechas de punta blanca) oscilan de un lado a otro, en dos planos perpendiculares fijos. (b) En una onda plana circularmente polarizada, los vectores eléctrico y magnético rotan alrededor de la dirección de movimiento, permaneciendo siempre perpendiculares y de la misma longitud constante. (c) Visto desde atrás, los diagramas muestran cómo los vectores eléctrico y magnético rotan a lo largo de la onda (caso de helicidad positiva); la figura inferior muestra la situación para la polarización circular, y la superior, para el caso general de la polarización elíptica, donde las puntas de las dos flechas describen elipses congruentes con sus ejes mayores perpendiculares. La función de onda de un solo fotón mostraría un comportamiento de este tipo.

y el polo sur representa el estado de helicidad negativa $|-\rangle$. Suponemos que el fotón viaja hacia arriba en la dirección de $|+\rangle$. Ahora, en lugar de marcar z/w sobre la esfera de Riemann, vamos a considerar su raíz cuadrada $q = (z/w)^{1/2}$ (no importa cuál), de modo que $q = 0,\infty$ dan $|+\rangle$, $|-\rangle$, respectivamente. Consideremos el radio de la esfera que pasa por q («vector de Stokes») y dibujemos el círculo máximo de la esfera que yace en el plano diametral perpendicular a dicha línea. Orientemos este círculo en un sentido dextrógiro alrededor del eje que pasa por q. Proyectemos luego este círculo ortogonalmente en el plano

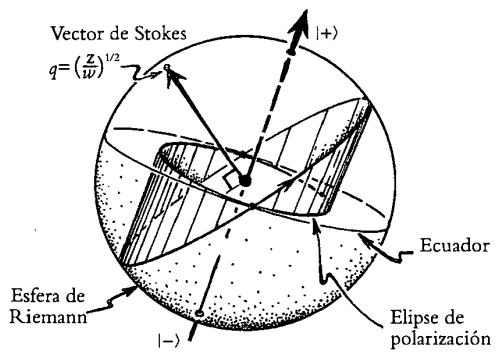


Fig. 22.13. Estados de polarización del fotón representados en la esfera de Riemann. Consideramos que el polo norte representa el estado de helicidad positiva $|+\rangle$ y el polo sur, el estado de helicidad negativa $|-\rangle$, y consideramos que el momento del fotón está en dirección norte. El estado de polarización general $w|+\rangle + z|-\rangle$ está representado por el punto $q = (z/w)^{1/2}$ en la esfera de Riemann. Consideremos el semidiámetro de la esfera desde q, llamado el «vector de Stokes», y tracemos el círculo máximo que yace en el plano diametral perpendicular al mismo. Orientemos este círculo a la derecha respecto al vector de Stokes. Luego proyectemos este círculo ortogonalmente en el plano ecuatorial de la esfera. Esto nos da la elipse de polarización requerida y la orientación correcta.

ecuatorial de la esfera. Obtenemos así la elipse de polarización requerida, junto con la orientación correcta. Véase la Fig. 22.13.[22.29]

22.10. Espín más alto: la imagen de Majorana

Como ejemplo adicional que ilustra la estrecha relación entre los números complejos aparentemente abstractos en mecánica cuántica y la geometría del espacio, consideremos los estados de espín de una partícula masiva —o un átomo— de espín $j = \frac{1}{2}n$. Como se ha comentado antes (en §22.8), este estado puede describirse por un espinor $\psi_{AB\ldots F}$ simétrico de n índices. Ahora bien, existe un teorema que afirma que todo espinor semejante tiene una «descomposición canónica» me-

[22.29] Verifique todo esto. ¿Por qué no me preocupo por el *signo* de q?

diante la cual es expresable unívocamente, salvo factores de escala y ordenamientos, como un producto simetrizado de espinores de 1-índice:[22.30]

$$\psi_{AB\ldots F} = \alpha_{(A}\beta_B \cdots \varphi_{F)}$$

donde recordamos de §12.7 que los paréntesis redondos alrededor de índices denotan simetrización. Utilizando la geometría de la Fig. 22.10, donde un espinor de índice único se representa geométricamente (salvo un factor complejo global) por un punto en la esfera de Riemann (i.e., por una dirección en el espacio), concluimos que el propio espinor $\psi_{AB\ldots F}$ puede representarse en la esfera de Riemann, salvo un factor de escala global, como un conjunto desordenado de n puntos en la esfera (i.e., n direcciones desordenadas en el espacio); véase la Fig. 22.14. Esta representación de un estado general de espín n se denomina descripción de *Majorana*. Fue encontrada originalmente en 1932 (aunque por un procedimiento[29] diferente que examinaremos brevemente en §22.11) por el brillante físico italiano Ettore Majorana. (A la temprana edad de treinta y un años, desapareció misteriosamente de un barco en la bahía de Nápoles, quizá por suicidio.)

Existe una base estándar de estados para espín $j = \frac{1}{2}n$. En la descripción de Majorana, estos se realizan como aquellos estados para los que todos los puntos de Majorana están en el polo norte o en el polo sur:

$$|\uparrow\uparrow\uparrow \ldots \uparrow\rangle,\ |\downarrow\uparrow\uparrow \ldots \uparrow\rangle,\ |\downarrow\downarrow\uparrow \ldots \uparrow\rangle,\ \ldots,\ |\downarrow\downarrow\downarrow \ldots \downarrow\rangle.$$

Estos $n + 1$ estados son los autoestados del observable $\boldsymbol{L}_3$ (siendo el eje x^3 la dirección «arriba») y por consiguiente son todos mutuamente ortogonales. Pueden distinguirse por los diferentes autovalores de espín, llamados valores m (§22.8), respectivamente: $j, j-1, j-2, \ldots, -j$. Veremos algo más sobre esto en §22.11.

Existe un aparato de medida estándar, conocido como aparato de Stern-Gerlach, que puede ser utilizado para medir este «valor m» para un átomo. Para que esto funcione, exigimos que el átomo posea un momento magnético (de modo que es un minúsculo imán), cuyo vector momento magnético es cierto múltiplo del vector de espín. Los átomos se hacen pasar primero a través de un campo magnético fuertemente inhomogéneo. Este desvía sus trayectorias de forma ligeramente diferente para cada valor m, puesto que m determina cómo está orientado el vector momento magnético de cada átomo con respecto al campo magnético inhomogéneo; véase la Fig. 22.15.

[22.30] Vea si puede demostrarlo utilizando el «teorema fundamental del álgebra» enunciado en la nota 4.2. *Sugerencia*: Considérese el polinomio $\psi_{AB\ldots F}\zeta^A\zeta^B \ldots \zeta^F$, donde las componentes de ζ^A son $\{1, z\}$.

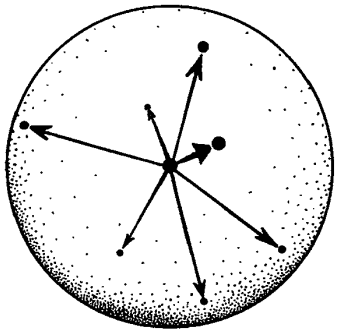


Fig. 22.14. Descripción de Majorana del estado de espín (proyectivo) general para una partícula masiva de espín $\frac{n}{2}$, que da n puntos desordenados en la esfera de Riemann. Podemos considerar que los vectores dirigidos desde el centro a dichos puntos contribuyen al espín $\frac{1}{2}$, de acuerdo con la receta de la Fig. 22.10. El producto simétrico de estos espines da el total. (En notación 2-espinorial, el estado de espín completo es el espinor n-valente simétrico, que factoriza $\psi_{AB\ldots F} = \alpha_{(A}\beta_B \ldots \varphi_{F)}$, donde $\alpha_A, \beta_A, \ldots, \varphi_A$, determinan los n puntos, como en la Fig. 22.10.)

Aunque los estados para cada valor m diferente son todos mutuamente ortogonales, las condiciones de ortogonalidad para las descripciones *generales* de Majorana son complicadas.[30] No obstante, podría comentarse que un estado de Majorana en el que interviene cierta dirección ↗ es necesariamente ortogonal al estado $|\swarrow\swarrow\swarrow \ldots \swarrow\rangle$, donde ↙ es diametralmente opuesto a ↗. Además, si ↗ interviene en la descripción de Majorana con multiplicidad r, entonces el estado es ortogonal a cualquier otro estado de espín $\frac{1}{2}n$, cuya descripción de Majorana incluye la dirección opuesta ↙ con multiplicidad de al menos $n - r + 1$.[22.31]

Estos resultados nos permiten interpretar físicamente las direcciones de Majorana. Las direcciones de Majorana son precisamente aquellas para las que una medida de Stern-Gerlach en dicha dirección tiene

[22.31] Trate de demostrar esto utilizando la geometría de §22.9. Aplique este resultado a la ortogonalidad de los diversos autoestados de $\boldsymbol{L}_3$.

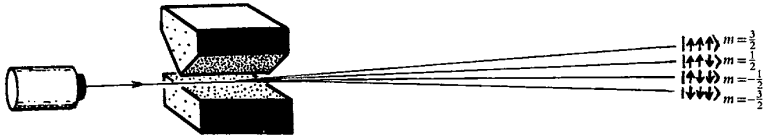


Fig. 22.15. Aparato de Stern-Gerlach, utilizado para medir el «valor m» del momento magnético de un átomo (acoplado a su espín). Los átomos atraviesan un campo magnético fuertemente inhomogéneo que desvía sus trayectorias de forma ligeramente diferente para cada valor de m.

probabilidad cero de encontrar el espín en la dirección exactamente opuesta. Para una dirección de Majorana de multiplicidad r, hay una probabilidad nula de que el valor m en dicha dirección sea algo comprendido entre $-j$ y $-j + r - 1$.[31]

Habría que señalar que el procedimiento esbozado antes para representar el estado de espín general para una partícula masiva no es muy familiar para muchos físicos. En su lugar, ellos adoptarían un procedimiento diferente que implica lo que se denomina *análisis armónico*. Este es un tema importante por muchas otras razones, y la sección siguiente ofrece una breve exposición de sus ideas más relevantes.

22.11. Armónicos esféricos

En §20.3 hemos examinado la teoría clásica de vibraciones (de pequeña amplitud y sin disipación). Nuestra discusión principal se centraba en sistemas de un número finito de grados de libertad. Pero también hemos considerado (brevemente) sistemas —como las vibraciones de un tambor o una columna de aire— para los que el número de grados de libertad se trata como si fuera infinito. Dichas vibraciones (en cualquiera de los dos casos) están compuestas de modos normales, cada uno de los cuales tiene su propia frecuencia de vibración denominada una frecuencia normal. Si el objeto vibrante es *compacto* (véanse §12.6 y las Figs. 12-14 para el significado de este término), sus modos constituirán una familia discreta, lo que proporciona un espectro discreto de frecuencias normales diferentes. En el caso concreto de una esfera S^2, los diferentes modos de vibración (que podemos visualizar como

los modos vibracionales de una burbuja de jabón, por ejemplo, o de un globo esférico) corresponden a lo que se denominan *armónicos esféricos*. ¿Qué tiene que ver esto con la mecánica cuántica del momento angular? Veremos esto enseguida.

Para clasificar estos armónicos buscamos autoestados del operador laplaciano ∇^2 definido en S^2. En §10.5 hemos visto el laplaciano 2-dimensional ordinario, definido en el plano euclídeo por $\nabla^2 = \partial^2/\partial x^2 + \partial^2/\partial y^2$. Sobre la esfera unidad S^2, esta expresión debe modificarse para tener en cuenta la métrica curva. Esta forma métrica es

$$ds^2 = g_{ab}dx^a dx^b = d\theta^2 + \operatorname{sen}^2\theta\, d\phi^2$$

en las coordenadas *polares esféricas* habituales (θ, ϕ), que es un etiquetado de puntos en S^2 para el que las coordenadas cartesianas (en el 3-espacio ordinario) de un punto son $x = \operatorname{sen}\theta\cos\phi$, $y = \operatorname{sen}\theta\operatorname{sen}\phi$, $z = \cos\theta$ (Fig. 22.16). Así pues, ϕ es esencialmente la *longitud* y $\frac{1}{2}\pi - \theta$ la *latitud* (todo en radianes). El laplaciano (con derivada covariante ∇_a; véase §14.3) es[22.32]

$$\nabla^2 = g^{ab}\nabla_a\nabla_b$$
$$= \frac{\partial^2}{\partial\theta^2} + \frac{\cos\theta}{\operatorname{sen}\theta}\frac{\partial}{\partial\theta} + \frac{1}{\operatorname{sen}^2\theta}\frac{\partial^2}{\partial\phi^2}.$$

Los autovalores posibles de ∇^2 resultan ser los números $-j(j+1)$ (para $j = 0, 1, 2, 3, \ldots$), de modo que:

$$\nabla^2\Phi = -j(j+1)\Phi,$$

donde Φ es la autofunción correspondiente.[32] Estas autofunciones son los *armónicos esféricos*, y es habitual exigir que los armónicos sean simultáneamente autofunciones del operador $\partial/\partial\phi$ (que conmuta con ∇^2). Los autovalores posibles de $\partial/\partial\phi$ son im, donde el entero m está en el intervalo $-j \le m \le j$:

$$\frac{\partial\Phi}{\partial\phi} = im\Phi.$$

Ejemplos de tales autofunciones son $\Phi = 1$ (para $j = m = 0$), $\Phi = \cos\theta$

[22.32] ¿Puede obtener esta expresión en polares esféricas?

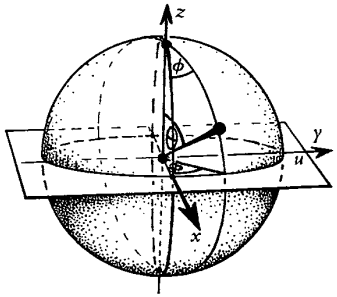


Fig. 22.16. Las coordenadas polares esféricas estándar θ y ϕ están relacionadas con las coordenadas cartesianas por $x = \text{sen}\,\theta\cos\phi$, $y = \text{sen}\,\theta\,\text{sen}\,\phi$, $z = \cos\theta$. Así pues, ϕ es básicamente una medida de longitud (marcada aquí en el polo norte y en el ecuador) y $\pi/2 - \theta$ es la latitud.

(para $j = 1, m = 0$), $\Phi = e^{\pm i\phi}\,\text{sen}\,\theta$ (para $j = 1, m = \pm 1$), $\Phi = 3(\cos\theta)^2 - 1$ (para $j = 2, m = 0$), etc.[33]

La extraordinaria similitud con los autovalores $j(j+1)$ para el operador de momento angular total $\mathbf{J}^2 = \boldsymbol{L}_1^2 + \boldsymbol{L}_2^2 + \boldsymbol{L}_3^2$, y con m para la componente $\boldsymbol{L}_3$, como se ha mencionado al final de §22.8 y en §22.10 respectivamente, no debería pasar desapercibida para el lector. De hecho, la dependencia angular de la función de onda, para una partícula con espín entero j, es necesariamente un armónico j-esférico. Además, los autoestados de $\boldsymbol{L}_3$ corresponden a armónicos que son autoestados de $\partial\Phi/\partial\phi$. De hecho, podemos «identificar»

$$\mathbf{J}^2 = -\nabla^2 \text{ y } \boldsymbol{L}_3 = -i\frac{\partial}{\partial\phi}$$

para el comportamiento angular de tales funciones de onda.[34]

Esto no nos da los casos «espinoriales» para los que j es un entero semiimpar (como, en consecuencia, lo es m). Para esto podemos generalizar los que se conocen como «armónicos esféricos ponderados por espín».[35] Estos no son simplemente funciones sobre la esfera S^2, sino que también tienen una dependencia de un vector tangente (espinorial) unidad en cada punto de S^2 (Fig. 22.17). (Pueden considerarse como funciones sobre la S^3 que representa la fibra de vectores espinoriales unidad tangentes a S^2 que nos proporciona el fibrado de Clifford, tal como se describe en §15.4.[36] No es este el lugar de entrar en los detalles del procedimiento, y se remite al lector a la literatura.)

De hecho, la descripción 2-espinorial de los estados de espín, tal como se ha introducido en §22.8 y se ha utilizado para la representación de Majorana de §22.10, está íntimamente relacionada con la teo-

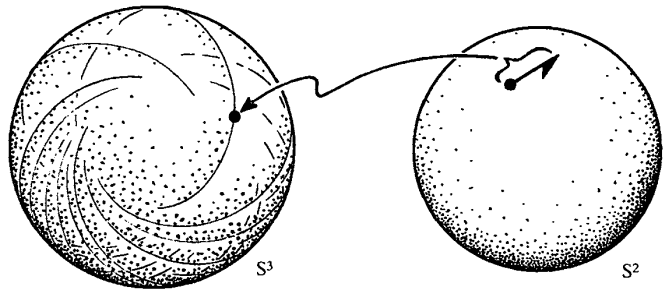


Fig. 22.17. Armónicos esféricos ponderados por espín. Las funciones ponderadas por espín en la esfera S^2 (dibujadas a la derecha) no son simplemente funciones en S^2, sino que también tienen una dependencia de un vector (espinorial) unidad tangente a S^2 en el punto en cuestión (representado aquí por una media flecha para indicar su naturaleza espinorial). Dichas funciones están representadas más adecuadamente en la S^3 a la izquierda, que es la fibra espín-vector de Clifford de la Fig. 15.10. (Una función de «peso de espín s» tiene una dependencia $e^{is\chi}$ del ángulo χ a través del cual el vector de espín rota dentro de su plano tangente a S^2. A medida que χ aumenta, el punto correspondiente en S^2 describe un círculo de Clifford.)

ría de los armónicos esféricos y los armónicos esféricos ponderados por espín. Cualquier tensor de espín simétrico de n índices $\psi_{AB\ldots F}$ corresponde explícitamente a una colección de armónicos esféricos (ponderados por espín) para $j = \frac{1}{2}n$. Para encontrarlos, tomamos dos 2-espinores ξ^A, η^A, con componentes

$$\{\xi^0, \xi^1\} = \{e^{i\phi/2}\cos\frac{\theta}{2}, e^{-i\phi/2}\text{sen}\,\frac{\theta}{2}\},$$

$$\{\eta^0, \eta^1\} = \{-e^{i\phi/2}\text{sen}\,\frac{\theta}{2}, e^{-i\phi/2}\cos\,\frac{\theta}{2}\},$$

de modo que ξ^A y η^A representan puntos diametralmente opuestos en S^2.[22.33] Para calcular cada armónico (ponderado por espín), tomamos la componente de $\psi_{AB\ldots F}$ con respecto a ξ^A y η^A (considerado como un sistema de referencia para variables espinoriales). Estas «componentes» son las cantidades

$$\psi_{A\ldots CD\ldots F}\xi^A \ldots \xi^C\eta^D \ldots \eta^F.$$

[22.33] Explique por qué los puntos son antípodas.

Si el número de ξs en esta expresión es igual al número de ηs ($= j$), entonces obtenemos armónicos ordinarios (en lugar de los ponderados por espín). (En general, el número de ξs y de ηs es $j + s$ y $j - s$, respectivamente, donde s es el «peso de espín».) Obtenemos (múltiplos de)[37] los armónicos esféricos *estándar*, que son autoestados de $\partial/\partial\phi$, si tomamos como $\psi_{AB\ldots F}$ de uno en uno, cada uno de los estados de base estándar mencionados en §22.10, a saber $|\downarrow \ldots \downarrow \uparrow \ldots \uparrow\rangle$. (Para estos, exactamente una de las $n + 1$ componentes independientes de $\psi_{AB\ldots F}$ es no nula.) Debemos tener en cuenta que estos estados de base están *simetrizados*. Por ejemplo, $|\downarrow\uparrow\uparrow\rangle$ es un múltiplo de $|\downarrow\rangle|\uparrow\rangle|\uparrow\rangle + |\uparrow\rangle|\downarrow\rangle|\uparrow\rangle + |\uparrow\rangle|\uparrow\rangle|\downarrow\rangle$. En este caso concreto, todas las componentes de ψ_{ABC} se anulan excepto para la única componente independiente $\psi_{011} = \psi_{101} = \psi_{110}$. Aunque mi descripción de estos temas es algo inapropiada debido a su brevedad,[38] transmite un esbozo de lo que está implicado, y el lector puede empezar a apreciar que los espinores proporcionan lo que en realidad es una ruta extraordinariamente efectiva (aunque poco convencional) a los armónicos esféricos.[22.34] Recordemos de §22.8 (cf. §13.7) que las cantidades tensores de espín $\psi_{AB\ldots F}$ proporcionan un espacio de representación irreducible $(n + 1)$-dimensional para el grupo de rotaciones SO(3), de modo que lo mismo se aplica al espacio de los armónicos esféricos (ponderados por espín) $j = 2n$.

La descripción de Majorana puede obtenerse inmediatamente de esta forma, y los espinores $\alpha_A, \beta_A, \ldots, \varphi_A$ en la descomposición $\psi_{AB\ldots F} = \alpha_{(A}\beta_B \ldots \varphi_{F)}$ corresponden a los ceros de los armónicos esféricos (ponderados por espín) que aparecen en la descripción anterior en la que solo hay ξs, y no ηs. De hecho, mediante estas consideraciones fue como Majorana encontró inicialmente su descripción. Es posible obtener algunas ideas valiosas sobre los armónicos esféricos utilizando el formalismo 2-espinorial. En muchos aspectos, la aproximación espinorial es más fácil de utilizar, pero no resulta muy familiar.

Los armónicos esféricos son importantes en muchas otras áreas, tales como en la física clásica, y en la mayoría de las aplicaciones no hay una conexión especial con el momento angular. (En tales situaciones es

[22.34] Calcule de esta forma, explícitamente, los armónicos esféricos ordinarios (salvo un factor global) para $j = 1, 2, 3$. Compruebe que son autoestados de ∇^2 y $\partial/\partial\phi$.

habitual utilizar la letra «ℓ» en lugar de «j», pues la última parece tener connotaciones de momento angular.) Pequeñas oscilaciones de una burbuja de jabón serían un ejemplo. Otro lo sería el análisis de la distribución de temperatura, sobre la esfera celeste, de la radiación de microondas (de 2,7 K) procedente de las profundidades del espacio, donde estamos particularmente interesados en algunos valores altos de ℓ, de 200 y más. Este análisis tiene gran importancia en cosmología, como veremos en §§27.7,10,11, §28.4 y, especialmente, en §28.10.

El contraste entre las manifestaciones cuánticas y clásicas de los armónicos esféricos es sorprendente y contraintuitivo. En un sistema cuántico en el que las coordenadas θ y ϕ tienen la interpretación angular estándar, el valor j (o ℓ) siempre tiene la interpretación como un momento angular, pero esto no es ni mucho menos cierto para un sistema clásico. En concreto, un sistema con momento angular cero, en mecánica cuántica, debe ser esféricamente simétrico porque una función de onda con $j = 0$ está compuesta solamente por el armónico esférico constante sobre la esfera; pero en física clásica, el momento angular cero (i.e., «no rotante») ¡no implica ciertamente simetría esférica!

Por el contrario, vemos que un sistema cuántico escogido aleatoriamente con un *gran* momento angular (alto valor j) tiene un estado definido por una descripción de Majorana consistente en $2j$ puntos salpicados más o menos aleatoriamente sobre la esfera S^2. Esto no guarda ningún parecido con el estado clásico de momento angular de un sistema con gran momento angular, ¡pese a la impresión común de que un sistema cuántico con valores altos de sus números cuánticos[39] debería aproximarse a un sistema clásico! Para un estado cuántico de apariencia clásica exigimos que los puntos de Majorana se arracimen alrededor de una dirección particular desde el centro de S^2, a saber, la dirección que es el eje (positivo) de giro clásico. ¿Por qué hay tal discrepancia entre estas dos imágenes? La respuesta es que la mayoría de los estados cuánticos «grandes» *no* se parecen a estados clásicos. El ejemplo más famoso es el hipotético gato de Schrödinger, que está en una superposición cuántica de vivo y muerto (véase §29.7). ¿Por qué no vemos realmente cosas como esta en un nivel clásico? Este es un aspecto de la *paradoja de la medida* que se discutirá en los capítulos 29 y 30.

El análisis armónico para espacios más generales que S^2 constituye una parte importante de muchas áreas de investigación científica. Es extraordinariamente valioso cuando se consideran pequeñas perturbaciones u oscilaciones de un sistema. No obstante, resulta oportuno hacer una advertencia. En un espacio que sea no compacto la situación puede hacerse mucho más complicada que la situación de S^2 que se ha considerado antes. Hemos visto algo de esto en el capítulo 9, cuando hemos pasado del análisis de Fourier (en el círculo compacto) a la transformada de Fourier (sobre la recta abierta no compacta). A veces hay una tendencia a creer que se puede trasladar el análisis de una forma compacta a una no compacta —digamos desde la esfera a un espacio hiperbólico— con solo unos pocos cambios de signo (y con las funciones trigonométricas reemplazadas por sus análogas hiperbólicas, de acuerdo con los procedimientos de §18.4). Por desgracia, la verdad puede ser mucho más complicada que esto. Tal «análisis armónico» capta solo una proporción muy pequeña de las funciones relevantes en el espacio hiperbólico, debido a que el sistema de armónicos no es completo.

22.12. Momento angular cuántico relativista

Abordemos ahora la cuestión del momento angular relativista. Recordemos las expresiones clásicas descritas en §18.7. Análoga a la combinación de masa-energía y momento lineal en un 4-vector p_a, existe una cantidad 6-tensorial antisimétrica M^{ab} que describe el momento *angular* y el movimiento del centro de masas de un objeto. ¿Cómo vamos a trabajar mecanocuánticamente con estos?[40]

Hemos visto en §§21.1-3 que las nociones cuánticas de energía y momento representan misteriosamente —o *son* (esencialmente)— los generadores de movimientos de traslación temporal y espacial del espaciotiempo. Análogamente, las componentes del 6-momento angular M^{ab} representan —son— los generadores de los movimientos rotacionales (Lorentz) del espacio de Minkowski $\mathbb{M}$. Junto con los movimientos traslacionales p_a, estos movimientos rotacionales dan lugar a todo el grupo de Poincaré (no reflexivo) (§18.2) —el análogo minkowskiano de los movimientos rígidos de la geometría euclídea.

Más explícitamente, los generadores de los movimientos traslacionales de Poincaré son las componentes p_0, p_1, p_2, p_3 del 4-momento p_a, donde la energía $E = p_0 = i\hbar\partial/\partial x^0$ genera traslación temporal y las tres componentes restantes (i.e., momento lineal) generan análogamente desplazamientos espaciales: $p_1 = i\hbar\partial/\partial x^1$, $p_2 = i\hbar\partial/\partial x^2$, $p_3 = i\hbar\partial/\partial x^3$, donde tenemos en mente que $(-p_1, -p_2, -p_3)$ son las componentes del 3-momento $\mathbf{p}$; véase §18.7. Los movimientos rotacionales de Poincaré en el 3-espacio están generados por las componentes $c^{-2}M^{23} = \boldsymbol{L}_1 = i\hbar\ell_1$, $c^{-2}M^{31} = \boldsymbol{L}_2 = i\hbar\ell_2$, $c^{-2}M^{12} = \boldsymbol{L}_3 = i\hbar\ell_3$, ya consideradas en §22.8, que definen la noción cuántica de momento angular ordinario. Estas son las componentes enteramente espaciales del 6-momento angular M^{ab}, y las 3 restantes componentes independientes[41] $c^{-2}M^{01}$, $c^{-2}M^{02}$, $c^{-2}M^{03}$, que generan las transformaciones de velocidad de Lorentz, se refieren al movimiento del centro de masas de acuerdo con §18.7 (véase la Fig. 18.16). (En esta sección no se ha supuesto $c = 1$.)

Puesto que el grupo de Poincaré es no abeliano, no todos sus generadores conmutan. Sus leyes de conmutación nos dan las leyes de conmutación para nuestros operadores cuánticos p_a y M^{ab}:

$$[p_a, p_b] = 0,$$

$$[p_a, M^{bc}] = i\hbar(g_a{}^b p^c - g_a{}^c p^b),$$

$$[M^{ab}, M^{cd}] = i\hbar(g^{bc}M^{ad} - g^{bd}M^{ac} + g^{ad}M^{bc} - g^{ac}M^{bd}).$$

Estas pueden parecer algo complicadas, pero tienen una importancia fundamental en la física relativista, pues definen el álgebra de Lie (§14.6) del grupo de Poincaré. Parecen algo más sencillas en notación diagramática, como se muestra en la Fig. 22.18.[22.35]

Recordemos que en el caso del momento angular no relativista hemos sido capaces de describir una base de estados en términos de los autovalores $j(j + 1)$ y m de los dos observables que conmutan $\mathbf{J}^2$ y $\boldsymbol{L}_3$; véanse §§22.8,11). Estos operadores proporcionan un conjunto *completo* conmutativo (en el sentido de que cualquier otro operador construido a partir de los generadores $\boldsymbol{L}_1, \boldsymbol{L}_2$ y $\boldsymbol{L}_3$, y que conmuta con $\mathbf{J}^2$ y $\boldsymbol{L}_3$,

[22.35] Demuestre que los conmutadores dados en §22.8 para el momento angular 3-dimensional están contenidos en estos.

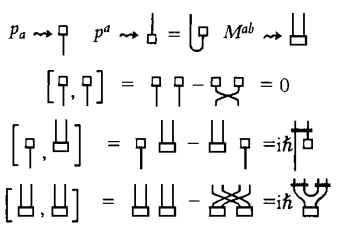


Fig. 22.18. Forma diagramática de los conmutadores cuánticos entre los 4-momentos y los 6-momentos angulares relativistas $[p_a, p_b] = 0$, $[p_a, M^{bc}] = i\hbar(g_a{}^b p^c - g_a{}^c p^b)$, $[M^{ab}, M^{cd}] = i\hbar(g^{bc}M^{ad} - g^{bd}M^{ac} + g^{ad}M^{bc} - g^{ac}M^{bd})$.

no da nada nuevo, puesto que él mismo debe ser función de estos dos). Una parte importante de la mecánica cuántica consiste en general en encontrar un conjunto completo conmutativo de este tipo para el sistema en consideración. Muy en especial nos gustaría poder hacer esto para operadores construidos a partir de las componentes de p_a y M^{ab}, y utilizar sus autovalores para clasificar partículas o sistemas relativistas.

¿Por qué estamos interesados en observables que conmutan? La razón es que si $\boldsymbol{A}$ y $\boldsymbol{B}$ son dos de estos —de modo que $\boldsymbol{AB} = \boldsymbol{BA}$—, entonces podemos encontrar estados $|\psi_{rs}\rangle$ que son autoestados de ambos simultáneamente, y el par de autovalores correspondientes (a_r, b_s) puede utilizarse para etiquetar dichos estados.[42] Si tenemos un conjunto completo de observables conmutativos $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}, \boldsymbol{D}, \ldots$ (cuyos autoestados llenan el espacio en consideración), entonces tenemos una familia de estados de base $|\psi_{rstu\ldots}\rangle$, que pueden etiquetarse por la correspondiente familia de autovalores $(a_r, b_s, c_t, d_u, \ldots)$.[22.36]

Para obtener un conjunto completo conmutativo, es habitual empezar encontrando los operadores de *Casimir*, que son operadores (escalares) que conmutan con *todos* los operadores del sistema en consideración. En el caso del momento angular 3-dimensional ordinario, hay solo *un* operador de Casimir (independiente),[43] a saber, $\mathbf{J}^2 = \boldsymbol{L}_1^2 + \boldsymbol{L}_2^2 + \boldsymbol{L}_3^2$. Una pregunta importante es: ¿cuáles son los operadores de Casimir para el sistema generado por p_a y M^{ab} que satisfacen las leyes de conmutación anteriores?

[22.36] Desarrolle los detalles de estas afirmaciones, donde puede suponer, por conveniencia, que los autoestados forman un sistema discreto y no continuo. Suponga primero que no hay autovalores degenerados, y demuestre luego cómo discurre el argumento cuando existen degeneraciones. *Sugerencia*: Exprese cada autovector de $\boldsymbol{A}$ en términos de autovectores de $\boldsymbol{B}$, y así sucesivamente.

Ahora el espín alrededor del centro de masas está definido por la cantidad

$$S_a = \frac{1}{2}\varepsilon_{abcd}M^{bc}p^d,$$

llamada *vector de espín de Pauli-Lubanski*, donde el ε_{abcd} antisimétrico de Levi-Civita se ha definido en §19.2, pero aquí tenemos $\varepsilon_{0123} = c^{-3}$, puesto que no se ha supuesto $c = 1$. (En la «notación del matemático» podríamos escribir $\boldsymbol{S} = *(\boldsymbol{M} \wedge \boldsymbol{p})$, utilizando ahora $\boldsymbol{p}$ para representar el 4-momento, en lugar del 3-momento anterior; cf. §11.6, §12.8, §19.2.) Hemos visto que una única partícula clásica sin estructura tiene $M^{ab} = x^a p^b - x^b p^a$, donde x^a es el vector de posición de un punto en la línea de universo de la partícula (véase el final de §18.7). Tomamos la misma expresión en el caso cuántico, de lo que se sigue que $S^a = 0$ para una partícula semejante. Pero S^a no tiene por qué anularse para un sistema total de dos o más de tales partículas. Además, para una única partícula con espín, el momento angular M^{ab} no tiene esta forma simple pues hay un término de *espín* adicional $\mu^{-2}\varepsilon^{abcd}S_c p_d$, suponiendo $\mu \neq 0$ (véase la nota 18.11). Encontramos que S^a es siempre ortogonal a p_a ($p_a S^a = 0$) y conmuta con p_a (i.e. $[S^a, p_b] = 0$), de modo que S^a, como p_a, es independiente del origen.[22.37]

Existen dos operadores de Casimir independientes (para el grupo de Poincaré), a saber,

$$p_a p^a = c^4\mu^2 \quad \text{y} \quad S_a S^a = \mu^2\mathbf{J}^2,$$

siendo μ la masa en reposo del sistema entero.[22.38] Encontramos que el «$\mathbf{J}^2$» definido en la segunda de las ecuaciones anteriores es, de hecho, $\mathbf{J}^2 = \boldsymbol{L}_1^2 + \boldsymbol{L}_2^2 + \boldsymbol{L}_3^2$, donde $\boldsymbol{L}_1$, $\boldsymbol{L}_2$ y $\boldsymbol{L}_3$ son las componentes del momento angular respecto al centro de masas en su sistema de reposo. Para completar el conjunto de operadores conmutativos, podemos escoger p_1, p_2, p_3, y una componente, por ejemplo S_3, del vector de espín que —junto con $p_a p^a$ y $S_a S^a$— nos dan un total de seis. (Aunque, en detalle, son posibles otras muchas elecciones, el número total de opera-

[22.37] Establezca las propiedades afirmadas en estas cuatro sentencias.
[22.38] Dé una razón sencilla por la que estos operadores mostrados deben conmutar con p_a y M^{ab}. *Sugerencia*: Eche una ojeada a §22.13.

dores independientes es siempre seis.)[44] Esto tiene una importancia considerable para las discusiones de §22.13 y §31.10.

La situación es así muy similar al caso no relativista, donde, para incluir traslaciones espaciales y temporales, podríamos escoger la energía E como un «operador de Casimir» para complementar la magnitud $\mathbf{J}^2$, y las tres componentes del momento, además de $\boldsymbol{L}_3$. Habría que advertir que en el caso relativista no obtenemos *directamente* $\mathbf{J}^2$, sino

$$\mathbf{J}^2 = -c^4(p_a p^a)^{-1} S_a S^a,$$

que nos da algo básicamente equivalente con tal de que $p_a p^a \neq 0$. De hecho, en la exposición anterior suponíamos que la masa en reposo μ no se anula. Si $\mu = 0$, no podemos expresar de este modo la magnitud del espín.

¿Cómo tratamos el caso sin masa $\mu = 0$? Recuperamos, en su lugar, la *helicidad* s, una cantidad que ya hemos visto, en el caso del fotón, en §§22.7,9. Esta se define por el requisito físico de que el vector de Pauli-Lubanski S^a sea proporcional al 4-momento p_a:[22.39]

$$S_a = s p_a.$$

La helicidad dextrógira está dada por $s > 0$, y la levógira por $s < 0$, aunque también está permitido $s = 0$. Ahora tenemos cuatro observables conmutativos independientes, que podemos considerar que son s, p_1, p_2, p_3. De hecho, resulta que la forma más clara, con mucho, de tratar el caso sin masa es acudir a la *teoría de twistores*. Llegaremos a esta en §33.6 (donde veremos que las «variables twistoriales» Z^0, Z^1, Z^2, Z^3 también pueden utilizarse como cuatro operadores conmutativos independientes).

22.13. El objeto cuántico aislado general

¿Cómo describe la mecánica cuántica un objeto aislado en general, tal como un átomo o una molécula? Estoy suponiendo que no hay fuerzas externas que actúan sobre el objeto y que este permanece localizado,

[22.39] ¿Cómo pueden S_a y p_a ser ortogonales y proporcionales?

aunque podría haber fuerzas internas actuando en su interior. Como característica importante de la descripción de tal objeto, dividimos esta descripción en (i) la caracterización externa del objeto como un todo y (ii) su funcionamiento detallado y su estructura geométrica interna.

Esta caracterización externa (i) se refiere a su masa-energía global, su momento, la posición y movimiento de su centro de masa, y su momento angular. Tomemos estas cantidades en su sentido relativista y utilicemos las p_a y M^{ab} de §22.12 para describir los parámetros externos. El funcionamiento interno (ii) se refiere a las partículas constituyentes, su naturaleza concreta, la naturaleza de las fuerzas que actúan entre ellas, y sus relaciones geométricas. Dichas relaciones se consideran dadas por ciertas coordenadas generalizadas q_r (§20.1) de naturaleza totalmente relativa[45] (por ejemplo, la distancia de cierta parte al centro de masa, o el ángulo que forman las diversas partes entre sí, o sus distancias mutuas). Por lo tanto, no cambian si el objeto entero es desplazado mediante un movimiento de traslación espacial o temporal, o rotado un cierto ángulo definido, o movido en alguna dirección con velocidad uniforme.

Debido a su naturaleza relativa, todas las coordenadas internas son invariantes por cualquier simetría del grupo de Poincaré. Se sigue que deben conmutar con p_a y M^{ab}. ¿Por qué? Supongamos que cierto operador de simetría $\boldsymbol{S}$ actúa sobre un sistema cuántico de acuerdo con

$$|\psi\rangle \mapsto \boldsymbol{S}|\psi\rangle,$$

y $\mathbf{Q}$ es cierto operador cuántico; entonces, la acción del operador de simetría sobre $\mathbf{Q}$ es[22.40]

$$\mathbf{Q} \mapsto \boldsymbol{S}\mathbf{Q}\boldsymbol{S}^{-1}.$$

Si $\mathbf{Q}$ es invariante por $\boldsymbol{S}$, entonces $\boldsymbol{S}\mathbf{Q}\boldsymbol{S}^{-1} = \mathbf{Q}$, de donde

$$\boldsymbol{S}\mathbf{Q} = \mathbf{Q}\boldsymbol{S}.$$

Así pues, tomando como $\boldsymbol{S}$ cada una de las componentes de p_a y M^{ab}, vemos que cada parámetro interno debe conmutar, de hecho, con p_a y M^{ab}.

En el contexto actual, esto significa que podemos *separar* la parte de

[22.40] Explique por qué. *Sugerencia*: Una ojeada a §22.4 puede ayudarle.

la función de onda que se refiere a los grados de libertad internos de la que se refiere a los parámetros externos del 4-momento y el 6-momento angular. En los tratamientos habituales suponemos que el sistema está en un autoestado del apropiado sistema completo de observables externos. En particular, la energía y el momento se considerarían dados por autovalores definidos, y sería habitual referir las cosas al sistema de referencia en el que el 3-momento es cero ($\mathbf{P} = 0$, en la notación de §21.5). Entonces, el momento angular puede tratarse de acuerdo con las discusiones no relativistas de §§22.8-11, de modo que podemos pedir que el sistema esté en un autoestado del momento angular total $\mathbf{J}^2$, y también de $\boldsymbol{L}_3$, si lo deseamos.

Los parámetros internos dependerán, por supuesto, de los detalles del sistema concreto bajo consideración. En ciertas circunstancias puede ser una buena aproximación considerar que los grados de libertad internos están bien descritos por pequeñas oscilaciones en torno al equilibrio. Entonces, el análisis clásico de §20.3 tiene relevancia. Recordemos de §20.3 que si tomamos un hamiltoniano de la forma

$$\mathcal{H} = \tfrac{1}{2}Q_{ab}q^a q^b + \tfrac{1}{2}P^{ab}p_a p_b,$$

donde Q_{ab} y P^{ab} son simétricos, definidos positivos y constantes, entonces, en el caso clásico, cada frecuencia normal $\omega/2\pi$ surge de un autovalor ω^2 de la matriz $\boldsymbol{W} = \boldsymbol{PQ}$ (i.e., de $W^a{}_c = P^{ab}Q_{bc}$).

Pero ¿qué pasa con la mecánica cuántica? Si recordamos la relación de Planck $E = h\nu = 2\pi\hbar\nu$, donde ν es la frecuencia, podríamos esperar una energía $E = \hbar\omega$ para una oscilación en dicho modo normal concreto. Quizá podríamos anticipar también valores más altos para la energía, puesto que clásicamente la amplitud de la oscilación podría llegar a ser tan alta como queramos (mientras no se perturbe la naturaleza de «pequeña oscilación» de la aproximación), y con mayor amplitud obtenemos mayor energía. Si suponemos que están implicados «armónicos superiores» —y recordamos de §9.1 que estos ocurren con frecuencias que son múltiplos enteros de la frecuencia básica $\omega/2\pi$—, entonces podríamos imaginar que los autoestados cuánticos permitidos para la energía serían:

$$0, \hbar\omega, 2\hbar\omega, 3\hbar\omega, 4\hbar\omega, \ldots$$

De hecho, esto no está muy lejos de la respuesta mecanocuántica correcta, pero resulta que existe una contribución adicional $\frac{1}{2}\hbar\omega$ a la energía, denominada *energía de punto cero*.[46] Los autoestados de la energía permitidos son entonces:

$$\tfrac{1}{2}\hbar\omega, \quad \tfrac{3}{2}\hbar\omega, \quad \tfrac{5}{2}\hbar\omega, \quad \tfrac{7}{2}\hbar\omega, \quad \tfrac{9}{2}\hbar\omega, \ldots$$

Esto procede de la discusión cuántica estándar del oscilador armónico 1-dimensional,[47] para el que el hamiltoniano es $\mathcal{H} = (m^2\omega^2q^2 + p^2)/2m$. Para cada modo por separado existe una contribución a la energía de uno de estos valores, por cada autovalor ω de la matriz $\boldsymbol{W}$.

Para un sistema cuántico general, estos valores serían solo aproximaciones, porque podrían empezar a cobrar importancia los términos de orden más alto. Sin embargo, hay varios sistemas que pueden aproximarse muy bien de esta manera. Y, lo que es más importante, resulta que la teoría cuántica de campos de fotones —o de cualquier otra partícula del tipo conocido como «bosón» (véanse §23.7 y §26.2)— puede tratarse como si el sistema de fotones fuera un conjunto de osciladores. Estos osciladores son exactamente del tipo armónico simple considerado antes (donde no hay términos de orden superior en el hamiltoniano), cuando los fotones están en un estado estacionario sin interacciones entre ellos.[48] En consecuencia, esta imagen de «oscilador armónico» proporciona un esquema aplicable con amplia generalidad. De todas formas, para proceder de forma más general es necesario un conocimiento detallado de las interacciones.

Por ejemplo, un átomo de hidrógeno consiste en un electrón en órbita alrededor del protón de su núcleo (normalmente considerado fijo, como buena aproximación, puesto que el protón es mucho más masivo que el electrón en un factor de aproximadamente 1836). Pero las reglas de la mecánica cuántica nos dicen que la órbita mecanocuántica no incluirá solo una única trayectoria clásica alrededor del núcleo, sino una superposición cuántica de muchas de estas. Dichas «órbitas cuánticas» serán soluciones estacionarias de la ecuación de Schrödinger, con un hamiltoniano que es básicamente el mismo que en el caso clásico, aunque «canónicamente cuantizado» de acuerdo con las reglas de §§21.2,3 (y de §23.8, donde sea necesario). Para un autoestado de

momento angular, encontramos funciones de onda que son armónicos esféricos en su dependencia angular (§22.11). En general, podríamos utilizar el autovalor E de la energía y el autovalor j del momento angular (junto con m, si es apropiado) como números cuánticos que etiquetan los diversos estados. En el caso del átomo de hidrógeno (si ignoramos los espines del electrón y el protón y consideramos una forma no relativista del hamiltoniano), encontramos que el autovalor E de la energía está determinado por el autovalor j del momento angular total, aunque accidentalmente j no está determinado por E. En una teoría más aproximada del átomo de hidrógeno (y para átomos más complicados), encontramos en general que E determina j, de modo que todos los estados diferentes están en realidad caracterizados únicamente por el valor de la energía.

En la primitiva teoría de Bohr para el átomo, propuesta en 1913, más de una década antes de la mucho más precisa y completa mecánica cuántica de Heisenberg, Schrödinger y Dirac, los valores permitidos del momento angular y la energía del hidrógeno se calculaban como si las órbitas fueran las clásicas órbitas elípticas de Kepler-Newton —dadas por la ley de la inversa del cuadrado para la atracción electrostática entre el núcleo y el electrón orbital—, pero con la «condición cuántica» de que el momento angular del electrón debía ser un múltiplo entero de $\hbar$. Tales «órbitas cuantizadas» se conocen a veces como *orbitales*; véase la Fig. 22.19. Este procedimiento funcionaba sorprendentemente bien,[49] pero no estaba apoyado por los soportes teóricos que proporcionó la mecánica cuántica posterior, que condujeron a resultados de mucha mayor generalidad y precisión. Los átomos más complicados, las moléculas sencillas, los efectos relativistas, la presencia del espín electrónico y nuclear, etc., pueden tratarse todos mediante el formalismo cuántico, aunque hay que utilizar técnicas de aproximación y computación numérica, en lugar de tratamientos matemáticos exactos.

La utilización que se ha hecho de la electrostática es también una aproximación, y deben permitirse transiciones de un estado estacionario a otro mediante la emisión/absorción de fotones. Esto requiere la teoría de Maxwell, aunque en su forma cuantizada, lo que hace necesario el formalismo de la *teoría cuántica de campos* (que será esbozada en

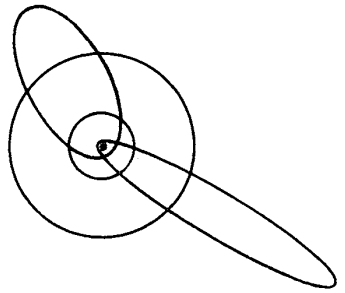


Fig. 22.19. «Átomo de Bohr», donde los electrones orbitales están vistos como si tuvieran básicamente órbitas elípticas de Kepler-Newton, de acuerdo con la ley inversa del cuadrado para la atracción electrostática, pero donde sus energías y momentos angulares están limitados por la «condición cuántica» de que los momentos angulares orbitales deben ser múltiplos enteros de $\hbar$. La idea se aplica más satisfactoriamente a órbitas circulares para el único electrón del hidrógeno.

el capítulo 26). También será necesario el electrón relativista de Dirac, del capítulo 24 para una precisión total. Un átomo en un autoestado de mínima energía, llamado *estado fundamental*, permanecerá en dicho estado (suponiendo que esté completamente aislado de perturbaciones ambientales); pero si está en un estado más energético —conocido como *estado excitado*—, habrá normalmente[50] una probabilidad de que caiga al estado fundamental, con la emisión de uno o más fotones. Por esta razón, se espera encontrar átomos o moléculas libres en sus estados fundamentales, o cerca de sus estados fundamentales. La frecuencia ν de un único fotón emitido cuando un átomo o molécula cae de un estado a otro está determinada, vía la $E = 2\pi\hbar\nu$ de Planck (véase §21.4) y la conservación de la energía, por la *diferencia* de energía E entre los dos estados.

Tales frecuencias han sido observadas hace tiempo en las *líneas espectrales*, cuya explicación había sido un persistente rompecabezas científico. La extraordinaria riqueza de información contenida en estas pautas de las líneas espectrales observadas queda explicada por la teoría cuántica de la forma mencionada antes. ¡Esto ha supuesto uno de los triunfos supremos de la física del siglo XX! La física clásica anterior (incluida la ley de Coulomb de la inversa del cuadrado para la atracción entre cargas positivas y negativas y las ecuaciones de Maxwell, que nos dicen cómo los electrones acelerados deben radiar energía electromagnética) predecía claramente que los electrones en órbita deberían caer catastróficamente al núcleo describiendo una espiral, para dar un estado *singular*, en un tiempo muy corto. Esta conclusión estaba en fla-

grante contradicción con los hechos observados. La mecánica cuántica no solo eliminó esta paradoja, sino que ofreció una teoría detallada de las líneas espectrales que ha proporcionado una herramienta extraordinariamente potente en muchas áreas de la ciencia, desde la ciencia forense hasta la física nuclear y la cosmología.

Como comentario final de considerable importancia, puede señalarse que la existencia de números cuánticos *discretos*, tales como los j y m del momento angular, o de los autoestados de la energía para el oscilador armónico o el átomo de hidrógeno, etc., aparecen en definitiva a partir de la *compacidad* de un espacio.[51] En el caso del momento angular, procede de la compacidad de la *esfera de direcciones espaciales*, que es el S^2 al que se aplica el análisis armónico de §22.11. Sin algo como la compacidad (o periodicidad), solo tendríamos soluciones de ecuaciones como $\nabla^2\Phi = -k\Phi$, en la que el autovalor k no está restringido. Resulta irónico que, en *ausencia* de tal compacidad, el formalismo general de la mecánica cuántica no habría proporcionado el sorprendente carácter *discreto* que dio origen a la teoría ¡y a partir del cual surgió originalmente el nombre mismo de «cuántica»!

Notas

Sección 22.1

22.1. Parece que son estos saltos los que estimularon la expresión ahora coloquial «salto cuántico». Para un físico, esta elección de palabras es muy extraña, pues los saltos cuánticos que ocurren en la reducción del estado cuántico tienden a ser extraordinariamente minúsculos, apenas detectables, ¡y quizá sucesos irreales!

22.2. Para una discusión general de los diferentes puntos de vista sobre la mecánica cuántica, véanse Rae (1994), Polkinghorne (2002), Home (1997), o DeWitt y Graham (1973).

Sección 22.2

22.3. Véanse el capítulo 29 y Everett (1957), Wheeler (1957), DeWitt y Graham (1973), Geroch (1984) y Deutsch (2000).

Sección 22.3

22.4. Como sucede con el caso de una sola partícula (§21.9), algunos autores podrían utilizar «$\|\psi\|^2$» en lugar de mi $\|\psi\|$.

22.5. Para una encantadora introducción al estudio de tales espacios, véanse Chen (2002) y Reed y Simon (1972).

22.6. En la literatura mecanocuántica se utiliza frecuentemente la notación $\mathbf{Q}^\dagger$ en lugar de la $\mathbf{Q}^*$ utilizada en la mayor parte de la literatura matemática relevante; véase §13.9.

Sección 22.4

22.7. Véase Dirac (1982) para la reimpresión más reciente. Véase Shankar (1994) para un tratamiento más reciente.

22.8. Véase Dirac (1966) para un argumento según el cual la imagen de Schrödinger no existe en la teoría cuántica de campos relativista.

22.9. La imagen de interacción, por ejemplo, se utiliza con frecuencia en cálculos en la «teoría de perturbaciones dependientes del tiempo». Véanse Shankar (1994), capítulo 18, y Dirac (1966).

Sección 22.5

22.10. Aquí estoy ignorando el hecho de que los autoestados no son normalizables, lo que podría descalificar la posición o el momento como *verdaderos* «observables» en algunas formulaciones.

22.11. Con más generalidad, sea o no un autovalor, decimos que q es el valor esperado de $\mathbf{Q}$ para el estado normalizado $|\psi\rangle$ si $q = \langle\psi|\,\mathbf{Q}\,|\psi\rangle$.

22.12. Véase Dirac (1982a). Parámetros complejos se utilizan en muchos campos, por ejemplo, Fortney (1997).

Sección 22.6

22.13. Un elegante tratamiento de las medidas de proyección puede encontrarse en Kraus (1983) y Nielsen y Chuang (2000).

22.14. Véase von Neumann (1955).

22.15. Véanse Lüders (1951) y también Penrose (1994).

Sección 22.7

22.16. Véase Elitzur y Vaidman (1993).

22.17. La idea original de medidas libres de interacción parece que se debe a Robert Dicke (cf. 1981). Tiene algunas aplicaciones muy sorprendentes, como en la detección de ondas gravitatorias; véase Braginsky

(1977). El extraordinario experimento mental de «comprobación de bombas» de Elitzur-Vaidman descrito aquí (véase también Penrose, 1994) podría llevar a otro tipo de aplicaciones.

22.18. Ahora hay evidencia de que al menos alguno de los varios tipos de neutrinos es *masivo*, y quizá lo sean *todos*. Aun así, la hipótesis de que no tiene masa proporciona una buena aproximación a su comportamiento. Volveré a esta cuestión en §25.3.

22.19. Este refinamiento del postulado de proyección se debe al parecer a Lüders (1951) y, en este caso, el punto $|\rho\text{–}\rangle$ en $\mathbb{P}\mathbf{H}^4$ se conoce como el «punto de Lüders».

22.20. Para un ejemplo más económico (e interesante), podríamos considerar una situación ligeramente diferente en la que como divisor de haz se utiliza la superficie de un medio refractivo, y donde, en lugar de estar polarizado, el haz incidente se envía a un *ángulo de Brewster* para el medio. El haz reflejado tiene entonces una polarización lineal específica, y el haz transmitido, la polarización lineal opuesta. El análisis es básicamente el mismo que antes (con polarización lineal en lugar de circular), pero ahora no necesitamos asegurar que el fotón incidente está polarizado, pues el mero hecho de que procede del exterior del medio (con un ángulo apropiado) y no del interior, es suficiente para asegurar que la medida nula produce el estado polarizado requerido; véanse Becker (1982) y Jackson (1998).

Sección 22.8

22.21. Una «partícula sin estructura» no tendría momento angular respecto a su centro, puesto que la expresión $\mathbf{M} = 2\mathbf{x} \wedge \mathbf{p}$ de §18.6 se anula cuando $\mathbf{x} = 0$. Pero, como se ha señalado en la nota 18.11, es necesario añadir al momento angular una cantidad que describe el «espín intrínseco» cuando existe la «estructura» definida por el espín de la partícula. Veremos esto más explícitamente en §22.12.

22.22. El lector avisado puede preguntarse si aquí hay implicadas sutilezas de signo, del tipo que hemos encontrado en §21.5, que surgen de la signatura de la métrica. Para un desarrollo detallado de la teoría del momento angular en mecánica cuántica, a partir del álgebra de Lie de SO(3), el lector debería ver las lúcidas exposiciones de Jones (2002) y Elliot y Dawber (1984). Una derivación alternativa, y algo más «física» (aunque más complicada) de la teoría del momento angular se da en Shankar (1994).

22.23. Véase Penrose y Rindler (1984).

22.24. Véase Geroch (University of Chicago Lectures; no publicado).

22.25. Witten (1959), Penrose (1960, 1968a), Geroch (1968, 1970), Penrose y Rindler (1984, 1986) y O'Donnell (2003).

22.26. La palabra «independiente» se utiliza en el sentido de que todas las componentes de $\psi_{AB\ldots F}$ pueden obtenerse algebraicamente a partir de este conjunto independiente, pero no de un conjunto más pequeño. Aquí esto aparece simplemente por simetría, de modo que el total de 2^n componentes se reduce a $n + 1$ componentes independientes (por ejemplo, ψ_{001}, ψ_{101} y ψ_{100} no son trivialmente componentes independientes, puesto que $\psi_{001} = \psi_{101} = \psi_{100}$).

22.27. Véase Shankar (1994).

Sección 22.9

22.28. Véase la nota 21.12; Nielsen y Chuang (2000) también discuten algunos aspectos de la ciencia de la información cuántica desde un punto de vista similar.

Sección 22.10

22.29. Véase Majorana (1932).

22.30. Véase Biedenharn y Louck (1981) para una revisión general. Para una interesante aplicación moderna, véase Swain (2004). Véase también la nota 22.31.

22.31. Véanse Penrose (1993, 1994) y Zimba y Penrose (1993).

Sección 22.11

22.32. En el contexto de los armónicos esféricos, se utiliza con frecuencia la letra ℓ, en lugar de la j que estoy utilizando aquí.

22.33. Para más detalles, véase cualquier texto de mecánica cuántica como, por ejemplo, Shankar (1994) o Arfken y Weber (2000).

22.34. Shankar (1994).

22.35. Véanse Newman y Penrose (1966) y Penrose y Rindler (1984).

22.36. Véase Goldberg *et al.* (1967).

22.37. Hay también propiedades de ortogonalidad y normalización (para fijar la escala global) de los armónicos esféricos que son importantes para utilizarlos y calcular con ellos. No obstante, estas cuestiones nos llevarían demasiado lejos, y se remite al lector a las siguientes exposiciones de la teoría de los armónicos esféricos: Groemer (1996) y Byerly (2003).

22.38. El lector que quiera profundizar más en el álgebra y la geometría espinorial debería tomar nota del hecho de que los índices espinoriales pueden «subirse» o «bajarse» de acuerdo con el esquema $\xi_1 = \xi^0$, $\xi_0 = -\xi^1$. Véanse Penrose y Rindler (1984) y Zee (2003), apéndice.

22.39. El término «número cuántico» se refiere normalmente a los posibles autovalores discretos de un observable cuántico significativo, tal como momento angular, carga, número bariónico, etc., que se utiliza para clasificar una partícula o un sistema cuántico sencillo. Véase §3.5.

Sección 22.12

22.40. En los capítulos 24-26 veremos que una propiedad mecanocuántica propiamente relativista (especial) requiere mucho más que las consideraciones básicas de esta sección, pero esto no afectará a la exposición actual.

22.41. Recuérdese la nota 22.6. Aquí la independencia tiene en cuenta la *anti*simetría de M^{ab}.

22.42. Hay una relación entre esto y el fenómeno de «separación de variables», que ocurre cuando una función general $f(\theta, \phi)$, pongamos por caso, puede escribirse como una suma $f(\theta, \phi) = \Sigma\, \lambda_{ij} g_i(\theta) h_j(\phi)$, donde $g_i(\theta)$ y $h_j(\phi)$ son autofunciones respectivas de operadores (conmutativos) apropiados $\boldsymbol{A}$ y $\boldsymbol{B}$. Los armónicos esféricos tienen esta propiedad. Véanse Groemer (1995) y Byerley (2003).

22.43. «Independiente» se refiere aquí a independencia funcional (compárese con la nota 22.26). Así, mientras que $2\mathbf{J}^2$, $(\mathbf{J}^2)^3$ y $\cos(\mathbf{J}^2)$ no son el mismo operador de Casimir que $\mathbf{J}^2$, estos no son independientes de $\mathbf{J}^2$.

22.44. Es necesaria cierta precaución respecto a la invariancia del «número de operadores conmutativos independientes». Estrictamente, esto se refiere a la dimensión de un espacio que se aplica a las soluciones *locales* de ecuaciones en derivadas parciales. En problemas mecanocuánticos es probable que haya requisitos de compacidad en el espacio de la solución (por ejemplo, el S^2 de §22.11) que restringen severamente los autovalores permitidos y confunden el recuento de los grados de libertad.

Sección 22.13

22.45. Aquí se ignoran cuestiones de relatividad general, de modo que «relativo» se toma en el sentido de la relatividad *especial*.

22.46. Sin embargo, existe libertad para añadir una constante al hamiltoniano,

como se permitía en §20.3, que simplemente redefine el cero de energía (cf. el ejercicio [24.2] de §24.3), de modo que a veces se considera que esta adición de $\frac{1}{2}\hbar\omega$ no tiene relevancia física directa.

22.47. Véase, por ejemplo, el clásico tratamiento de Dirac en *Los principios de la mecánica cuántica*, Dirac (1982a).

22.48. Las cantidades $\eta = (2m\hbar\omega)^{-1/2}(p + imq)$, en la imagen de Heisenberg §22.4 desempeña el papel de operadores de creación de §26.2.

22.49. En particular, explicaba la hasta entonces incomprensible fórmula de Balmer para las frecuencias de las líneas espectrales del hidrógeno $\nu = R(N^{-2} - M^{-2})$, donde R es una constante (conocida como la constante de Rydberg-Ritz) y donde $M > N > 0$ son enteros.

22.50. Puede haber «reglas de selección», surgidas de leyes de conservación, que prohíben algunas de estas transiciones.

22.51. Compárese con la nota 22.44.

23

El entrelazado mundo cuántico

23.1. Mecánica cuántica de sistemas de muchas partículas

En los dos capítulos anteriores hemos visto cuán misterioso es el comportamiento de las partículas cuánticas individuales, con o sin espín, y qué extraño y maravilloso formalismo matemático se ha desarrollado para tratar este comportamiento. Y puesto que nuestro formalismo nos ha descrito el comportamiento cuántico de partículas individuales, no sería irrazonable esperar que también nos dijera cómo describir sistemas que contienen varias partículas, que quizá interaccionen mutuamente de formas diversas. De algún modo, esto es verdad —hasta cierto punto—, puesto que el formalismo general de §21.2 es suficientemente general para ello, pero surgen nuevos aspectos cuando en el sistema hay presente más de una partícula. Lo que resulta nuevo es el fenómeno del *entrelazamiento cuántico*, por el que un sistema de más de una partícula debe tratarse como una sola unidad holística; diferentes manifestaciones de este fenómeno se nos presentan en el comportamiento cuántico de forma más misteriosa incluso de lo que ya hemos visto. Además, las partículas que son idénticas están siempre *automáticamente* entrelazadas unas con otras, aunque observaremos que esto puede suceder de dos modos completamente distintos, dependiendo de la naturaleza de la partícula.

Volvamos a lo que hemos establecido en los dos capítulos anteriores para las matemáticas de un sistema cuántico. La aproximación del hamiltoniano cuántico, que nos da la ecuación de Schrödinger para la evolución del vector de estado cuántico, se sigue aplicando cuando hay muchas partículas que pueden estar en interacción y pueden estar girando, exactamente igual que se aplicaba para una única partícula sin espín. Lo único que necesitamos es un hamiltoniano apropiado para incorporar todos estos aspectos. No tenemos una función de onda independiente para cada partícula; en su lugar, tenemos *un* vector de estado que describe el sistema entero. En una representación en el espacio de posiciones, este único vector de estado puede seguir considerándose como una función de onda Ψ, pero sería una función de todas las coordenadas de posición de *todas* las partículas, de modo que en realidad es una función en el espacio de *configuración* del sistema de partículas (véase §12.1), y también podría depender de ciertos parámetros discretos que etiquetan los estados de espín (por ejemplo, si utilizamos una descripción 2-espinorial $\Psi_{AB\ldots F}$ para describir una partícula giratoria, como en §22.8, entonces los «parámetros discretos» etiquetarían las diferentes componentes individuales). La ecuación de Schrödinger nos dirá cómo evoluciona Ψ en el tiempo, de modo que Ψ tendrá que depender también de la variable temporal t.

Una característica digna de mención de la teoría cuántica estándar es que para un sistema de muchas partículas existe solo una coordenada temporal, mientras que cada una de las partículas independientes incluidas en el sistema cuántico tiene su propio conjunto independiente de coordenadas de posición. Esta es una característica curiosa de la mecánica cuántica no relativista si nos gusta considerarla como un tipo de aproximación límite a cierta teoría *relativista* «más completa», ya que en un esquema relativista deberíamos tratar el tiempo esencialmente de la misma forma que tratamos el espacio. Puesto que cada partícula tiene sus propias coordenadas espaciales, también debería tener su propia coordenada temporal. Pero no es así como trabaja la mecánica cuántica ordinaria. Hay solo un tiempo para todas las partículas.

Cuando pensamos en física de una forma ordinaria «no relativista», esto puede parecer plausible, puesto que en la física no relativista el tiempo es externo y absoluto, y simplemente «marcha» de una forma universal, con independencia de los contenidos concretos del universo en cualquier instante. Pero, desde la introducción de la relatividad, sabemos que dicha imagen solo puede ser una aproximación. Lo que es el «tiempo» para un observador es una mezcla de espacio y tiempo para otro, y viceversa. La teoría cuántica ordinaria exige que cada partícula

debe llevar individualmente su propia coordenada espacial. Por consiguiente, también debería llevar su propia coordenada temporal en una teoría cuántica adecuadamente relativista. De hecho, este punto de vista ha sido adoptado en ocasiones por varios autores,[1] y se remonta hasta finales de la década de 1920, pero no parece que se haya desarrollado en una teoría relativista generalizada. Una dificultad básica en permitir que cada partícula tenga su propio tiempo independiente es que entonces cada partícula parece ir a su aire en una dimensión temporal independiente, y se necesitan ingredientes adicionales para traernos de nuevo a la realidad.

En §26.6 introduciré la aproximación de «integrales de camino» de la teoría cuántica relativista, que se basa en un formalismo lagrangiano relativista, y no en un formalismo hamiltoniano, y evita el problema «un tiempo/muchos espacios»; sin embargo, como veremos más adelante, aparecen nuevos y serios problemas, como siempre parecen serlo, no importa qué procedimiento (conocido) se utilice. Más aún, veremos enseguida que la propia ecuación de Schrödinger ordinaria no es inmune a las dificultades de «volver a la realidad». En mi opinión, esta simple asimetría espaciotemporal del enfoque de Schrödinger oculta algo profundo que sigue estando ausente de nuestra imagen cuántica de las cosas; pero esto no debería preocuparnos ahora. Volveré a ello más tarde. Por el momento ignoraré estas cuestiones y presentaré las cosas simplemente desde el punto de vista de la teoría cuántica no relativista, donde puede considerarse aplicable la noción de un tiempo externo universal. Pero la cuestión de la relatividad no desaparecerá, y tendremos que volver a ella al final de este capítulo, en §23.10.

¿Cómo vamos a tratar entonces los sistemas de muchas partículas de acuerdo con la imagen de Schrödinger estándar no relativista? Como se ha descrito en §21.2, tendremos un único hamiltoniano en el que deben aparecer todas las variables de momento lineal para todas las partículas del sistema. Cada uno de estos momentos queda reemplazado, según la receta de cuantización de la representación (de Schrödinger) en el espacio de posiciones, por un operador de derivada parcial con respecto a la coordenada de posición relevante de dicha partícula concreta. Todos estos operadores tienen que actuar sobre algo, y para

que su interpretación sea consistente, todos deben actuar sobre el mismo objeto. Este objeto es la función de onda. Como se ha dicho antes, tenemos, de hecho, *una* función de onda Ψ para el sistema entero, y esta función de onda debe ser en realidad una función de todas las diferentes coordenadas de posición de *todas* las partículas por separado.

23.2. La enormidad del espacio de estados de muchas partículas

Esto suena bastante inocuo, pero ¿lo es? Hagamos una pausa para digerir la enormidad de este último requisito aparentemente simple. Si cada partícula hubiera de tener su propia función de onda por separado, entonces para n partículas escalares (i.e., no giratorias) deberíamos tener n diferentes funciones complejas de la posición. Aunque esto nos obligaría a forzar un poco nuestra imaginación, es algo que quizá pudiéramos tratar en el caso de un número n pequeño de partículas. (Al hacer estas consideraciones, estoy ignorando el tiempo; simplemente tomamos todo en un solo instante.) A efectos de visualización, nuestra imagen no sería diferente de la de un *campo* en el espacio con n componentes diferentes, cada una de las cuales describiría un «campo» independiente. (Cada uno de estos campos independientes representaría la función de onda de una partícula individual.) Quizá deberíamos decir $2n$ componentes si estamos hablando de componentes reales, porque las funciones de onda son complejas. Un campo electromagnético tiene, después de todo, 6 componentes reales —es decir, 6 funciones de 3 variables (análogas a las tres funciones de onda escalares complejas)— ¡y un campo de vectores eléctrico y magnético no supone una tensión tan terrible para la imaginación!

¿Cómo vamos a contar la «libertad» en un campo escalar complejo, tal como la función de onda para una partícula escalar en el 3-espacio? ¿Cuál es el «número» de tales posibles campos diferentes? Recordemos que, de acuerdo con la notación de §16.7, la expresión $\infty^{a\infty^b}$ denota la libertad disponible para un campo (suave) libremente escogido con a componentes reales en un espacio de b dimensiones reales. Así, para un campo escalar complejo, $a = 2$ (puesto que un número

complejo cuenta como dos números reales), de modo que la libertad sería $\infty^{2\infty^1}$. Esto es si se toma el campo solo en un instante —i.e., t es constante—, de modo que estamos considerando el 3-espacio ordinario, lo que da $b = 3$ (en lugar del valor espaciotemporal $b = 4$). También podríamos considerar el espaciotiempo, pero en este caso tenemos ecuaciones de campo que restringen la libertad. En el caso de la función de onda, dicha restricción es la ecuación de Schrödinger, que reduce la libertad a lo que puede especificarse libremente, como datos iniciales, en un 3-espacio inicial, de modo que seguimos teniendo $\infty^{2\infty^1}$ como libertad en el campo.

Como consideración secundaria, podemos examinar el caso de un campo de Maxwell libre sin fuentes (cargas) de las que preocuparnos. Aquí tenemos 6 componentes reales en un 3-espacio ordinario, de modo que si tomamos el campo solo en algún instante t fijo e ignoramos las ecuaciones de Maxwell, obtenemos la libertad $\infty^{6\infty^1}$. Pero las ecuaciones de Maxwell implican que debe haber dos ligaduras en el 3-espacio de datos iniciales, a saber, la anulación de la divergencia de los vectores del campo eléctrico y magnético.[23.1] Esto reduce en 2 el número efectivo de componentes libres en la 3-superficie de datos iniciales, de modo que la libertad es realmente $\infty^{4\infty^1}$.

Consideremos ahora la descripción mecanocuántica de n partículas escalares. Si la descripción consistiera simplemente en n funciones de onda diferentes, entonces la libertad sería $\infty^{2n\infty^1}$, puesto que esta es la libertad al escoger n números complejos por punto en el 3-espacio. Pero en el caso de una función de onda cuántica *real* que describe a n partículas escalares, tenemos una función compleja de $3n$ variables reales. Esto es similar a un campo escalar complejo en un espacio de $3n$ dimensiones, de modo que la libertad es en su lugar $\infty^{2\infty^{3n}}$, que es *inimaginablemente mayor.*

Probablemente no resulta tan fácil apreciar la enormidad de este aumento cuando está oculto en todos aquellos infinitos. Así que permítanme considerar un universo «de juguete» que tiene solo 10 puntos. Podemos etiquetar estos puntos, **0**, **1**, **2**, **3**, **4**, **5**, **6**, **7**, **8**, **9**. La fun-

[23.1] ¿Puede explicar esta anulación? Recordemos la noción 4-dimensional de «divergencia» descrita en §19.3; aquí necesitamos la versión del 3-espacio. *Sugerencia*: Véase el ejercicio [19.2].

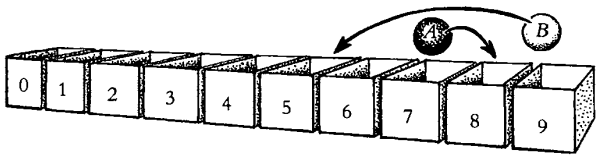


Fig. 23.1. Imaginemos un «universo de juguete» con solo 10 posibles posiciones para las partículas, ilustradas aquí por diez cajas. Se muestran dos partículas distinguibles A y B, cada una de las cuales puede ocupar cualquiera de las cajas, independientemente de las otras.

ción de onda de una partícula escalar en este universo consistiría en un número complejo en cada uno de los 10 puntos, i.e., 10 números complejos $z_0, z_1, z_2, \ldots, z_9$. El espacio de todas estas funciones de onda sería el espacio de Hilbert 10-complejo dimensional (20-real dimensional) $\mathbf{H}^{10}$. Si normalizamos la función de onda de modo que la suma de los cuadrados de los módulos de estas zs sea la unidad, entonces $|z_6|^2$ representaría la probabilidad de que una medida de posición encuentre la partícula en **6**, y lo mismo para los demás.

Este modelo discreto no es realmente tan absurdo. En situaciones físicas reales podríamos tener lo que es, en efecto, una secuencia de 10 cajas, en una de las cuales podría haber un electrón. Véase la Fig. 23.1. Los experimentadores pueden construir cosas de esta naturaleza, llamadas *puntos cuánticos*, que son importantes para la posibilidad teórica de construir *computadores cuánticos* que utilizarían la enormidad de los tamaños de los tipos de espacios de funciones de onda que voy a considerar a continuación.

Supongamos ahora que existen *dos* partículas en nuestro universo. Es mejor que ambas no sean partículas del mismo tipo, por una razón a la que llegaré más adelante, así que las llamaremos partícula A y partícula B. Cada una de las dos partículas podría estar en 10 diferentes lugares alternativos, de modo que existen 100 diferentes colocaciones posibles para el par (admitiendo que ambas puedan estar en la misma caja). Ahora necesitamos 100 números complejos diferentes, digamos $z_{00}, z_{01}, \ldots, z_{09}, z_{10}, z_{11}, \ldots, z_{19}, z_{20}, \ldots, z_{99}$, para definir la función de onda, estando asignado un número complejo a cada par de colocaciones. Si normalizamos de modo que la suma de los cuadrados de los

módulos de todas estas zs sea la unidad, entonces $|z_{38}|^2$, por ejemplo, representaría la probabilidad de encontrar que la partícula A está en **3** y la partícula B está en **8**. Ahora tratemos con $\mathbf{H}^{100}$. Si tuviéramos *tres* partículas diferentes, una partícula A, una partícula B y una partícula C, entonces la función de onda consistiría en 1.000 números complejos $z_{000}, z_{001}, \ldots, z_{999}$, y nuestro espacio de estados sería $\mathbf{H}^{1.000}$. Si las reglas hubieran sido tener meramente tres funciones de onda individuales, entonces el espacio de estados habría sido solo $\mathbf{H}^{30}$. Para *cuatro* partículas diferentes, tendríamos $\mathbf{H}^{10.000}$, mientras que para cuatro funciones de onda individuales tendríamos simplemente $\mathbf{H}^{40}$, y así sucesivamente.

Volviendo a la notación «$\infty^{a\infty^{3n}}$» que he utilizado antes, tomemos nota del hecho de que el «∞^3» superior se refiere al «número de puntos» en el 3-espacio euclídeo $\mathbb{E}^3$. Este número es ahora reemplazado por 10, el número real de puntos en nuestro universo *de juguete*, de modo que $\infty^{a\infty^{3n}}$ se convierte en ∞^{a10^n} (que denota el número de puntos en un espacio $(a \times 10^n)$-real dimensional). Así pues, en lugar de $\infty^{2\infty^{3n}}$ para la libertad en una función de onda escalar para n-partículas en $\mathbb{E}^3$, ahora tenemos $\infty^{2\times 10^n}$ para la libertad en una función de onda de n-partículas para nuestro universo de juguete. El espacio de Hilbert complejo es ahora $\mathbf{H}^{10^n}$ para la función de onda de n-partículas de nuestro universo de juguete, frente a $\mathbf{H}^{10n}$ para n independientes funciones de onda complejas de 1-partícula. Así pues, nuestra función de onda de n-partículas está definida en un espacio 2×10^n-dimensional (es decir, un espacio de Hilbert 10^n-complejo dimensional), en lugar de un mero espacio $20n$-dimensional para n funciones de onda independientes. Para solo 8 partículas, por ejemplo, esto es 200.000.000 dimensiones en lugar de unas meras 160.

23.3. Entrelazamiento cuántico; desigualdades de Bell

¿Qué está haciendo toda esta información extra? Está expresando lo que se conoce como las relaciones de «entrelazamiento» entre las partículas. ¿Cómo hay que entenderlas? Los entrelazamientos entre partículas, una noción que hizo explícita por primera vez Schrödinger (1935b), son los que conducen a los fenómenos extraordinariamente

enigmáticos, pero realmente observados, conocidos como efectos *Einstein-Podolski-Rosen* (EPR).[2] Sin embargo, son aspectos bastante sutiles del mundo cuántico muy difíciles de demostrar experimentalmente de un modo convincente. Llama la atención que tengamos que dirigirnos a algo tan esotérico y oculto a la vista cuando, para sistemas de muchas partículas, ¡casi *toda* la «información» en la función de onda concierne a tales cuestiones! Este es un enigma al que volveré más adelante (§23.6). En mi opinión, está tratando de decirnos algo sobre las nuevas direcciones en que debería moverse nuestro formalismo cuántico actual. Sea como fuere, nos está diciendo ciertamente algo del poder potencial de la *computación cuántica*,[3] un tema de investigación actual muy activa que pretende explotar los enormes recursos de «información» que yacen ocultos en estas relaciones de entrelazamiento.

Así pues, ¿qué es el entrelazamiento cuántico? ¿Qué son los efectos EPR? Estará más claro si consideramos solo una situación de dimensión finita, lo que podemos hacer si nos concentramos solamente en estados de espín. La situación EPR más sencilla es la que consideró David Bohm (1951). En esta, imaginamos un par de partículas de espín 1/2, pongamos por caso, la partícula P_I y la partícula P_D, que parten de un estado combinado de espín 0 y luego viajan en sentidos opuestos, una hacia la izquierda y la otra hacia la derecha, hasta detectores respectivos I y D separados a una gran distancia (véase la Fig. 23.2). Supongamos que cada uno de los detectores puede medir el espín en cierta dirección de la partícula que se le aproxima, y esta dirección solo se decide una vez que las dos partículas están bien separadas una de otra. El problema consiste en ver si es o no posible reproducir las predicciones de la mecánica cuántica utilizando un modelo en el que las partículas se consideran entidades de tipo clásico, independientes y desconectadas, incapaces cada una de ellas de comunicarse con la otra una vez que se han separado.

Como consecuencia de un notable teorema debido al físico norirlandés John S. Bell, resulta que no es posible reproducir de este modo las predicciones de la teoría cuántica. Bell obtuvo unas desigualdades[4] que relacionan las probabilidades conjuntas de los resultados de dos medidas físicamente separadas. Tales desigualdades deben ser necesariamente satisfechas por cualquier modelo en el que las dos partículas se

Fig. 23.2. Experimento mental EPR-Bohm. Un par de partículas de espín 1/2, P_I y P_D se originan en un estado combinado de espín 0, y luego viajan en direcciones opuestas, a izquierda y derecha, hacia detectores respectivos I y D ampliamente separados. Cada detector está preparado para medir el espín de la partícula que se le acerca, pero en una dirección que solo se decide una vez que las partículas están en vuelo. El teorema de Bell nos dice que no hay manera de reproducir las predicciones de la mecánica cuántica con un modelo en el que las dos pueden actuar como objetos independientes de tipo clásico que no pueden comunicarse después de haberse separado.

comportan como entidades independientes una vez que se han separado físicamente, pero las predicciones de la mecánica cuántica las violan. Así pues, la violación de la desigualdad de Bell indica la presencia de efectos esencialmente cuánticos —siendo estos los efectos de los entrelazamientos cuánticos entre partículas físicamente separadas— que no pueden ser explicados por ningún modelo que trate a las partículas como objetos reales desconectados e independientes.

En la literatura hay ejemplos particularmente sorprendentes de este tipo de violación de la desigualdad de Bell.[5] Algunos de estos, conocidos como «desigualdades de Bell sin probabilidades»,[6] son particularmente notables en cuanto que implican solo resultados sí/no, y no tenemos que preocuparnos por probabilidades, o, más bien, nos preocupamos solo por los casos extremos *definidos* con probabilidades 0 («nunca») y 1 («siempre»). Aquí daré únicamente dos versiones explícitas de esta contradicción tipo desigualdad de Bell entre partículas cuánticas y partículas individuales. Ambas incluyen un par de partículas de espín 1/2 que viajan independientemente hacia un detector I a la izquierda y hacia otro detector D a la derecha. La primera, que sigue un razonamiento debido a Henry Stapp (1971, 1979), es un ejemplo directo de la versión original de Bohm para EPR, como se ha mencionado antes, y en la que necesitamos examinar valores de probabilidad. La segunda, debida a Lucien Hardy (1992), es «casi» una versión sin probabilidades, pero con un ligero matiz extra.

Antes de dar estos ejemplos en detalle, necesitaré un poco más de notación (de Dirac). Supongamos que tenemos un sistema cuántico

que consiste en dos partes $|\psi\rangle$ y $|\phi\rangle$, que pueden considerarse mutuamente independientes. Entonces, si deseamos considerar el estado cuántico que consiste en ambas partes juntas, escribimos esto

$$|\psi\rangle\,|\phi\rangle.$$

Este es todavía un único estado, y sería legítimo escribir una ecuación tal como $|\chi\rangle = |\psi\rangle\,|\phi\rangle$ que exprese este hecho. El tipo de producto que se está utilizando aquí es lo que los algebristas denominan un *producto tensorial*, y satisface las leyes $(z|\psi\rangle)\,|\phi\rangle = z(|\psi\rangle\,|\phi\rangle) = |\psi\rangle(z|\phi\rangle)$, $(|\theta\rangle + |\psi\rangle)\,|\phi\rangle = |\theta\rangle\,|\phi\rangle + |\psi\rangle\,|\phi\rangle$, $|\psi\rangle(|\theta\rangle + |\phi\rangle) = |\psi\rangle\,|\theta\rangle + |\psi\rangle\,|\phi\rangle$. La operación del producto tensorial se denota habitualmente en la literatura matemática por $\otimes$, y el producto $|\psi\rangle\,|\phi\rangle$ podría denotarse mediante $|\psi\rangle\otimes|\phi\rangle$.

En cualquier caso, es práctico utilizar el símbolo $\otimes$ en conexión con los *espacios* (de Hilbert) a los que pertenecen tales productos. Así, si $|\psi\rangle$ pertenece a $\mathbf{H}^p$ y $|\phi\rangle$ pertenece a $\mathbf{H}^q$, entonces $|\psi\rangle\,|\phi\rangle$ pertenece a $\mathbf{H}^p \otimes \mathbf{H}^q$. La dimensión de $\mathbf{H}^p \otimes \mathbf{H}^q$ es el producto de las dimensiones de sus dos factores, de modo que podríamos escribir legítimamente $\mathbf{H}^p \otimes \mathbf{H}^q = \mathbf{H}^{pq}$. Estoy admitiendo que o bien uno, o bien ambos, p, q puedan ser ∞, en cuyo caso tomamos que el producto es también ∞. Solo una parte muy pequeña de $\mathbf{H}^p \otimes \mathbf{H}^q$ consiste en elementos de la forma $|\psi\rangle\,|\phi\rangle$ (suponiendo $p, q > 1$), donde $|\psi\rangle$ pertenece a $\mathbf{H}^p$ y $|\phi\rangle$ pertenece a $\mathbf{H}^q$. Estos son estados *desentrelazados*. Un elemento *general* de $\mathbf{H}^p \otimes \mathbf{H}^q$ sería una combinación lineal de estos estados desentrelazados (que puede implicar una suma infinita, si ambos p y q son infinitos).[7] Sin embargo, deberíamos tener en cuenta que la noción misma de entrelazamiento depende de la separación *concreta* de nuestro espacio de Hilbert entero $\mathbf{H}^{pq}$ en algo de la forma $\mathbf{H}^p \otimes \mathbf{H}^q$. (Ninguna separación de un espacio de Hilbert general $\mathbf{H}^{pq}$ es preferida a cualquier otra. Algebraicamente, habrá siempre muchas maneras de expresar $\mathbf{H}^n$ como un producto tensorial, siempre que n sea un número compuesto.) En situaciones donde uno está interesado en la noción de «entrelazamiento», la separación concreta de interés físico es algo razonablemente obvio, muy particularmente cuando hay partículas «individuales» separadas por una gran distancia, que es de lo que trata EPR.

A veces es útil utilizar una formulación de índices abstractos para operaciones tales como esta (véase §12.8). El vector ket $|\psi\rangle$ podría escribirse ψ^α, con un superíndice abstracto, y su correspondiente (conjugado complejo) vector bra $\langle\psi|$ por $\bar{\psi}_\alpha$, con un subíndice abstracto. El paréntesis completo $\langle\psi|\phi\rangle$ sería $\bar{\psi}_\alpha\phi^\alpha$, y una expresión para $\langle\psi|\mathbf{Q}|\phi\rangle$ sería $\bar{\psi}_\alpha Q^\alpha{}_\beta\phi^\beta$. El producto tensorial $|\psi\rangle|\phi\rangle$ de $|\psi\rangle$ por $|\phi\rangle$ podría entonces escribirse $\psi^\alpha\phi^\beta$. Los estados desentrelazados siempre se separan de esta forma. Pero un estado general (probablemente entrelazado) sería una entidad de la forma $\phi^{\alpha\beta}$. Más adelante, en este mismo capítulo, veremos un uso concreto para este tipo de notación.

23.4. Experimentos EPR tipo Bohm

Volvamos ahora a la versión de Bohm de EPR. Consideremos el estado inicial, inmediatamente antes de que se haga una medida sobre él. Las dos partículas de espín 1/2 separadas deben constituir, tomadas en conjunto, un estado de espín 0. Esto se debe a que el momento angular se conserva y las partículas parten de un estado combinado de espín 0. Por consiguiente, necesitamos una combinación de estados base para el espín de cada partícula tal que el espín total sea cero. Esto se consigue mediante el estado $|\Omega\rangle$, de espín 0, dado por

$$|\Omega\rangle = |\uparrow\rangle|\downarrow\rangle - |\downarrow\rangle|\uparrow\rangle$$

(donde sigo sin preocuparme por la normalización de mis estados).[23.2],[23.3] En la literatura esto se suele ver escrito de alguna forma tal como $|\uparrow\mathbf{I}\rangle|\downarrow\mathbf{D}\rangle - |\downarrow\mathbf{I}\rangle|\uparrow\mathbf{D}\rangle$, donde la notación hace explícito qué estado se refiere a la partícula izquierda y cuál se refiere a la partícula derecha. En mi opinión esto no es necesario porque (i) la notación está discriminando solo la parte de espín de la función de

[23.2] Si $|\uparrow\rangle$ y $|\downarrow\rangle$ están normalizados, ¿por qué factor hay que multiplicar $|\Omega\rangle$ para hacerlo normalizado? (Puede suponer que $\||\alpha\rangle|\beta\rangle\| = \|\alpha\|\,\|\beta\|$.)

[23.3] ¿Puede ver rápidamente por qué esto tiene espín 0? *Sugerencia*: Una forma es utilizar la notación de índices para demostrar que cualquier combinación antisimétrica semejante debe ser esencialmente un escalar, teniendo en cuenta que el espacio de espín es 2-dimensional.

onda, y no la posición o el momento o cualquier otra cosa de la partícula, de modo que la dirección del espín determina lo que nos interesa, y (ii) puesto que los productos tensoriales no conmutan, podemos decir inequívocamente qué «lado» del producto es cada uno. Mi convenio es que el término de la izquierda del producto se refiere a la partícula de la izquierda y el término de la derecha se refiere a la partícula de la derecha. Los lectores que encuentren esto confuso pueden reintroducir **I** y **D** en los kets a lo largo de toda la discusión, si así lo desean.

Este es un claro ejemplo de un estado entrelazado, puesto que no puede reescribirse en la forma $|\alpha\rangle|\beta\rangle$ con $|\alpha\rangle$ localizado en **I** y $|\beta\rangle$ localizado en **D**.[23.4] Tratemos de ver qué implicaciones tiene este estado entrelazado. Ahora voy a imaginar que estoy situado a la izquierda, en **I**, y voy a realizar una medida del espín de la partícula izquierda P_I en la dirección «arriba» ↑ (sí si ↑; NO si ↓). Esto proyectaría el estado entero $|\Omega\rangle$ en $|\uparrow\rangle\ |\downarrow\rangle$ si obtengo la respuesta sí, y lo proyectaría en $(-)\,|\downarrow\rangle|\uparrow\rangle$ si obtengo NO. El resultado estaría ahora desentrelazado, excepto que la evolución **U** estándar nos diría que es probable que P_I esté ahora irremediablemente entrelazada con mi propio aparato de medida **I**. Lo que puede establecerse claramente es que si obtengo la respuesta sí, entonces a mi colega, situado en el detector derecho **D**, se le presentará P_D con un estado de espín $|\downarrow\rangle$, mientras que si mido NO, entonces a mi colega se le presentará $|\uparrow\rangle$. Al realizar posteriormente una medida «arriba» sobre P_D, mi colega obtendrá necesariamente un resultado opuesto al mío.

No hay nada especial aquí en la elección arriba/abajo, pues sea cual sea la dirección que decida medir, digamos ↙, entonces si mi colega escoge la misma dirección ↙, su resultado será contrario al mío. Esto debería estar claro a partir de la invariancia rotacional del espín 0, pero es instructivo realizar un cálculo algebraico directo para verificar que (donde $\propto$ significa «igual, salvo un factor no nulo»; véase §12.7)

$$|\Omega\rangle \propto |\swarrow\rangle|\nearrow\rangle - |\nearrow\rangle|\swarrow\rangle,$$

[23.4] ¿Por qué no? Encuentre una manera de hacerlo, sin embargo, si $|\alpha\rangle$ y $|\beta\rangle$ no están localizados así.

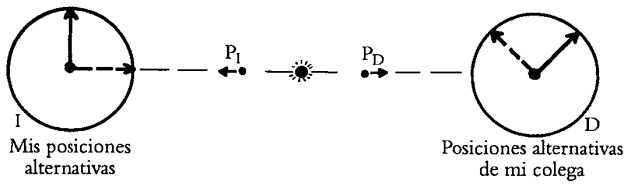


Fig. 23.3. Disposiciones de polarización para la versión de Stapp de EPR-Bohm; ejemplo de las desigualdades de Bell. Inicialmente, tomamos las medidas del espín en uno y otro lado en las direcciones dadas por las flechas sólidas, pero, por un cambio de idea, una u otra, o ambas, podrían rotarse para estar en la dirección de las flechas de trazos. Las probabilidades cuánticas conjuntas no pueden modelarse por ningún esquema de tipo clásico con pares de partículas que se comporten como entidades independientes incomunicadas sin saber por adelantado las direcciones de las medidas del espín propuestas.

donde ↗ es la dirección opuesta a ↙. (Nota: si $|\swarrow\rangle = a|\uparrow\rangle + b|\downarrow\rangle$, entonces $|\nearrow\rangle \propto \bar{b}|\uparrow\rangle - \bar{a}|\downarrow\rangle$.)[23.5]

De todo esto también deducimos cuáles serían las probabilidades conjuntas para SS, SN, NS, NN (abreviando SÍ por S y NO por N), si mi colega y yo elegimos direcciones *diferentes* en las que medir el espín. Supongamos que yo elijo ↖ y mi colega ↗, donde el ángulo entre ↖ y ↗ es θ. Entonces, utilizando el valor para la probabilidad dado en §22.9 (véase la Fig. 22.11), encontramos las probabilidades conjuntas

$$\text{acuerdo} = \frac{1}{2}(1 - \cos\theta), \quad \text{desacuerdo} = \frac{1}{2}(1 + \cos\theta)$$

(donde «acuerdo» significa SS o NN y «desacuerdo» significa SN o NS).

Consideremos ahora el ejemplo de Stapp. Las cosas están dispuestas de modo que mi propio aparato puede orientarse para medir el espín en la dirección ↑, que se toma en vertical hacia arriba, o en la dirección →, que es una dirección horizontal (perpendicular a ↑). El aparato de mi colega está orientado para medir el espín en la dirección ↗, que yace en el plano de las direcciones ↑ y →, formando un ángulo de 45° con cada una de ellas, o en la dirección ↖, que yace en el mismo plano, pero que está a 45° de ↑ y 135° de → (Fig. 23.3). Existen tres posibilidades para las que mi dirección de medida está a 45° de la de mi colega, y hay solo una para la que el ángulo entre ellas es de 135°. En los casos de 45° obtenemos una probabilidad de acuerdo algo por debajo del 15 %, mientras que para 135° obtenemos algo más del 85 %.

[23.5] Confirme el comentario del paréntesis, y verifique por cálculo directo de la expresión anterior para $|\Omega\rangle$. *Sugerencia*: Véase el ejercicio [22.26].

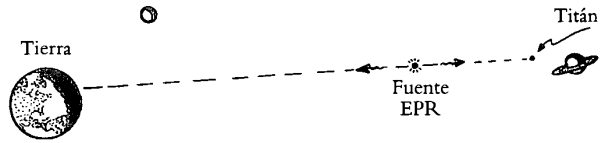


Fig. 23.4. El autor, situado en la Tierra, se imagina ser el receptor de una de las componentes de una serie de pares de partículas EPR; la otra componente es recibida por un colega en Titán. La fuente de los pares está aproximadamente equidistante entre los dos receptores. Incluso para partículas que viajan a la velocidad de la luz habría unos 45 minutos para decidir las orientaciones de los detectores.

Permitamos que la decisión sobre cuál de las dos posibles medidas voy a realizar no se tome realmente hasta que las partículas estén en vuelo, y lo mismo se aplica a mi colega. Muy bien; situemos a mi colega en Titán (una de las lunas de Saturno) y pongamos la fuente de las partículas en algún lugar entre nosotros dos, de modo que incluso a la velocidad de la luz ¡tendríamos unos tres cuartos de hora para tomar nuestra decisión! Véase la Fig. 23.4. Las partículas no tienen ninguna forma de «saber» en qué dirección vamos a orientar mi colega y yo (independientemente) nuestros aparatos de medida.

Supongamos que he elegido ↑ y mi colega ha elegido ↗ cuando cada uno de nosotros recibe una corriente de partículas aparentemente orientadas de forma aleatoria. Llegan de una en una, y cada una de ellas es un miembro de un par EPR-Bohm enviado desde la fuente central, una a mí y otra a mi colega. Cuando comparamos nuestras notas (quizá algunos años más tarde, cuando regrese mi colega), encontramos que hay un acuerdo un poco menor del 15 % entre nuestros resultados, en conformidad con lo anterior.

Ahora bien, si las partículas no tienen conocimiento previo de cómo vamos a orientar nuestros aparatos de medida, y se comportan como entidades independientes sin comunicación (de tipo clásico) entre ellas, no debería suponer ninguna diferencia para las medidas reales de mi

colega el que yo hubiera cambiado repentinamente de opinión en el último minuto y midiera la dirección → en lugar de la ↑. Si lo hubiera hecho así, entonces —puesto que el ángulo entre las direcciones sigue siendo de 45° grados— aún tendría que haber solo un 15 % de acuerdo entre las medidas que yo obtuviera ahora y las originales de mi colega. Por el contrario, supongamos que fuera *mi colega* el que hubiera cambiado de idea en el último minuto y midiese ↖ en lugar de ↗, pero que yo no hubiera cambiado. El cambio de mi colega no debería haber afectado a mis propias medidas ↑ originales. Una vez más, habríamos encontrado que las nuevas medidas ↖ de mi colega habían estado algo por debajo del 15 % en acuerdo con mis propias medidas ↑ originales.

Pero supongamos que *ambos* hubiéramos decidido cambiar las orientaciones en el último minuto, de modo que mis propias medidas serían → y las de mi colega ↖. Ahora el ángulo entre ellas ha pasado a ser 135°, de modo que las predicciones de la mecánica cuántica nos dicen que el acuerdo debería ser entonces algo más del 85 %. ¿Es eso compatible con que los pares de partículas proporcionen las probabilidades conjuntas correctas para cada uno de los posibles pares de orientaciones de los detectores que acabamos de considerar? Bien, veámoslo. Los pares de partículas tienen que estar listos para encontrar cualquiera de las cuatro combinaciones posibles de posiciones de los detectores, y para dar las probabilidades mecanocuánticas correctas en cada caso. Recordemos cuáles son estas. Cabría esperar que los resultados de mi aparato alterado en posición → no presenten más de un 15 % de acuerdo con la posición ↗ original de mi colega. Esta, a su vez, no debería tener más de un 15 % de acuerdo con mi posición ↑ original, y esta no debería tener más de un 15 % de acuerdo con la posición alterada ↖ de mi colega. Si un par de partículas concreto va a dar acuerdo en el caso →, ↖, entonces no puede estar en desacuerdo en los tres casos →, ↗ y ↑, ↗ y ↑, ↖. (Tres desacuerdos deben dar desacuerdo, no acuerdo.) De modo que al menos en uno de estos tres posibles pares de posiciones debe haber acuerdo. Pero esto sucede en menos de un 15 % de los casos, para cada par de posiciones posible. Hay solo tres de estas, de modo que esto no permite más de 15 % + 15 % + 15 % = 45 % de acuerdo cuando llegamos al caso →, ↖. (De hecho, el porcentaje de acuerdo resulta un poco menos de esto, porque he contado efectivamente tres veces el caso en el que hay acuerdo en los tres pares de posiciones.) Pero 45 % no está nunca cerca de 85 %, ¡de modo que tenemos una flagrante contradicción con nuestras hipótesis «tipo clásico» para los pares de partículas!

Algunos podrían lamentarse de que este razonamiento se ha expresado en términos de medidas hipotéticas que «podrían haber sucedido pero no han sucedido» (los hechos «contrafácticos» del filósofo). Pero esto no es importante. La cuestión clave es que se ha supuesto que las partículas se comportan independientemente unas de otras una vez que han dejado la fuente, y dan las correctas probabilidades cuánticas conjuntas cualquiera que sea la combinación de posiciones de detectores que encuentran. El punto importante es que las partículas tienen que imitar las predicciones de la mecánica cuántica. Hemos encontrado que estas no pueden separarse en predicciones independientes para las dos partículas por separado. La única forma en que las partículas pueden proporcionar consistentemente las respuestas mecanocuánticas correctas es que estén, de algún modo, «conectadas» entre sí, hasta que una u otra de ellas sea medida realmente. Esta misteriosa «conexión» entre ellas es el entrelazamiento cuántico.

Por supuesto, ningún experimento de esta naturaleza se ha realizado sobre tales distancias. Pero se han llevado a cabo realmente muchos experimentos tipo EPR de una naturaleza esencialmente similar (normalmente utilizando fotones, y no partículas de espín 1/2, pero las diferencias no son importantes). ¡Las predicciones de la mecánica cuántica (antes que las del sentido común) han sido sistemáticamente confirmadas! Aunque, por supuesto, no se han observado todavía entrelazamientos cuánticos directos sobre distancias Tierra-Saturno, algunos experimentos recientes han confirmado violaciones de la desigualdad de Bell sobre distancias de más de 15 kilómetros.[8]

23.5. El ejemplo EPR de Hardy: casi libre de probabilidad

Llegamos ahora al bello ejemplo de Lucien Hardy.[9] Una vez más, mi colega y yo estamos preparados para hacer medidas de espín: yo elijo

entre las medidas ↑ y → (verticalmente hacia arriba y en dirección horizontal hacia la derecha), como antes, pero ahora mi colega también elige entre ↑ y →, de forma completamente independiente de mi propia elección. El nuevo aspecto crucial es que la fuente de pares de partículas no las emite ahora en un estado combinado de espín 0, sino en un estado particular de espín 1. Considero que este estado inicial es el que tiene la descripción de Majorana $|{\leftarrow}{\nearrow}\rangle$ (§22.10, Fig. 22.14), donde la dirección de ↗ yace en el cuadrante generado por las direcciones perpendiculares ↑ y →, y tiene una pendiente hacia arriba de 4/3 (de modo que el ángulo θ entre → y ↗ satisface $\cos\theta = 3/5$), y donde ← es opuesta a →; véase la Fig. 23.5. Podemos expresar este estado como[23.6]

$$|{\leftarrow}{\nearrow}\rangle = |{\leftarrow}\rangle|{\nearrow}\rangle + |{\nearrow}\rangle|{\leftarrow}\rangle,$$

ignorando un factor global, y tiene la característica importante de que, aunque no es ortogonal a

$$|{\downarrow}\rangle|{\downarrow}\rangle$$

(con ↓ opuesto a ↑), *sí* es ortogonal a cada uno de los[23.7]

$$|{\downarrow}\rangle|{\leftarrow}\rangle,\ \ |{\leftarrow}\rangle|{\downarrow}\rangle,\ \ |{\rightarrow}\rangle|{\rightarrow}\rangle.$$

Estas relaciones de ortogonalidad son responsables, respectivamente, de las siguientes propiedades «SÍ/NO» (0), (1) (2) y (3):

(0) a veces obtengo NO para una medida ↑ cuando mi colega obtiene NO para una medida ↑;

(1) si obtengo NO para una medida ↑, entonces mi colega debe obtener SÍ para una medida →;

(2) si mi colega obtiene NO para una medida ↑, entonces debo obtener SÍ para una medida →;

(3) nunca obtengo SÍ para una medida → cuando mi colega obtiene SÍ para una medida →.

Podría mencionarse, en relación con (0), que la probabilidad mecanocuántica real para que los dos obtengamos una respuesta NO, dado

[23.6] ¿Por qué?

[23.7] Vea si puede demostrarlas. *Sugerencia*: Utilice las descripciones de coordenadas y/o geométricas de §22.9.

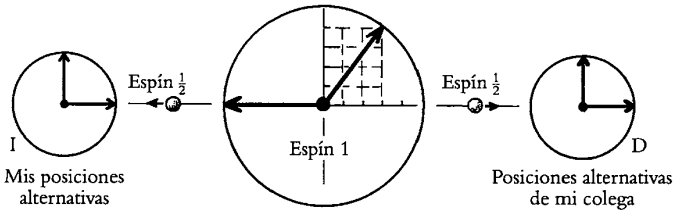


Fig. 23.5. Versión de Hardy de EPR «casi» sin probabilidades. El estado inicial del espín 1 es $|{\leftarrow}{\nearrow}\rangle = |{\leftarrow}\rangle|{\nearrow}\rangle + |{\nearrow}\rangle|{\leftarrow}\rangle$, donde la dirección ↗ está en el cuadrante generado por la vertical ↑ y la horizontal →, con una pendiente hacia arriba de 4/3. Cada detector mide el espín de la partícula que se le aproxima o bien verticalmente, o bien horizontalmente.

que ambos decidimos realizar la medida ↑, es exactamente 1/12 en este experimento.[23.8] Nótese que 1/12 = 8,33 %, mientras que el valor óptimo de Hardy, con algunos ligeros ajustes, está muy próximo[10] a 9,017 %.

Debería dejar completamente claro por qué no hay forma de conseguir los resultados (0), ..., (3), si las dos partículas son entidades separadas y no comunicadas, y sin ningún conocimiento previo de los experimentos que se van a realizar con ellas. Debido a (0), cada una de las dos partículas (supuestas ahora incomunicadas y sin conocimiento previo) deben estar preparadas conjuntamente para proporcionar una respuesta NO de vez en cuando (de hecho, 1/12 de las veces), ante la eventualidad de que mi colega y yo pudiéramos realizar simultáneamente medidas ↑. Además, la preparación de las partículas debe haber sido cuidadosa para haber preconvenido (cuando estaban juntas) que en aquellas ocasiones en que podrían hacerlo (i.e., dar respuestas NO simultáneas a nuestras medidas ↑ simultáneas) también deberían dar *decididamente* la respuesta SÍ a una medida → si solo uno u otro de nosotros cambiase a esta dirección, para no violar así (1) o (2). Pero esa misma decisión las lleva a poner a (3) en grave peligro, porque mi colega y yo podríamos casualmente realizar a la vez medidas →, en una ocasión semejante, y obtener así el resultado prohibido SÍ, SÍ.

[23.8] Demuéstrelo.

23.6. Dos misterios del entrelazamiento cuántico

A mi modo de ver, el entrelazamiento cuántico presenta dos misterios completamente diferentes, y creo que la respuesta a cada uno de ellos tiene un carácter completamente distinto (aunque interrelacionado). El primer misterio es el propio fenómeno. ¿Cómo debemos entender el entrelazamiento cuántico y darle sentido en términos de ideas que podamos comprender, de modo que podamos llegar a aceptarlo como algo que forma una parte importante de la actuación de nuestro universo real? El segundo misterio es de algún modo complementario del primero. Puesto que, según la mecánica cuántica, el entrelazamiento es un fenómeno tan ubicuo —recordemos que la inmensa mayoría de los estados cuánticos son realmente estados entrelazados—, ¿por qué es algo que apenas advertimos en nuestra experiencia directa del mundo? ¿Por qué no se nos presentan continuamente estos efectos ubicuos del entrelazamiento? No creo que este segundo misterio haya recibido toda la atención que merece, pues la perplejidad de la gente se ha concentrado casi enteramente en el primero.

Empecemos por abordar este segundo misterio. Volveré al primero a su debido tiempo. Un enigma al que hay que hacer frente es el hecho de que los entrelazamientos tienden a difundirse. Parecería que finalmente todas las partículas en el universo deben entrelazarse entre sí. ¿O ya están todas entrelazadas entre sí? ¿Por qué no experimentamos precisamente un revoltijo entrelazado, sin el más mínimo parecido con el mundo (casi) clásico que realmente percibimos? La evolución de Schrödinger de un sistema no nos ayuda en esto. Tiende a empeorar las cosas cada vez más, pues a medida que pasa el tiempo cada vez más partes del universo se van entrelazando con cualquier sistema del que partamos. En términos del espacio de Hilbert **H**, pienso que se acepta generalmente que la ecuación de Schrödinger (proceso **U**) no nos sacará por sí sola de nuestras dificultades. Si empezamos a salvo en una parte relativamente desentrelazada de **H**, entonces la evolución de Schrödinger nos sumergirá (normalmente) casi de inmediato en las profundidades del entrelazamiento y no nos proporcionará por sí misma ninguna ruta, ni siquiera una guía, para salir de este enorme piélago de estados entrelazados (véase la Fig. 23.6).

Pese a todo, parece que en la vida cotidiana lo llevamos muy bien, sin notar siquiera estos entrelazamientos. ¿A qué se debe esto? Si no vamos a obtener ninguna ayuda del proceso **U** de la teoría cuántica, entonces debemos dirigirnos a su otro ingrediente esencial: el proceso **R**. De hecho, ya hemos visto algo de cómo podría ayudarnos esto en nuestras consideraciones de los efectos EPR. Recordemos que antes imaginaba la realización de una medida sobre un miembro de un par EPR, cuyo otro miembro se estaba acercando a mi colega en el planeta Titán. Si hago mi medida primero, entonces, por el hecho de realizarla, este mismo acto deja la partícula de mi colega libre de su entrelazamiento con la mía, y a partir de ahí (hasta que sea medida por mi colega) dicha partícula poseerá un vector de estado por sí misma, sin ser molestada por ninguna responsabilidad adicional hacia su compañera, independientemente de lo que yo pudiera hacer con esta posteriormente. Así pues, parece que son las *medidas* las que cortan drásticamente estos entrelazamientos. ¿Puede ser cierto esto? ¿Es **R** la respuesta general al segundo enigma que presenta el fenómeno del entrelazamiento cuántico?

Creo que esto debe ser cierto, al menos si pensamos en cómo se utiliza la mecánica cuántica en la práctica. Esto tiene relevancia para la forma en que montamos cualquier experimento cuántico, tal como los experimentos mentales que acabamos de considerar. Recordemos que en nuestras consideraciones EPR exigíamos un número de pares de partículas que estuvieran dispuestas en un estado cuántico particular: de espín 0 en el ejemplo de Stapp y de espín 1 en el ejemplo de Hardy.

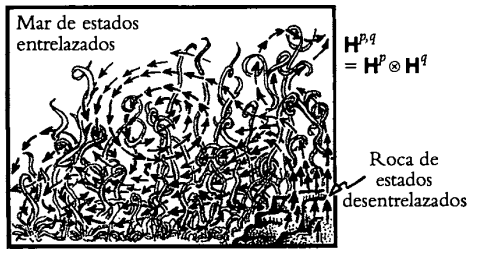


Fig. 23.6. La evolución de Schrödinger, lejos de un estado inicial desentrelazado (ilustrado por la roca abajo a la derecha), lleva casi siempre a entrelazamientos crecientes (ilustrados por el mar lleno de algas dispersas). Entonces, ¿por qué los objetos de la experiencia ordinaria aparecen como cosas independientes?

¿Cómo podríamos asegurar, utilizando solo procesos **U**, que nuestras partículas no estuvieran ya terriblemente entrelazadas con todas las otras cosas que nos rodean? Creo que algo de la naturaleza de una «medida» es siempre una parte esencial del *montaje* de un experimento cuántico, para asegurar que el estado no esté contaminado por enjambres de estos entrelazamientos indeseados. Con esto no quiero decir que el experimentador monte deliberadamente una «medida» para conseguirlo. Mi idea es que la propia naturaleza está activando continuamente efectos de procesos **R**, sin que haya ninguna intención deliberada por parte de un experimentador ni ninguna intervención de un «observador consciente».

Aquí estoy entrando en aguas controvertidas, y más tarde tendré que volver a mi propia posición sobre estas cuestiones (en §§30.9-13). Pero ¿cómo se trata el tema en la mecánica cuántica «convencional»? Parece que, «en la práctica», los físicos suponen siempre que estos supuestos entrelazamientos con el mundo exterior pueden ser ignorados. De lo contrario, no se podría confiar en la mecánica clásica ni en la mecánica cuántica convencional. Parece existir la idea de que todos los entrelazamientos se «promediarán» de alguna manera, de modo que no necesitan ser considerados en la práctica en ninguna situación real. Pero no conozco ninguna demostración remotamente convincente de que este sea el caso. Más que promediar, el resultado probable sería que todas las cosas se parecerían cada vez menos al universo que conocemos, en donde hay objetos individuales que tienen posiciones más o menos bien definidas que no están condicionadas por otras muchísimas ocurrencias en otros lugares del universo. No veo ninguna salida a este dilema si tenemos que verlo como un problema aislado de la paradoja **U**/**R** que yace en el centro de la interpretación de la mecánica cuántica.

Con independencia de cómo examinemos la cuestión de este entrelazamiento generalizado con el resto del universo, no podemos divorciarla de la pregunta más amplia de cuál es la razón de que, por una parte, los procedimientos **U** funcionen tan extraordinariamente bien para sistemas suficientemente simples, mientras que por otra parte tenemos que abandonar **U** de vez en cuando e intercalar, abruptamente pero a hurtadillas, el proceso **R**. ¿Por qué?, así como, ¿cuándo? y

¿cómo? Este es el *problema de la medida* o (más exactamente, creo) la *paradoja* de la medida, en palabras del premio Nobel Tony Leggett. Volveré a la cuestión en el capítulo 29.

Aún no he acabado con los otros enigmas que nos presenta el entrelazamiento. Algunos de estos tienen que ver con la falta de compatibilidad entre la medida de un sistema entrelazado y los requisitos de la relatividad, puesto que una medida de una parte de un par entrelazado parecería afectar a la otra *simultáneamente*, lo que, como hemos visto en el capítulo 17, es algo que no deberíamos aceptar si nos vamos a atener a los principios relativistas. Antes de intentar afrontar este problema debería abordar otro aspecto del entrelazamiento. Es un aspecto aún más ubicuo que los que hemos abordado en los párrafos anteriores. Es tan ubicuo que ni siquiera las medidas lo cortan, como quiera que elijamos considerar la paradoja de la medida. Más aún, es un rasgo característico de la mecánica cuántica que parece ser independiente de los otros que he estado abordando hasta ahora. Me refiero a la forma extraordinaria en que la mecánica cuántica trata los sistemas de *partículas idénticas*.

23.7. Bosones y fermiones

Recordemos (§23.2) nuestro entretenimiento con un «universo de juguete» en el que había 10 posiciones distintas abiertas a cada partícula, etiquetadas **0**, **1**, **2**, ..., **9**. Cuando consideraba que este universo podría estar habitado por más de una partícula, he tenido cuidado en exigir que no se considerase que las partículas eran «del mismo tipo», y me he referido a ellas como una «partícula *A*» y una «partícula *B*», etc., en lugar de, pongamos por caso, «dos electrones», o algo similar. La razón para esto es que la mecánica cuántica trata las partículas reales de la naturaleza mediante procedimientos que son claramente diferentes de los de nuestra discusión anterior. De hecho, en este punto debemos hacer una distinción entre dos de estos procedimientos muy *diferentes*. Uno de estos procedimientos se aplica a partículas conocidas como *bosones* y el otro a las conocidas como *fermiones*. Los bosones resultan ser partículas con *espín entero* (i.e., donde el espín, en unidades de $\hbar$, toma uno

de los valores 0, 1, 2, 3, ...), y los fermiones, partículas con espín semientero (valores $\frac{1}{2}, \frac{3}{2}, \frac{5}{2}, \frac{7}{2}, \ldots$). (Esta asociación se sigue de un famoso teorema matemático, en el contexto de la teoría cuántica de campos, conocido como el *teorema espín-estadística*; véase §26.2.) Las partículas compuestas, tales como los núcleos o los átomos enteros o, de hecho, los hadrones individuales tales como los protones o los neutrones (considerados compuestos de quarks) pueden también ser tratadas, con un grado adecuado de aproximación, como bosones o fermiones individuales. Así, los fotones son bosones, y también lo son los mesones (piones, kaones, etc.) y las partículas responsables de las interacciones débiles (las partículas W y Z) y las interacciones fuertes (los gluones). Las evidentemente compuestas partículas-α (2 protones, 2 neutrones), los deuterones (1 protón, 1 neutrón), etc., también se comportan muy aproximadamente como bosones. Por el contrario, los electrones, los protones, los neutrones y sus quarks constituyentes, los neutrinos, los muones y muchas otras partículas son fermiones. Podemos tomar nota del hecho de que las funciones de onda de los fermiones son *objetos espinoriales*, en la terminología de §11.3 (compárese con §22.8), mientras que las de los bosones no lo son.

Para ver qué distingue realmente a los bosones de los fermiones, volvamos a nuestro universo de juguete con solo 10 puntos en él, etiquetados **0**, **1**, ..., **9**. Recordemos que el análogo apropiado de una función de onda sería simplemente una colección de números complejos $z_0, z_1, \ldots, z_9$ para una sola partícula, $z_{00}, z_{01}, \ldots, z_{99}$ para un par de partículas, $z_{000}, z_{001}, \ldots, z_{999}$ para tres partículas, etc. Para un par de bosones, sin embargo, el requisito es que la colección de números complejos z_{ij} debería ser *simétrica* en sus índices:

$$z_{ij} = z_{ji},$$

de modo que $z_{38} = z_{83}$, por ejemplo. Así, por lo que respecta a esta «función de onda», no supone ninguna diferencia cuál de las partículas es la que está en **3** y cuál es la que está en **8**. Hay solamente un *par de partículas* que ocupan los dos puntos **3** y **8**. Nótese que ambos miembros del par de bosones pueden estar perfectamente en el mismo lugar; por ejemplo, z_{33} es el factor de peso complejo para que ambos

bosones ocupen simultánemente el punto **3**. Vemos que existen solo $\frac{1}{2}(10 \times 11) = 55$ maneras distinguibles de colocar el par (desordenado) de partículas en los 10 puntos, y solo se necesita este número de números complejos (i.e., $\mathbf{H}^{55}$ en lugar de $\mathbf{H}^{100}$). Con tres bosones idénticos, tenemos simetría en los tres argumentos

$$z_{ijk} = z_{jik} = z_{jki} = z_{kji} = z_{kij} = z_{ikj},$$

de modo que ahora tenemos $\frac{1}{6}(10 \times 11 \times 12) = 220$ números complejos para definir el estado: un elemento de $\mathbf{H}^{220}$ en lugar de $\mathbf{H}^{1.000}$. Para n bosones idénticos, el número es $(9 + n)!/9!n!$, que es el número de números complejos independientes $z_{ij\ldots m}$ necesarios para que sea totalmente *simétrico* en los índices (véanse §§12.4,7 y §14.7 para la notación):

$$\mathbf{z}_{ij\ldots m} = \mathbf{z}_{(ij\ldots m)}.$$

Consideremos ahora los fermiones. La diferencia es que para los fermiones se requiere que la función de onda sea *antisimétrica* en sus argumentos,

$$z_{ij} = -z_{ji},$$
$$z_{ijk} = -z_{jik} = z_{jki} = -z_{kji} = z_{kij} = -z_{ikj},$$
$$\mathbf{z}_{ij\ldots m} = \mathbf{z}_{[ij\ldots m]},$$

de modo que tenemos $\frac{1}{2}(10 \times 9) = 45$ números complejos para dos fermiones idénticos, $\frac{1}{6}(10 \times 9 \times 8) = 120$ números complejos para tres fermiones idénticos, y $10!/n!(10 - n)!$ para n fermiones idénticos.[23.9] La diferencia en el recuento surge del hecho de que ahora no está permitido tener dos fermiones en el mismo punto porque la antisimetría implica que los factores de peso complejos $z\ldots$ deben anularse cuando eso ocurre: $z_{33} = 0$, $z_{474} = 0$, etc.

Nótese que cuando tenemos más de 5 fermiones idénticos en nuestro universo de juguete, los números empiezan a reducirse de nue-

[23.9] Explique todos estos números tanto en el caso de los bosones como de los fermiones.

vo. Cuando llegamos a 10 fermiones, hay solo un estado posible, y no podemos tener más de 10 fermiones idénticos en conjunto en nuestro universo de juguete. Vemos en este modelo una manifestación de un principio importante en la mecánica cuántica llamado *principio de exclusión de Pauli*. Según este, dos fermiones idénticos no pueden estar en el mismo estado (lo que es simplemente una característica de la antisimetría de la función de onda fermiónica). El hecho de que los materiales sólidos no colapsen sobre sí mismos depende, en última instancia, de este principio. La materia sólida ordinaria está básicamente compuesta de fermiones: electrones, protones y neutrones. Estos tienen que «mantenerse fuera del camino de los otros», debido al principio de Pauli.

En el caso de los bosones, las cosas son de otra manera. Hay una ligera tendencia de los bosones a «preferir» estar en el mismo estado. (Esto resulta como un efecto puramente estadístico, cuando comparamos el recuento de diferentes estados bosónicos con los diferentes estados clásicos correspondientes.) Cuando la temperatura es muy baja, este efecto puede ser importante y puede tener lugar un fenómeno conocido como *condensación de Bose-Einstein*, cuando la mayoría de las partículas relevantes se reúnen en el mismo estado. Los superfluidos son ejemplos de esto, e incluso los láseres se basan en ello. En un superconductor, los electrones tienen una forma de «aparearse» y estos *pares de Cooper* tienen la capacidad de comportarse como si fueran bosones individuales. Algunos de los usos prácticos más impresionantes y contraintuitivos de la mecánica cuántica proceden de este tipo de fenómeno «colectivo».

23.8. Los estados cuánticos de los bosones y los fermiones

Aunque he expuesto los requisitos de simetría y antisimetría de los bosones y los fermiones solo en referencia a nuestro «universo de juguete», los requisitos de simetría bosón/fermión para una colección de bosones o fermiones reales en el espacio ordinario son básicamente los mismos. La función de onda será una función de un número de puntos en el espacio, etiquetados $\mathbf{u}, \mathbf{v}, \ldots, \mathbf{y}$, así como de varios parámetros

discretos $u, v, \ldots, y$, respectivamente, para resumir el grupo de índices (espinoriales o tensoriales) de cada partícula. Preguntemos, en primer lugar, qué aspecto tendrá la función de onda ψ para un *par* de bosones idénticos. El requisito es que la función $\psi = \psi(\mathbf{u}, u; \mathbf{v}, v)$ debería ser *simétrica* bajo intercambio de las partículas:

$$\psi(\mathbf{u}, u; \mathbf{v}, v) = \psi(\mathbf{v}, v; \mathbf{u}, u).$$

Para tres bosones idénticos, nuestra función de onda debería ser simétrica bajo permutaciones de las tres partículas:

$$\psi(\mathbf{u}, u; \mathbf{v}, v; \mathbf{w}, w) = \psi(\mathbf{v}, v; \mathbf{u}, u; \mathbf{w}, w) = \psi(\mathbf{v}, v; \mathbf{w}, w; \mathbf{u}, u) = \ldots,$$

y así sucesivamente.

En el caso de los fermiones, estas relaciones se reemplazan por *antisimetría* bajo intercambio de las partículas:

$$\psi(\mathbf{u}, u; \mathbf{v}, v) = -\psi(\mathbf{v}, v; \mathbf{u}, u).$$
$$\psi(\mathbf{u}, u; \mathbf{v}, v; \mathbf{w}, w) = -\psi(\mathbf{v}, v; \mathbf{u}, u; \mathbf{w}, w) = \psi(\mathbf{v}, v; \mathbf{w}, w; \mathbf{u}, u) = \ldots,$$

y así sucesivamente. Nótese que en cada caso el estado de espín (caracterizado por las variables discretas $u, v, \ldots$) debe ser transportado con la partícula en estos intercambios. La consecuencia de esto es que cuando se aplica el principio de exclusión de Pauli consideramos los estados idénticos solo si los estados de espín son también idénticos, además de ser idénticas sus localizaciones. Esto es importante en química, por ejemplo, donde dos electrones pueden compartir el mismo orbital con tal de que sus espines sean opuestos (véanse §24.8 y la Fig. 24.2).

Este es un lugar donde resulta práctica la notación de índices (abstractos) para estados mencionada en §23.3 (y también puede utilizarse una notación diagramática, como se ha descrito en §12.8, ilustrada en la Fig. 26.1). Por consiguiente, podríamos utilizar la notación ψ^α para la función de onda de una partícula concreta a la que ha sido asignada la etiqueta «α», y ϕ^β para la función de onda de una segunda partícula a la que se ha asignado la etiqueta «β», y así sucesivamente. Si las partículas no son idénticas, entonces la función de onda para el par de ellas sería el estado producto (tensorial)

$$\psi^\alpha \phi^\beta,$$

mientras que si son bosones idénticos, el estado es (sin preocuparnos de los factores de normalización)

$$\psi^\alpha\phi^\beta + \phi^\alpha\psi^\beta.$$

(Una puntualización sobre la formulación de índices abstractos: tenemos multiplicación conmutativa, por ejemplo, $\phi^\alpha\psi^\beta = \psi^\beta\phi^\alpha$. La no conmutatividad del producto tensorial se trata mediante el ordenamiento de índices, de modo que $|\phi\rangle|\psi\rangle \neq |\psi\rangle|\phi\rangle$ se expresa como $\phi^\alpha\psi^\beta \neq \psi^\alpha\phi^\beta$.) Podemos escribir este estado simetrizado (ignorando un factor 2) como

$$\psi^{(\alpha}\phi^{\beta)},$$

utilizando nuestra notación de paréntesis redondos para la simetrización (§12.7 y §22.8). Esto tiene la ventaja de que podemos escribir inmediatamente el estado cuántico para n bosones idénticos, cuyos estados individuales serían ψ^α, ϕ^β, ..., χ^κ, como el producto simetrizado

$$\psi^{(\alpha}\phi^{\beta}\dots\chi^{\kappa)}.$$

Podemos hacer exactamente lo mismo para los fermiones, donde para estados individuales ψ^α, ϕ^β, ..., χ^κ, la colección de todos los n fermiones idénticos tendría el estado antisimetrizado (§12.4)

$$\psi^{[\alpha}\phi^{\beta}\dots\chi^{\kappa]}.$$

Nótese que todos estos estados de muchas partículas están técnicamente *entrelazados* (como vemos, en particular, para la descripción de un par de fermiones idénticos que es la combinación $\psi^\alpha\phi^\beta - \phi^\alpha\psi^\beta$). No obstante, es un tipo suave de entrelazamiento, pues la superposición es entre estados que son «físicamente indistinguibles», ya que solo se aplica a partículas idénticas. Los estados $\psi^{(\alpha}\phi^{\beta}\dots\chi^{\kappa)}$ y $\psi^{[\alpha}\phi^{\beta}\dots\chi^{\kappa]}$, para bosones y fermiones respectivamente, son lo más próximo que podemos tener a estados «desentrelazados» y podríamos adoptar la postura alternativa de llamar a tales estados «desentrelazados». (El estado bosónico general de n-partículas, en esta notación, sería algún $\Psi^{\alpha\beta\dots\kappa} = \Psi^{(\alpha\beta\dots\kappa)}$, que no se separa de esta manera. Análogamente, un estado fermiónico general de n-partículas es un $\Phi^{\alpha\beta\dots\kappa} = \Phi^{[\alpha\beta\dots\kappa]}$ «sin separación».) En términos de kets, podríamos concebir una notación «producto cuña» $|\psi\rangle \wedge |\phi\rangle \wedge \dots \wedge |\chi\rangle$ para trabajar con estos requisitos de simetría y antisimetría,[11] donde tenemos en cuenta que los términos conmutan o anticonmutan dependiendo de los «grados» de los factores individuales (véase §11.6).

Aunque el tipo de «entrelazamiento» que ocurre con bosones o fermiones idénticos es relativamente «inocuo» (y, de hecho, sirve para reducir más que incrementar el gran número de alternativas abiertas a un sistema cuántico), tiene al menos una consecuencia importante para un efecto que se extiende sobre grandes distancias físicas. La naturaleza bosónica «entrelazada» de los fotones que llegan a la Tierra desde lados opuestos de una estrella relativamente próxima ha sido utilizada para medir los diámetros de tales estrellas, utilizando un método que se debe a Hanbury Brown y Twiss (1954, 1956). Cuando su método fue propuesto por primera vez, encontró una gran oposición por parte de muchos (incluso distinguidos) físicos cuánticos, que argumentaban que «los fotones solo pueden interferir consigo mismos, y no con otros fotones»; pero ellos habían pasado por alto el hecho de que los «otros fotones» eran parte de un todo bosón entrelazado.

23.9. Teleportación cuántica

Para terminar este capítulo, volvamos a los enigmas que son inherentes a la interpretación de los efectos EPR. Más concretamente, recordemos el aparente conflicto con la relatividad especial que supone el hecho de que la «comunicación» entre pares EPR parece no respetar los propios requisitos de Einstein según los cuales no debería estar permitido el envío de señales más rápidas que la luz. Para ilustrar estas cuestiones, expondré una consecuencia adicional bastante misteriosa del entrelazamiento cuántico conocida como *teleportación cuántica*. En mi opinión, esta consecuencia nos lleva en una dirección que bien podría ser la que tenemos que explorar si queremos entender adecuadamente los efectos EPR en general. Pese a todo, observaremos que esta dirección conduce a un territorio en el que muchas personas serían, sin duda, muy reacias a entrar, ¡y con razón, tal como veremos!

¿Qué significa el término «teleportación»? El nombre, sin duda, re-

cuerda a *Star Trek* y las imágenes del capitán Kirk y algunos miembros de su tripulación enviados a la superficie de un planeta inexplorado. Aquí se adopta la idea de que para transmitir con éxito la «identidad» real de una persona es necesario que en la superficie del planeta se proyecte fielmente un *estado cuántico* real, y no solo un listado clásico de posiciones de partículas, etc. Una perspectiva semejante tiene la ventaja filosófica de que el procedimiento de teleportación no podría utilizarse para *duplicar* un individuo, lo que podría plantear delicadas interrogantes concernientes a cuál de los dos representa la continuación del «flujo de consciencia» del individuo.[12] ¿Por qué no es posible copiar un estado cuántico desconocido? La cuestión ha sido convincentemente abordada en la literatura,[13] pero a partir de consideraciones elementales podemos ver que una posibilidad semejante conduciría a una contradicción con los principios de la mecánica cuántica **U**/**R** estándar. A menos que uno esté dispuesto a destruir el original, no se puede hacer una copia exacta, y ciertamente no se pueden hacer dos copias exactas de un estado cuántico.

¿Por qué no? Si se pudiese, entonces repitiendo el proceso podríamos tener 4 copias, luego 8, luego 16, etc. Supongamos que el estado es solo un simple estado de espín $|\nearrow\rangle$ de una partícula masiva de espín 1/2. Entonces, después de copiarlo muchas veces tendríamos $|\nearrow\rangle|\nearrow\rangle \ldots |\nearrow\rangle = |\nearrow\nearrow \ldots \nearrow\rangle$, que para un momento angular suficientemente grande podría medirse de una forma clásica y se obtendría así la dirección espacial $\nearrow$. De esta manera, deberíamos haber obtenido una medida de cuál es realmente el estado (salvo un factor de proporcionalidad). Pero los procedimientos estándar **U**/**R** de la mecánica cuántica no nos permiten hacer eso. Las únicas medidas sobre un estado $|\nearrow\rangle$ que **R** nos permite realizar vienen dadas por algún operador hermítico (o normal), y estos proporcionan simplemente preguntas: «¿Está el espín en alguna dirección $\searrow$? Responder sí o NO». Después de la medida, el estado estará o bien en la dirección pedida $\searrow$ (SÍ), o bien en su opuesta $\nwarrow$ (NO). Podemos realizar realmente otras medidas si consideramos el estado de espín entrelazado con otras cosas (y pronto veremos su valor). Pero si el estado que se está examinando va a considerarse desentrelazado del mundo externo, entonces lo máximo que podemos hacer es realizar una medida directa sobre él. Todo lo que podemos

obtener del estado es una simple respuesta SÍ/NO, i.e., solo un *bit* (dígito binario) de información. Podemos rotar el aparato de medida en cualquier ángulo que queramos, pero el sistema no nos dirá la dirección $\nearrow$ en la que el estado apunta realmente. Ciertamente, esa dirección se distingue por el hecho de que es la única para la que la respuesta SÍ llega con certeza (probabilidad 1), pero no podemos saber por adelantado qué dirección es esta. (Si quien ha preparado el estado cuántico nos dijera que la dirección es $\nearrow$, entonces podríamos copiarlo, pero no es así como se ha planteado el problema; el que estamos examinando y el que nos proponemos copiar es un estado cuántico previamente *desconocido*.)

Lo que pretende conseguir la teleportación cuántica es el *envío* de un estado cuántico de un lugar a otro, digamos desde la nave espacial *Enterprise* de Kirk a la superficie del planeta inexplorado. La mecánica cuántica no pone ningún impedimento para lograr esto; en realidad, podríamos transportar el objeto cuántico corporalmente de un lugar a otro de una forma ordinaria. Pero la idea es que en la situación concreta imaginada las condiciones son demasiado «ruidosas» para el transporte fiable de un objeto cuántico o de una señal cuántica de cualquier tipo. La transmisión de información clásica ordinaria va a ser todo lo que las condiciones permiten. Sin embargo, no es posible transmitir un estado cuántico utilizando solo la señalización clásica. La razón de esto debería ser clara, porque las señales clásicas, por su propia naturaleza, pueden copiarse. Si pudieran utilizarse para transmitir un estado cuántico, entonces los estados cuánticos también podrían copiarse, y ya hemos visto antes que esto no debería ser posible. Lo que necesitamos hacer es «preparar el terreno» primero. Puesto que la imagen «planeta inexplorado» es poco apropiada para esto, permítanme recabar la ayuda de mi colega de Titán; es a este colega a quien intento transmitir un estado cuántico desconocido de espín 1/2.

La «preparación del terreno» que se requiere es que cada uno de nosotros debe estar en posesión de un miembro de un par de partículas EPR de espín 1/2. Podemos suponer que las partículas empezaron juntas en un estado de espín 0, igual que en la versión original de Bohm de EPR. Nuestra suposición era que resulta ahora poco fiable enviar estados cuánticos a través del espacio comprendido entre la Tie-

rra y Titán. Pero digamos que hace cinco años, antes de que mi colega partiera para Titán, cada uno de nosotros tomamos nuestras respectivas partículas del par entrelazado mencionado antes. Cada partícula se mantuvo perfectamente aislada de cualquier perturbación externa. Si nuestras partículas aún permanecen sin perturbar en el momento en que vuelva mi colega, entonces, al juntar las dos, se recuperaría de nuevo el estado de espín 0.

Supongamos ahora que un amigo me presenta otra partícula de espín 1/2, mantenida aislada una vez más de perturbaciones externas. Se me pide que transmita inmediatamente el estado de espín de dicha partícula intacto a mi colega en Titán. Teniendo en cuenta que ahora las condiciones no permiten que un estado cuántico pueda ser enviado a través del espacio desde aquí a Titán, solo se me permite enviar una señal de radio clásica. Pero antes de hacerlo llevo la partícula de mi amigo a donde he almacenado mi partícula EPR y junto estas dos partículas. Cada una de las partículas tiene espín 1/2, de modo que juntando sus estados constituirían un sistema 4-dimensional (solo $\mathbf{H}^4$, si no fuera por el entrelazamiento entre mi partícula y la partícula de mi colega en Titán). Ahora realizo una medida sobre el par de partículas juntas (la mía y la de mi amigo), que distingue los cuatro estados ortogonales (llamados *estados de Bell*)

$$\begin{aligned}
&(0):\ |\uparrow\rangle|\downarrow\rangle - |\downarrow\rangle|\uparrow\rangle,\\
&(1):\ |\uparrow\rangle|\uparrow\rangle - |\downarrow\rangle|\downarrow\rangle,\\
&(2):\ |\uparrow\rangle|\uparrow\rangle + |\downarrow\rangle|\downarrow\rangle,\\
&(3):\ |\uparrow\rangle|\downarrow\rangle + |\downarrow\rangle|\uparrow\rangle.
\end{aligned}$$

Transmito el resultado de esta medida a mi colega en Titán mediante una señal clásica ordinaria, codificada (digamos) por los números 0, 1, 2, 3 correspondientes, respectivamente, a aquel de los cuatro estados anteriores que revele mi medida. Al recibir mi mensaje, mi colega saca el otro miembro del par EPR, cuidadosamente protegido hasta entonces de las perturbaciones externas, y realiza sobre él la rotación siguiente

(0): dejarlo como está,
(1): 180° alrededor del eje x,

(2): 180° alrededor del eje y,
(3) : 180° alrededor del eje z.

Puede comprobarse directamente que esto consigue la «teleportación» acertada del estado cuántico de mi amigo a mi colega en Titán.[23.10]

Lo que resulta particularmente sorprendente en la teleportación cuántica es que, enviando a mi colega simplemente 2 bits de información clásica (uno de los números 0, 1, 2, 3, que podría haber sido codificado como 00, 01, 10, 11, respectivamente), he transmitido la «información» de un punto en la esfera de Riemann entera (recordemos la Fig. 22.10). En términos clásicos ordinarios, esto hubiera necesitado la información contenida en la elección sin restricciones de un punto en un continuo: ¡estrictamente $\aleph_0$ bits, para una precisión perfecta! (Véase §§16.3,4.) ¿Cómo hemos conseguido esta hazaña? En este punto debería mencionar que se han realizado experimentos reales que confirman las predicciones de la teleportación mecanocuántica (sobre distancias del orden de metros, no intervalos Tierra-Saturno, por supuesto),[14] de modo que debemos tomar estas cosas en serio. No solo eso, sino que el tema floreciente de la criptografía cuántica depende de cosas de esta naturaleza general, así como lo hacen muchas de las ideas de la computación cuántica.

Echemos una ojeada a la Fig. 23.7. Esta es un diagrama espaciotemporal en el que están indicadas mi línea de universo, la de mi amigo y la de mi colega, y, lo que es más importante, la de todas las partículas de relevancia para la historia, junto con la señal clásica que envío a mi colega en Titán. De algún modo, la «información» de la dirección del espín de la partícula de mi amigo (indicada por $|\swarrow\rangle$) se ha transportado a Titán, pese al hecho de que solo 2 bits de información han sido transmitidos realmente. ¿Cómo han llegado todos los otros $\aleph_0$ bits a mi colega?

Algunos podrían refugiarse en el punto de vista de que los estados cuánticos «no son objetos reales» puesto que «no son medibles», o algo similar. Yo mismo encuentro difícil de entender este tipo de perspecti-

[23.10] Confírmelo con los convenios adecuados concernientes a ejes de coordenadas, etc.

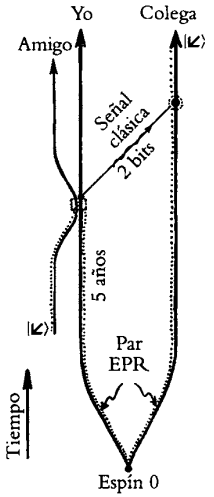


Fig. 23.7. «Teleportación cuántica» que demuestra la propagación acausal del entrelazamiento. Un diagrama espacio-temporal ilustra el proceso por el que el estado cuántico desconocido ($|K\rangle$) de espín 1/2 que me ha dado mi amigo puede ser transportado a mi colega en Titán por la mera transmisión de 2 bits de información clásica, con tal de que mi colega y yo compartamos ya un par EPR. El enlace de entrelazamiento acausal se muestra como un camino de puntos.

va del mundo. En efecto, la dirección de un estado de espín 1/2 está diciendo algo muy definido sobre el mundo. Está diciendo que si alguien (en el planeta Titán, en este ejemplo) elige medir el espín en dicha dirección concreta, entonces se obtendrá la respuesta sí con certeza. Además, mi amigo, o un amigo de mi amigo, podría muy bien haber *preparado* la partícula original para que tuviese un espín en alguna dirección preasignada, y sabría con certeza el resultado de una medida realizada en Titán en esa misma (u opuesta) dirección. Eso suena real para mí. (No se desanimen por el hecho de que mis ejemplos son un poco descabellados; ¡la clave está en el principio!)

Miremos de nuevo la Fig. 23.7. Algo real se ha transmitido desde mi amigo a mi colega, pero el canal clásico (de solo 2 bits) es demasiado estrecho para ofrecer un ámbito para que pasen los $\aleph_0$ bits restantes. Pese a todo, hay un vínculo de conexión. Consta de un pequeño tramo que va desde mi amigo hasta mí, uno largo desde mí al origen de nuestro par EPR, y otro tramo largo desde ese punto de origen hasta mi colega en Titán. Este es, de hecho, el *único* vínculo de conexión entre nosotros que es capaz de soportar la cantidad de «información» que se requiere. El problema, por supuesto, es que contiene un tramo ¡que se extiende a cinco años en el pasado!

23.10. Cuanlazamiento

Debo dejar muy claro que no estoy tratando de apoyar la idea de que la información ordinaria puede propagarse hacia atrás en el tiempo ni de que puedan utilizarse efectos EPR para enviar información clásica a velocidad mayor que la de la luz; véase *infra*. Esto llevaría a todo tipo de paradojas con las que no deberíamos tener trato en absoluto (volveré a estas cuestiones en §30.6). La información, en el sentido ordinario, no puede viajar hacia atrás en el tiempo. Estoy hablando de algo completamente diferente que a veces se conoce como *información cuántica*. Ahora bien, hay una dificultad con este término, a saber, la aparición de la palabra «información». En mi opinión, el calificativo «cuántica» no hace lo suficiente por debilitar la asociación con la información ordinaria, de modo que propongo que adoptemos un nuevo término[15] para ello:

CUANLAZAMIENTO.

En este libro, al menos, llamaré *cuanlazamiento* a lo que normalmente se denomina «información cuántica». El término sugiere «mecánica cuántica» y sugiere «entrelazamiento», algo muy adecuado. Es de lo que trata el cuanlazamiento. El cuanlazamiento tiene también mucho que ver con la información, pero no es información. No hay modo de enviar una señal ordinaria por medio del cuanlazamiento únicamente. Esto mismo queda claro a partir del hecho de que pueden utilizarse canales de cuanlazamiento dirigidos hacia el pasado tanto como canales dirigidos al futuro. Si el cuanlazamiento fuera información transmisible, entonces sería posible enviar mensajes al pasado, y no lo es. Pero el cuanlazamiento puede ser utilizado en unión con canales de información ordinaria para facilitar que estos consigan cosas que la sola señalización ordinaria no puede conseguir. Es algo muy sutil. En cierto sentido, la computación cuántica y la criptografía cuántica, y ciertamente la teleportación cuántica, dependen de forma crucial de las propiedades del cuanlazamiento y su interrelación con la información ordinaria.

Por lo que puedo entender, los enlaces del cuanlazamiento están siempre limitados por los conos de luz, igual que los enlaces de infor-

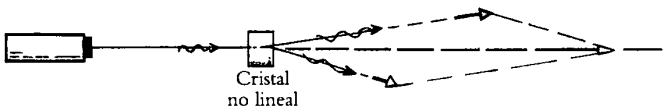


Fig. 23.8. Subconversión paramétrica. Un fotón, que emerge de un láser e incide sobre un «cristal no lineal» adecuado, produce un par de fotones entrelazados. Este entrelazamiento se manifiesta en la naturaleza EPR de los estados de polarización correlacionados de los fotones secundarios, pero también en el hecho de que sus estados de 3-momento deben sumar el del fotón incidente.

mación ordinarios, pero los enlaces de cuanlazamiento tienen la característica novedosa de que pueden zigzaguear hacia atrás y hacia delante en el tiempo[16] para conseguir una «propagación de género espacio» efectiva. Puesto que el cuanlazamiento no es información, esto no permite enviar señales reales más rápidas que la luz. Hay también una asociación entre cuanlazamiento y geometría espacial ordinaria (a través de las conexiones entre la esfera de Riemann y el espín, tal como se representan en las Figs. 22.10, 22.14 y 22.16), asociación que queda reflejada espacialmente en una inversión de la dirección temporal, con interesantes implicaciones.[17] Nos llevaría demasiado lejos explorar estas implicaciones en detalle.

Uno de los usos más directos de la idea de cuanlazamiento está en ciertos experimentos en los que se produce un par de fotones entrelazados mediante el proceso conocido como *subconversión paramétrica* (véase la Fig. 23.8). Esto ocurre cuando un fotón producido por un láser entra en un tipo concreto de cristal que lo convierte en un par de fotones. Estos fotones emitidos están entrelazados de diversas maneras. La suma de sus momentos debe ser igual al momento del fotón incidente, y sus polarizaciones están también mutuamente relacionadas en una forma EPR, como los ejemplos que se han dado antes.

En un experimento particularmente sorprendente, uno de los fotones (fotón A) pasa a través de un agujero de una forma particular cuando se dirige hacia su detector D_A. El otro fotón (fotón B) pasa a través de una lente que está situada de modo que lo enfoca, oportunamente, en su detector D_B. La posición del detector D_B se mueve ligeramente cada vez que se emite cada par fotónico. La situación se ilustra en la Fig. 23.9a. Cuando quiera que D_A registra la recepción del

fotón A y al mismo tiempo D_B registra la recepción de B, la posición de D_B es anotada. Esto se repite muchas veces, y poco a poco se forma una imagen por el detector D_B, contando solo las posiciones de B cuando D_A registra simultáneamente. La forma del agujero por el que pasa A se forma poco a poco en D_B, ¡incluso si el fotón B nunca pasa directamente por dicho agujero! Es como si D_B «viera» la forma del agujero mirando hacia atrás en el tiempo al punto de emisión C en el cristal, y luego hacia delante en el tiempo disfrazado de fotón A. Puede hacer esto porque el proceso de «ver» en esta situación se consigue por cuanlazamiento. Este salto hacia atrás y hacia delante en el tiempo es precisamente el tipo de cosas que se permiten hacer al cuanlazamiento. Incluso la intensidad y la posición de la lente pueden entenderse en términos de cuanlazamiento. Para obtener la posición de la lente, consideremos un espejo colocado en el punto de emisión C. La lente (una lente positiva) se coloca de modo que la imagen del agujero, tal como refleja en este espejo en C, se focaliza en el detector D_B. Por supuesto, no hay ningún espejo real en C, pero los enlaces de cuanlazamiento actúan como si se reflejasen en un espejo, aunque se reflejan en el tiempo tanto como en el espacio.[23.11]

Por si el lector encuentra este experimento rocambolesco, debería dejar claro que este es un efecto real. Ha sido confirmado con éxito en experimentos[18] realizados en la Universidad de Baltimore, Maryland. Se han realizado también otros experimentos relacionados que implican subconversión paramétrica y que pueden entenderse mejor en términos de cuanlazamiento.[19]

Por otra parte, el tipo general de situación que se ilustra en la Fig. 23.9a podría considerarse como algo que no es «esencialmente mecanocuántico», pues se podría imaginar un dispositivo en C que simplemente expulsa partículas clásicas por pares en las direcciones apropiadas y, aparte de la focalización, podrían obtenerse resultados similares. Podemos remediar esto utilizando una modificación del montaje de Elitzur-Vaidman que se ha ilustrado en la Fig. 22.6 (reflejado horizontalmente); véase la Fig. 23.9b. Ahora hay solo un fotón cada vez. Puede

[23.11] Intente dar una explicación más completa de esto utilizando ideas de cuanlazamiento u otras.

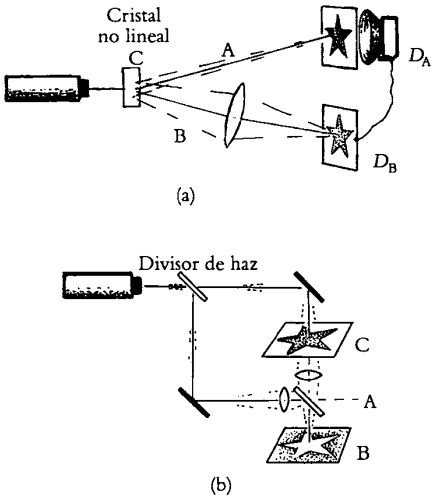


Fig. 23.9. Transmisión de una imagen vía efectos cuánticos. (a) Fotones entrelazados A, B son producidos por subconversión paramétrica en C. El fotón A tiene que pasar por un agujero de una forma especial para llegar al detector D_A, mientras que B atraviesa una lente, colocada de modo que su foco está en el detector D_B. Las posiciones de los detectores se mueven de manera gradual, adecuadamente al unísono, y cuando ambos registran, se anota la posición de D_B. Si se repite esto muchas veces, una imagen de la forma del agujero se forma poco a poco en D_B, contando solo las posiciones de B cuando D_A también registra. (Esto se ilustra aquí esquemáticamente teniendo, en lugar de D_B, una placa fotográfica fija que solo se activa cuando D_A registra.) El entrelazamiento se ilustra porque la posición de la lente está determinada como si C fuera un «espejo» que refleja hacia atrás el fotón tanto en el tiempo como en la dirección. (b) Esquema alternativo utilizando una adaptación del test de la bomba de Elitzur-Vaidman de la Fig. 22.6 (que debe reflejarse en una línea horizontal). La placa fotográfica en B recibe el fotón solo cuando el fotón «haya sido detenido» por la plantilla en C, ¡pero realmente ha tomado la ruta inferior!

registrarse en la placa fotográfica solo si se destruye la interferencia cuando la ruta alternativa para el fotón evitara el agujero en C.

Consideremos ahora de nuevo un efecto EPR ordinario, como los ejemplos de Stapp y Hardy que se han comentado antes. En la aplicación ordinaria de los procesos **R** cuánticos, uno imagina un sistema de referencia concreto en el que existe una coordenada temporal t que proporciona secciones de tiempo paralelas, cada una de las cuales co-

rresponde a un valor t constante en el espaciotiempo. El proceder normal consiste en adoptar el punto de vista (no relativista) según el cual, cuando se mide un miembro de un par EPR, el estado del otro miembro se reduce simultáneamente, de modo que una medida posterior ve un estado reducido (desentrelazado) más que un estado entrelazado. Este es el tipo de descripción que puede utilizarse, por ejemplo, en mis ejemplos EPR concretos. Supongamos que, desde el punto de vista de un sistema de referencia estacionario con respecto al Sol, es la medida de mi *colega* en Titán la que se hace primero, unos quince minutos antes de mi propia medida aquí en la Tierra. Así, en esta imagen de las cosas, es la medida de mi colega la que reduce el estado, y yo realizo posteriormente una medida sobre una partícula con un estado desentrelazado. Pero podríamos imaginar que, en su lugar, la situación global viene descrita desde la perspectiva de un observador O que pasa a gran velocidad (digamos $\frac{2}{3}c$) en la dirección que va desde mi colega en Titán hasta mí. Desde el punto de vista de O, yo fui quien hizo primero la medida sobre el par EPR, reduciendo con ello el vector de estado, y mi colega quien midió el estado desentrelazado reducido (Fig. 23.10) (véanse §18.3 y la Fig. 18.5b). Las probabilidades conjuntas son las mismas de las dos maneras, pero O tiene una imagen de la realidad «diferente» de la que mi colega y yo teníamos antes. Si consideramos **R** como un proceso *real*, entonces parece que estamos en conflicto con el principio de la relatividad especial, porque existen dos puntos de vista incompatibles respecto a quién de nosotros efectuó la reducción del estado y quién de nosotros observó el estado reducido tras la reducción.

Podemos deducir de esto que los efectos EPR, pese a su naturaleza aparentemente acausal, no pueden utilizarse directamente para transmitir acausalmente información ordinaria, que cabría imaginar que pudiera influir en el comportamiento de un receptor separado del transmisor con una separación de género espacio. Siempre se puede escoger un sistema de referencia en el que es el «suceso de recepción» el que ocurre primero, y luego el «transmisor» solo tiene el estado reducido para examinar. Y entonces es «demasiado tarde» para utilizar el entrelazamiento para enviar una señal porque ya ha sido destruido por la reducción del estado.

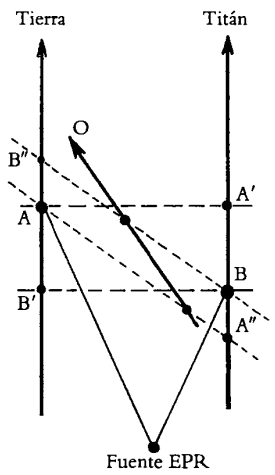


Fig. 23.10. ¿Conflicto entre la relatividad y la objetividad de la reducción de estado? Diagrama espaciotemporal de una situación EPR, con detectores en la Tierra y Titán y fuente más próxima a Titán que a la Tierra. Desde la perspectiva de un sistema inercial, estacionario con respecto al Sol, el detector en Titán registra primero (en B) y esto reduce el estado simultáneamente en B′ en la Tierra. Solo más tarde tiene lugar la detección en la Tierra (en A) de un estado ahora desentrelazado (simultáneo con A′ en Titán). Sin embargo, para un observador O, que viaja hacia la Tierra desde Titán a gran velocidad, la detección tiene lugar primero en la Tierra (en A, simultánea con A″ en Titán, de acuerdo con las líneas de simultaneidad «en pendiente» de O) y Titán recibe el estado desentrelazado reducido (en B, simultáneo con B″ en la Tierra).

¿Cuál es la perspectiva de cuanlazamiento sobre estas cuestiones?[20] Véase §30.3. En esta imagen no es correcto pensar que es una de las dos medidas (la mía o la de mi colega) la que ha efectuado la reducción y que la otra (la de mi colega o la mía) es la que mide el estado reducido. Los dos sucesos de medida están en pie de igualdad, y consideramos que el cuanlazamiento proporciona una conexión entre estos dos sucesos que los correlaciona. No supone ninguna diferencia qué suceso se vea como el pasado del otro, pues el cuanlazamiento puede considerarse igualmente propagándose hacia el pasado o propagándose hacia el futuro. Al no ser capaz directamente de transportar información, el cuanlazamiento no respeta las restricciones normales de la causalidad relativista. Simplemente produce ligaduras sobre las probabilidades conjuntas de los resultados de medidas diferentes.

Aunque el cuanlazamiento es una idea útil para «dar sentido» a este tipo de experimento cuántico enigmático, no estoy seguro de hasta dónde pueden llevarse estas ideas, ni con qué precisión pueden ser delineados los efectos del cuanlazamiento. La idea de cuanlazamiento no resuelve ciertamente la cuestión de la medida cuántica, ya que nos dice poco, si es que nos dice algo, sobre las circunstancias en las que **R** domina sobre **U**. Esta cuestión será abordada con más detalle en los capí-

tulos 29 y 30, especialmente en §30.12, pero, en mi opinión, el papel preciso del cuanlazamiento no está todavía muy claro. Hay una conexión más prometedora con algunas ideas de la teoría de twistores, las cuales serán examinadas brevemente en §33.2.

Notas

Sección 23.1

23.1. Véanse Eddington (1929b), Mott (1929) y Dirac (1932).

Sección 23.3

23.2. Véanse Einstein *et al.* (1935), Schrödinger (1935) y también Afriat (1999).

23.3. No obstante, lo que nos dice esto en detalle sobre la computación cuántica es una cuestión sutil; véanse Josza (1998) y Josza y Linden (2002).

23.4. Véase Bell (1987). Quizá la versión más clara y más ampliamente citada de esta desigualdad es la debida a Clauser *et al.* (1969). Toma la forma $|E(A, B) - E(A, D)| + |E(C, B) + E(C, D)| \leq 2$, donde $E(x, y)$ es el valor esperado de acuerdo ($E = 1$ para completo acuerdo y $E = -1$ para completo desacuerdo) entre los resultados de medidas alternativas A, C para una componente del par EPR, y B, D para la otra.

23.5. Véanse Bohm (1951), Redhead (1987) y Afriat (1999). Para una reciente y extraordinaria confirmación experimental de efectos EPR, véase Tittel *et al.* (1998). Para varios ejemplos concretos relevantes para esto (Heywood y Redhead, 1983 y Stairs, 1983), véanse Kochen y Specker (1967), Peres (1991, 1995), Conway y Kochen (2002), Penrose (1994), sección 5.3, Penrose (2000b) y Zimba y Penrose (1993).

23.6. Véanse Stapp (1971, 1979) y Hardy (1992, 1993).

23.7. Véase, por ejemplo, Hannbuss (1997). Para una discusión general de estas cuestiones en entrelazamiento, véase Nielsen y Chuang (2000).

Sección 23.4

23.8. Véase la nota 23.5.

Sección 23.5

23.9. Véase Hardy (1992, 1993).

23.10. Véase Hardy (1993).

Sección 23.8

23.11. Un esquema de este tipo se adoptaba efectivamente en mi libro *Sombras de la mente* (1994), §5.15, pero sin utilizar «cuñas» explícitamente.

Sección 23.9

23.12. Véase Penrose (1989).
23.13. Véase Wooters y Zurek (1982).
23.14. Véase Jennewein *et al.* (2002).

Sección 23.10

23.15. Véase Penrose (2002).
23.16. Véanse Josza (1998) y Peres (2000).
23.17. Véase Penrose (1998a).
23.18. Véase Shih *et al.* (1995).
23.19. Véase Gisin *et al.* (2003), por ejemplo, para hacerse una idea de esta importante área.
23.20. Compárese con Aharonov y Vaidman (2001), Cramer (1988), Costa de Beauregard (1995) y Werbos y Dolmatova (2000).

24

El electrón y las antipartículas de Dirac

24.1. Tensión entre la teoría cuántica y la relatividad

Las consideraciones de §23.10 tan solo empiezan a tocar algunas de las cuestiones profundas acerca de la relación entre los principios de la mecánica cuántica y los de la relatividad. De hecho, al presentar en los tres capítulos precedentes el modo detallado en que opera la teoría cuántica, he adoptado una posición no muy relativista, aparentando ignorar las importantes lecciones que nos han enseñado Einstein y Minkowski sobre la interdependencia del tiempo y el espacio (como se ha descrito en el capítulo 17). De hecho, esto es bastante habitual en la teoría cuántica. El enfoque estándar adopta una «imagen de la realidad» en la que el tiempo se trata de forma diferente del espacio. Como se ha señalado en el capítulo 22, existe una única coordenada temporal externa, pero hay muchas coordenadas espaciales, pues cada partícula requiere su propio conjunto. Es habitual considerar que esta simetría es una característica «provisional» de la teoría cuántica no relativista, que sería meramente una aproximación a algún esquema relativista mucho más completo. En este capítulo y en los dos siguientes empezaremos a ser testigos de las cuestiones profundas que se plantean cuando tratamos seriamente de unir los principios de la teoría cuántica con los de la relatividad especial. (La unión más ambiciosa con la relatividad *general* de Einstein —donde la gravitación y la curvatura espaciotemporal entran también en la imagen— requiere mucho más, y todavía no hay consenso sobre las líneas de búsqueda más prometedoras. Abordaré algunas de estas en los capítulos 28 y 30-33.)

Una característica particular de la combinación de la teoría cuán-

tica y la relatividad especial es que la teoría resultante se convierte no solo en una teoría de partículas cuánticas, sino en una teoría de *campos* cuánticos. La razón para esto puede reducirse al hecho de que la introducción de la relatividad implica que las partículas individuales ya no se conservan, sino que pueden ser creadas y destruidas en unión con sus *antipartículas*. Este comentario necesita alguna explicación. ¿Por qué hay esta necesidad de «antipartículas» en una teoría cuántica relativista? ¿Por qué la presencia de antipartículas nos lleva de una teoría cuántica de partículas a una teoría cuántica de campos? Este capítulo está dirigido básicamente a responder a estas dos preguntas, pero en especial a la primera, y con particular referencia a las maravillosas ideas de Dirac sobre la descripción matemática de los electrones.

La teoría cuántica de campos propiamente dicha será estudiada en el capítulo 26, y vislumbraremos parte de la tensión permanente entre la relatividad especial y la teoría cuántica que ha guiado muy eficazmente la disciplina de la física de partículas hacia esquemas matemáticos cada vez más elaborados. Nosotros mismos nos veremos tentados a hacer un viaje largo y fascinante. Cuando la tensión puede resolverse de manera adecuada, como sucede con el modelo estándar de la física de partículas que se examinará en el capítulo 25, se ve que la teoría resultante presenta un acuerdo muy notable con los hechos observacionales.

Pese a todo, en muchos aspectos, esta tensión ha permanecido y nunca ha sido completamente resuelta. Estrictamente hablando, la teoría cuántica de campos (al menos en la mayoría de los ejemplos no triviales y muy relevantes de esta teoría que conocemos) es *matemáticamente inconsistente* y se necesitan varios «trucos» para hacer cálculos que tengan sentido. Saber si estos trucos son meramente procedimientos provisionales que nos facilitan el ir avanzando dentro de un marco matemático que quizá sea fundamentalmente defectuoso en un nivel profundo, o si estos trucos reflejan verdades profundas que tienen realmente una importancia genuina para la propia naturaleza, es una cuestión delicada. La mayor parte de los intentos recientes para avanzar en la física fundamental consideran que muchos de estos «trucos» son fundamentales. Veremos varios ejemplos de tan ingeniosos esquemas en este y en posteriores capítulos. Algunos de ellos parecen estar desvelando verdaderamente algunos de los secretos de la naturaleza. Por otra parte, ¡muy bien podría suceder que la naturaleza esté mucho menos en sintonía con algunos de los otros!

24.2. ¿Por qué las antipartículas implican campos cuánticos?

La predicción teórica de las antipartículas en una teoría cuántica relativista parece haber desvelado uno de los secretos genuinos de la naturaleza, ahora bien apoyado por la observación. Veremos algo de las razones teóricas para las antipartículas hacia el final del presente capítulo, y más concretamente en §24.8. Por el momento, en lugar de abordar esta cuestión, centremos nuestra atención en la segunda de las dos preguntas planteadas antes, a saber, ¿por qué la presencia de antipartículas nos lleva de una teoría cuántica de partículas a una teoría cuántica de campos? Aceptemos por ahora que existe una antipartícula para cada tipo de partícula, y tratemos de entender las consecuencias de este hecho notable.

La propiedad clave de una antipartícula (al menos, la antipartícula de una partícula masiva) es que la partícula y la antipartícula pueden juntarse y aniquilarse mutuamente, convirtiéndose su masa combinada en energía de acuerdo con la expresión $E = mc^2$ de Einstein; recíprocamente, si se introduce energía suficiente en un sistema localizado en una región apropiadamente pequeña, entonces hay una fuerte posibilidad de que esta energía pudiera servir para crear una partícula junto con su antipartícula. Así pues, con este potencial para la producción de antipartículas existe siempre la posibilidad de que más partículas entren en la imagen, apareciendo cada partícula junto con su antipartícula. Así pues, nuestra teoría relativista no puede ser solo una teoría de partículas individuales, ni de ningún número fijo de partículas cualesquiera. (Como veremos en especial en los capítulos 25 y 26, en teoría cuántica, si puede suceder algo potencialmente —por ejemplo, la producción de numerosos pares partícula-antipartícula—, entonces esta posibilidad potencial aporta realmente su contribución al estado cuántico.) Por consiguiente, cuando uno trata de llegar a una teoría de partículas relativistas se ve impulsado a ofrecer una teoría en la que existe un potencial para la creación de un número ilimitado de partículas.

Esto nos lleva fuera del marco de los capítulos 21-23, pero en el capítulo 26 veremos que la teoría cuántica de campos nos permite acomodar tal comportamiento. De hecho, de acuerdo con un punto de vista habitual, las entidades primarias en tal teoría son campos cuánticos, y las propias partículas aparecen meramente como «excitaciones del campo». Pese a todo, encontraremos que esta no es la única manera de considerar la teoría cuántica de campos. En el enfoque de diagramas de Feynman, que abordaremos en los capítulos 25 y 26, se adopta una fuerte perspectiva «tipo partícula» sobre los procesos básicos que van a constituir la teoría cuántica de campos, donde, de hecho, puede crearse o destruirse un número ilimitado de partículas.

Resulta instructivo desarrollar un poco más las razones subyacentes en la creación de partículas, como una característica de una teoría cuántica relativista razonable. Por el momento, estoy suponiendo todavía que existen las antipartículas. En esencia, la razón para esperar la creación de partículas se reduce a la famosa $E = mc^2$ de Einstein. La energía es básicamente intercambiable con la masa (siendo c^2 meramente una «constante de conversión» entre las unidades de energía y las de masa que se estén utilizando). Si se dispone de energía suficiente, entonces la masa de una partícula puede conjurarse a partir de dicha energía.

Sin embargo, tener los medios para producir la masa de la partícula no es por sí solo suficiente para crear la propia partícula. Es probable que haya varios números cuánticos (aditivos) *conservados*, tales como la carga eléctrica (u otros, por ejemplo, el número bariónico), que se supone que no pueden cambiar en un proceso físico. Crear simplemente una partícula cargada a partir de pura energía, por ejemplo, representaría una violación de la conservación de carga (y lo mismo sería aplicable a otras cantidades conservadas, tales como el número bariónico, etc.). Sin embargo, con la hipótesis de que para cada tipo de partícula existe una correspondiente *antipartícula*, para la cual cada número cuántico aditivo tiene el signo invertido, una partícula puede crearse junto con su antipartícula a partir de pura energía (véase la Fig. 24.1). Todos los números cuánticos aditivos se conservarán en este proceso.

La masa en reposo de la antipartícula es, por el contrario, la misma que la de la partícula original (puesto que la masa en reposo no es aditiva). Necesitamos energía suficiente —al menos el doble de la masa-

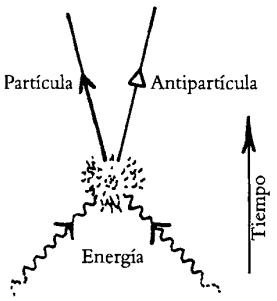


Fig. 24.1. Una partícula y su antipartícula pueden crearse a partir de energía. Todos los números cuánticos aditivos conservados de la partícula tienen signo inverso para la antipartícula, para asegurar la conservación de estas cantidades en el proceso de creación.

energía en reposo de la propia partícula— para crear ambas, la partícula y su antipartícula, en este proceso. Recíprocamente, si una partícula de un tipo dado choca con otra partícula que es del tipo de su antipartícula, entonces es posible que se aniquilen mutuamente con producción de energía. Una vez más, la energía tiene que ser al menos el doble de la masa-energía en reposo de la partícula individual. En el proceso de creación o de aniquilación, la energía puede tener un valor mayor que este, porque es probable que la partícula y la antipartícula estén en movimiento relativo y habrá una energía en este movimiento —la energía cinética— que se sume al total. En cualquier caso, vemos que la presencia de antipartículas nos obliga a apartarnos de la teoría cuántica de partículas individuales, como se ha descrito en los capítulos 21-23.

24.3. Positividad de la energía en la mecánica cuántica

Volvamos ahora al camino que, en última instancia, nos lleva al requisito de antipartículas en una teoría cuántica relativista. Tendremos que examinar el marco de la teoría cuántica desde una perspectiva algo más profunda que antes. En primer lugar, recordemos la forma básica de la ecuación de Schrödinger

$$i\hbar\frac{\partial\psi}{\partial t} = \mathcal{H}\psi.$$

Supongamos que exigimos que nuestro sistema cuántico tenga un valor definido E para su energía, de modo que ψ es un *autoestado de la*

energía con valor propio E; es decir (puesto que $\mathcal{H}$ es el operador que define la energía total del sistema), exigimos

$$\mathcal{H}\psi = E\psi.$$

Según el proceso **R** mecanocuántico (§§22.1,5), un estado ψ semejante sería el resultado de haber realizado una medida en un sistema planteándole la pregunta «Cuál es tu energía», para la que hemos recibido la respuesta concreta «E». La ecuación de Schrödinger nos dice ahora

$$i\hbar\frac{\partial\psi}{\partial t} = E\psi.$$

Las soluciones de esta ecuación tienen la forma[24.1]

$$\psi = C\,e^{-iEt/\hbar},$$

donde C es *independiente* de t (i.e., una función compleja de las variables espaciales solamente).

Ahora bien, es importante que el valor de la energía E sea un número positivo. Los estados de energía negativa son «malas noticias» en mecánica cuántica, por varias razones (su presencia lleva a inestabilidades catastróficas).[1][24.2] Cuando la energía E es positiva, el coeficiente $-iE/\hbar$ de t en el exponente (en $e^{-iEt/\hbar}$) es un múltiplo *negativo* de i. Recordemos de §9.5 (y véase la nota 9.3) que de las funciones $\psi(t)$ de esta naturaleza, o combinaciones lineales de tales funciones, se dice (de forma algo confusa) que son de *frecuencia positiva*.

Recordemos también que en §9.3 hemos abordado la separación de una función $f(x)$ (de una variable real x) en sus partes de frecuencia positiva y negativa de una manera en apariencia completamente diferente, a saber, en términos de la geometría de la esfera de Riemann.[2] Allí hemos tratado esto solo como una pieza elegante de las matemáticas puras. La recta real podía considerarse enrollada una vez alrededor del ecuador de la esfera de Riemann, y se entendía que la parte de frecuencia positiva de la función f era esa parte que se extendía —holomórficamente (véase §7.1)— al hemisferio sur, mientras que, análogamente, la parte de frecuencia negativa se extendía al hemisferio norte. Pero ahora tenemos que llegar a una razón *física* notable para la gran importancia de esta noción. Cualquier función de onda que se autorrespete, aunque no tenga que ser un estado propio de energía, debería ser expresable como combinación lineal de autoestados de energía, y cada autovalor de la energía debería ser positivo. Así pues, la dependencia temporal de cualquier función de onda decente debería tener esta propiedad crucial de frecuencia positiva. Creo que esta notable relación entre un requisito físico esencial, por una parte, y una elegante propiedad matemática, por la otra, es un maravilloso ejemplo de la relación profunda, sutil y misteriosa entre las ideas matemáticas más sofisticadas y el funcionamiento interno de nuestro universo real.

En la mecánica cuántica no relativista, este requisito de frecuencia positiva suele darse automáticamente, como una característica natural de la teoría, siempre que el hamiltoniano proceda de un problema físico razonable donde las energías clásicas son positivas. Por ejemplo, en el caso de una única partícula (sin espín) libre y no relativista de masa (positiva) μ, tenemos el hamiltoniano $\mathcal{H} = p^2/2\mu$ (recordemos §20.2 y §21.2). La expresión p^2 y con ella el propio hamiltoniano $\mathcal{H}$, es lo que se denomina «definida positiva».[3] Clásicamente, esto se debe a que p^2 es una suma de cuadrados, y esta no puede ser negativa: $p^2 = \mathbf{p}\cdot\mathbf{p} = (p_1)^2 + (p_2)^2 + (p_3)^2$. En mecánica cuántica debemos hacer el reemplazamiento de $\mathbf{p}$ por $-i\hbar\boldsymbol{\nabla}$, donde $\boldsymbol{\nabla} = (\partial/\partial x^1, \partial/\partial x^2, \partial/\partial x^3)$, y ahora la expresión «definida positiva» se refiere a los *autovalores* del operador $-\nabla^2$ (para estados normalizables, i.e., elementos de un espacio de Hilbert **H** apropiado), y de nuevo estos no pueden ser negativos, esencialmente por la misma razón que en el caso clásico.[24.3]

[24.1] Compruebe que esta es realmente una solución.

[24.2] Explique por qué añadir una constante K al hamiltoniano tiene simplemente el efecto de que todas las soluciones de la ecuación de Schrödinger se multiplican por el mismo factor. Encuentre este factor. ¿Afecta esto sustancialmente a la dinámica cuántica? Supongamos que estamos interesados en el efecto gravitatorio de un sistema cuántico. ¿Por qué no podemos simplemente «renormalizar» de este modo la energía en tales circunstancias?

[24.3] Aquí la ecuación de Schrödinger es $\partial\psi/\partial t = (i\hbar/2\mu)\nabla^2\psi$. Confirmando, primero, que para un autoestado de la energía con energía E tenemos $-\nabla^2\psi = A\psi$, donde $A = 2\mu\hbar^{-2}E$, utilice el *teorema de Green* $\int\psi\nabla^2\psi\,d^3x = -\int\boldsymbol{\nabla}\psi\cdot\boldsymbol{\nabla}\psi\,d^3x$ para demostrar que A debe ser positiva para un estado normalizable. (Recíprocamente, es cierto de hecho que, para A positiva, existen muchas soluciones de $-\nabla^2\psi = A\psi$, que decrecen adecuadamente en el infinito, de modo que la norma $\|\psi\|$ permanece finita[4] y podemos normalizar a $\|\psi\| = 1$ si queremos.) Demuestre cómo se obtiene el teorema de Green a partir del teorema fundamental del cálculo exterior.

24.4. Dificultades con la fórmula de la energía relativista

Consideremos ahora una partícula cuántica *relativista*. En este caso, el hamiltoniano se obtiene a partir de la expresión relativista para la energía, que ya no es $p^2/2\mu$ sino

$$[(c^2\mu)^2 + c^2p^2]^{1/2}.$$

Esta expresión procede directamente de la ecuación $(c^2\mu)^2 = E^2 - c^2p^2$ de §18.7, donde μ es ahora la masa en reposo de la partícula. El lector al que le preocupe que esta expresión no se parece mucho a $p^2/2\mu$ debería remitirse al ejercicio [18.20]. Eso nos dijo, a partir del desarrollo en serie de potencias de $[(c^2\mu) + c^2p^2]^{1/2}$, que nuestra expresión relativista incorpora como primer término la famosa $E = mc^2$ de Einstein. Este término es la contribución a la energía que procede de la masa en reposo de la partícula, y se añade a la energía cinética del movimiento de la partícula. El segundo término nos da, de hecho, el hamiltoniano newtoniano (energía cinética) $p^2/2\mu$.

¡El lector puede así tranquilizarse por nuestra elección del hamiltoniano relativista! De todas formas, sería bastante complicado (y no muy iluminador) intentar utilizar esta expresión en una serie de potencias real para nuestro hamiltoniano, debido especialmente a que la serie clásica ni siquiera converge cuando $p^2 > \mu^2$. Pese a todo, encontraremos que la raíz cuadrada (potencia un medio) en la expresión explícita $[(c^2\mu)^2 + c^2p^2]^{1/2}$ lleva sus propias y profundas dificultades en relación con el mantenimiento del requisito de frecuencia-positiva. Tratemos de comprender algo de la importancia de esto.

Para no recargar innecesariamente nuestras expresiones, volvamos a unidades para las que la velocidad de la luz es la unidad

$$c = 1,$$

de modo que nuestro hamiltoniano relativista (incluyendo la energía en reposo) es ahora

$$\mathcal{H} = (\mu^2 + p^2)^{1/2}.$$

Debemos tener en cuenta que en mecánica cuántica el p^2 es realmente el operador de derivada parcial de segundo orden $-\hbar^2\nabla^2$, de

modo que necesitaremos una considerable sofisticación matemática si vamos a asignar realmente un significado consistente a la expresión $(\mu^2 - \hbar^2\nabla^2)^{1/2}$, ¡que es la raíz cuadrada de un operador de derivada parcial! (Para apreciar esta dificultad, pensemos en tratar de asignar un significado a algo como $\sqrt{(1 - \mathrm{d}^2/\mathrm{d}x^2)}$, por ejemplo.)[24.4]

Hay una dificultad más seria con esta expresión de raíz cuadrada, porque contiene una *ambigüedad de signo* implícita. En física clásica, tales cosas no nos preocuparían, porque las magnitudes en consideración son funciones ordinarias con valor real, y podemos imaginar que podríamos mantener los valores *positivos* separados de los negativos. Sin embargo, esto no es tan fácil en mecánica cuántica. La razón para ello es, en parte, que las funciones de onda cuánticas son complejas, y las dos raíces cuadradas de una expresión compleja no tienden a separarse claramente en «positivas» y «negativas» de una forma globalmente consistente (§5.4). Habría que considerar esto en relación con el hecho de que la mecánica cuántica trabaja con operadores que actúan sobre funciones complejas, y cosas como las raíces cuadradas pueden llevar a ambigüedades esenciales que no se resuelven simplemente diciendo «tómese la raíz positiva».

Existe otra forma de expresar esta dificultad. En la mecánica cuántica hay que considerar que todos los diversos objetos posibles que «pudieran» intervenir en una situación física pueden contribuir al estado cuántico, y por lo tanto todas estas alternativas tienen una influencia sobre lo que suceda. Cuando hay implicado algo como una raíz cuadrada, cada una de las dos raíces tiene que considerarse como una «posibilidad», de modo que incluso una «energía negativa no física» tiene que ser considerada como una «posibilidad física». En cuanto existe la potencialidad de tal estado de energía negativa, entonces se abre la posibilidad de una transición espontánea de energía positiva a negativa, lo que puede conducir a una inestabilidad catastrófica. En el caso de una partícula libre no relativista no tenemos este problema de la posibilidad de energía negativa porque la cantidad definida positiva $p^2/2\mu$ no tiene esta incómoda raíz cuadrada. Sin embargo, la expresión relativista

[24.4] Haga algunas sugerencias utilizando la transformada de Fourier (§9.4), o una serie de potencias, o integrales de contorno, o cualquier otra cosa.

$(\mu^2 + p^2)^{1/2}$ es más problemática porque no tenemos normalmente una regla nítida para descartar raíces cuadradas negativas.

Sucede que en el caso de una única partícula libre (o un sistema de partículas no interactuantes) esto no provoca realmente una dificultad, porque podemos restringir nuestra atención a superposiciones de ondas planas de energía positiva que son soluciones de la ecuación de Schrödinger libre, precisamente las consideradas en §21.5, y no hay transiciones a estados de energía negativa. Sin embargo, en presencia de *interacciones* ya no es así. Incluso para una única partícula cargada relativista en un campo electromagnético de fondo fijo, la función de onda no puede mantener, en general, la condición de ser de frecuencia positiva. En esto empezamos a percibir la tensión entre los principios de la mecánica cuántica y los de la relatividad.

Como veremos en §24.8, el gran físico Paul Dirac fue el primero en encontrar una forma de resolver esta tensión particular. Como primer paso, hizo una propuesta ingeniosa y profundamente intuitiva —su ahora famosa ecuación para el electrón— que se deshacía de la molesta raíz cuadrada de una forma maravillosa e inesperada. Esto llevó después a un punto de vista muy original en el que las energías negativas se eliminan, al estar asumidos sus efectos por lo que entonces era una predicción sorprendente: la existencia de *antipartículas*. Para comprender todo esto, volvamos a esa característica esencial de la relatividad general en la que se origina la raíz cuadrada.

24.5. La no invariancia de $\partial/\partial t$

Recordemos la razón subyacente en nuestra aparente necesidad de adoptar el hamiltoniano $(\mu^2 + p^2)^{1/2}$ en el caso relativista. Esta se reduce, en última instancia, al hecho de que la ecuación de Schrödinger hace uso del operador $\partial/\partial t$ (i.e., «tasa de cambio con respecto al tiempo»), mientras que, en relatividad, $\partial/\partial t$ no es un objeto invariante porque tiempo y espacio no pueden considerarse por separado, sino que son simplemente aspectos particulares de un «espaciotiempo» combinado. Así pues, no es «relativísticamente invariante» considerar $\partial/\partial t$ como un objeto fundamental. Ahora bien, como hemos visto en §21.3,

el $\partial/\partial t$ en la ecuación de Schrödinger procede del «truco de cuantización» general por el que el 4-momento p_a espaciotemporal estándar (i.e., la energía E, y el 3-momento lineal negativo $-\mathbf{p}$) se reemplazan por los operadores diferenciales $i\hbar\partial/\partial x^a$ (i.e., la energía E por $i\hbar\partial/\partial t$ y $-\mathbf{p}$ por $i\hbar\boldsymbol{\nabla}$). La «no invariancia relativista» de $\partial/\partial t$ está así íntimamente relacionada con la no invariancia de la energía. De la misma forma que el tiempo y el espacio se mezclan en la teoría de la relatividad, también se mezclan la energía y el momento lineal (como hemos visto en §18.7).

Recordemos también que la $E = mc^2$ de Einstein (con el convenio de que $c = 1$) nos dice que la energía es masa y la masa es energía, de modo que también la masa es «no invariante». Sin embargo, esto se refiere al concepto de «masa» aditiva m (componente temporal del 4-vector energía-momento) que no es intrínseca de una partícula, sino que es la masa medida en algún sistema de referencia que no tiene por qué compartir la velocidad de dicha partícula. Cuanto mayor sea la velocidad de la partícula, mayor será su masa «percibida» (lo que, de hecho, es la razón de que m no sea una cantidad invariante). La masa en reposo μ de una partícula es invariante, pero el problema con la masa en reposo es que no es aditiva y no se conserva en transformaciones de partículas, de modo que se trata de una pobre elección para algo que va a igualarse con un hamiltoniano. Más aún, μ viene dada como una raíz cuadrada de una expresión en la energía y el momento, a saber (tomando $c = 1$),

$$\mu^2 = p_a p^a = m^2 - p^2 \quad \text{i.e., } \mu = (m^2 - p^2)^{1/2},$$

que reexpresa la expresión en la raíz cuadrada para la masa-energía $m = E(= \mathcal{H})$ que teníamos antes, a saber, $m = (\mu^2 + p^2)^{1/2}$.

De todas formas, podríamos jugar con la idea de utilizar esta energía en reposo invariante μ, o su cuadrado μ^2, en una ecuación tipo Schrödinger, en lugar de la componente de energía no invariante m. El truco de cuantización (i.e., m reemplazado por $i\hbar\partial/\partial t$ y $\mathbf{p}$ por $-i\hbar\boldsymbol{\nabla}$) aplicado a la energía en reposo al cuadrado, a saber, $\mu^2 = m^2 - p^2$, nos proporciona $(i\hbar)^2$ multiplicado por el operador[5]

$$\Box = \left(\frac{\partial}{\partial t}\right)^2 - \boldsymbol{\nabla}^2$$

$$= \left(\frac{\partial}{\partial t}\right)^2 - \left(\frac{\partial}{\partial x}\right)^2 - \left(\frac{\partial}{\partial y}\right)^2 - \left(\frac{\partial}{\partial z}\right)^2$$

en coordenadas minkowskianas (t, x, y, z). Este se denomina *operador de onda* o *D'Alembertiano*, y tiene un significado invariante. (Recordemos que $(\partial/\partial x)^2$ significa el operador de derivada segunda $\partial^2/\partial x^2$, etc.) Aunque la ecuación de Schrödinger convencional no nos permite emplear directamente este operador (por las razones que se han indicado antes, la ecuación de Schrödinger requiere el «$\partial/\partial t$» de primer orden y no el «$(\partial/\partial t)^2$» de segundo orden), podemos esperar de todas formas que la ecuación de segundo orden $(i\hbar)^2\Box\psi = \mu^2\psi$ (donde $(i\hbar)^2\Box$ se obtiene a partir de μ^2 por el truco de cuantización, y la μ en la ecuación *es* realmente la masa en reposo) debería tener significado como ecuación de onda para una partícula relativista. Esta ecuación puede reescribirse como

$$(\Box + M^2)\psi = 0,$$

donde $M = \mu/\hbar$, que sí tiene significado en la teoría cuántica relativista. Hoy día esta ecuación se suele conocer como la «ecuación de Klein-Gordon», aunque parece que fue el propio Schrödinger el primero en proponer esta ecuación relativísticamente invariante, lo que hizo incluso antes de que estableciera su ahora más famosa «ecuación de Schrödinger» (como se ha descrito en §21.3).[6]

En el contexto de la moderna teoría cuántica de campos puede utilizarse realmente la ecuación de Klein-Gordon, si se interpreta de la forma adecuada, para describir partículas masivas sin espín, en particular aquellas partículas conocidas como *mesones* (partículas de masa intermedia, tales como los piones o los kaones). Pero esta interpretación requiere la herramienta completa de la teoría cuántica de campos, que estaba solo en una forma embrionaria cuando Dirac sugirió por primera vez su ecuación de aspecto muy diferente para el electrón en 1928. Dirac había argumentado a favor de una ecuación en la que la derivada temporal $\partial/\partial t$ aparece en una forma de primer orden (como aparece en la ecuación de Schrödinger) más que en la forma de segundo orden $(\partial/\partial t)^2$ que tiene en el operador de onda $\Box$. Sus razones estaban relacionadas con las antes indicadas, pero más concretamente él partía del requisito de que la función de onda de una partícula debería proporcionar una expresión para una densidad de probabilidad de encontrar la partícula en cualquier lugar escogido, cualitativamente similar a la $\bar{\Psi}\Psi$ de la mecánica cuántica no relativista estándar (§21.9), que debería ser definida positiva de modo que esta probabilidad nunca pueda hacerse negativa. Esto no es exactamente lo mismo que el requisito de que la energía sea definida positiva, pero es un requisito complementario de importancia esencialmente similar.[7]

24.6. La raíz cuadrada de Clifford-Dirac de un operador de ondas

Mediante una ingeniosa y soberbiamente intuitiva resolución del conflicto aparentemente irresoluble entre las demandas de la relatividad y su necesidad de una $\partial/\partial t$ de primer orden, Dirac se las arregló para encontrar una ecuación que *es* de primer orden en $\partial/\partial t$. Lo hizo tomando explícitamente la raíz cuadrada del operador de onda $\Box$ de una forma que es relativísticamente invariante de un modo sutil. Para ello introdujo ciertas cantidades adicionales no conmutativas. Tales cantidades son legítimas en mecánica cuántica porque deben tratarse como operadores lineales que actúan sobre la función de onda, al modo de los operadores no conmutativos de posición y momento que hemos encontrado originalmente en §21.2. Como veremos enseguida, lo realmente notable es que estos operadores no conmutativos, que el propio Dirac se vio obligado a introducir, describen en realidad los grados de libertad del *espín* físico de los fermiones más fundamentales de la naturaleza (véase §23.7), a saber, los electrones y los protones que se conocían en la época de Dirac, y los neutrones, muones, quarks y muchas otras partículas de espín 1/2 que se conocen hoy día.

De hecho, al encontrar sus cantidades de «espín» no conmutativas, Dirac redescubrió (un ejemplo de) las *álgebras de Clifford* que hemos visto en §11.5. Parece que no conocía el trabajo anterior de William Kingdon Clifford, ni el hecho de que Clifford (1877), e incluso Hamilton antes que él, ya habían advertido que pueden utilizarse elementos de dichas álgebras para «tomar la raíz cuadrada» de laplacianos, siendo el operador de onda $\Box$ un tipo particular de laplaciano en el que la dimensión es 4 y la signatura $+---$. De hecho, como sabía el propio Clifford, William Rowan Hamilton ya había demostrado alrededor de

1840 que podía obtenerse una raíz cuadrada del laplaciano tridimensional ordinario mediante el uso de *cuaterniones*:[8]

$$\left(\mathbf{i}\frac{\partial}{\partial x}+\mathbf{j}\frac{\partial}{\partial y}+\mathbf{k}\frac{\partial}{\partial z}\right)^2=-\left(\frac{\partial}{\partial x}\right)^2-\left(\frac{\partial}{\partial y}\right)^2-\left(\frac{\partial}{\partial z}\right)^2=-\nabla^2$$

(véase §11.1). El procedimiento de Clifford generalizaba esto a dimensiones más altas.[9] No es sorprendente que Dirac desconociera los descubrimientos que había hecho Clifford más de medio siglo antes porque este trabajo no era ni mucho menos bien conocido en los años veinte, ni siquiera para muchos especialistas en álgebra. Incluso si Dirac hubiera conocido las álgebras de Clifford, esto no habría oscurecido el brillo de su comprensión de que tales entidades son importantes para la mecánica cuántica de un electrón giratorio, lo que constituye un avance fundamental e inesperado en el conocimiento físico.

En el caso de Dirac, lo que se necesita es tomar la raíz cuadrada del operador de onda, siendo este el laplaciano 4-dimensional (lorentziano) de relevancia para la geometría de Minkowski:

$$\Box=\left(\frac{\partial}{\partial t}\right)^2-\nabla^2.$$

Utilizamos así elementos $\gamma_0, \ldots, \gamma_3$ de un álgebra de Clifford «lorentziana», que satisfacen

$$\gamma_0^2=1,\ \gamma_1^2=-1,\ \gamma_2^2=-1,\ \gamma_3^2=-1.$$

En un álgebra de Clifford estándar (signatura $+ + \ldots +$), cada uno de estos cuadrados sería -1. Aquí sigo el que parece ser el convenio estándar entre los físicos con respecto a los signos, en el que las γs espaciales retienen los originales cuadrados negativos de Clifford.[10] Sin embargo, la γ_0 temporal tiene cuadrado positivo. Es en este sentido en el que el álgebra de Clifford de Dirac es «lorentziana». Para γs diferentes, sigue siendo válida la anticonmutación de Clifford (§11.5):

$$\gamma_i\gamma_j=-\gamma_j\gamma_i\ \ (i\neq j).$$

El hecho clave del que hizo uso Dirac es que el operador de onda es el cuadrado de un operador de primer orden definido con la ayuda de estos elementos de Clifford[24.5]

[24.5] Compruébelo.

$$\Box=(\gamma_0\partial/\partial t-\gamma_1\partial/\partial x-\gamma_2\partial/\partial y-\gamma_3\partial/\partial z)^2.$$

Podemos escribir esto de forma más concisa en notación vectorial, con $\boldsymbol{\gamma}=(\gamma_1,\gamma_2,\gamma_3)$, como

$$\Box=(\gamma_0\partial/\partial t-\boldsymbol{\gamma}\cdot\boldsymbol{\nabla})^2,$$

o de forma aún más concisa, como

$$\Box=\not{\partial}^2,$$

donde la cantidad

$$\begin{aligned}\not{\partial}&=\gamma_0\partial/\partial t-\boldsymbol{\gamma}\cdot\boldsymbol{\nabla}\\&=\gamma^a\partial/\partial x_a\end{aligned}$$

(con $\gamma^a=g^{ab}\gamma_b$) se denomina *operador de Dirac*. Esta práctica notación «barrada» fue introducida por Richard Feynman; con más generalidad, un vector A^a podría representarse por el elemento del álgebra de Clifford

$$\not{A}=\gamma_aA^a.$$

24.7. La ecuación de Dirac

Volvamos ahora a nuestra «ecuación de onda» $(\Box+M^2)\psi=0$. Utilizando el operador de Dirac $\not{\partial}$, podemos ahora factorizar la cantidad $\Box+M^2$ que aparece en esta ecuación:

$$\begin{aligned}\Box+M^2&=\not{\partial}^2+M^2,\\&=(\not{\partial}-\mathrm{i}M)(\not{\partial}+\mathrm{i}M),\end{aligned}$$

donde $M=\mu/\hbar$. La *ecuación de Dirac* para el electrón es entonces $(\not{\partial}+\mathrm{i}M)\psi=0$, i.e.,

$$\not{\partial}\psi=-\mathrm{i}M\psi,$$

o, reintroduciendo $\hbar$ y escribiéndola en términos de la masa en reposo μ,

$$\hbar\not{\partial}\psi=-\mathrm{i}\mu\psi.$$

Es evidente de la factorización anterior que cuando quiera que sea válida esta ecuación también debe serlo la ecuación de onda $(\Box + M^2)\psi = 0$. (Esto se aplicaría también a la «ecuación anti-Dirac» $(\not\partial - iM)\psi = 0$, pero con convenios estándar, que se referiría a una partícula con masa negativa $-\hbar M$.) Así pues, las funciones de onda que satisfacen la ecuación de Dirac anterior deben satisfacer también la «ecuación de onda» que gobierna las partículas relativistas de masa en reposo $\hbar M$.

La ecuación de Dirac tiene la ventaja sobre la ecuación de onda de que es de primer orden en $\partial/\partial t$. De hecho, la ecuación de Dirac puede reescribirse de la forma de una ecuación de Schrödinger[24.6]

$$i\hbar\frac{\partial\psi}{\partial t} = (i\hbar\boldsymbol{\gamma}_0\boldsymbol{\gamma}\cdot\boldsymbol{\nabla} + \boldsymbol{\gamma}_0\mu)\boldsymbol{\psi},$$

donde $-i\hbar\boldsymbol{\gamma}_0\boldsymbol{\gamma}\cdot\boldsymbol{\nabla} + \boldsymbol{\gamma}_0\mu$ desempeña el papel de un operador hamiltoniano. La discriminación del operador $\partial/\partial t$ no es, por supuesto, relativísticamente invariante, pero la ecuación de Dirac completa $\partial\psi = -iM\psi$ *sí* es relativísticamente invariante. (Para ver esto se necesita un examen cuidadoso del intercambio entre los elementos del álgebra de Clifford y las transformaciones de Lorentz.)[24.7] Para los físicos de la época supuso un choque considerable el aprender que existen entidades relativísticamente invariantes que yacen fuera del marco estándar del cálculo vector-tensorial (capítulos 12 y 14). Lo que Dirac había iniciado efectivamente era un potente y nuevo formalismo, hoy día conocido como *cálculo espinorial*,[11] un cálculo que va más allá de lo que había sido el cálculo vector-tensorial convencional.

El «precio» que al parecer tenemos que pagar por esta notable eliminación de la incómoda raíz cuadrada, aunque reteniendo la invariancia relativista, es la aparición de estos extraños elementos no conmutativos γ_a del álgebra de Clifford. ¿Qué es lo que significan? Bien, tenemos que pensar en estos objetos como operadores que actúan sobre la función de onda. Puesto que estos operadores concretos son objetos nuevos, que no surgen directamente de las variables cuánticas de

[24.6] Demuéstrelo.
[24.7] Explíquelo. *Sugerencia*: ¡Puede servirle de ayuda la ecuación que se ha dado en el ejercicio [22.18]!

posición y momento (no conmutativas) para una partícula que habíamos estado considerando antes, deben referirse a (y actuar sobre) ciertos nuevos grados de libertad de nuestra partícula. Debemos preguntar para qué propósitos físicos pueden servir estos nuevos grados de libertad. Con la perspectiva que nos ofrece nuestra terminología actual, vemos la respuesta en el propio nombre «espinor»: los nuevos grados de libertad describen el *espín* del electrón.[12] Recordemos lo que se ha dicho en §11.5: «Un espinor puede considerarse como un objeto sobre el que actúan como operadores los elementos del álgebra de Clifford». En la ecuación de Dirac, los elementos de Clifford actúan sobre la función de onda ψ. Así pues, la propia ψ debe ser un espinor. Tiene grados de libertad extra (de una naturaleza que examinaremos de inmediato), además de la mera dependencia de posición y tiempo de una función de onda escalar ordinaria, ¡y estos grados de libertad extra describen efectivamente el espín del electrón!

¡Ahora empezamos a ver que el precio que hemos tenido que pagar por poder factorizar el operador de onda mediante el uso de elementos de Clifford ha sido una ganga increíble! No solo nos da una teoría que describe con exactitud el espín del electrón, sino que cuando añadimos al hamiltoniano el término estándar que da la interacción con un campo electromagnético de fondo —un término que introduce la electrodinámica precisamente de acuerdo con las «recetas gauge»[13] de §19.4 y §21.9—, observamos que el electrón de Dirac responde correctamente al campo electromagnético, como debería hacerlo el electrón cargado, incluidos algunos términos sutiles que aparecen del movimiento relativista del electrón.

Pero no solo es el comportamiento del electrón como partícula cargada el que se describe correctamente; además, el electrón de Dirac responde como una partícula en posesión de un *momento magnético* de una cantidad muy concreta, a saber,

$$\hbar^2 e/4\mu c,$$

donde $-e$ es la carga del electrón y μ es su masa. Esto quiere decir que el electrón de Dirac no solo está eléctricamente cargado, sino que también se comporta como un pequeño imán cuya intensidad viene dada por el valor anterior. Resulta notable que este valor preciso de

Dirac para el momento magnético del electrón está muy próximo al valor realmente observado, con un error de una parte entre mil. La mejor determinación moderna del momento magnético del electrón difiere del valor original de Dirac, que se ha dado amtes, en el factor multiplicativo

1,001 159 652 118 8 ...

Incluso esta pequeña discrepancia se explica ahora, hasta la *precisión explícita* anterior, a partir de efectos correctores que provienen de la electrodinámica cuántica, uno de cuyos ingredientes fundamentales es la ecuación de Dirac. ¡El acuerdo con la naturaleza que se revela en la pequeña y sutil ecuación de Dirac $\partial\!\!\!/\psi = -iM\psi$ es realmente extraordinario!

24.8. La ruta de Dirac al positrón

Pero no hemos acabado ni mucho menos con esta historia; simplemente he descrito sus más desnudos inicios. Prosigamos observando una aparente anomalía en las matemáticas de la ecuación de Dirac con respecto al espín del electrón. Esta aparente anomalía tiene que ver con el número de componentes independientes que van a encontrarse en el espinor de Dirac ψ. Resulta que la ψ de Dirac tiene *cuatro* componentes independientes, mientras que en una visión superficial cabría esperar *dos*, porque una partícula de espín 1/2 tiene solo dos estados de espín independientes (véase §22.8). Tratemos de entender este problema con más detalle.

En 1925, menos de tres años antes de que Dirac encontrara su ecuación (en 1928), George Uhlenbeck y Samuel Goudsmit habían llegado a la conclusión de que el electrón debe poseer un espín mecanocuántico, construido a partir de dos estados de espín básicos. En 1927, Wolfgang Pauli mostró cómo se podían representar las transformaciones de estos estados de espín bajo rotación mediante el uso de lo que ahora llamamos las «matrices de Pauli» (véanse §22.8 y la imagen de la esfera de Riemann de los estados para espín 1/2 descrita en la Fig. 22.10). Las matrices de Pauli (que son básicamente cuaterniones, con

un factor i) son también elementos de un álgebra de Clifford, aunque para el grupo de rotación 3-dimensional.[24.8]

De hecho, hay una fuerte necesidad física para los dos estados de espín del electrón. La propia disciplina de la química, tal como la conocemos, depende de ello. En un átomo, los electrones que rodean el núcleo están obligados a describir órbitas en ciertos estados concretos conocidos como «orbitales» (véase §22.13). Por el principio de exclusión de Pauli, parecería que cada orbital no puede estar ocupado por más de *un* electrón, y pese a todo encontramos que siempre está permitido un segundo electrón en cada uno de los orbitales. Los dos electrones pueden coexistir y seguir satisfaciendo el principio de exclusión porque sus estados *no* son idénticos, sino que tienen espines opuestos. Pero no puede haber más de dos electrones en cualquier orbital porque solo hay dos estados de espín independientes para el electrón. La noción química de «enlace covalente» depende del mismo fenómeno, pues los dos electrones compartidos coexisten en el mismo estado porque sus espines son opuestos; véase la Fig. 24.2.

La descripción de Pauli del electrón es una entidad de dos componentes $\psi_A = (\psi_0, \psi_1)$, correspondiente al hecho de que las matrices de Pauli son 2×2. Pero encontramos que los elementos de Clifford de Dirac $(\boldsymbol{\gamma}_0, \boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2, \boldsymbol{\gamma}_3)$ requieren matrices 4×4 para representar las leyes

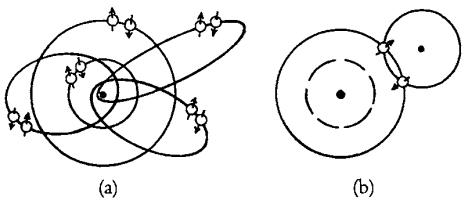


Fig. 24.2. Evidencia del espín 1/2 del electrón. (a) En un átomo, dos electrones, pero ninguno más, pueden ocupar el mismo orbital. Esto se consigue si sus estados de espín son opuestos, de modo que no se viola el principio de exclusión de Pauli. (b) El «enlace covalente» de la química implica un par de electrones de espín opuesto que comparten orbitales de dos átomos separados.

[24.8] Explique este comentario en relación con la conexión entre cuaterniones y elementos de Clifford que se ha explicado en §11.5.

de multiplicación de Clifford.[24.9] Así pues, el electrón de Dirac es una entidad de 4-componentes, en lugar de tener solo las 2 componentes de un «espinor de Pauli» que describe los 2 estados de espín independientes que posee una partícula no relativista de espín 1/2, tal como se ha descrito en §22.8.

De hecho, existen solo 2 componentes de espín para una partícula descrita por la ecuación de Dirac, a pesar de que hay 4 componentes para la función de onda. *Matemáticamente*, la razón para esto está estrechamente relacionada con el hecho de que la ecuación de Dirac $\partial\!\!\!/\psi = iM\psi$ es una ecuación de primer orden, y su espacio de soluciones está generado por solo la mitad de las soluciones que en el caso de la ecuación de ondas de segundo orden $(\Box + M^2)\psi = 0$. (Esta ecuación es también satisfecha por la ecuación «anti-Dirac» $\partial\!\!\!/\psi = +iM\psi$, que es la ecuación de Dirac para la masa en reposo negativa $-M$.) *Físicamente*, este «recuento»[14] de soluciones de la ecuación de Dirac debe tener en cuenta el hecho de que los grados de libertad de la antipartícula del electrón, a saber, el positrón, están también ocultos en las soluciones de la ecuación de Dirac. No obstante, sería erróneo considerar que dos de las componentes de la ecuación de Dirac se refieren al electrón y las otras dos al positrón (compárese con §25.2). Las cosas son mucho más sutiles que esto, como enseguida veremos.

Recordemos que una de nuestras tareas principales, que nos lleva a la consideración de la ecuación de Dirac, ha consistido en ver lo que había que hacer con las soluciones indeseadas de frecuencia negativa (i.e., energía negativa) de la ecuación de Schrödinger. Pero resulta que las soluciones de la ecuación de Dirac *no* están restringidas a ser de frecuencia positiva, a pesar de toda nuestra (o más bien de Dirac) astucia y trabajo duro para eliminar la raíz cuadrada en el hamiltoniano. Como sucede con los intentos previos que se han descrito antes, la presencia de interacciones, tales como un campo electromagnético de fondo, hará que una onda inicialmente de frecuencia positiva recoja partes de frecuencia negativa.

Pero el ingenio de Dirac no iba a retroceder en esta fase. Cuando

[24.9] Demuestre por qué las matrices 2×2 no pueden satisfacer todas las condiciones; encuentre un conjunto de matrices 4×4 que sí lo hagan.

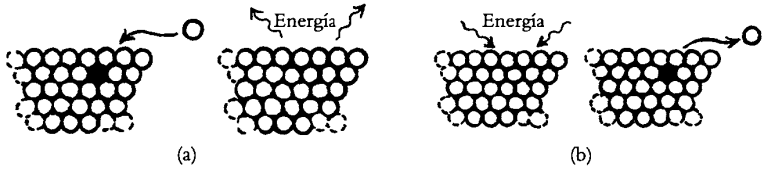


Fig. 24.3. Los positrones como «agujeros» en el «mar de Dirac» de estados electrónicos de energía negativa. Dirac propuso que casi todos los estados de energía negativa del electrón están llenos, y el principio de Pauli impide que un electrón caiga en uno de estos estados. El estado ocasionalmente desocupado —un «agujero» en este mar de energía negativa— aparecería como un antielectrón (positrón), teniendo así energía positiva. (a) Un electrón que cae en tal agujero sería interpretado como la aniquilación del electrón y el positrón, con la liberación de energía, la suma de las contribuciones positivas del electrón y el positrón. (b) A la inversa, el suministro de energía suficiente al mar de Dirac podría producir un par electrón-positrón. (Las partículas son solo esquemáticas, y la estructura reticular mostrada no tiene ninguna relevancia real para el mar de Dirac.)

finalmente se convenció de que las soluciones de frecuencia negativa no podían ser eliminadas matemáticamente, razonó como sigue. ¿Cuál es, después de todo, el peligro de las soluciones de frecuencia negativa? El problema estaría en que si existieran estados de energía negativa, entonces un electrón podría caer en uno de esos estados con la emisión de energía; y si hubiera un número ilimitado de tales estados, entonces habría una inestabilidad catastrófica, pues todos los electrones se hundirían en estados de energía negativa cada vez más bajos, con la emisión de cada vez más energía, sin límite alguno. Pero, razonó Dirac, los electrones satisfacen el principio de Pauli, y no está permitido que una partícula semejante ocupe un estado si dicho estado ya está ocupado. Por eso hizo la sorprendente sugerencia de que ¡todos los estados de energía negativa deberían estar ya ocupados! Este océano de estados ocupados de energía negativa se conoce ahora como el «mar de Dirac». Así pues, de acuerdo con la «idea loca» de Dirac, imaginamos que en realidad los estados de energía negativa ya están llenos y, por el principio de Pauli, no hay lugar para que un electrón caiga en tal estado.

Pero, como Dirac razonó posteriormente, ocasionalmente podría haber algunos estados de energía negativa que estén desocupados. ¿Qué sucedería entonces? Un «agujero» semejante en el mar de Dirac de es-

tados de energía negativa aparecería precisamente como una partícula de energía positiva (y, así, una partícula de masa positiva), cuya carga sería la opuesta a la carga del electrón. Tal estado vacío de energía negativa podría ahora ser ocupado por un electrón ordinario, de modo que el electrón podría «caer» en dicho estado con emisión de energía (normalmente en forma de radiación electromagnética, i.e., fotones). Esto daría como resultado que el «agujero» y el electrón se aniquilarían mutuamente a la manera en que ahora consideramos que una partícula y su antipartícula sufren una aniquilación mutua (Fig. 24.3a). Recíprocamente, si inicialmente no hubiera presente un agujero, pero en el sistema entrara una cantidad de energía suficiente (digamos en forma de un fotón), entonces un electrón podría ser expulsado de uno de los estados de energía negativa para dejar un agujero (Fig. 24.3b). El «agujero» de Dirac es la antipartícula del electrón, ahora conocida como *positrón*.

En un primer momento, Dirac no se atrevió a afirmar que esta teoría predecía realmente la existencia de antipartículas de los electrones; inicialmente (en 1929) pensó que los «agujeros» podrían ser protones, que en aquella época eran las únicas partículas masivas conocidas con una carga positiva. Pero no pasó mucho tiempo antes de que quedase claro[15] que la masa de cada agujero tenía que ser igual a la masa del electrón, y no a la masa de un protón, que es unas 1.836 veces mayor. En 1931, Dirac llegó a la conclusión de que los agujeros deben ser «antielectrones» —partículas previamente desconocidas que ahora llamamos positrones—. Un año después de la predicción teórica de Dirac, Carl Anderson anunció el descubrimiento de una partícula que tenía realmente las propiedades que Dirac había predicho: ¡se había encontrado la primera antipartícula!

Notas

Sección 24.3

24.1. Técnicamente, la catástrofe se evita si la energía está lo que se denomina «acotada inferiormente», lo que significa que es mayor que cierto valor dado E_0 que podría ser negativo. En tal circunstancia podemos

«renormalizar» la energía sumando $-E_0$ al hamiltoniano, y los autovalores de la energía resultantes serán ahora todos positivos.

24.2. Hay una sutileza en el tratamiento del punto ∞, pues es probable que f sea singular allí. El tratamiento hiperfuncional de §9.7 es apropiado; véase Bailey *et al.* (1982).

24.3. Estrictamente, deberíamos decir semidefinida positiva, puesto que el espectro (continuo) de autovalores llega hasta, e incluye, el cero.

24.4. Véanse Shankar (1994) para las aplicaciones a la mecánica cuántica, y Arfken y Weber (2000) para una discusión general.

Sección 24.5

24.5. Algunos definen este operador con el signo contrario, normalmente porque adoptan la signatura $+\,+\,+\,-$ y no la $+\,-\,-\,-$ que he utilizado aquí.

24.6. Véanse Pais (1986), Miller (2003) y Dirac (1983).

24.7. Estos dos requisitos se combinan como ingredientes esenciales de la demostración del teorema CPT, al que llegaremos en §25.4.

Sección 24.6

24.8. Véase Trautman (1997) para una referencia sobre estas ideas acerca de la «raíz cuadrada».

24.9. Véanse Clifford (1882), pp. 778-815 y también Lounesto (2001), para un tratamiento general.

24.10. Este convenio parece estar en contra del convenio habitual del matemático (véanse Harvey, 1990; Budinich y Trautman, 1988; Lounesto, 2001; Lawson y Michelson, 1990) y también con el mío propio (véase Penrose y Rindler, 1986, apéndice). Si se adopta la signatura $+\,-\,-\,-$ para el espaciotiempo, como aquí, la ecuación definitoria para un álgebra de Clifford general es $\boldsymbol{\gamma}_i\boldsymbol{\gamma}_j - \boldsymbol{\gamma}_j\boldsymbol{\gamma}_i = -2g_{ij}$.

Sección 24.7

24.11. Véanse Clifford (1981), Cartan (1966), Van der Waerden (1929), Infeld y Van der Waerden (1933), Laporte y Uhlenbeck (1931), Penrose (1960), Penrose y Rindler (1984, 1986) y O'Donnell (2003). En la notación 2-espinorial de §22.8 esto lleva a la forma «zigzag» de la ecuación de Dirac que se da en §25.2.

24.12. Al parecer, el término «espinor» fue introducido por Paul Ehrenfest en una carta a Bartel van der Waerden.

24.13. El término añadido es $ie\not{A}$, donde $\not{A} = g^{ab}A_a\gamma_b$ y A_a es el potencial electromagnético, lo que equivale a reemplazar el operador $\not{\partial}$ por $\not{\partial} - ie\not{A}$.[24.10]

Sección 24.8

24.14. El recuento de soluciones, para ecuaciones relativistas, se realiza más fácilmente mediante el método de «conjuntos exactos» en el cálculo de 2-espinores (véase Penrose y Rindler, 1984).

24.15. Este trabajo fue realizado por Igor Tamm, Hermann Weyl y J. Robert Oppenheimer; véase Oppenheimer (1930) para un ejemplo del razonamiento que hay detrás de ello. Hay algunas sutilezas en el camino hacia el positrón que nos llevarían bastante lejos; véase Zee (2003) para un tratamiento completo, riguroso y encantador de todo esto.

[24.10] Explique por qué esta es la «prescripción gauge» estándar.

25

El modelo estándar de la física de partículas

25.1. Los orígenes de la moderna física de partículas

La ecuación de Dirac para el electrón supuso un momento crucial para la física en muchos aspectos. En 1928, cuando Dirac propuso su ecuación, las únicas partículas conocidas para la ciencia eran los electrones, los protones y los fotones. Las ecuaciones de Maxwell libres describen el fotón —como fue previsto efectivamente por Einstein en 1905— en un primer trabajo cuyas ideas fueron desarrolladas por Einstein, Bose y otros, hasta que en 1927 Jordan y Pauli proporcionaron un esquema matemático global para describir los fotones libres de acuerdo con la teoría de Maxwell para el campo libre cuantizado. Además, tanto el protón como el electrón parecían estar muy bien descritos por la ecuación de Dirac. La interacción electromagnética, que describe la influencia de los fotones sobre los electrones y los protones, estaba excelentemente tratada mediante la receta de Dirac, a saber, mediante la idea *gauge* (tal como fue introducida básicamente por Weyl en 1918; véase §19.4), y el propio Dirac ya había empezado a construir en 1927 una formulación de una teoría completa de los electrones (o protones) en interacción con los fotones (i.e., la electrodinámica cuántica).[1] Así pues, todas las herramientas básicas parecían estar más o menos a punto para la descripción de todas las partículas conocidas de la naturaleza, junto con sus más manifiestas interacciones.

Pese a todo, la mayoría de los físicos de la época no estaban tan locos como para pensar que esto pudiera llevar pronto a una «teoría de todo». En efecto, eran conscientes de que ni las fuerzas necesarias para mantener unidos a los núcleos —las que ahora llamamos fuerzas *fuer-*

tes— ni los mecanismos responsables de la desintegración radiactiva —ahora llamados fuerzas *débiles*— podrían acomodarse sin importantes avances adicionales. Si los electrones y los protones estilo Dirac, que solo interaccionaban electromagnéticamente, fueran los únicos ingredientes de los átomos, incluyendo sus núcleos, entonces todos los núcleos ordinarios (excepto el simple protón que constituye el núcleo del hidrógeno) se desintegrarían al instante debido a la repulsión electrostática de las cargas positivas predominantes. ¡Tenía que estar actuando alguna otra cosa desconocida, algo que explicara una fuerte influencia atractiva dentro del núcleo! En 1932, Chadwick descubrió el neutrón, y se comprendió por fin que el modelo protón/electrón para el núcleo, que había sido popular hasta entonces, debía ser reemplazado por otro en el que los protones y los neutrones estarían presentes, y donde una fuerte interacción protón-neutrón mantendría el núcleo unido. Pero esta fuerza fuerte no era lo único que faltaba en el conocimiento de la época. La radiactividad del uranio se conocía desde las observaciones de Henri Becquerel en 1896, y se presentaba como el resultado de otra interacción —la fuerza *débil*— diferente de la fuerte y de la electromagnética. Incluso un neutrón, si se dejara libre, sufriría una desintegración radiactiva en un período de unos quince minutos. Uno de los misteriosos productos de la radiactividad era el evasivo neutrino, propuesto como hipótesis provisional por Pauli en 1929, aunque no fue observado directamente hasta 1956. Fue el estudio de la radiactividad el que finalmente llevaría a los físicos a alcanzar una influencia y notoriedad desacostumbradas, hacia el final de la Segunda Guerra Mundial, y en el período que le siguió...

Las cosas han avanzado mucho desde los inicios de la comprensión de la física de partículas, tal como estaban en el primer tercio del siglo XX. Ahora que nos embarcamos en el siglo XXI, disponemos de una imagen mucho más completa conocida como el *modelo estándar* de la física de partículas. Este modelo parece acomodar casi todo el comportamiento observado concerniente al vasto conjunto de partículas que hoy se conocen. Al fotón, el electrón, el protón, el positrón y el neutrón se han añadido el muón, los diversos neutrinos, los piones (predichos por Yukawa en 1934), los kaones, las lambdas, las sigmas y la celebradamente predicha partícula omega menos. El antiprotón fue directa-

mente observado en 1955, y el antineutrón, en 1956. Hay nuevos tipos de entidades conocidas como quarks, gluones, bosones W y Z; hay vastas hordas de partículas cuya existencia es tan fugaz que no han sido nunca directamente observadas, por lo que se suelen conocer como «resonancias». El formalismo de la teoría moderna exige asimismo entidades transitorias llamadas partículas «virtuales», y también cantidades conocidas como «fantasmas» que están aún más alejadas de la observabilidad directa. Existe un número desmesurado de partículas propuestas —aún no observadas— que son predichas por ciertos modelos teóricos, aunque no son en absoluto consecuencias del armazón general de la física de partículas aceptada: a saber, «bosones X», «axiones», «fotinos», «squarks», «gluinos», «monopolos magnéticos», «dilatones», etc. Está también la misteriosa partícula de Higgs —todavía no observada en el momento de escribir esto—, cuya existencia, de una u otra forma, es esencial para la física de partículas actual, donde la partícula de Higgs se considera responsable de las masas de todas las partículas.

25.2. La imagen zigzag del electrón

En este capítulo presentaré una breve guía del modelo estándar de la física de partículas de hoy, aunque mi aproximación particular a ella podría juzgarse un poco «no estándar» en ciertos temas. Empecemos de una forma ligeramente no estándar, reexaminando la ecuación de Dirac en términos de la «notación 2-espinorial» que se ha expuesto en §22.8. Como se ha comentado en §24.8, la descripción de «espinor de Pauli» para una partícula de espín 1/2 es una cantidad ψ_A con 2-componentes. (Las componentes son ψ_0 y ψ_1.) Cuando consideramos la relatividad, de acuerdo con §22.8, necesitamos también cantidades con índices *primados* A', B', C', ..., que son el resultado de la conjugación compleja aplicada a índices no primados. Resulta[2] que el espinor de Dirac ψ, descrito antes, con sus 4 componentes, puede representarse como un par de 2-espinores[3] α_A y $\beta_{A'}$, uno con índice no primado y otro con índice primado:

$$\psi = (\alpha_A, \beta_{A'}).$$

La ecuación de Dirac puede escribirse entonces como una ecuación que acopla estos dos 2-espinores, cada uno de los cuales actúa como un tipo de «fuente» para el otro, con una «constante de acoplamiento» $2^{-1/2}M$ que describe la intensidad de la «interacción» entre los dos:

$$\nabla^{A}_{B'}\alpha_{A} = 2^{-1/2}M\beta_{B'},\ \ \nabla^{B'}_{A}\beta_{B'} = 2^{-1/2}M\alpha_{A'}.$$

Los operadores $\nabla^{A}_{B'}$ y $\nabla^{B'}_{A}$ son solo traducciones en 2-espinores del operador gradiente ordinario ∇. No nos preocuparemos de todos estos índices, los $2^{-1/2}$ y la forma exacta de estas ecuaciones. Los presento aquí solo para mostrar cómo puede introducirse la ecuación de Dirac en el armazón general del cálculo 2-espinorial, y que cuando se hace esto se revelan algunas ideas nuevas respecto a la naturaleza de la ecuación de Dirac.[4]

A partir de la forma de estas ecuaciones, vemos que el electrón de Dirac puede considerarse compuesto de dos ingredientes α_A y $\beta_{B'}$. Es posible obtener una especie de interpretación física de estos ingredientes. Formamos una imagen en la que existen dos «partículas», una descrita por α_A y la otra por $\beta_{A'}$, ambas sin masa,[25.1] y donde cada una de ellas se está convirtiendo continuamente en la otra. Llamémoslas partícula «zig» y partícula «zag»: α_A describe a la zig y $\beta_{A'}$ describe a la zag. Al no tener masa, cada una de ellas debería viajar a la velocidad de la luz, pero podemos considerarlas más bien haciendo un «vaivén» hacia atrás y hacia delante, donde el movimiento hacia delante de la zig se está transformando continuamente en el movimiento hacia atrás de la zag, y viceversa. De hecho, esta es una realización del fenómeno conocido como *zitterbewegung*. Según este, la medida del movimiento instantáneo del electrón es siempre la velocidad de la luz, debido al movimiento de vaivén del electrón, incluso si el movimiento promediado global del electrón da una velocidad menor que la de la luz.[5] Cada ingrediente tiene un *espín* de magnitud $\frac{1}{2}\hbar$ alrededor de su dirección de

[25.1] Con referencia a la ecuación del neutrino de Weyl, dada en §25.3, explique por qué es razonable adoptar el punto de vista de que cada uno de los α_A y $\beta_{A'}$ describen partículas sin masa, acopladas por una interacción que convierte a cada una de ellas en la otra.

movimiento, espín que es levógiro en el caso de la zig y dextrógiro en el de la zag. (Esto tiene que ver con el hecho de que la α_A de la zig tiene un índice no primado, que está asociado con la helicidad negativa, mientras que la $\beta_{B'}$ de la zag tiene un índice primado, que indica helicidad positiva. Todo esto tiene relevancia para la exposición de §§33.6-8, pero por el momento no es oportuno entrar en detalles.) Notemos que aunque la velocidad sigue invirtiéndose, la dirección del espín permanece constante en el sistema en reposo del electrón (Fig. 25.1). En esta interpretación, la partícula zig actúa como fuente para la partícula zag y la partícula zag como fuente para la partícula zig, y la intensidad del acoplamiento está determinada por M.

En la Fig. 25.2 he dado una ilustración diagramática de cómo va a contribuir este proceso al «propagador de Feynman» completo (véase §26.7), al modo de los diagramas de Feynman[6] a los que llegaremos

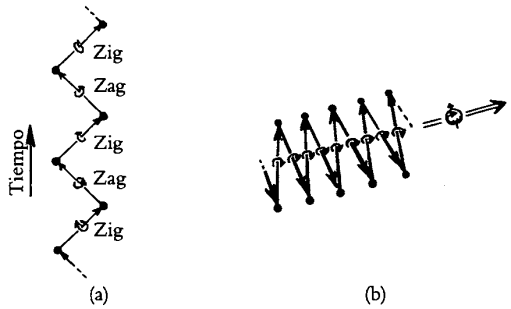


Fig. 25.1. Imagen zigzag del electrón. (a) El electrón (u otra partícula masiva del espín $\frac{1}{2}$) puede verse en el espaciotiempo oscilando entre una partícula zig sin masa levógira (helicidad $-\frac{1}{2}$, descrita por el 2-espinor no primado α_A o, en la notación más habitual del físico, por la parte proyectada por $\frac{1}{2}(1-\gamma_5)$) y una partícula zag sin masa dextrógira (helicidad $+\frac{1}{2}$, descrita por el 2-espinor primado $\beta_{B'}$, la parte proyectada por $\frac{1}{2}(1+\gamma_5)$). Cada una es fuente de la otra, con la masa en reposo como constante de acoplamiento. (b) Desde una perspectiva 3-espacial, en el «sistema de reposo» del electrón hay una inversión continua de la velocidad (siempre la velocidad de la luz), pero la dirección del espín permanece constante. (Por claridad, la figura no está totalmente dibujada en el sistema en reposo del electrón, y el electrón se desplaza lentamente hacia la derecha.)

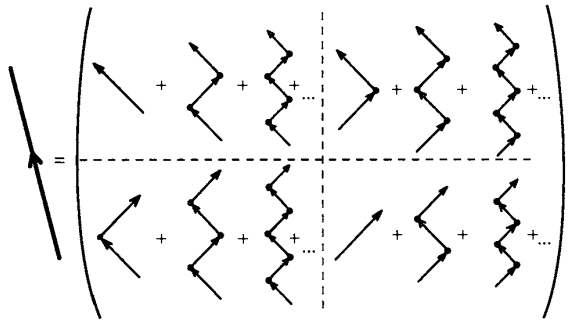


Fig. 25.2. Cada proceso zigzag contribuye por separado, como parte de una superposición cuántica infinita, al «propagador» total a la manera de un diagrama de Feynman. El propagador convencional de una sola línea de Feynman se dibuja a la izquierda y representa la matriz entera de sumas infinitas de zigzags finitos, dibujados a la derecha.

con más detalle en el próximo capítulo. Cada proceso zigzag constituyente es de longitud finita, pero la totalidad de ellos, al incluir zigzags de longitud siempre creciente, contribuyen a la propagación entera del electrón, de acuerdo con la matriz 2×2 que se muestra en la Fig. 25.2. Típicamente, una partícula zig se convierte en una zag, y la zag se convierte luego en una zig, que después se convierte de nuevo en una zag, y así sucesivamente para un tramo finito. En el proceso total, encontramos que el ritmo promedio al que esto sucede está relacionado (inversamente) con el parámetro de acoplamiento M; de hecho, este ritmo es esencialmente la *frecuencia de De Broglie* del electrón (véase §21.4).

Sin embargo, debo hacer una advertencia acerca de cómo vamos a interpretar los diagramas de Feynman. Podemos considerar legítimamente que el proceso se representa como una descripción espaciotemporal de lo que está sucediendo; pero, en el nivel cuántico, debemos adoptar el punto de vista de que, incluso para una única partícula, existen muchísimos procesos de este tipo que ocurren simultáneamente. Cada uno de estos procesos individuales debe verse como parte de una gigantesca superposición cuántica de muchísimos procesos diferentes. El estado cuántico real del sistema consiste en la superposición entera.

Un diagrama de Feynman individual representa meramente una componente del mismo.

Por consiguiente, mi descripción anterior del movimiento del electrón consistente en este vaivén hacia atrás y hacia delante, donde una zig se está transformando continuamente en una zag y vuelta de nuevo, debe tomarse apropiadamente en este sentido. El movimiento real está compuesto de un número enorme de tales procesos individuales (de hecho, infinitos de ellos) todos superpuestos, y podemos considerar que el movimiento percibido del electrón es un tipo de «promedio» de aquellos. Incluso esto describe tan solo al electrón libre. Un electrón real estará experimentando continuamente interacciones con otras partículas (tales como los fotones, los cuantos del campo electromagnético). Todos estos procesos de interacción deberían incluirse también en la superposición global.

Teniendo esto en cuenta, planteemos la cuestión de si estas partículas zig y zag son o no «reales». ¿O quizá son solo artificios del formalismo matemático concreto que he estado adoptando aquí para la descripción de la ecuación de Dirac para el electrón? Esto plantea una pregunta más general: ¿qué justificación física hay para permitir que uno se vea arrastrado por la elegancia de cierta descripción matemática y luego trate de considerar que dicha descripción corresponde a una «realidad»? En el caso presente deberíamos empezar cuestionando la importancia (y la elegancia) del propio formalismo 2-espinorial, solo como técnica *matemática*. Debería advertir al lector que en realidad no es este el formalismo más utilizado por los físicos que se interesan por la ecuación de Dirac y sus implicaciones, como sucede en la electrodinámica cuántica (QED), que es la teoría con más éxito de las teorías cuánticas de campos.[7]

La mayoría de los físicos utilizan lo que se denomina el formalismo del «espinor de Dirac» (o 4-espinor), en el que se evitan los índices espinoriales. En lugar del 2-espinor «α_A», utilizan el 4-espinor «$(1 - i\gamma_5)\psi$» (llamado «la parte de helicidad levógira del electrón de Dirac», o algo similar, en lugar de utilizar mi «partícula zig»).[8] Aquí, la cantidad γ_5 es el producto

$$\gamma_5 = -i\gamma_0\gamma_1\gamma_2\gamma_3,$$

y tiene la propiedad de que anticonmuta con todo elemento del álgebra de Clifford, y $(\gamma_5)^2 = 1$.[25.2] Análogamente, utilizan $(1 + \gamma_5)\psi$ en lugar de β_A, (la parte de helicidad dextrógira). Se podría decir que esto es solo una cuestión de notación, y de hecho es posible hacer una traducción entre los formalismos 2-espinorial y 4-espinorial. La imagen «zigzag» que he presentado aquí es ciertamente una descripción válida (aunque no del todo habitual) en cualquiera de los dos formalismos, pero está más directamente sugerida por los 2-espinores que por los 4-espinores.

De modo que ¿son reales estas zigs y zags? Por mi parte, diría que sí; son tan reales como lo es el propio «electrón de Dirac» —como descripción matemática idealizada y muy adecuada de uno de los ingredientes más fundamentales del universo—. Pero ¿es esta «realidad» *real*? En §§1.3,4 he abordado esta cuestión general de la realidad matemática y física, y la relación entre ambas. Al final del libro, en §34.6, retomaré esta cuestión.

25.3. Interacciones electrodébiles, asimetría de reflexión

Cada una de las partículas zig y zag tiene la misma carga eléctrica (como debe ser, puesto que la carga se conserva y cada partícula se transforma continuamente en la otra). En la imagen de diagramas de Feynman, la interacción con el campo electromagnético que experimentan las partículas cargadas está representada por el añadido de una línea que representa un fotón. Esto se muestra en la Fig. 25.3a, de acuerdo con los procedimientos convencionales en los que la trayectoria del electrón se representa como una línea única 4-espinorial de Dirac al modo usual, y en la Fig. 25.3b utilizando esta (poco convencional) descripción «zigzag».

Nótese que tanto la levógira (zig) como la dextrógira (zag) intervienen igualmente en la interacción electromagnética. Resulta, no obstante, que hay otra interacción física —la interacción *débil*— según la cual las cosas son completamente *des*iguales, en el sentido de que solo

[25.2] Demuestre ambas cosas.

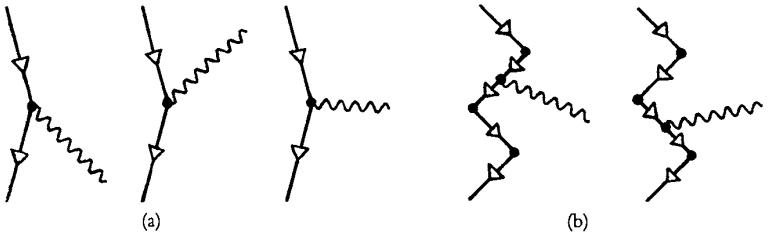


Fig. 25.3. (a) Diagrama de Feynman (en forma convencional, sin zigzags) de un electrón en interacción con un cuanto de campo electromagnético, o fotón. Mientras que en la figura de la izquierda podemos ver el proceso como la absorción de un fotón, la figura central como la emisión de uno y la figura derecha como una influencia electrostática, estos procesos deben considerarse como si fueran el mismo, y se conocen como una interacción con un fotón «virtual» (fuera de la capa). (b) Lo mismo, con los zigzags representados. El fotón (virtual) interacciona igualmente con la zag y la zig. En todas estas figuras, la propagación de carga eléctrica está representada por la flecha blanca en forma de triángulo. Todas las figuras aquí ilustradas apuntan al pasado, porque estamos considerando electrones, que están cargados negativamente.

la zig del electrón toma parte en estas interacciones débiles, y la zag no participa en absoluto (véase la Fig. 25.4). Las interacciones débiles están mediadas por análogos del fotón, llamados bosones W y Z. Como se ha comentado antes, estas interacciones son responsables de la desintegración radiactiva, por la que, por ejemplo, un núcleo de uranio U^{238} se desintegrará espontáneamente, en aproximadamente 5×10^9 años, en promedio, en un núcleo de torio y un núcleo de helio (una partícula α), o por la que un neutrón libre se desintegrará en un protón, un electrón y un antineutrino en un tiempo medio de aproximadamente quince minutos; véase la Fig. 25.5. Estos procesos se conocen como «desintegración β», en cuyo contexto el electrón suele recibir el nombre de «partícula β» (por razones históricas).

Durante muchos años, las interacciones débiles se habían tratado como simples procesos puntuales —ilustrados como el punto simple de desintegración en la Fig. 25.5— según un esquema que se remonta a 1933, que había sido propuesto por el destacado físico italiano Enrico Fermi. Pero más adelante esto tropezó con dificultades teóricas, que fueron finalmente resueltas en la teoría electrodébil de Weinberg, Salam, Ward y Glashow, de la que veremos algo en §25.5. Como parte de

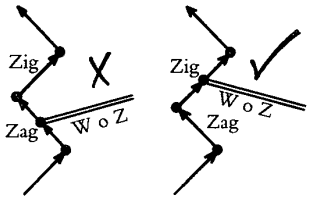


Fig. 25.4. En el caso de interacciones débiles, por el contrario, solo la zig de una partícula débilmente interactuante interacciona con el bosón W o el Z. (Pero para lo que se clasifica como una «antipartícula», sería la zag la que interacciona débilmente.)

estas ideas más nuevas, se advirtió que, en lugar de la interacción puntual de Fermi, habría un «bosón gauge» intermedio —las partículas W y Z recién mencionadas— que mediaría el proceso de interacción débil, según el cual la desintegración β de la Fig. 25.5 se interpreta como se muestra en la Fig. 25.6. Pero ¿cuál es la importancia de la asimetría zig/zag? En 1956, los físicos sufrieron una gran conmoción cuando Tsung Dao Lee y Chen Ning Yang hicieron una sorprendente propuesta[9] acerca de la desintegración β —y sobre las interacciones débiles en general— según la cual estas no deberían ser invariantes por reflexión, propuesta que fue espectacularmente confirmada experimentalmente por Chien-Shiung Wu y sus colaboradores poco después, en enero de 1957. Según esto, la *reflexión especular* de un proceso de interacción débil *no* sería en general un proceso de interacción débil permitido, de modo que las interacciones débiles muestran *quiralidad*. En particular, el experimento de Wu examinó la pauta de emisión de electrones del cobalto 60 radiactivo, encontrando una relación con evidente asimetría especular entre la distribución de electrones emitidos y

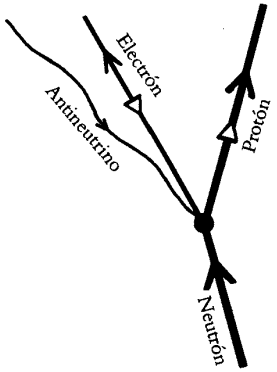


Fig. 25.5. La desintegración β de un neutrón en un protón, un electrón y un antineutrino, que tarda aproximadamente 15 minutos (en promedio) en el caso de un neutrón libre. La flecha invertida en el antineutrino indica que es una «antipartícula» en el esquema de clasificaciones de leptones. Como en la Fig. 25.4, la flecha blanca en las líneas del electrón y el protón indica carga eléctrica.

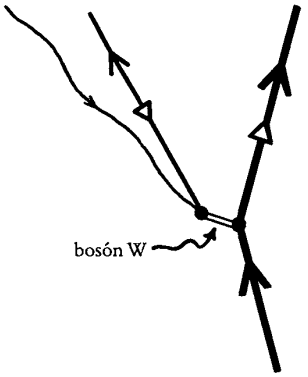


Fig. 25.6. Las interacciones débiles no son «puntuales», como sugeriría la Fig. 25.5 (teoría de Fermi original), sino que ocurren a través de la intermediación de un «bosón vectorial» ($W^{\pm}$ o Z), aquí una partícula W.

las direcciones del espín de los núcleos de cobalto (véase la Fig. 25.7). Esto era sorprendente porque ¡nunca antes se había observado un fenómeno asimétrico por reflexión en un proceso físico básico!

En términos de nuestras zigs y zags, la asimetría quiral surge del hecho de que, en un espejo, una zig se ve como una zag y una zag como una zig. Recordemos que la zig tiene una helicidad levógira mientras que la zag es dextrógira. Cada una de estas se transforma en la otra bajo reflexión especular. (En la terminología más convencional, γ_5 cambia de signo bajo reflexión, de modo que se intercambian los papeles de las partes de helicidad izquierda y derecha de la función de onda del electrón $(1 - \gamma_5)\psi$ y $(1 + \gamma_5)\psi$.) Así, la no invariancia de las interacciones débiles bajo simetría de reflexión se manifiesta en el hecho de que solo la parte zig del electrón interviene en interacciones débiles. Lo mismo puede decirse del neutrón cuando sufre desintegración β espontánea,

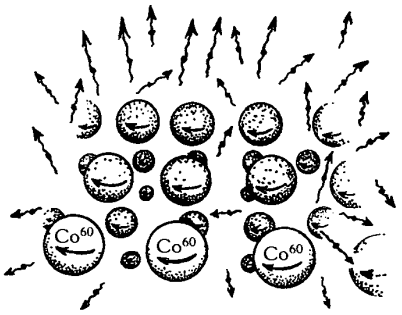


Fig. 25.7. El experimento de Wu examinaba la pauta de emisión de electrones del cobalto 60 radiactivo, encontrando una relación con evidente asimetría especular entre la distribución de electrones emitidos y las direcciones del espín de los núcleos de cobalto. Aquí salen más electrones hacia la parte superior de la figura que hacia la inferior.

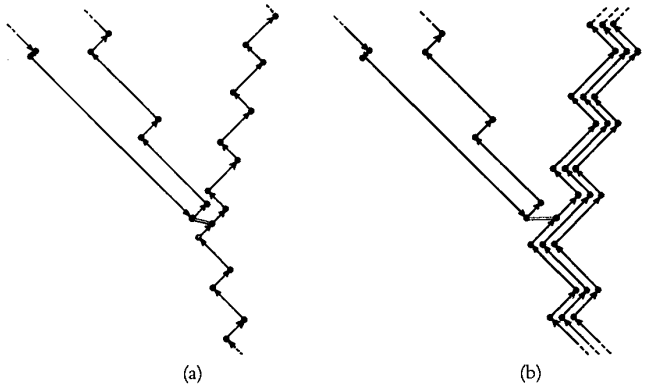


Fig. 25.8. El proceso de desintegración β de la Fig. 25.5 expresado en términos de zigzags. (a) Un neutrón y un protón pueden describirse con un buen grado de aproximación, como una partícula de Dirac, de modo que los zigzags no son inadecuados. Como en la Fig. 25.4, solo las partes zig del neutrón y el protón intervienen en un proceso de desintegración débil, aunque con el antineutrino zag (levógiro), permitiéndose una masa pequeña por la presencia de la minúscula zig en la parte superior izquierda. (b) Sin embargo, el neutrón y el protón se consideran compuestos, formado cada uno de ellos por tres quarks, que se consideran individualmente partículas de Dirac, de modo que la imagen zigzag es apropiada para ellos. (No se muestran las flechas de carga ni los gluones que conectan los quarks.)

y también del protón resultante. Un neutrón y un protón también pueden describirse, con un alto grado de aproximación, mediante la ecuación de Dirac, por lo que la descripción zigzag se hace apropiada para cada uno de ellos. Una vez más, son solo las partes zig del neutrón y del protón las implicadas en los procesos de desintegración débil, y esto se ilustra en la Fig. 25.8a. Más apropiado —de acuerdo con la imagen moderna— es considerar el neutrón y el protón como partículas compuestas, cada una de ellas hecha de tres quarks. Los propios quarks se consideran descritos individualmente por la ecuación de Dirac, de modo que la imagen zigzag se hace apropiada también para cada uno de ellos; en la Fig. 25.8b se representa la desintegración β del neutrón en estos términos.

También el neutrino es objeto de especial interés a este respecto. Al menos con muy buena aproximación puede tratarse como una partí-

cula *sin masa*. (Su masa es, en cualquier caso, extraordinariamente minúscula en relación con la masa de un electrón, y ciertamente no más que 6×10^{-6} de la masa del electrón.) Si ponemos $M = 0$ en la versión 2-espinorial de la ecuación de Dirac, las ecuaciones se *desacoplan* para dar

$$\nabla^{A}_{B'}\alpha_{A} = 0, \quad \nabla^{B'}_{A}\beta_{B'} = 0.$$

Cada una podría existir en ausencia de la otra (y cualquiera de estas, por sí misma, se conoce como la «ecuación de Weyl»[10] para el neutrino). Pero solo la versión zig (dada por la α_{A} *no primada*, sujeta a $\nabla^{A}_{B'}\alpha_{A} = 0$) interviene en interacciones débiles, o podría crearse en un proceso de interacción débil. Así pues, los neutrinos son partículas con helicidad levógira.

¿Poseen masa realmente los neutrinos? Parece que ahora existe una buena evidencia experimental de que al menos dos de los tres tipos de neutrinos deben ser masivos. Estos tres tipos son el «neutrino electrónico» ν_{e} (que es el único implicado en la desintegración β ordinaria, siendo su antipartícula $\bar{\nu}_{e}$ la que se emite en la desintegración del neutrón; véase la Fig. 25.5), el «neutrino muónico» ν_{μ} y el «neutrino tauónico» ν_{τ}. Observaciones en el detector japonés Superkamiokande indican claramente que las diferencias entre las masas de estos tres tipos de neutrinos, aunque pequeñas (alrededor de 10^{-7} de la masa de un electrón), no pueden ser nulas, debido al hecho de que tienen una tendencia a saltar de uno a otro («oscilaciones de neutrinos»), algo que no puede darse con masa nula. Supongo que todavía es posible que ν_{e} (o concebiblemente alguna «combinación lineal» cuántica adecuada de los tres tipos de ellos) pudiera tener masa nula, pero todavía no hay evidencia definitiva sobre estas cuestiones. Un neutrino sin masa podría ser totalmente zig, pero con una masa pequeña la imagen se parecería más a la mostrada en la Fig. 25.9a, donde la zig saltaría ocasional y momentáneamente a la zag y volvería de nuevo. Sin embargo, visto con respecto a un sistema de referencia que se mueve con el neutrino, los aspectos zig y zag parecerían contribuir por igual a su movimiento global (Fig. 25.9b).

Aquí son necesarias algunas palabras de clarificación. Cuando antes he comentado que son las partículas zig (i.e., levógiras) las que in-

tervienen en interacciones débiles, y no las partículas zag, he dado por supuesto que sabemos cómo distinguir una «partícula» de una «antipartícula». Con la antipartícula las cosas son al revés. En el caso de la antipartícula del electrón, el positrón, podemos presentar de nuevo una descripción «zigzag» en la que la zig es levógira y la zag dextrógira, pero la zig del positrón es la antipartícula de la zag del electrón, y viceversa. Así pues, en el caso de un positrón es la *zag* dextrógira (la antipartícula de la zig del electrón) la que interviene en interacciones débiles, y no la zig. Un comentario similar sería aplicable al antiprotón y al antineutrón y, de hecho, al antiquark. También sería aplicable al antineutrino que, si no tuviera masa, sería enteramente zag.

Ahora bien, esto podría causar cierta confusión porque no he dado ningún criterio para decidir si una entidad tipo partícula (con espín 1/2) va a considerarse como una «partícula» o como una «antipartícula» para que podamos así saber si es su zig o su zag la que interviene en interacciones débiles. Aunque en el capítulo anterior he presentado la noción de una antipartícula solo en términos de la idea original de Dirac de un «agujero» en el «mar de estados de energía negativa», una antipartícula no debería considerarse realmente como un tipo de entidad completamente independiente de una partícula. En el contexto de la moderna teoría cuántica de campos no es necesario presentar las cosas al modo original de Dirac (aparentemente asimétrico). Las antipartículas son tan «partículas» como las partículas de las que son las «antis». Más aún, la noción de antipartículas se refiere a los bosones (partículas de espín entero) tanto como a los fermiones, mientras que solo los fermiones están sujetos al principio de Pauli (véanse §§23.7,8), de modo

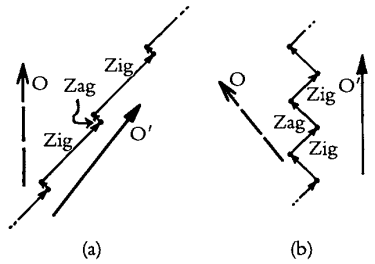


Fig. 25.9. (a) Un neutrino sin masa podría ser totalmente zig; pero con una masa pequeña debemos imaginar el ocasional «salto» momentáneo a una zag y vuelta. La imagen se muestra desde la perspectiva del sistema en reposo del laboratorio O. (b) Visto con respecto a un segundo sistema O′, que se mueve con el neutrino, los aspectos zig y zag parecen contribuir igualmente al movimiento global.

que la perspectiva del «mar de Dirac» para las antipartículas no puede aplicarse a los bosones. El pión cargado positivamente (el mesón π^+), por ejemplo, que es un bosón, tiene una antipartícula que es el pión cargado negativamente (mesón π^-). De hecho, varios bosones son sus propias antipartículas. Un fotón es un ejemplo de esto; también lo es el pión neutro (mesón π^0). Hasta donde se sabe (y según la teoría estándar), cada partícula de la naturaleza tiene una antipartícula.

25.4. Conjugación de carga, paridad e inversión temporal

La operación que reemplaza cada partícula por su antipartícula se conoce como C (que significa *conjugación de carga*). Una interacción física que es invariante bajo el reemplazamiento de partículas por sus antipartículas (y viceversa) se denomina C invariante. La operación de reflexión espacial (reflexión en un espejo) se conoce como P (que viene de «paridad»). De acuerdo con la exposición anterior en §25.3, las interacciones débiles ordinarias no son invariantes bajo P ni bajo C por separado, pero resulta que son invariantes bajo la operación combinada CP(= PC). Podemos considerar que CP es la operación realizada por un espejo inusual en el que cada partícula se refleja como su antipartícula. Vemos que la CP envía el zig de una partícula al zag de su antipartícula, y viceversa. Hay otra operación que se discute habitualmente en relación con estas, que es la de *inversión temporal*, conocida como T. Una interacción es invariante bajo T si no se ve alterada si la vemos desde la perspectiva de una dirección temporal que es la inversa de la normal. Existe un famoso teorema en teoría cuántica de campos, conocido como el *teorema CPT*, que afirma que toda interacción física es invariante si se le aplican las tres operaciones C, P y T. Por supuesto, un teorema es «solo un fragmento de matemáticas», de modo que su validez física depende de la validez física de sus hipótesis. Esta cuestión adquirirá importancia más adelante (§30.2), cuando plantee una cuestión crucial que quizá nos lleve a cuestionar las conclusiones —y, por consiguiente, las hipótesis— del teorema CPT. Sin embargo, no hay razón para esperar ninguna dificultad de este tipo en relación con las interacciones débiles ordinarias. Por consiguiente, la invariancia CP de las in-

teracciones débiles ordinarias también implica su invariancia bajo T (simetría de inversión temporal).

Habría que comentar que en el momento de escribir esto existe exactamente un proceso físico (un proceso débil «no ordinario», observado por primera vez por Fitch y Cronin en 1964) que se sabe que es no invariante bajo CP. Es también no invariante bajo T (pero, hasta donde puede decirse, invariante bajo CPT, de acuerdo con el teorema CPT). Se trata de la desintegración del mesón K^0 (que puede hacerse en 2 piones o en 3 piones, y en el que se plantea una cuestión sofisticada concerniente a que K^0 salta a su antipartícula $\bar{K}^0$, teniendo lugar una oscilación entre ambas).

El teorema CPT nos ofrece una perspectiva alternativa sobre las antipartículas que es diferente de utilizar el «mar de Dirac» y es más satisfactoria porque también puede aplicarse a bosones. Suponiendo CPT, podemos considerar C —el intercambio de partículas con sus antipartículas— como equivalente a PT, de modo que podemos considerar la antipartícula de cierta partícula como la «reflexión espaciotemporal» (PT) de dicha partícula. Ignorando en esto el aspecto de reflexión espacial, obtenemos la interpretación de una antipartícula como la partícula que viaja *hacia atrás en el tiempo*. De hecho, así es como a Richard Feynman le gustaba interpretar las antipartículas. Ofrece una forma muy conveniente y consistente de tratar antipartículas dentro del contexto de los diagramas de Feynman. (La idea le había sido sugerida a Feynman por John A. Wheeler y, había sido propuesta antes, independientemente, por Stückelberg en 1942.) A su modo diferente, ¡es una idea tan «loca» como lo era el mar de Dirac!

En un diagrama de Feynman, las partículas que no son sus propias antipartículas tienen que tener líneas en los diagramas que están dirigidas en cierto sentido, que puede indicarse uniendo a cada línea un tipo adecuado de flecha. Podríamos considerar que esta flecha apunta hacia el futuro —cuando la línea representa a la propia partícula—, pero, en este caso, cuando apunta hacia el pasado obtenemos la antipartícula de la partícula. Esta perspectiva sobre las antipartículas tiene la gran ventaja de que pone de manifiesto que muchos procesos de partículas con aspecto muy diferente son básicamente el mismo proceso, aunque visto desde «ángulos» diferentes en el espaciotiempo. Como

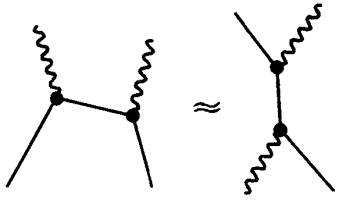


Fig. 25.10. Cruce de simetría. Procesos que difieren solo con respecto a las ordenaciones temporales en varios lugares, pero sin que se vea afectada la tipología del diagrama, son matemáticamente equivalentes (mediante prolongación analítica, §7.4). Esto está ilustrado por una equivalencia semejante entre la aniquilación del par partícula-antipartícula, como se ilustra a la izquierda, y el proceso de dispersión de Compton a la derecha (dibujado sin zigzag).

ejemplo, en la Fig. 25.10 (aunque sin preocuparnos por los zigzags), he representado la aniquilación electrón-positrón en un par de fotones, mostrando que este es «esencialmente el mismo» proceso (i.e., reorganizado en el espaciotiempo) que la dispersión Compton de un electrón por un fotón. (Enseguida veremos que también tenemos que admitir que las líneas de las partículas puedan apuntar en direcciones de género espacio, describiendo lo que se conoce como una «partícula virtual», ¡pero esto ya es suficiente confusión por el momento!)

Volvamos ahora al problema de decidir, para una entidad de espín 1/2, si es su zig o su zag la que toma parte en las interacciones débiles. Necesitamos una regla clara que nos diga si va a contar como una «partícula» o como una «antipartícula». La regla que se utiliza decreta que aquellas partículas conocidas como «leptones» (electrones, muones, tauones y sus correspondientes neutrinos ν_e, ν_μ y ν_τ), y también los quarks que componen los protones, neutrones (y otros hadrones), deben contar como «partículas». Estas tienen zigs que experimentan interacciones débiles. Las «antis» de todas ellas cuentan como «antipartículas», y en esos casos son las zags las que experimentan interacciones débiles. La situación se complica por el hecho de que existen también entidades (con masa) de espín 1 implicadas en interacciones débiles,[11] a saber, los bosones W y Z. Estos son los mediadores de las interacciones débiles, y desempeñan papeles similares a los fotones que median en las interacciones electromagnéticas (pues los fotones son los cuantos del campo electromagnético). Tales partículas se denominan a veces «cuantos gauge», por razones a las que pronto llegaremos. Existen *dos*

bosones W diferentes, etiquetados W^+ y W^- (antipartículas uno del otro), que tienen cargas eléctricas respectivas 1 y −1 (en unidades dadas por la carga del positrón), mientras que hay solo una Z^0 descargada (su propia antipartícula). Cada una de estas toma parte en interacciones débiles, con una línea en el diagrama de Feynman que se une en cualquiera de los extremos a una parte zig de un leptón o quark o a una parte zag de un antileptón o antiquark (véase la Fig. 25.11). A lo largo de cada proceso de interacción débil, la carga eléctrica se conserva, y también lo hace el número leptónico. De hecho, existen tres tipos diferentes de número leptónico (números electrónico, muónico y tauónico), cada uno de los cuales se conserva por separado en el modelo estándar de interacciones débiles, contando como cero los números leptónicos de las W y la Z^0. Para comprobar estas cuatro leyes de conservación, en un diagrama de Feynman, todo lo que tenemos que hacer es asegurar que cada uno de los cuatro tipos de flechas en las líneas sigue un camino continuo y consistentemente orientado a través del diagrama.

25.5. El grupo de simetría electrodébil

Todo esto suena sin duda algo complicado para una teoría fundamental. Sí, *es* complicado, aunque existe una pauta subyacente que todavía no he explicado; sin embargo, apenas he empezado a describir, y en términos muy cualitativos, lo que es nuestra comprensión actual de la física de partículas de acuerdo con menos de la mitad de lo que se denomina el «modelo estándar». Más aún, mis comentarios han tenido hasta ahora un carácter más bien «botánico» concerniente a las diferentes partículas involucradas en las interacciones débiles (y electromagnéticas). De hecho, en el modelo estándar las interacciones débil y electromagnética se consideran unificadas en la que se denomina teoría *electrodébil*, donde existe una simetría especial que relaciona W^+, W^-, Z^0 y el *fotón* γ, según el grupo $SU(2) \times U(1)$, o más correctamente[12] $U(2)$. (Véase §13.9, si es necesario recordar qué son estos grupos.) Es esta simetría (oculta) la que suministra la pauta subyacente antes mencionada.

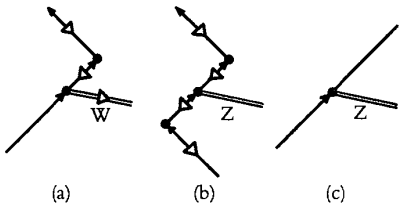


Fig. 25.11. Ilustraciones de interacciones entre una partícula zig y un bosón gauge en interacción débil. (a) Las W^+ y W^- cargadas (antipartículas una de otra) inducen un cambio en la carga eléctrica de la zig (para asegurar la conservación de la carga eléctrica), mientras que (b) la Z^0 no cargada no lo hace (y Z^0 es su propia antipartícula). (c) El zig del neutrino puede interaccionar con la Z^0 no cargada.

Más adelante en este mismo capítulo explicaré el papel de esta simetría de forma algo más completa. Esta simetría también interrelaciona las partes zig de varios leptones y quarks. La idea tiene como consecuencia que, desde una perspectiva más primitiva, las W^+, W^-, Z^0 y γ pueden, en cierto sentido, ser «rotadas una en otra» continuamente, de modo que diversos conjuntos de *combinaciones lineales* (cuánticas) de dichas partículas ¡están en pie de igualdad con las propias partículas individuales!

Tal como se ha descrito, esta «simetría» parece muy extraña y sutil, debido especialmente a que el electromagnetismo puro es invariante por reflexión, al estar involucradas por igual las partes zig y zag de las fuentes, mientras que las interacciones débiles son lo más *no* invariantes por reflexión que pueden serlo, al involucrar solo las partes zig de las partículas. Más aún, el fotón parece estar claramente discriminado entre todos los bosones de la teoría por ser una partícula *sin masa*. De hecho, la masa del fotón, si no es nula, tendría que ser menor que 10^{-20} de la masa de un electrón por buenas razones observacionales, y por lo tanto es menor que aproximadamente 5×10^{-26} de la masa medida de los bosones W y Z. Además, los bosones W están cargados eléctricamente, mientras que, por el contrario, el fotón no lleva una carga débil.

En la Fig. 25.12 he dado la lista de todos los posibles vértices de Feynman con 3 puntas que incluyen solo bosones gauge (i.e., W^+, W^-, Z^0 o γ). Hay solo dos de ellos. Pero el hecho de que existan siquiera es una expresión de una *no linealidad* en el campo gauge libre, que apare-

ce porque el grupo gauge es no abeliano, y esto se aplica a U(2). (La pura electrodinámica procede del grupo gauge abeliano U(1) y, en consecuencia, no hay un análogo diagrama de Feynman de 3 puntas que incluya solo fotones. Estos habrían dado lugar a no linealidades en el campo de Maxwell libre de fuentes. Un enunciado análogo se aplica a vértices con n puntas, con $n > 2$.) Véanse §15.8 y §19.2. Por la naturaleza limitada del conjunto de diagramas en la Fig. 25.12, parecería que no puede haber simetría completa entre todos los bosones gauge.

¿Cómo reconciliar estas patentes desviaciones de la simetría con el objetivo requerido de una teoría simétrica unificada? Lo primero que hay que entender es que en realidad hay más simetría oculta en los diagramas de Feynman que la que se manifiesta a primera vista, y de hecho muestran simetría U(2) si se consideran de la forma apropiada. En primer lugar, consideremos los dos diagramas de la Fig. 25.12. Para hacernos una idea mejor de la simetría aquí subyacente, pensemos en una matriz hermítica 2×2 (véase §13.9). Podemos imaginar que sus dos elementos diagonales *reales* son análogos a Z^0 y γ, y que los dos elementos restantes fuera de la diagonal —*complejos conjugados* uno de otro— son análogos a W^+ y W^-. El carácter de números reales de los elementos diagonales corresponde a que Z y γ son iguales que sus antipartículas respectivas (líneas sin flechas en la Fig. 25.12), mientras que el carácter de complejos conjugados de los elementos fuera de la diagonal corresponde al hecho de que W^+ y W^- son antipartículas una de otra (inversión de la dirección de la flecha al pasar de una a otra). Una transformación general U(2) de esta matriz hermítica (que debemos tener en cuenta que incluye tanto premultiplicación por la matriz U(2) como posmultiplicación por la inversa de dicha matriz) «remueve» los elementos de esta matriz hermítica, de formas muy concretas, pero su carácter hermítico se conserva siempre. De hecho, esta analogía está

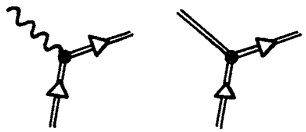


Fig. 25.12. Vértice con 3 partículas bosones gauge electrodébiles que puede ocurrir teóricamente, debido a la naturaleza no abeliana del grupo gauge. (Las líneas de los fotones son onduladas; las líneas de la Z^0 son dobles y sin flecha; las líneas de las $W^\pm$ son dobles y con flecha blanca.)

muy próxima a la forma en que actúa U(1) en la teoría electrodébil (la única complicación es que debemos admitir una combinación lineal de los elementos diagonales con la traza, en esta identificación, relacionada con el «ángulo de Weinberg» al que llegaremos en §25.7). La asimetría que parecemos ver en el mundo real con respecto a estas partículas ocurre en la teoría electrodébil simplemente porque la naturaleza escoge ciertas combinaciones particulares —i.e., superposiciones cuánticas particulares de estos elementos— para realizarlas como partículas libres reales.

Pero ¿qué pasa con la otra asimetría aparentemente más manifiesta en nuestros diagramas de Feynman, por la que las Z^0 y $W^\pm$ pueden unirse solo a las líneas zig de las partículas mientras que la γ se une indiscriminadamente a zig o a zag? De nuevo, esto es cuestión de qué superposiciones nos permite encontrar la naturaleza como partículas libres. Por ejemplo, habría una superposición concreta de Z^0 y γ —llamémosla Y— que ve solo la parte *zag* de una partícula. (Aproximadamente: «restar» Z^0 de γ para eliminar la interacción zig, dejando solo la parte zag.) Podríamos recuperar nuestra γ original a partir de Z^0 e Y, pero podría haber muchas otras superposiciones posibles que igualmente podrían haber desempeñado el papel del fotón si la naturaleza hubiera elegido las cosas de otra manera.

Por lo tanto, una cuestión clave es: ¿qué criterios adopta la naturaleza al permitirnos encontrar ciertas superposiciones particulares como partículas libres y no otras? La respuesta básica es que en el caso de una partícula libre necesitamos que sea un autoestado de *masa*, de modo que necesitamos saber qué es lo que determina la masa de las partículas en general. Aquí no podemos esperar simetría completa bajo U(2); en otras palabras, la masa implica algún tipo de *ruptura de simetría*. ¿Cómo se produce esto en el modelo estándar? La idea, al menos tal como se presenta normalmente, es que la asimetría que realmente observamos hoy en las interacciones entre partículas es el resultado de una ruptura *espontánea* de simetría que se supone que ha ocurrido en las primitivas etapas del universo. Antes de dicho período las condiciones eran muy diferentes de las que existen hoy, y la teoría electrodébil estándar afirma que a las temperaturas extraordinariamente altas del universo primitivo la simetría U(2) era *exactamente* válida, de modo que

los W^+, W^-, Z^0 y γ serían completamente equivalentes a muchos otros conjuntos de superposiciones cuánticas de dichas partículas, y donde el fotón γ estaría en pie de igualdad con todo tipo de otras combinaciones que pudieran aparecer de esta manera. Pero según esta teoría, cuando la temperatura del universo se enfrió (hasta por debajo de aproximadamente de 10^{16}K, aproximadamente 10^{-12} segundos después del big bang; véanse §§28.1-3), las W^+, W^-, Z^0 y γ concretas que hoy observamos fueron «congeladas» por este proceso de ruptura espontánea de simetría. Así, de este proceso salieron cuatro partículas reales a partir de la variedad completamente simétrica de posibilidades iniciales. Solo tres de ellas adquirieron masa y se conocen como las W y la Z; la otra sigue siendo sin masa y se denomina fotón. En la versión inicial «pura» y no rota de la teoría, cuando había simetría completa U(2), las W, Z y γ tendrían que ser todas efectivamente carentes de masa. Como un aspecto fundamental de esta propuesta de ruptura de simetría, es necesario que intervenga otra partícula/campo, conocida como la (partícula de) *Higgs*. La partícula (campo) de Higgs se considera la responsable de asignar masa a todas estas partículas (incluyendo la propia partícula de Higgs), así como a los quarks que componen las otras partículas del universo.

¿Cómo lo hace? Los detalles completos de este notable e ingenioso conjunto de ideas deben quedar, por desgracia, fuera del alcance de este libro, aunque daré algunos de sus ingredientes más tarde, en §26.11 y §28.1. Por el momento, creo que puedo describir mejor el papel del campo de Higgs (de forma muy incompleta) volviendo a la descripción «zigzag» del electrón de Dirac, como se ha mostrado en la Fig. 25.2. Recordemos que el electrón se consideraba en términos de una oscilación continua entre una zig levógira (α_A) y una zag dextrógira ($\beta_{B'}$), cada una de las cuales carece en sí misma de masa. Se consideraba que hay una «constante de acoplamiento» $2^{-1/2}M$ que gobierna el ritmo de «salto» entre las partes α_A y $\beta_{B'}$ del espinor de Dirac. El punto de vista «Higgs» consiste en considerar $2^{-1/2}M$ como un *campo* —esencialmente, el campo de Higgs— que entra como una interacción donde previamente teníamos la constante de acoplamiento $2^{-1/2}M$. (Véase la Fig. 25.13). Se supone que uno de los efectos del acto de ruptura espontánea de simetría en el universo muy primitivo es que este campo de

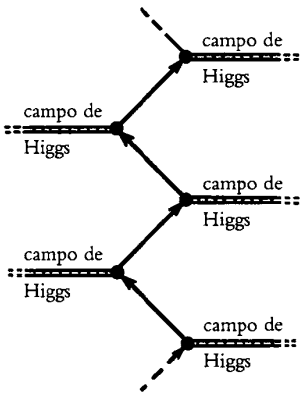


Fig. 25.13. En la imagen zigzag de una partícula de Dirac, los vértices pueden verse como interacciones con el campo de Higgs (constante).

Higgs se asienta para tener un valor constante en todo lugar. Este valor fijaría una escala global para la determinación de las masas de todas las partículas, cuyos diferentes valores están escalados por cierto factor numérico que depende de los detalles de cada partícula concreta.

Pospondré mi valoración de esta extraordinaria colección de ideas hasta §§28.1-3, donde encaja con más propiedad. Pero como quiera que las consideremos, la teoría unificada resultante de fuerzas débiles y electromagnéticas —la teoría electrodébil—[13] ha tenido un éxito notable. Entre sus predicciones estaba la propia existencia de la Z^0 (y también de $W^\pm$, aunque la existencia de $W^\pm$ ya había sido inferida sobre la base de ideas anteriores) y algunos valores bastante concretos para las masas de $W^\pm$ y Z^0 (aproximadamente 80 y 90 GeV, respectivamente).[14] Las $W^\pm$ y Z^0 fueron observadas en experimentos en el CERN (en Ginebra, Suiza) en 1983, y los valores predichos de las masas fueron confirmados con una precisión, pues los valores modernos observados son de aproximadamente 81,4 y 91,2 GeV, respectivamente. También se han confirmado muchas otras predicciones de distinto tipo, y en el momento de escribir esto la teoría electrodébil se halla en excelente forma desde el punto de vista observacional.

25.6. Partículas fuertemente interactuantes

Ahora bien, ¿qué hay de las interacciones *fuertes*? La teoría moderna que las describe constituye la otra «mitad» del modelo estándar, y se conoce como *cromodinámica cuántica* o QCD. Este puede parecer un nombre extraño, puesto que el griego *khroma*, de donde procede el nombre, significa «color», y podemos preguntarnos qué lugar ocupa el «color» en una teoría de interacciones fuertes que gobiernan las fuerzas nucleares. La respuesta es que la noción de «color» a la que aquí nos referimos es enteramente caprichosa y no tiene nada que ver con el concepto ordinario de color, que está relacionado con la frecuencia de la luz visible.[15] Para explicar lo que podría ser la noción de «color» en la física de partículas (nuclear), será conveniente retroceder un poco y considerar el sorprendente conjunto de partículas conocido como *hadrones*, de los que los neutrones y los protones son ejemplos particulares.

El nombre «hadrón» procede del griego *hadros*, que significa «voluminoso». Los hadrones son las más masivas entre las partículas básicas de la naturaleza, y participan en interacciones fuertes (cuya intensidad proporciona una gran contribución de energía a esta masa). La familia de los hadrones incluye a los fermiones conocidos como «bariones» y también a los bosones conocidos como «mesones». Todos los hadrones se consideran compuestos de *quarks*, en la teoría convencional; sobre estos hablaré enseguida. En particular, los hadrones conocidos como *bariones* son los «nucleones» (neutrones o protones) ordinarios y sus primos más pesados, llamados «hyperones» (descubiertos en chaparrones de rayos cósmicos y en aceleradores de partículas). Los *mesones* originales fueron una notable predicción teórica del físico japonés Hideki Yukawa en 1934, basado en su análisis de las fuerzas nucleares, y son los *piones* (mesones-π) los que fueron finalmente encontrados en 1947 por C.F. Powell en trazas de rayos cósmicos. Ahora se conocen también muchos otros primos mesónicos del pión.

El término «barión» procede del griego *barys*, que significa «pesado», en contraste con «leptón», de *leptos*, que significa «pequeño». Los *leptones* son el electrón y sus partículas hermanas, el muón y el tauón, junto con sus correspondientes neutrinos; las antipartículas de estos se conocen como *antileptones*. Tanto los leptones como los bariones son

fermiones de espín 1/2, pero los leptones se distinguen de los bariones por el hecho de que no se implican directamente en interacciones fuertes, que es quizá la «razón» principal por la que los leptones tienden a ser mucho menos masivos que los bariones (aunque el tauón es una excepción, al ser más masivo que el protón o el neutrón).

Desde finales de la década de 1940 se han descubierto muchos hadrones, en rayos cósmicos y en aceleradores: Λ^0, $\Sigma^\pm$, Σ^0, Ξ^-, Ξ^0, Δ^{++}, $\Delta^\pm$, Δ^0, Ω^-, ρ^0, $\rho^\pm$, ω^0, η^0, $K^\pm$, K^0, y numerosas versiones más pesadas de muchas de estas partículas con espín superior (indicado aquí por la adición de asteriscos a los símbolos, por ejemplo, Ξ^{*-}) que se conocen como «recurrencias de Regge» (véase la Fig. 31.6). Esto hubiera sido totalmente desconcertante si no fuera por el hecho de que se observó que se agrupaban en ciertas familias, llamadas *multipletes*. Se obtuvo una buena comprensión (por parte de Murray Gell-Mann y Yuval Ne'eman en 1961) de la naturaleza de estos multipletes sobre la base de que tales multipletes proporcionan representaciones del grupo SU(3) o, más correctamente, $SU(3)/\mathbb{Z}_3$ (véanse §13.6 para la noción de «representación» y §13.2 para la interpretación de la noción de «grupo cociente» que está implicada en el uso de «/»; aquí «$\mathbb{Z}_3$» representa el grupo cíclico con 3 elementos, que surge de modo natural como un subgrupo normal de SU(3);[25.3] véase también §5.5).

La mejor manera de entender lo que está implícito en estas representaciones es hacer la hipótesis (lo que hicieron explícitamente Zweig y Gell-Mann en 1963) de que cada hadrón está construido a partir de ciertas entidades básicas de espín 1/2, que Gell-Mann bautizó como «quarks» (tres tipos) y «antiquarks» (tres tipos). Se supone que cada barión consta de tres de estos quarks, y cada mesón de un quark y un antiquark, donde los tres tipos de quark —conocidos como tres *sabores*— se denominan (de forma bastante poco imaginativa) «up» (arriba), «down» (abajo) y «strange» (extraño). Una característica misteriosa de los quarks es que tienen que poseer carga eléctrica fraccionaria (en unidades de la carga del protón), teniendo los quarks up, down y strange valores de carga 2/3, −1/3 y −1/3, respectivamente.

[25.3] Encuentre el subgrupo normal. *Sugerencia*: Considere el determinante de una matriz 3×3.

Quizá debido fundamentalmente a estos valores aparentemente implausibles de las cargas eléctricas de los quarks —y el hecho relacionado de que nunca se observan quarks aislados (y las partículas observadas tienen siempre valores enteros para sus cargas; véase §5.5)— originalmente los quarks no se consideraron como partículas reales, sino que simplemente se suponía que proporcionaban un «libro de contabilidad» conveniente para las diferentes representaciones de $SU(3)/\mathbb{Z}_3$. No obstante, la contabilidad solo funcionaba si los quarks se trataban como entidades que satisfacían la «estadística equivocada» para partículas de espín 1/2. Es decir, había que fingir que los quarks son «bosones» para que los multipletes saliesen bien, y no los fermiones que parecía exigir el teorema espín-estadística (véanse §23.7 y §26.2).

Para entender este último punto, consideremos dos ejemplos. El más claro de los dos es el que ofrece el decuplete de 10 partículas de espín 3/2 que condujo a la predicción de la partícula Ω^- por parte de Gell-Mann y Ne'eman en 1962 (todas las demás partículas del multiplete eran ya conocidas), predicción que fue confirmada en 1964:[16]

$$\Delta^{++}, \Delta^{+}, \Delta^{0}, \Delta^{-},$$
$$\Sigma^{*+}, \Sigma^{*0}, \Sigma^{*-},$$
$$\Xi^{*0}, \Xi^{*-},$$
$$\Omega^{-}.$$

Este conjunto puede entenderse si consideramos que cada partícula está hecha de tres quarks, de varios sabores, donde d representa «down», u representa «up» y s representa «strange»:[25.4]

uuu, uud, udd, ddd,
uus, uds, dds,
uss, dss,
sss.

Ahora bien, esto funciona solo porque los tres quarks están en un estado simétrico. Por ejemplo, uud no se distingue de udu. Más aún, estados con dos quarks del mismo tipo, tales como uuu y uud, no desaparecen idénticamente, que es lo que harían en el caso de un estado antisimétrico para el que fuera válido el principio de Pauli. El hecho de que el espín sea 3/2 significa que los espines de los tres quarks (cada uno de valor 1/2) están alineados, de modo que hay simetría completa por lo que se refiere al aspecto de espín del estado. Si los quarks se comportaran como fermiones, entonces tendríamos antisimetría, y no simetría, bajo intercambio de los quarks, lo que es incompatible con esta imagen.[25.5]

Un comentario similar (aunque más elaborado) se aplica a la situación más complicada que surge para el octete de 8 partículas de espín 1/2 al que pertenecen el protón ordinario (N^+) y el neutrón ordinario (N^0):[17]

$$\begin{array}{ccccc} & & N^+ & & N^0 \\ & & & \Sigma^0 & \\ \Sigma^+ & & & \Lambda^0 & & \Sigma^- \\ & & \Xi^0 & & \Xi^-. \end{array}$$

Aquí debemos pensar que Σ^0 y Λ^0 ocupan básicamente el «mismo espacio» en el centro de una configuración hexagonal. Esta configuración ocurre cuando consideramos que el espín es ahora 1/2, de modo que podemos pensar que dos de los espines de los quarks son paralelos y uno de ellos es antiparalelo. Resulta que hay solo dos formas linealmente independientes de disponer esto para la composición de quarks uds que está representada en el centro (correspondiente al par Σ^0, Λ^0); no hay ninguna en absoluto para uuu, ddd y sss, lo que explica una configuración hexagonal antes que una triangular; y hay solo una para cada una de las demás.[25.6]

[25.4] Compruebe que los valores de la carga indicados por los superíndices en la primera tabla son correctos.

[25.5] Explique esto en detalle utilizando la descripción 2-espinorial con índices para los espines de los quarks que se ha descrito en §22.8 y utilizando un nuevo «índice SU(3)» 3-dimensional que toma tres valores u, d, s.

[25.6] Vea si puede explicar todo esto en detalle. Es necesario tener cuidado en el tratamiento de los índices de espín 2-espinoriales, si desea utilizarlos. Una antisimetría en un par de ellos permite que el par sea eliminado (como cuando se representa un estado de espín 0 en términos de un par de partículas de espín 1/2, como en §23.4). Pese a todo, hay también una simetría (oculta), porque hay solo dos estados de espín independientes para cada quark.

25.7. «QUARKS COLOREADOS»

¿Cómo podemos tratar los quarks como partículas reales si tienen la relación «espín-estadística» equivocada (véanse §23.7 y §26.2)? La forma de tratar este problema[18] en el modelo estándar consiste en exigir que cada sabor de quark se dé también en tres (denominados) «colores», y que cualquier partícula real, compuesta de quarks, debe ser completamente antisimétrica en el grado de libertad de color. Esta antisimetría pasa por alto los propios estados de quarks, de modo que la antisimetría entre quarks (fermiónicos) individuales se convierte efectivamente en simetría en una partícula de tres-quarks.[25.7] Los colores no se manifiestan nunca en partículas libres, de modo que el color es esencialmente «inobservable». Cualquier partícula libre debe tener un «color neutro». Por ejemplo, no tenemos tres versiones diferentes de la partícula Δ^+, dependiendo de qué color sea el quark-d en «uud». La antisimetría en el grado de libertad de color, para partículas libres reales, lo garantiza.[25.8]

A veces los «colores» se escogen como «rojo», «blanco» y «azul», lo que me choca por ser a la vez confuso (porque no considero el blanco un color) y revelador de cierto tipo de patriotismo equivocado. A veces los «colores» se denominan «rojo», «verde» y «azul», lo cual es mejor, pero puesto que la asociación entre «color de quark» y los receptores de color en el ojo no tiene en ningún caso una justificación científica, utilizaré en su lugar «rojo» (R), «amarillo» (Y) y «azul» (B). Esta elección de terminología tiene la ventaja de que puedo «mezclar» mis colores con más facilidad, pero debemos notar que «naranja», «verde» y «púrpura» (considerando estos como superposiciones cuánticas de los R,Y y B originales) servirían exactamente igual que el conjunto original. De hecho, aquí hay una simetría que va más allá de la simple permutación de los colores. Tiene que haber un completo SU(3) 8-real-dimensional de simetría de color, en el que R,Y y B proporcionan meramente un conjunto de elementos básicos para

[25.7] Utilice índices para explicar este comentario, donde hay un nuevo índice de *color* SU(3) 3-dimensional, además de un índice de sabor 3-dimensional del ejercicio [25.5].
[25.8] Explíquelo.

el espacio vectorial sobre el que actúan las matrices SU(3) (véase §13.9).

En esta etapa, la introducción de tales grados de libertad de «color» aparentemente inobservables podría parecer bastante artificial, puesto que ahora tenemos nueve quarks básicos (junto con sus diversas antipartículas y superposiciones cuánticas):

$$d_R, d_Y, d_B; \quad u_R, u_Y, u_B; \quad s_R, s_Y, s_B;$$

ninguno de los cuales puede ser directamente observado. De hecho, la situación en el modelo estándar es en realidad «el doble de mala» que esta, porque ha habido que introducir tres sabores más, llamados (asimismo de forma poco imaginativa) «charm» (encanto) (c), «bottom» (fondo) (b) y «top» (cima) (t), de modo que tenemos también:

$$c_R, c_Y, c_B; \quad b_R, b_Y, b_B; \quad t_R, t_Y, t_B;$$

lo que da en total dieciocho quarks independientes, ninguno de los cuales es directamente observable.

Si satisfacer la relación espín-estadística fuera el único beneficio de esta proliferación de hipotéticas partículas inobservables, entonces el esquema se vería decididamente artificial. Pero la inobservabilidad completa del color del quark «libre» ¡trae realmente una recompensa generosa! En efecto, esta inobservabilidad y la (íntimamente relacionada) naturaleza totalmente intacta de la simetría de color SU(3) nos proporciona la capacidad de utilizar esta simetría directamente como base para la maravillosa idea de una conexión gauge, como se ha descrito en §§15.1,8. Recordemos que así es como se describe la interacción electromagnética, siendo U(1) el grupo gauge en este caso (véanse §19.4, §21.9 y §24.7). De hecho, la simetría gauge U(1) del electromagnetismo se considera exacta e intacta.[19] Recordemos también que en la misma base de la idea de fibrado, descrita en el capítulo 15, está la presencia de un grupo de simetría exacta que actúa sobre las fibras. El «grupo de color» hadrónico SU(3) de las interacciones fuertes ofrece precisamente una simetría exacta de este tipo, y la analogía con el grupo gauge electromagnético U(1) es muy estrecha. La generalización del electromagnetismo, basada en una conexión gauge para un grupo no abeliano tal como SU(2) o SU(3), se denomina teoría de *Yang-Mills*.[20]

Esta es, de hecho, la base de la QCD (cromodinámica cuántica). Como en el caso del electromagnetismo, podemos utilizar una cantidad similar al potencial electromagnético A_a para modificar la derivada $\partial/\partial x^a$, cuando actúa sobre campos de quarks, y obtener una noción adecuada de «operador derivada covariante» (como el $\partial/\partial x^a - ieA_a$ del electromagnetismo) que nos proporciona una conexión fibrada (véanse §15.8 y §19.4). Puesto que el espacio de color es 3-dimensional, tenemos algo más complicado que lo que teníamos para el caso electromagnético 1-dimensional, y es conveniente introducir índices para trabajar con estos grados de libertad extra. Una diferencia crucial entre el caso electrodinámico y el de las interacciones fuertes es que mientras que U(1) es abeliano (i.e., conmutativo; véase §13.1), el grupo de color SU(3) es no abeliano, y en consecuencia la teoría se conoce como una *teoría gauge no abeliana*. Esto lleva a aspectos especiales complicados e interesantes. Para los detalles completos de lo que aquí está involucrado, remito al lector a la literatura,[21] pero la idea esencial de cómo se manifiestan las interacciones fuertes es básicamente la que acabo de describir.

Los «bosones gauge» de la QCD (los análogos SU(3) de los fotones) son cantidades conocidas como *gluones*. En las descripciones mediante diagramas de Feynman, las líneas gluónicas se unen a las líneas de quarks de la misma forma que las líneas fotónicas se unen a las líneas de una partícula cargada (Fig. 25.14a). El carácter no abeliano de SU(3) se manifiesta en el hecho de que las propias líneas gluónicas poseen «carga de color», de modo que pueden darse diagramas de Feynman gluónicos con tres puntas (o más), algo que no sucede en el caso electromagnético abeliano (Fig. 25.14b).

Así pues, el papel principal del grupo SU(3), en el modelo estándar, ha pasado de la «simetría de sabor» que tuvo en los años sesenta y setenta a la «simetría de color» del modelo estándar actual. De hecho, en este modelo estándar los tres sabores de d, u y s no se agrupan ahora de una forma fundamental ni mucho menos. En su lugar, los agrupamientos ofrecen tres *generaciones* de dobletes (d,u), (s,c), (b,t). La idea de tener tres generaciones se aplica también a los leptones, siendo estas generaciones las del electrón, el muón y el tauón (y sus correspondientes neutrinos).

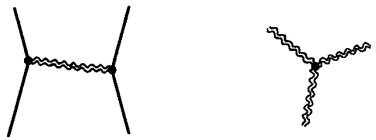


Fig. 25.14. Los gluones son los «bosones gauge» de la QCD. (a) El intercambio de gluones entre quarks (aquí dibujados sin zigzags) subyace en las fuerzas nucleares y al confinamiento de quarks. (b) Puesto que la teoría gauge es no abeliana, las propias líneas de los gluones poseen «carga de color», de modo que pueden ocurrir diagramas de Feynman para los gluones con tres puntas (como en la Fig. 25.12).

En el modelo estándar global existen interrelaciones complicadas entre las interacciones fuerte y electrodébil. En particular, se dan ciertas «rotaciones» entre las entidades básicas que se reconocen mediante interacciones fuertes y las que se reconocen mediante interacciones débiles. Un ejemplo ocurre con el mesón K^0, que puede producirse en colisiones protón-protón de alta energía. Decimos que K^0 es un *autoestado propio de interacciones fuertes*. No obstante, cuando se desintegra el propio K^0 lo hace débilmente, y por ello tiene que ser considerado como una combinación lineal cuántica de los dos estados propios K_L (K largo) y K_S (K corto) de interacciones *débiles*. (El K_L se desintegra normalmente en tres piones en aproximadamente 5×10^{-8} segundos, mientras que el K_S se desintegra normalmente en dos piones en la escala de tiempo mucho más corta de 10^{-10} segundos.) Cada uno de los K_L y K_S es una combinación lineal de K^0 y su antipartícula $\bar{K}^0$, a la que K^0 puede «saltar» mediante interacciones débiles pero no fuertes. La «rotación» entre los estados base de interacción fuerte ($K^0, \bar{K}^0$), y los estados base de interacción débil, (K_L, K_S), tiene lugar a través de un ángulo (abstracto) conocido como *ángulo de Cabibbo* (que es de aproximadamente 0,26 radianes). Este mismo ángulo interviene en las interrelaciones entre interacciones fuertes y débiles en general.

De manera algo similar, existe un ángulo conocido como *ángulo de Weinberg* o *ángulo de mezcla débil* (§25.5) que interviene en la interrelación entre interacciones débiles y electromagnéticas, y forma una parte integral de la teoría electrodébil. De hecho, algunas de las confirmaciones más impresionantes de la teoría electrodébil proceden de

diversos tipos (aparentemente independientes) de determinación observacional de este ángulo que dan respuestas en estrecho acuerdo. Sin embargo, tal como está la teoría, existe una diferencia entre los papeles de los ángulos de Cabibbo y de Weinberg. En efecto, se considera que las interacciones débil y electromagnética están unificadas, y puede adoptarse el punto de vista de que el ángulo de Weinberg fue algo que se «congeló» cuando se «rompió» la simetría U(2) de la teoría electrodébil, aproximadamente 10^{-12} segundos después del big bang (§28.1); pero el ángulo de Cabibbo no tiene un estatus semejante en el modelo estándar, puesto que el modelo no hace ninguna afirmación acerca de cómo podrían unificarse las interacciones electrodébil y fuerte. Se supone que el grupo de simetría básico[22] del modelo estándar completo es $SU(3) \times SU(2) \times U(1)/\mathbb{Z}_6$.

25.8. ¿Más allá del modelo estándar?

Por otra parte, respecto al ángulo de Cabbibo se *podría* adoptar una perspectiva correspondiente a la que se adopta respecto al ángulo de Weinberg, pero esto requeriría algo que está más allá del modelo estándar actual de la física de partículas. Necesitaríamos un modelo en el que las interacciones fuerte y débil estuvieran unidas bajo algún grupo de simetría mayor que incluyera a SU(3) y SU(2). Tal teoría se conoce como una *teoría de gran unificación* o GUT. No hay ninguna GUT comúnmente aceptada, pero ha habido muchos intentos (los más importantes basados en SU(5), o SO(10), o el grupo excepcional E_8; véase §13.2). En §31.14 veremos que la teoría de cuerdas tiene también algo que decir sobre estas cuestiones. En §28.2 se considerarán algunas consecuencias notables de ciertos modelos GUT.

En cualquier caso, el modelo estándar no es evidentemente la «respuesta final» para la física de partículas porque, pese a su indudable éxito, contiene muchas características no explicadas y «flecos sueltos». Incluye unos 17 parámetros inexplicados que sencillamente hay que tomar de la observación (tales como los ángulos de Cabibbo y de Weinberg, las masas de los quarks y leptones, y algunas otras características). Existe también la asimetría bastante extraña entre los papeles de

SU(3) y U(2), en cuanto que SU(3) se considera exacta mientras que U(2) está seriamente rota. De hecho, en mi opinión, parece haber algo extraño en tomar U(2) como un «grupo gauge», lo que parecería exigir una simetría intacta y exacta (véase el capítulo 15, especialmente el párrafo final de §15.8).

En este punto resulta pertinente mencionar otro desarrollo distinto de la idea GUT que aborda esta cuestión concreta de una forma novedosa. Para mí tiene un atractivo especial por razones que se harán evidentes en §33.13. Se trata de una propuesta que se debe al matrimonio chino-británico Chan Hong Mo y Tsou Sheung-Tsun. En su esquema, cada grupo de simetría de partículas (no abeliano) tiene un correspondiente grupo *dual*, que es el mismo grupo abstracto que el original pero que desempeña un papel opuesto. Recordemos el *dual* $^*\boldsymbol{F}$ del tensor de Maxwell $\boldsymbol{F}$, que se ha introducido en §19.2. Podríamos imaginar una conexión gauge U(1) «dual» que tiene $^*\boldsymbol{F}$ como su curvatura fibrada (véase §15.8) en lugar de $\boldsymbol{F}$. La idea consiste en hacer algo parecido para los restantes grupos de simetría del modelo estándar SU(2) y SU(3); pero puesto que estos grupos son *no abelianos*, no es posible considerar simplemente las correspondientes curvaturas duales como curvaturas fibradas de nuevo,[25.9] y se requiere algo más sofisticado (donde hay que considerar cantidades «dependientes del camino»).

Uno de los aspectos más atractivos de este esquema es que el grupo y el grupo dual desempeñan papeles cualitativamente diferentes, al ser uno de ellos exacto, como el SU(3) de la QCD (o el U(1) del electromagnetismo), y estar el otro roto como el SU(2) de la teoría electrodébil, y donde se espera «confinamiento» para el grupo exacto (que es lo que impide que los quarks «con carga de color» escapen al ancho mundo). (Esta propiedad se relaciona con un trabajo anterior de 't Hooft y Weinberg.)[23] En el esquema de Chan-Tsou habría un nuevo SU(2) exacto (dual del roto que interviene actualmente en la teoría electrodébil) que remitiría a una simetría no descubierta hasta ahora, que relaciona análogos de quarks que serían constituyentes leptónicos «2-coloreados» confinados. (Estas subpartículas serían muy pesadas, que es la

[25.9] ¿Puede ver cuál es la dificultad? *Sugerencia*: Escriba expresiones para la curvatura gauge, identidades de Bianchi, etc.

razón de que no hubieran sido detectadas todavía y de que los leptones aparezcan como partículas puntuales a las energías hoy alcanzables.) En consecuencia, debe haber también un SU(3) roto (dual del SU(3) de color), y se considera simplemente que este es el «SU(3)» de las 3 generaciones de quarks y de leptones que parecen tan enigmáticas en el modelo estándar, tal como se entiende convencionalmente. El esquema de Chan-Tsou hace también claras predicciones concernientes a los (aproximadamente) 17 parámetros libres del modelo estándar, pues calcula 14 de ellos a partir de 3 parámetros ajustables. Esto me parece un paso adelante definitivo, siempre que se confirmen las predicciones del esquema. Tal como están las cosas, parecen prometedoras.

Menos claro me resulta, en la actitud convencional hacia el modelo estándar, el hecho de tomar realmente el grupo SU(2) como un grupo gauge, ya que está seriamente roto. Algunos podrían considerar que este SU(2) refleja algún tipo de «simetría oculta» que es realmente exacta y que actúa solo «potencialmente» como un grupo gauge, y el SU(2) de la teoría electrodébil es un tipo de manifestación externa de ello. (Quizá esto no está tan alejado de la idea de Chan-Tsou, aunque no tan explícita.) La perspectiva convencional sobre el SU(2) de la teoría electrodébil parece consistir en que esta es (o, mejor, era) realmente exacta y que se ha roto en procesos extremos que tuvieron lugar en el universo primitivo. Echaremos una ojeada a algunas de las desagradables implicaciones de esto en el capítulo 28. Mientras tanto, en el capítulo siguiente veremos algunas de las ideas matemáticas exóticas pero esenciales que yacen detrás de la forma de tratar en la actualidad la ruptura de simetría en el modelo estándar.

Notas

Sección 25.1

25.1. Véase Pais (1986), pp. 334 y 336, refs. 25, 26.

Sección 25.2

25.2. No he entrado en los detalles de cómo la ecuación de Dirac, que se ha descrito en §24.7, puede transcribirse en la forma 2-espinorial que

se da aquí. El lector interesado puede consultar Zee (2003), apéndice. Weyl introdujo los 2-espinores en Weyl (1929). Véanse Van der Waerden (1929), Infeld y Van der Waerden (1933), Penrose y Rindler (1984), pp. 221-223, y Zee (2003), apéndice.

25.3. Estos son los *espinores reducidos* (o semiespinores) mencionados en §11.5.

25.4. Véanse Infeld y Van der Waerden (1933), Laporte y Uhlenbeck (1931) y Penrose y Rindler (1984, 1986).

25.5. Véanse Schrödinger (1930), Dirac (1982a), Huang (1949) y, para una interesante perspectiva moderna, Hestenes (1990).

25.6. Aquellos lectores que ya estén familiarizados con los diagramas de Feynman pueden encontrar confuso mi ordenamiento temporal vertical. En la comunidad de la QFT es más habitual mostrar el tiempo en aumento hacia la derecha. Mi preferencia por mostrar el aumento del tiempo hacia arriba está de acuerdo con buena parte de la comunidad relativista, puesto que es consistente con la mayor parte de los diagramas espaciotemporales (véase en especial el capítulo 17).

25.7. ¡De hecho, la vida de esos físicos podría haber sido mucho más simple con el uso del formalismo 2-espinorial en QED! Véanse Geroch (notas de sus lecciones en la Universidad de Chicago, no publicadas) y también §34.3.

25.8. Mis propios convenios hubieran sido escribir aquí $(1 \pm i\gamma_5)\psi$, en lugar de $(1 \pm \gamma_5)\psi$ (véase Penrose y Rindler 1984, 1986, apéndice), como harían otros autores. Aquí me estoy acomodando a lo que parece ser estándar en la comunidad física.

Sección 25.3

25.9. Quizá esto haya sido estimulado en parte por una sugerencia de Martin Block (y transmitida por Richard Feynman); véase la fascinante exposición por Martin Gardner en *The New Ambidextrous Universe* (W. H. Freeman, 1990), capítulo 22.

25.10. Esta ecuación fue propuesta por Weyl en 1929 y también había sido considerada por Dirac antes de que diera con su «ecuación de Dirac para el electrón»; Dirac (1928) y Dirac (1982b). Pauli se había opuesto vehementemente a la ecuación de Weyl debido a su falta de invariancia bajo reflexión espacial. Por desgracia, Weyl murió un año antes de que se propusiera la no invariancia por reflexión en las interacciones débiles, que reivindicaba su propuesta. Zee (2003) discute ambas ecuaciones.

Sección 25.4

25.11. Las partículas masivas de espín 1 pueden describirse como si tuvieran *tres* ingredientes, una «zig» levógira (helicidad 1), una «zag» dextrógira (helicidad-1) y una, digamos, «zog» sin rotación (helicidad 0). (El 2-espinor zig y el 2-espinor zag tienen dos índices no primados y dos índices primados, respectivamente, mientras que el 2-espinor zog tiene un índice de cada tipo.) Podemos adoptar el punto de vista de que es solo la partícula zog la que media las interacciones débiles.

Sección 25.5

25.12. El grupo podría expresarse como $SU(2) \times U(1)/\mathbb{Z}_2$, donde el «$/\mathbb{Z}_2$» significa «dividir por un subgrupo Z_2». Sin embargo, hay más de un subgrupo semejante, de modo que esta notación no es del todo explícita. La notación «U(2)» escoge automáticamente el correcto. (Agradezco a Florence Tsou esta observación.) Parece que la razón de que el grupo de simetría electrodébil no se conozca convencionalmente como «U(2)» es que esto no se extiende fácilmente a la simetría del modelo estándar completo, que también incorpora al grupo de simetría fuerte SU(3), siendo el grupo completo una versión de $SU(3) \times SU(2) \times U(1)/Z_6$; véase §25.7.

25.13. La teoría electrodébil fue desarrollada por Stephen Weinberg, Sheldon Glashow y Abdus Salam a finales de la década de 1960, trabajo que les valió a los tres el premio Nobel. Véase Weinberg (1967), Salam y Ward (1959) y Glashow (1959); para una referencia general sobre la teoría electrodébil, véase Zee (2003), o Halzen y Martin (1984) y Kaku (1993).

25.14. GeV son giga electrón-voltios. Giga es un prefigo griego que indica multiplicación por 10^9; y un electrón-voltio es una medida de energía, en concreto, la energía que ganará un electrón no ligado cuando atraviese una diferencia de potencial de 1 voltio. Es aproximadamente $1{,}6 \times 10^{-19}$ J.

Sección 25.6

25.15. La luz visible cae entre las longitudes de onda λ = 400-700 nanometros; puede pasarse de longitudes de onda a frecuencias ν mediante la relación $\nu = \frac{c}{\lambda}$.

25.16. Véase Gell-Mann y Ne'eman (2000) para la teoría; el artículo de V. E. Barnes sobre la observación de la Ω^-, originalmente publicado en 1964, está en la misma obra, en pp. 88-92.

25.17. En la terminología de la moderna física de partículas, «N^+» y «N^0» parecen haber reemplazado a «p» y «n» para denotar el protón y el neutrón, respectivamente. Esto es consistente con la notación para otras partículas en cuanto que (N^+, N^0) constituye un *doblete* en el esquema de clasificación SU(3), como (Ξ^-, Ξ^0), etc., y nos permite llamar genéricamente un *nucleón* como «N».

Sección 25.7

25.18. Véase Han y Nambu (1965).

25.19. Véase Weinberg (1992).

25.20. C. N. Yang y R. L. Mills descubrieron esta teoría en 1954, aunque la idea básica había sido descubierta antes (y rechazada porque las partículas gauge tenían que carecer de masa) por Wolfgang Pauli, en los años posteriores a la Segunda Guerra Mundial, y Ronald Shaw, en 1955. Para una historia exhaustiva de estas cuestiones, presentada en su Conferencia Nobel, véase Abdus Salam (1980). El truco que se utiliza ahora para evitar el problema «sin masa» es el «mecanismo de Higgs» para la ruptura de simetría, al que se ha aludido en §25.5 y se discutirá en §26.11.

25.21. Para los detalles técnicos, véanse Aitchison y Hey (2004), vol. 2, o Zee (2003). Véase Chan y Tsou (1993) para una visión general de las ideas de las teorías gauge.

25.22. Véase la nota 25.12. Puede encontrarse una visión general del modelo estándar en cualquier buen libro de texto de teoría cuántica de campos, como por ejemplo, Zee (2003).

Sección 25.8

25.23. La idea de Chan-Tsou se expone en Chan y Tsou (2002); se basa en una propiedad desarrollada en 't Hooft (1978).

26

Teoría cuántica de campos

26.1. El estatus fundamental de la QFT en la teoría moderna

En el capítulo anterior nos hemos familiarizado brevemente con el modelo estándar del siglo XX para la física de partículas. Es un modelo matemático en notable acuerdo con los hechos observacionales en un amplio abanico de fenómenos, e incluye algunos ingeniosos ingredientes matemáticos que parecen estar en profunda armonía con los modos de la naturaleza. Pese a todo, tal como he presentado las cosas, la estructura matemática de este modelo podría parecer algo complicada y arbitraria. Por supuesto, gran parte de esta estructura ha estado motivada por hechos brutos de la física de partículas que los físicos han tenido que manejar tal como se los ha presentado la naturaleza. Así debería ser en cualquier teoría científica seria. Pero también hay poderosas razones *teóricas* subyacentes en las elecciones concretas que se han hecho para la estructura del modelo estándar. De hecho, el poder predictivo de la teoría depende crucialmente de la consistencia matemática de tales soportes teóricos.

La fuerza impulsora teórica es una continuación de la historia que hemos comenzado en el capítulo 24: ¿cómo encontrar una teoría cuántica para la física de partículas que sea compatible con los requisitos de la teoría de la relatividad especial de Einstein? En ese capítulo hemos visto la importancia de la introducción por parte de Dirac de las antipartículas en una teoría cuántica relativista, lo que nos llevaba al marco de una teoría cuántica de campos. De hecho, el modelo estándar es un ejemplo particular de una teoría cuántica de campos interactuantes,

impulsado básicamente por algunos potentes requisitos de consistencia difíciles de satisfacer en tales teorías. Para apreciar algo de la fuerza que hay tras estos requisitos de consistencia (que continúan impulsando las teorías especulativas más modernas, tales como la teoría de cuerdas), necesitaremos examinar algo de la estructura de la teoría cuántica de campos (QFT). Esto nos ayudará asimismo a apreciar el significado de los diagramas de Feynman que hemos visto en la sección anterior. Además, obtendremos también otra perspectiva sobre las antipartículas, algo más amplia que las que hemos visto en los capítulos 23 y 24.

La teoría cuántica de campos constituye la base esencial subyacente en el modelo estándar, así como prácticamente en cualquier otra teoría física que intente sondear los fundamentos de la realidad física. Por lo tanto es necesario que podamos captar una idea de ese magnífico e imponente esquema de cosas, un esquema que surgió en buena medida de aquellas extraordinarias intuiciones de Paul Dirac con las que nos hemos familiarizado en el capítulo 24. Habría que señalar que el propio Dirac fue el principal iniciador de la QFT, aunque también hubo contribuciones iniciales importantes por parte de Jordan, Heisenberg y Pauli. Sin embargo, tal como estaba, y para muchos problemas de interés, dicha teoría no era capaz de ofrecer respuestas *finitas*, en lugar del «∞» que casi siempre aparecía. Se necesitaron potentes desarrollos adicionales por parte de Bethe, Tomonaga, Dyson, Schwinger, y especialmente Feynman, para hacer la teoría viable para QFT adecuadas conocidas como «renormalizables». Aportaciones más recientes de Ward, Weinberg, Salam, Wilson, Veltman y 't Hooft, entre otros, nos han llevado a un tipo muy apropiado de teorías renormalizables, que han dado un aporte vital para lo que es ahora el modelo estándar de la física de partículas (capítulo 25), del que pueden obtenerse realmente respuestas consistentes.[1] (¡Los requisitos teóricos son tan rígidos que parece casi secundario que dichas respuestas estén realmente en un excelente acuerdo con el experimento!) El problema básico ha sido siempre el evitar los infinitos de una manera adecuada, y ha sido este impulso, junto con importantes aportaciones procedentes de la observación, lo que ha llevado la teoría en las direcciones apropiadas y fructíferas.

De hecho, la QFT parece subyacer en prácticamente todas las teo-

rías físicas que intentan ofrecer seriamente una imagen del funcionamiento del universo en sus niveles más profundos. Muchos físicos (quizá incluso la mayoría de ellos) adoptarían el punto de vista de que el armazón de la QTF está «aquí para quedarse», y que la culpa de cualquier inconsistencia (que habitualmente son los infinitos que surgen de integrales divergentes, o de series divergentes, o de ambas cosas) la tiene el esquema concreto al que se está aplicando la QFT, y no el armazón de la propia QFT. Tales esquemas suelen estar especificados por un lagrangiano sujeto a ciertos principios de simetría. En §§26,6,10 veremos cómo se aplican en general las ideas lagrangianas a la QFT.

Muchos intentos modernos por eliminar los infinitos de la QFT consideran que la *gravedad* altera profundamente el comportamiento del espaciotiempo a escalas extraordinariamente minúsculas, y con ello suministra los «cortes» que podrían hacer finitas las expresiones que por el momento parecen divergentes (véase §31.1, en particular). Pese a todo, queda la pregunta de si la propia QFT podría necesitar modificación cuando se introducen los principios de la relatividad *general* de Einstein (véase el capítulo 30). Sin embargo, a juzgar por las actividades de la inmensa mayoría de los investigadores actuales en estas áreas, la QFT en su forma presente no suele ser cuestionada, y será importante que entendamos de la mejor forma posible sus sofisticadas ideas. Por supuesto no podré entrar en muchos detalles en mis descripciones de este magnífico, profundo y difícil esquema de cosas, a veces fenomenológicamente preciso, pero a menudo frustrantemente inconsistente. Pero intentaré, brevemente, transmitir algo de su sabor, aunque de forma muy incompleta, antes de volver finalmente a aquellos aspectos que suministran la fuerza impulsora teórica subyacente en el modelo estándar.

26.2. Operadores de creación y aniquilación

Una de las primeras ideas de la QFT fue el procedimiento que lleva el nombre más bien equívoco de «segunda cuantización». Mediante este procedimiento intentamos que la propia función de onda ψ de una

partícula se convierta en un «operador» que actúa sobre cierto vector de estado fantasma, que puede denotarse por «$|0\rangle$», oculto en el extremo derecho (compárese con §21.3 y la imagen de Heisenberg de §22.4). Denotaré a este «operador función de onda» por la letra griega *mayúscula* $\boldsymbol{\Psi}$, correspondiente a la letra griega ψ que denota nuestra función de onda de una partícula. Como en la mecánica cuántica de partículas ordinaria, $\boldsymbol{\Psi}$ puede considerarse una función de la posición $\mathbf{x}$ de la partícula en el 3-espacio, i.e., $\boldsymbol{\Psi} = \boldsymbol{\Psi}(\mathbf{x})$ o, alternativamente, de su 3-momento $\mathbf{p}$, si se prefiere una representación de momentos, i.e., $\tilde{\boldsymbol{\Psi}} = \tilde{\boldsymbol{\Psi}}(\mathbf{p})$.

¿Cómo se interpreta este extraño «operador función de onda» $\boldsymbol{\Psi}$ (o $\tilde{\boldsymbol{\Psi}}$)? Ahora no representa el estado cuántico real, sino que describe la operación que «crea» una nueva partícula que tiene esta función de onda dada[2] ψ, introduciéndola en el estado que hay previamente, estado «previo» que está representado por la expresión que sigue inmediatamente a la derecha del símbolo $\boldsymbol{\Psi}$ (o $\tilde{\boldsymbol{\Psi}}$). Un operador semejante se conoce como un *operador creación*.

Normalmente se toma el vector de estado fantasma $|0\rangle$ en el extremo derecho como el «estado vacío», que representa la completa ausencia de partículas de cualquier tipo. Una serie de estos operadores creación crea entonces una serie de partículas, añadidas una a una al vacío, de modo que

$$\boldsymbol{\Psi\Phi} \ldots \boldsymbol{\Theta}|0\rangle$$

es el estado que resulta de introducir sucesivamente partículas con funciones de onda

$$\theta, \ldots, \phi, \psi.$$

Puesto que cualquier tipo concreto de partícula será o un fermión o un bosón, hay que tener en cuenta este hecho. En particular, hay que incorporar el principio de Pauli que nos impide introducir dos fermiones, uno tras otro, en el mismo estado. El principio de Pauli se expresa en este formalismo por la propiedad $\boldsymbol{\Psi}^2 = 0$ (i.e., $\boldsymbol{\Psi\Psi} = 0$) para cualquier función de onda fermiónica ψ, lo que nos dice que si tratamos de introducir dos veces esta función de onda fermiónica concreta en el estado, entonces obtenemos cero, que no es un vector de estado

admisible. Esta regla del «principio de Pauli» es solo un ejemplo concreto de la propiedad de anticonmutación

$$\boldsymbol{\Psi\Phi} = -\boldsymbol{\Phi\Psi},$$

donde $\boldsymbol{\Psi}$ y $\boldsymbol{\Phi}$ son operadores creación que describen el mismo tipo de fermión. Para operadores creación $\boldsymbol{\Theta}$ y $\boldsymbol{\Xi}$ que describen el mismo tipo de bosón, tenemos la propiedad de conmutación[26.1]

$$\boldsymbol{\Theta\Xi} = \boldsymbol{\Xi\Theta}.$$

Así pues, vemos que los operadores creación satisfacen las reglas de un álgebra de Grassmann (graduada), como se ha descrito en §11.6, donde los operadores creación de fermiones se consideran de grado impar y los operadores de creación de bosones, de grado par.

De acuerdo con la discusión de §24.3, las funciones de onda que se introducen en un estado para la creación de una partícula deben ser de frecuencia positiva. Las cantidades de frecuencia negativa desempeñan también un papel en el formalismo, a saber, como operadores aniquilación. La conjugada compleja $\bar{\psi}$ de la función de onda de frecuencia positiva ψ es una cantidad de frecuencia negativa. Está asociada con el operador aniquilación $\boldsymbol{\Psi}^*$, que es el conjugado hermítico[3] del operador creación $\boldsymbol{\Psi}$ (véase §13.9). La interpretación de $\boldsymbol{\Psi}^*$ es que representa la eliminación de una partícula del estado total (siendo dicho estado total aquel que, como antes, está descrito por cualquier cosa que esté a la derecha de $\boldsymbol{\Psi}^*$ en la expresión). Puesto que nuestro estado vacío fantasma $|0\rangle$, a la derecha, no contiene ninguna partícula, la acción de cualquier operador aniquilación directamente sobre él debe dar cero:

$$\boldsymbol{\Psi}^*|0\rangle = 0.$$

Esto no significa, por supuesto, que los operadores aniquilación den siempre cero, porque podríamos tener algunas partículas al principio. Por ejemplo, una expresión como $\boldsymbol{\Psi}^*\boldsymbol{\Phi\Theta}|0\rangle$ no tiene por qué ser cero. Esto es válido incluso si ninguno de los estados $\boldsymbol{\Phi}$ y $\boldsymbol{\Theta}$ es el mismo que el $\boldsymbol{\Psi}$ que estamos eliminando. De hecho, no deberíamos con-

[26.1] Explique por qué esto da estados con las simetrías correctas para bosones y fermiones, como se ha descrito en §23.8.

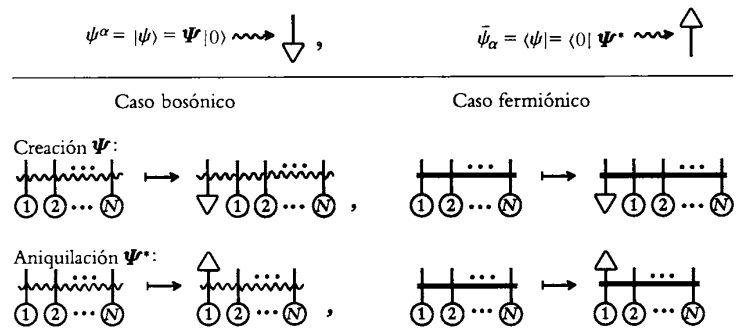


Fig. 26.1 Forma diagramática de la acción de un operador creación $\boldsymbol{\Psi}$ en el caso bosónico $\phi_1^{(\beta}\phi_2^{\gamma} \ldots \phi_N^{\nu)} \mapsto \psi^{(\alpha}\phi_1^{\beta}\phi_2^{\gamma} \ldots \phi_N^{\nu)}$ y en el caso fermiónico $\phi_1^{[\beta}\phi_2^{\gamma} \ldots \phi_N^{\nu]} \mapsto \psi^{[\alpha}\phi_1^{\beta}\phi_2^{\gamma} \ldots \phi_N^{\nu]}$; y de un operador de aniquilación $\boldsymbol{\Psi}^*$ en el caso bosónico $\phi_1^{(\alpha}\phi_2^{\beta} \ldots \phi_N^{\mu)} \mapsto \bar{\psi}^{(\alpha}\phi_1^{\alpha}\phi_2^{\beta} \ldots \phi_N^{\mu)}$ y en el caso fermiónico $\phi_1^{[\alpha}\phi_2^{\beta} \ldots \phi_N^{\mu} \mapsto \bar{\psi}^{[\alpha}\phi_1^{\alpha}\phi_2^{\beta} \ldots \phi_N^{\mu]}$.

siderar que el operador $\boldsymbol{\Psi}^*$ está eliminando simplemente de dicho estado la función de onda específica ψ de la partícula.[4] En general, es poco probable que la función de onda concreta ψ intervenga exactamente como parte del estado. En su lugar, lo que hace $\boldsymbol{\Psi}^*$ es formar un producto escalar con la parte del estado que se refiere al tipo de partícula que se está eliminando. (En la Fig. 26.1 —sobre todo para diversión de los expertos— indico la forma diagramática de lo que entiendo por esto, tanto en el caso fermiónico como en el bosónico, y también presento diagramas que representan el proceso de creación tanto como el de aniquilación.)[26.2] De acuerdo con esto, resulta que los operadores creación y aniquilación (para el mismo tipo de partícula) deben satisfacer reglas de (anti)conmutación

$$\boldsymbol{\Psi}^*\boldsymbol{\Phi} \pm \boldsymbol{\Phi\Psi}^* = \mathrm{i}^k\langle\psi|\phi\rangle\boldsymbol{I},$$

donde el signo «más» se refiere a fermiones y el signo «menos» a bosones, $\boldsymbol{I}$ representa el operador *identidad*, $\langle\ |\ \rangle$ representa la expresión del

[26.2] Dé sentido a todo esto (y verifique esta ley de conmutación para la creación y la aniquilación de partículas de un tipo dado) remitiendo a la notación de índices de §23.8 o la notación diagramática de la Fig. 12.17, o ambas, utilizando expresiones como $\bar{\psi}_\alpha\psi^{[\alpha}\phi^{\beta} \ldots \chi^{\kappa]}$. Ordene todos los factoriales que conservan la normalización del estado, tanto en el caso fermiónico como en el bosónico.

producto escalar ordinario en el espacio de Hilbert para partículas individuales (habiéndose considerado el caso sin espín en §22.3, y existiendo una generalización apropiada para partículas con espín),[5] y donde i^k representa uno de los 1, i, −1, −i, dependiendo del espín (y no les molestaré con cuál). Tenemos, asimismo, las siguientes reglas de (anti)conmutación para dos operadores creación (que se han dado también antes), y para dos operadores aniquilación (signo más para fermiones, menos para bosones):

$$\boldsymbol{\Psi\Phi} \pm \boldsymbol{\Phi\Psi} = 0, \quad \boldsymbol{\Psi^*\Phi^*} \pm \boldsymbol{\Phi^*\Psi^*} = 0.$$

Habría que señalar que el teorema espín-estadística, que hemos mencionado brevemente en §23.7, exige que tengamos reglas de anticonmutación (signo más en lo anterior, y con ello fermiones) para partículas con espín semientero (1/2, 3/2, 5/2, ...) y reglas de conmutación (signo menos, y con ello bosones) para partículas con espín entero (0, 1, 2, 3, ...). Las razones para ello van más allá del alcance de esta exposición.[6] Sin embargo, las cuestiones esenciales tienen que ver con la positividad de la energía (en el caso de los fermiones) y positividad del número de partículas (en el caso de los bosones), junto con propiedades combinatorias de los índices espinoriales relevantes.[7]

26.3. Álgebras de dimensión infinita

Es un hecho notable que en el caso de los fermiones estas reglas de anticonmutación tienen exactamente la misma forma algebraica que las que definen un álgebra de Clifford, como se ha descrito en §11.5.[26.3] La única diferencia esencial es que las álgebras de Clifford ordinarias son de dimensión finita, mientras que el espacio de operadores de creación y aniquilación para un campo fermiónico es de dimensión infinita, pues esa es la dimensión del espacio de funciones de onda de una partícula. Sin embargo, el lector debería ser advertido de que los espacios de dimensión infinita, aunque análogos a los de dimensión finita,

[26.3] Explique esta estructura de álgebra de Clifford, detallando más explícitamente el papel del producto escalar. (Tome las leyes definitorias para un álgebra de Clifford en la forma $\gamma_p\gamma_q + \gamma_q\gamma_p = -2g_{pq}\boldsymbol{I}$.)

pueden tener propiedades muy diferentes, y con frecuencia es mucho más difícil trabajar con ellos.

Es interesante que el formalismo de la QFT incluye también versiones de dimensión infinita de algunos otros tipos de estructura algebraica de dimensión finita que hemos considerado antes en este libro. El producto escalar $\langle\ |\ \rangle$, por ejemplo, es realmente una versión en dimensión infinita del producto escalar hermítico que se ha considerado en §13.5. (cf. §22.3). De hecho, en QFT resulta que no solo estamos interesados en la «hermiticidad» (unitariedad), sino que encontramos que las formas simétricas («pseudoortogonalidad»), las formas antisimétricas (simplécticas) y las estructuras complejas también desempeñan sus papeles.[8] Las versiones ordinarias en dimensión finita de las formas pseudoortogonal y simpléctica se han considerado en §§13.8,10 y las estructuras complejas ordinarias en dimensión finita en §12.9.

Es de particular importancia, para la QFT, la forma en que aparece una estructura compleja (de dimensión infinita). Ya hemos visto que los números complejos, las funciones holomorfas y los espacios vectoriales complejos tienen un papel fundamental en la teoría cuántica (y por consiguiente en la estructura básica de nuestro mundo). Pero la estructura compleja concreta de dimensión infinita que entra en este momento en el estudio de la QFT parece tener un significado algo diferente (aunque interrelacionado) de estos ejemplos primitivos de la magia de los números complejos. Va más allá de la mera afirmación de que los espacios de Hilbert de la teoría cuántica son complejos (i.e., que las superposiciones cuánticas tienen lugar con coeficientes complejos). Tratemos de entender de qué se trata.

Recordemos cómo se ha introducido en §12.9 la noción de *estructura compleja*. Un espacio vectorial complejo de n dimensiones puede considerarse como un espacio vectorial real de $2n$ dimensiones, donde existe una operación $\boldsymbol{J}$, que satisface $\boldsymbol{J}^2 = -1$, cuya acción sobre el $2n$-espacio real equivale a «multiplicación por i» en el n-espacio complejo. La versión de dimensión infinita que tiene relevancia para la QFT tiene que ver con el paso de un campo clásico a un campo cuántico. Hasta ahora he expresado las cosas en un lenguaje de función de onda/partícula. Pero tenemos que saber también cómo pasar directamente de un campo clásico a un campo cuántico, puesto que con los

campos clásicos no tenemos a mano una imagen de partículas clásicas que podamos «cuantizar» según los procedimientos de los capítulos 21-23.

Es útil tener en mente el campo electromagnético como modelo. Aquí la linealidad de las ecuaciones de Maxwell (§19.2) facilita las cosas. El espacio $\mathcal{F}$ de soluciones de las ecuaciones de Maxwell libres (con condiciones de decaimiento apropiado en el infinito para hacer que converjan las integrales relevantes) es un espacio vectorial real de dimensión infinita.[26.4] Utilizando procedimientos relacionados con los descritos en §§9.2,3,5, podemos expresar cada solución $\boldsymbol{F}$ de las ecuaciones de Maxwell[9] como una *suma* de una solución $\boldsymbol{F}^+$ de frecuencia positiva y una solución $\boldsymbol{F}^-$ de frecuencia negativa

$$\boldsymbol{F} = \boldsymbol{F}^+ + \boldsymbol{F}^-.$$

Esta separación en frecuencias positiva y negativa es crucial para la construcción de la QFT apropiada (recordemos los comentarios sobre esta cuestión en §24.3 y §26.2). La operación $\boldsymbol{J}$, tal como se aplica a este espacio vectorial real de dimensión infinita $\mathcal{F}$, lo transforma en un espacio vectorial complejo (de dimensión infinita) y, al hacerlo, proporciona una forma de englobar esta separación en frecuencias positiva/negativa. $\boldsymbol{J}$ lo hace actuando sobre cada campo de Maxwell libre $\boldsymbol{F}$ de la siguiente manera:

$$\boldsymbol{J}(\boldsymbol{F}) = \mathrm{i}\boldsymbol{F}^+ - \mathrm{i}\boldsymbol{F}^-.$$

Los autoestados de $\boldsymbol{J}$ con autovalor i son los campos (complejos) de frecuencia positiva, y los autoestados con autovalor –i son los campos de frecuencia negativa.[26.5] Los campos de frecuencia positiva suministran las funciones de onda de un único fotón que los operadores de creación van a introducir. Existe también una expresión explícita para el producto escalar que puede utilizarse para normalizar estados cuando sea necesario (que incluye una integral sobre cualquier 3-superficie de género espacio de una expresión que incluye las componentes del campo

[26.4] Explique qué significan en este espacio la suma y la multiplicación por una constante escalar.

[26.5] Demuéstrelo. (No se preocupe por sutilezas como «condiciones de decaimiento».)

de Maxwell multiplicadas por componentes del potencial de Maxwell;[10] compárese con §21.9 y §22.3 para el caso escalar). Otros campos clásicos pueden tratarse de un modo análogo, pero cuando las «ecuaciones de campo libre» son no lineales (como sucede en la relatividad general) pueden aparecer dificultades profundas. Podemos calificar los campos no lineales como campos «autointeractivos», y podemos atribuir tales dificultades a los problemas asociados con la cuantización en presencia de interacciones, a lo que llegaremos enseguida.

26.4. Antipartículas en QFT

Pero antes de hacerlo volvamos a la cuestión de las antipartículas. En el capítulo 24, y de nuevo en §26.1, he señalado la importancia del concepto de antipartícula en QFT. ¿Cómo intervienen las antipartículas en el formalismo QFT actual? Como se ha comentado en §25.3, algunas partículas son sus propias antipartículas, mientras que la mayoría de las partículas no lo son. Desde el punto de vista matemático, la cuestión es si la operación de conjugación compleja, tal como se aplica directamente a las cantidades de campo clásicas (o a la función de onda de 1-partícula), da o no una cantidad del mismo tipo que antes. En el caso de un campo escalar, esto se expresa normalmente (pero no muy adecuadamente) como la cuestión de si el campo clásico es o no un campo real. Los campos complejos se consideran campos cargados, en donde el ángulo de fase complejo (el factor $e^{i\theta}$) debe tratarse de acuerdo con las recetas de «campo gauge» para las interacciones electromagnéticas, tal como se han descrito en §15.8 y §19.4. El conjugado complejo de un campo semejante tiene la carga opuesta, y por ello no es una «cantidad del mismo tipo» (de modo que, por ejemplo, no tendría sentido sumar el campo a su conjugado complejo). En tales situaciones, la partícula y su antipartícula son ciertamente diferentes. Sin embargo, la naturaleza compleja —o naturaleza «cargada»— del campo clásico no es ni mucho menos toda la historia. El mesón K^0 no cargado, por ejemplo, difiere de su antipartícula, mientras que el mesón π^0 no cargado es el mismo que su antipartícula. En ambos casos el campo clásico sería un escalar real.

¿Qué pasa con los campos espinoriales (o fermiónicos)? En el caso del electrón de Dirac, su carga es suficiente para caracterizar su complejo conjugado como poseedor de un carácter diferente del campo original. Pero en el caso de los neutrinos, si tienen masa, hay más de una posibilidad. Por ejemplo, en la situación de lo que se conoce como un campo espinorial *de Majorana*, el neutrino (con masa) sería su propia antipartícula. (En las descripciones dadas en el capítulo 23, el zig del neutrino sería el zag del antineutrino, y viceversa.) De acuerdo con el conocimiemto actual,[11] todos los neutrinos difieren de sus antineutrinos.

Así que ¿cómo trata el formalismo QFT las antipartículas? Consideremos el caso en que la operación de conjugación compleja sobre el campo clásico —o sobre la función de onda de 1-partícula ψ— da una cantidad de un carácter diferente, de modo que la partícula cuántica difiere de su antipartícula. (La situación en que partícula y antipartícula son la misma se ha tratado en nuestra anterior exposición, en §26.2.) Ahora bien, una función de onda ψ ordinaria debería tener frecuencia positiva, pero también podemos considerar cierta cantidad ϕ, del mismo tipo que ψ, aunque de frecuencia negativa. Entonces la conjugada compleja $\bar{\phi}$ de ϕ sería una función de onda de un tipo diferente de ψ, aunque tanto $\bar{\phi}$ como ψ son ahora de frecuencia positiva. La cantidad $\bar{\phi}$ proporcionaría una función de onda para un estado de 1-antipartícula. El correspondiente operador creación para la antipartícula en dicho estado sería $\bar{\boldsymbol{\Phi}}$, y el operador aniquilación, $\bar{\boldsymbol{\Phi}}^*$.

Intentemos tomar contacto con el «mar» original de Dirac (descrito en §24.8) considerando que el estado «en la sombra», a la derecha de todos los operadores, en el caso de Dirac, es un estado diferente del estado de vacío habitual $|0\rangle$, que, recordemos, está totalmente desprovisto de partículas o antipartículas. En su lugar, vamos a considerar que el nuevo estado «vacío» es el propio «mar» de Dirac, denotado por $|\Sigma\rangle$, que va a estar completamente lleno de todos los estados electrónicos de energía negativa, pero nada más. Imaginemos ahora la situación de un único positrón, que en la imagen original de Dirac está descrito por un único «agujero» en los estados electrónicos de energía negativa. Todos los demás estados de energía negativa estarán llenos excepto este estado particular que falta, que está dado por alguna ϕ de frecuencia

negativa. La descripción en teoría cuántica de campos, utilizando este vacío $|\Sigma\rangle$, sería el resultado del operador aniquilación $\boldsymbol{\Phi}^*$ actuando sobre $|\Sigma\rangle$, puesto que tal operador elimina de este vacío el estado de energía negativa ϕ, dando el estado total $\boldsymbol{\Phi}^*|\Sigma\rangle$.[26.6]

Si fuéramos a utilizar la descripción que utiliza el vacío más usual $|0\rangle$, entonces pensaríamos que en lugar de eliminar el estado electrónico ϕ de energía negativa estamos insertando el estado positrónico con función de onda $\bar{\phi}$. Esto se consigue aplicando el operador creación $\bar{\boldsymbol{\Phi}}$ a $|0\rangle$, lo que da el estado total $\bar{\boldsymbol{\Phi}}|0\rangle$. Este no *parece* el mismo que el $\boldsymbol{\Phi}^*|\Sigma\rangle$ que teníamos con la descripción del «mar» de Dirac, pero en cierto sentido los estados $\bar{\boldsymbol{\Phi}}|0\rangle$ y $\boldsymbol{\Phi}^*|\Sigma\rangle$ son básicamente equivalentes. Tanto el operador $\bar{\boldsymbol{\Phi}}$ como el $\boldsymbol{\Phi}^*$ suponen introducir la misma cantidad algebraica en el estado total, a saber, ese vector particular definido por $\bar{\phi}$ en el espacio de Hilbert de funciones de onda de 1-partícula. La diferencia entre los operadores $\bar{\boldsymbol{\Phi}}$ y $\boldsymbol{\Phi}^*$ está meramente en la forma algebraica[26.7] en que se estima que este vector del espacio de Hilbert actúa sobre el estado total. Puesto que siempre podemos utilizar leyes de anticonmutación para mover $\bar{\boldsymbol{\Phi}}$ y $\boldsymbol{\Phi}^*$ al extremo derecho, la forma en que se estima que $\bar{\phi}$ actúa sobre el estado total se reduce a su acción sobre el estado vacío escogido, $|0\rangle$ o $|\Sigma\rangle$ según sea el caso, en el extremo derecho. Esta información puede considerarse parte de la especificación de lo que realmente significa este estado vacío.

26.5. Vacíos alternativos

Aquí habría que hacer algunos comentarios importantes acerca de esta situación en la que parece haber elecciones alternativas para nuestro «estado vacío». Veremos que la cuestión de los «vacíos alternativos» tiene una considerable importancia en la moderna QFT. Consideremos

[26.6] Explique por qué podemos eliminar un estado específico de este modo, a pesar de mis anteriores matizaciones sobre lo que realmente hace un operador aniquilación. (*Sugerencia*: Véase el ejercicio [26.2].)

[26.7] Remitiéndose al ejercicio [26.2] y la Fig. 12.18, muestre esta diferencia algebraica en la notación de índices o en la notación diagramática.

el álgebra $\mathcal{A}$, consistente en todos los operadores $\boldsymbol{A}$ que pueden construirse a partir de expresiones algebraicas, o expresiones en series de potencias convergentes, de los operadores creación y aniquilación. Dos propuestas para un «estado vacío», digamos $|0\rangle$ o $|\Sigma\rangle$, pueden tener la propiedad de que no haya ningún elemento de $\mathcal{A}$ que pueda aplicarse bien a $|0\rangle$, o bien a $|\Sigma\rangle$ para obtener el otro. En tales casos, los estados $|0\rangle$ y $|\Sigma\rangle$ tienen que considerarse pertenecientes a espacios de Hilbert diferentes, y entonces observamos que hay expresiones de la forma

$$\langle\Sigma|\boldsymbol{A}|0\rangle \text{ o } \langle 0|\boldsymbol{A}|\Sigma\rangle,$$

donde $\boldsymbol{A}$ pertenece a $\mathcal{A}$, que nos dan respuestas infinitas o respuestas sin ningún sentido. Esto nos prohíbe construir una teoría cuántica consistente en la que aparezcan a la vez los estados $|0\rangle$ y $|\Sigma\rangle$. (Recordemos de la exposición de §22.5 que cantidades como $\langle\Sigma|\boldsymbol{A}|0\rangle$ son expresiones del tipo general que tendríamos que utilizar para computar probabilidades; véase §22.3 para la notación.)

Lo que aquí se plantea es una cuestión profunda en QFT y desempeña un papel vital en los enfoques modernos de la física de partículas. La «elección del estado vacío» es una cuestión de importancia comparable (y complementaria) a la elección del álgebra $\mathcal{A}$ de operadores creación y aniquilación que define en cierto sentido la *dinámica* de la QFT. En el caso de los electrones libres, los dos vacíos que hemos considerado, a saber $|0\rangle$ (que no contiene partículas ni antipartículas) y $|\Sigma\rangle$ (en la que todos los estados de energía negativa están llenos), pueden considerarse, en cierto sentido, esencialmente equivalentes pese al hecho de que $|0\rangle$ y $|\Sigma\rangle$ nos dan espacios de Hilbert diferentes. Podemos considerar que la diferencia entre el vacío $|\Sigma\rangle$ y el vacío $|0\rangle$ es simplemente una cuestión de dónde trazamos una línea que defina el «cero de carga».

De hecho, podríamos pensar que el mar de Dirac es físicamente diferente de un vacío propiamente dicho, porque el mar de electrones de energía negativa nos proporcionaría una carga eléctrica enorme, en realidad infinita. Para que el estado $|\Sigma\rangle$ tenga sentido físico, tenemos que «renormalizar» la carga, de modo que el valor de la carga total infinita del «mar» (de hecho, negativamente infinita, pues negativa es la

carga del electrón) cuenta como cero. Una situación análoga resultaría si consideráramos la *masa* del mar de Dirac, pues podríamos estar interesados en su influencia gravitatoria (activa). La energía negativa total infinita del mar de Dirac proporcionaría (por $E = mc^2$) una masa infinita negativa, lo que es físicamente absurdo, igual que en el caso de la carga eléctrica infinita. Una vez más, si vamos a tomar el mar de Dirac en serio debemos renormalizar también la masa del vacío, «añadiendo una densidad de masa infinita» a la del mar, de modo que el total nos proporcione el valor cero que requerimos para la densidad de masa del vacío observado.

El lector puede haberse formado la impresión de que esta cuestión de los «vacíos alternativos» y de la aparente necesidad de «renormalizar» cosas tales como la carga y la masa, sumando una constante infinita para tener respuestas físicas razonables, es un mero artificio de la extraña idea del «mar» que Dirac consideró necesario introducir. Sin embargo, descubriremos que estas dos características no son ni mucho menos específicas del extraordinario «mar» de Dirac. Parecen ser ubicuas en todas las aproximaciones serias a una teoría realista de la física de partículas, al menos tal como hoy están dichas aproximaciones. El modelo estándar hace un uso fundamental de la renormalización y de los vacíos alternativos. Lejos de ser una anomalía de la historia, el mar de Dirac es un tipo de modelo que debemos tener presente cuando tratamos de avanzar, al menos dentro del esquema de cosas de que hoy disponemos. Aunque satisfechos de forma incompleta, los dos criterios gemelos, basados en la observación y la consistencia matemática, han tomado una ruta que hasta ahora ha sido dependiente de las ideas de renormalización y vacíos no únicos.

26.6. Interacciones: lagrangianos e integrales de camino

Nos hemos visto llevados en estas direcciones por las dificultades que surgen cuando intentamos tratar *interacciones* dentro del marco de la QFT. De hecho, hasta ahora mi exposición se ha centrado básicamente solo en el caso de campos libres; y aunque no he examinado todos los detalles, espero que el lector me crea cuando digo que las cosas si-

guen de una forma esencialmente libre de dificultades cuando no hay interacciones. Pueden construirse estados en los que existen superposiciones de números diferentes de partículas y antipartículas, e incluso números ilimitados de tales partículas. Estos estados se obtienen actuando sobre $|0\rangle$ con un elemento arbitrario de $\mathcal{A}$, i.e., una expresión en operadores creación y aniquilación (un polinomio o una serie de potencias, prestando la debida atención a las cuestiones de convergencia en el último caso). El espacio de tales estados se conoce como *espacio de Fock* (por el físico ruso V.A. Fock, que fue uno de los primeros en estudiar estas cosas), y puede considerarse como lo que se denomina una *suma directa*[12] (véase §13.7) de espacios de Hilbert con un número creciente de partículas. El número de partículas en un estado puede ser ilimitado, como sucede con los *estados coherentes*, que son, en cierto sentido bien definido,[13] los más «clásicos» de entre los estados de campo cuántico. Son estados de la forma

$$e^{\boldsymbol{\Xi}}|0\rangle,$$

donde $\boldsymbol{\Xi}$ es el *operador campo* asociado a la configuración de campo concreta $\boldsymbol{F}$ (que tomaremos como un campo de Maxwell real libre con propiedades apropiadas de decaimiento en el infinito para asegurar que existe una norma finita). Definimos $\boldsymbol{\Xi}$ como la suma de los operadores creación y aniquilación (no normalizados) correspondientes a las partes de frecuencia positiva y frecuencia negativa de $\boldsymbol{F}$, respectivamente.

Recordemos de §26.2 que los operadores de creación y aniquilación satisfacen ciertas relaciones de conmutación. Se sigue de ello que, en general, las diversas componentes de los operadores de campo no conmutan entre sí. Por ejemplo, en el caso del campo electromagnético las componentes del operador que define el campo magnético $\mathbf{B}$ y las que definen el potencial electromagnético $\mathbf{A}$ (véanse §§19.2,4) satisfacen reglas de conmutación *canónicas* (como las que existen entre la posición y el momento de una partícula; véase §21.2).[14] Se sigue que las relaciones de incertidumbre de Heisenberg (véase §21.11) deben ser válidas entre estas cantidades, lo que impone un límite a la precisión con que pueden medirse simultáneamente.

¿Cómo tratamos entonces las interacciones? Fundamental para el

armazón general de la QFT moderna es el *lagrangiano* (véase §20.1), que en muchos aspectos resulta mucho más apropiado que un hamiltoniano cuando estamos interesados en una teoría relativista. Como recordamos de §21.2, §23.1 y el capítulo 24, los procedimientos de cuantización Schrödinger/hamiltoniano estándar no se llevan muy bien con la simetría espaciotemporal de la relatividad. Sin embargo, a diferencia del hamiltoniano, que está asociado con una elección de coordenada temporal, el lagrangiano puede tomarse como una entidad relativísticamente invariante (véase §20.5). ¿Cómo construimos una QFT partiendo de un lagrangiano? La idea básica, como tantas de las ideas subyacentes al formalismo de la teoría cuántica, se remonta a Dirac,[15] aunque la persona que la introdujo como base para la teoría cuántica relativista fue el brillante físico norteamericano Richard Feynman.[16] Por consiguiente, se conoce habitualmente como la formulación en términos *de integrales de camino de Feynman* o *suma sobre historias* de Feynman. Es también la base de los diagramas de Feynman que hemos visto en el capítulo 25.

La idea básica consiste en una perspectiva diferente sobre el principio mecanocuántico fundamental de la superposición lineal compleja que hemos encontrado antes, y que se ha hecho particularmente explícito en §22.2. Aquí consideramos dicho principio aplicado no solo a estados cuánticos específicos, sino a historias espaciotemporales enteras. Estas historias son «posibles trayectorias clásicas alternativas» (en el espacio de configuración). La idea es que en el mundo cuántico, en lugar de haber solo una «realidad» clásica representada por una trayectoria semejante (una historia), existe una gran *superposición compleja* de todas estas «realidades alternativas» (historias alternativas superpuestas). Según esto, debe asignarse a cada historia un factor de ponderación complejo, al que llamamos una *amplitud* (§22.5) si el total se normaliza a módulo unidad, de modo que el módulo al cuadrado de una amplitud nos da una probabilidad. Normalmente estamos interesados en amplitudes para ir de un punto a a un punto b en el espacio de configuración.

El papel mágico que desempeña el lagrangiano consiste en decirnos qué amplitud hay que asignar a cada una de estas historias; véase la Fig. 26.2. Si conocemos el lagrangiano $\mathcal{L}$, entonces podemos obtener

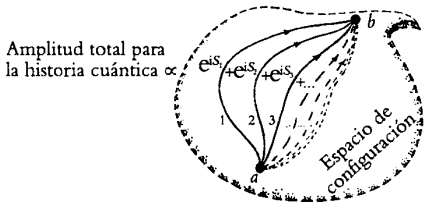


Fig. 26.2. En la aproximación de integrales de camino a la teoría cuántica y en la QFT consideramos superposiciones cuánticas de historias clásicas alternativas, siendo una historia una trayectoria en el espacio de configuración, tomada aquí entre puntos fijos *a* y *b*. La amplitud asignada a un camino semejante es $e^{iS/\hbar}$ (multiplicado por una constante determinada), donde la acción *S* es la integral del lagrangiano a lo largo del camino, como en §20.1 (Fig. 20.3). La amplitud total que se obtiene de *a* a *b* es la suma de estas.

la *acción* S para dicha historia (pues la acción es la integral de $\mathcal{L}$ para esa historia clásica, según la receta dada en §20.5; véase la Fig. 20.3). La amplitud compleja que debe asignarse a dicha historia particular viene entonces dada por la fórmula engañosamente simple

$$\text{amplitud} \propto e^{iS/\hbar}.$$

La engañosa simplicidad de esta fórmula se debe en parte a que la «amplitud» no es aquí realmente un número complejo (que, tal como está escrita, debería tener módulo unidad), sino un tipo de densidad. Si tuviéramos solo una familia discreta de historias clásicas alternativas, numeradas 1, 2, 3, 4, ..., por ejemplo, entonces cabría imaginar que a la *n*-ésima historia podría asignársele como amplitud un genuino número complejo α_n, cuyo módulo al cuadrado $|\alpha_n|^2$ podría interpretarse como la probabilidad de dicha historia, de acuerdo con las reglas de la medida cuántica (§22.5); y deberíamos normalizar, para tener $\sum |\alpha_n|^2 = 1$, con la suma extendida sobre todas las alternativas clásicas, para dar la probabilidad total 1. Pero aquí tenemos un infinito *continuo* de alternativas clásicas. Nuestra «amplitud» anterior debe así ser considerada como una «densidad de amplitud», y necesitamos algo como $\int |\alpha(X)|^2 dX = 1$, donde tenemos que integrar sobre el espacio de estados clásicos para conseguir nuestro requisito de que la probabilidad total resulte ser 1. Esto no sería especialmente molesto; hemos hecho cosas de este tipo

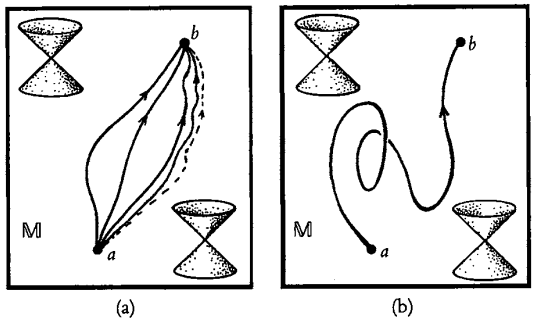


Fig. 26.3. (a) En el caso de una única partícula sin estructura, una historia clásica es una curva en el espaciotiempo (aquí el espacio de Minkowski 𝕄) tomada entre sucesos fijos *a*, *b*. (b) La curva no tiene por qué ser una línea de universo suave admisible clásicamente, con tangentes siempre de género tiempo futuro; puede ir hacia atrás y hacia delante en el tiempo.

en §21.9, en el caso de una función de onda, para la mecánica cuántica ordinaria de una partícula puntual (donde $|\psi(\boldsymbol{x})|^2$ nos daba la densidad de probabilidad de encontrar la partícula en el punto $\boldsymbol{x}$). Pero las malas noticias aquí son que el «espacio de caminos clásicos» resultará ser casi con certeza de *dimensión infinita*. Es un problema de un orden de magnitud diferente dar sentido a las diversas cantidades involucradas —y asegurar que se obtengan respuestas finitas al final— cuando tenemos que definir todas las cosas que necesitamos para trabajar en un espacio de dimensión infinita.

La ilustración más accesible de una integral de camino es el caso de una única partícula puntual moviéndose en un campo de fuerzas (de modo que el espacio de configuración es ahora el propio espacio). Aquí consideramos todas las diversas historias que parten de un punto espaciotemporal *a* y acaban en otro punto espaciotemporal *b*, como en la Fig. 26.3a. Estas historias se consideran caminos espaciotemporales continuos que van de *a* a *b*. No exigimos que el camino sea «legal» de acuerdo con las reglas de la relatividad especial (i.e., que esté obligado a yacer dentro de los conos de luz, como exige la relatividad clásica; véase §17.8), ni siquiera exigimos que todo el camino vaya hacia el futuro. ¡La «historia» puede oscilar hacia arriba y hacia abajo en el tiempo si se quie-

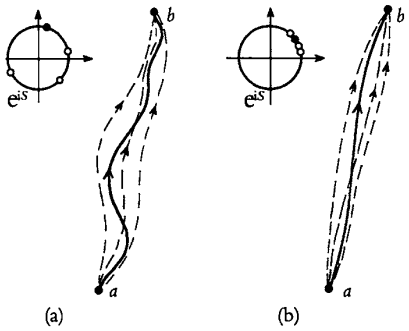


Fig. 26.4. El «principio de Hamilton» cuántico. (a) Una historia para la que S no es estacionaria (y es grande comparado con $\hbar$). Los valores de $e^{iS/\hbar}$ para historias próximas tienden a variar mucho alrededor del círculo unidad, y por consiguiente hay mucha cancelación en la suma. (b) Una historia para la que S es estacionaria (y grande). Para historias vecinas, los valores de $e^{iS/\hbar}$ no cambian mucho, de modo que la cancelación es mucho menor.

re! (Fig. 26.3b)[26.8] Supongamos que tenemos un lagrangiano $\mathcal{L}$, que describe (de acuerdo con §20.1) la energía cinética de la partícula menos la energía potencial debida al campo de fuerzas. Para cada historia habrá una acción S, donde S es la integral del lagrangiano a lo largo del camino (recordemos la Fig. 20.3). En mecánica clásica, nuestro amigo Joseph L. Lagrange nos habría dicho que buscásemos una historia particular para la cual la integral de acción fuera estacionaria (principio de Hamilton; véase §20.1), lo que nos daría el movimiento de la partícula real compatible con el movimiento clásico bajo la fuerza dada. En el enfoque de integrales de camino de la mecánica cuántica vamos a adoptar un punto de vista diferente. Se supone que *todas* las historias «coexisten» en superposición cuántica, y a cada historia se le asigna una amplitud $e^{iS/\hbar}$. ¿Cómo vamos a tomar contacto con el requisito de Lagrange, quizá solo en un sentido aproximado, según el cual debería haber una historia particular singularizada para la que la acción es estacionaria?

La idea es que las contribuciones de aquellas historias dentro de nuestra superposición que están muy alejadas de una historia de «acción estacionaria» se cancelarán prácticamente con las contribuciones de historias vecinas (Fig. 26.4a). Esto se debe a que los cambios que resultan en S cuando varía la historia producirán ángulos de fase $e^{iS/\hbar}$ que se distribuyen sobre toda la circunferencia del reloj, y por ello se cancelarán en promedio. (Esto se aplica, en particular, a las contribuciones muy «no físicas» procedentes de las historias exageradamente acausales de la Fig. 26.3b). Solo si la historia está muy próxima a una para la cual la acción es estacionaria (así dice el argumento), su contribución empezará a verse reforzada por las de sus vecinas, en lugar de ser cancelada por ellas (Fig. 26.4b), porque en este caso habrá un gran agrupamiento de ángulos de fase en la misma dirección.[26.9]

Esta es una idea muy bella. De acuerdo con la filosofía de «integral de camino», no solo deberíamos obtener la historia clásica como la mayor contribuyente a la amplitud total —y, por consiguiente, a la probabilidad total—, sino también *correcciones cuánticas* menores a este comportamiento clásico, que proceden de historias que no son del todo clásicas y dan contribuciones que no se cancelan completamente y que a menudo pueden ser experimentalmente observables. Aunque mis descripciones anteriores han sido expresadas en términos de una partícula puntual que se mueve en un campo de fuerzas, las ideas se aplican con toda generalidad, y pueden aplicarse a la dinámica de campos tanto como al movimiento de partículas. Una vez más, las «historias de campos» que representan soluciones clásicas de las ecuaciones de campo deberían emerger como aquellas que proporcionan las contribuciones principales, y habrá también correcciones cuánticas que provengan de las historias casi clásicas.

26.7. Integrales de camino divergentes: la respuesta de Feynman

Al menos eso es lo que *se supone* que sucede. Pero ¿lo hace? ¿Están justificadas matemáticamente las crudas descripciones que he presentado antes? Incluso si no lo están, y seguimos adelante ignorando sutilezas matemáticas, ¿obtenemos buenas respuestas físicas que estén de acuerdo con el experimento?

[26.8] Dé una «interpretación física» de la historia de la Fig. 26.3b en términos de creación y aniquilación de partículas.

[26.9] Trate de hacer estos enunciados un poco más precisos con referencia a cambios de primer orden en el camino, utilizando símbolos «O» (como en §14.5), y relacionándolo con la exposición dada en §20.1 concerniente al significado de «acción estacionaria». (Suponga que S es grande en unidades de $\hbar$.)

Solo puedo dar respuestas muy contradictorias a estas preguntas. La cuestión de la justificación matemática es especialmente molesta, y la respuesta más justa que se debería dar en este punto sería: «No; no tal como están hoy las cosas». Incluso el caso de una única partícula puntual, que se ha descrito antes, es decididamente problemático. El espacio de caminos es ciertamente de dimensión infinita,[26.10] y se requiere una «medida» adecuada (la versión en dimensión infinita de un volumen) para manejarlo. Resulta que esta medida está fuertemente sesgada en favor de historias que no son siquiera suaves, así que tenemos que preocuparnos por lo que el lagrangiano significa siquiera en tales circunstancias. Todo diverge, tal como está la definición.

Estas divergencias son ciertamente graves desde el punto de vista matemático, y quizá prefiramos recurrir a la filosofía «euleriana» que en §4.3 nos ha llevado a especular sobre el sentido en el que podríamos confiar en la suma «absurda»

$$1 + 2^2 + 2^4 + 2^6 + 2^8 + \ldots = \frac{1}{3},$$

obtenida al sustituir $x = 2$ en $1 + x^2 + x^4 + x^6 + x^8 + \ldots = (1 - x^2)^{-1}$. De hecho, el enfoque de integrales de camino parece depender por completo de una fe en que las expresiones incontroladamente divergentes que se presentan tienen realmente (como la expresión divergente anterior) un significado «platónico» más profundo que aún no percibimos adecuadamente. Parece que nos vemos obligados a admitir que debe de haber algo de esta naturaleza, porque, desde el lado físico, se nos presentan con cierta frecuencia respuestas de extraordinaria precisión cuando (si se me permite evocar una inverosímil metáfora) ¡desbrozamos nuestro camino a través de las matemáticas con gran sensibilidad y precisión! Por ejemplo, estos procedimientos de cálculo dan el factor de corrección 1,001 159 652 188, mencionado en §24.7, al valor original de Dirac para el momento magnético del electrón, lo que proporciona un acuerdo entre teoría y observación[17] con una discrepancia menor que 10^{-11}.

Es realmente notable lo lejos que puede llevarnos la aplicación de la sensibilidad matemática/física, hasta respuestas excelentes en muchos

[26.10] ¿Por qué?

casos. Un valioso primer paso para dar sentido a estas integrales de camino,[18] en el caso de partículas cuánticas individuales, es el reemplazamiento de la colección incontrolada de historias por lo que se denomina el *propagador de Feynman*.[19] Esto nos da la interpretación matemática de una de las líneas en un diagrama de Feynman (tal como los que se encuentran en el capítulo 25).

Más concretamente, consideremos una suma sobre historias para la que cierta partícula libre va a empezar en un punto p y terminar en algún otro punto q en el espaciotiempo. En principio, tenemos que hacer la suma (integral) de todos los $e^{iS/\hbar}$ para caminos que empiezan en p y terminan en q, pero esto es por supuesto incontroladamente divergente, tal como está. Por el contrario, podemos suponer que la suma $K(p, q)$ tiene un tipo de existencia matemática («euleriana/platónica»), y preguntamos qué propiedades formales algebraicas y diferenciales debería tener esta suma, si existiera. Estas propiedades (incluyendo una adecuada condición de «frecuencia positiva»; véase §24.3) determinan unívocamente la forma de $K(p, q)$ (si somos razonablemente afortunados en el ejemplo que hemos escogido), y esto nos da el propagador de Feynman que buscamos. De hecho, es más habitual (aunque en modo alguno esencial)[20] describir estas cosas en el espacio de momentos más que en el espacio de posición, pues las descripciones en el espacio de momentos son significativamente más simples.

En el caso de una partícula de Dirac (por ejemplo, un electrón), el propagador en el espacio de momentos toma la forma $i(\not{P} - M + i\varepsilon)^{-1}$, donde $\not{P} = \boldsymbol{\gamma}^a P_a$ (véanse §§24.6,7), siendo la cantidad P_a el 4-momento que tiene la partícula para el camino escogido bajo consideración. La cantidad «ε» se toma como un número real positivo muy pequeño, lo que es un artificio dispuesto para garantizar los requisitos de frecuencia positiva/negativa del propagador de Feynman. Resulta que en el límite $\varepsilon \to 0$ tenemos una singularidad en el propagador —un valor infinito— cuando la «masa en reposo» $(P_a P^a)^{1/2}$ que resulta tener la partícula para ese camino escogido toma el valor M, que es la masa en reposo real observada de la partícula.[26.11] Para una partícula clásica,

[26.11] Explique cómo aparece esta singularidad, reescribiendo primero $(\not{P} - M + i\varepsilon)$ como un cociente en el que el denominador es $P_a P^a - M^2 - \varepsilon^2$.

exigiríamos que esta «masa en reposo» sí tome este valor, i.e., que $P_a P^a = M^2$; pero con la suma mecanocuántica sobre historias debemos admitir que la partícula experimente valores del momento para los que la masa en reposo resulta «errónea». Sin embargo, debido a la singularidad que acabamos de mencionar, encontramos que la amplitud se hace muy grande cuando $P_a P^a$ se hace muy próxima al valor M^2, de modo que el valor clásico para la masa da la contribución dominante. Esta es una característica que no es específica de una partícula de Dirac, sino que se aplica con toda generalidad.

26.8. Construyendo diagramas de Feynman; la matriz S

Lo que se ha descrito en la sección precedente es el primer paso para obtener un diagrama de Feynman. Esto necesita alguna explicación adicional. Lo que hemos encontrado es una única línea (o segmento de línea) de un diagrama semejante. Una línea concreta en un diagrama de Feynman sería normalmente tan solo una parte de una expresión complicada, que incluye otras líneas de partículas y varios vértices donde se juntan las líneas. Las contribuciones a la amplitud total procedentes de los *vértices* suelen ser[21] factores sencillos, que incluyen una constante de acoplamiento escalar (tal como la carga eléctrica) que gobierna la intensidad de la interacción, quizá un término tal como γ_a, que es necesario para «hacer que los índices empalmen», y un término «función delta» (§9.7) para asegurar que las únicas contribuciones no nulas a la amplitud total aparecen cuando en cada vértice hay conservación del 4-momento.[22] Habrá términos de varios tipos que aparecen a partir de las diferentes variedades de líneas en el diagrama (dependiendo del valor del espín y de la masa en reposo de la partícula que la línea representa). Los *infinitos* en la expresión (aparte de aquellos de los términos función delta, que normalmente se considera que solo proporcionan ligaduras que aseguran la conservación del 4-momento) surgen cuando el momento P_a adquiere los valores concretos que se esperan para caminos clásicos (por ejemplo, $P_a P^a = M^2$). Esto tiene sentido porque esperamos que el comportamiento clásico domine la integral de camino. La presencia de estas singularidades (los infinitos aparte de los de las funciones delta) está así íntimamente relacionada con el requisito de que el comportamiento clásico proporcione, en un sentido aproximado, la contribución principal a la amplitud mecanocuántica. Pese a todo, hay un peligro que acecha en estas singularidades, como veremos enseguida.

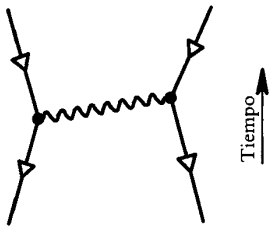


Fig. 26.5. Dispersión de Möller de electrones: la manifestación cuántica más primitiva de la fuerza electrostática (de Coulomb) entre dos partículas cargadas. La fuerza electrostática proviene aquí del «intercambio» de un único fotón (línea ondulada) entre dos electrones. El fotón está necesariamente «fuera de la capa» y por consiguiente es virtual, como se sigue de la conservación del 4-momento en cada vértice.

Solo para subrayar la necesidad de estas expresiones singulares, debería señalar que no podemos considerar la condición $P_a P^a = M^2$ como una ligadura (similar a la conservación de momento en los vértices) debido a la existencia de procesos básicos tales como el de la Fig. 26.5 en el que dos electrones «intercambian un fotón» (fotón que está indicado por una línea ondulada, como en §§25.3-5). Esta es la manifestación mecanocuántica básica de la repulsión electrostática (coulombiana) entre las dos partículas cargadas negativamente (dispersión de Möller). Las dos líneas entrantes (en la parte inferior del diagrama) representan los dos electones en su estado inicial, y las dos líneas salientes (en la parte superior[23] del diagrama) representan los electrones en su estado final. Estos se toman como «cosas dadas» —que proporcionan los momentos *externos*— y no deben «integrarse» al calcular la amplitud final.

Para estos estados externos (y solo para estos), tconsideramos que los momentos satisfacen la relación clásica $P_a P^a = M^2$. Cuando se satisface esta relación decimos que la masa de una partícula está *en la capa*, siendo esta la «capa de masas», que es la versión en el espacio de momentos del hiperboloide en forma de cuenco mostrado en la Fig. 18.7. Véase la Fig. 26.6. Las partículas reales (las que se observan realmente como partículas libres) están siempre en la capa. Sin embargo, con respecto a las líneas *internas* de un diagrama de Feynman, no esperamos

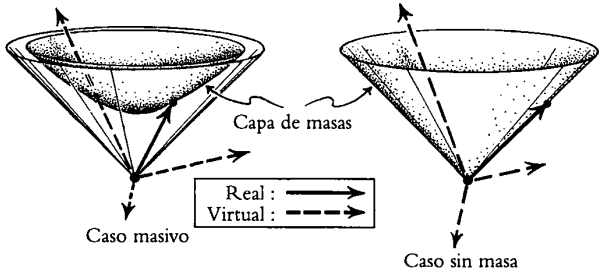


Fig. 26.6. La capa de masas en el espacio de momentos. (Compárese con las Figs. 18.7 y 18.17.) Para partículas reales (libres) de masa en reposo M, el 4-momento p^a yace en la capa de masas (de modo que p^a es de género tiempo futuro o nulo futuro con $p_a p^a = M^2$), pero en el interior de un diagrama de Feynman puede haber partículas virtuales «fuera de la capa».

que se cumpla este requisito de estar en la capa. En particular, el fotón intercambiado en el diagrama de Feynman de la Fig. 26.5 no puede estar en la capa (i.e., su 4-momento no satisface $P_a P^a = M^2$) cuando quiera que haya una interacción no trivial.[26.12] Tales partículas fuera de la capa se conocen como partículas *virtuales*, y solo pueden ocurrir en el interior de un diagrama de Feynman. El fotón intercambiado de la Fig. 26.5 es virtual, y no puede «escapar» para ser observado a grandes distancias.

El proceso de la Fig. 26.5 es bastante especial en cuanto que el estado de la línea interna (fotón virtual) está completamente determinado por las líneas externas. En este caso, la «integración sobre estados internos» que generalmente se requiere es completamente trivial, pues consiste en un único término. Sin embargo, en procesos más complicados, tales como los que se muestran en las Figs. 26.7a,b, donde se intercambian *dos* fotones, hay cierta libertad en los 4-momentos de las líneas internas.[26.13] La idea es que en tales casos (y en multitud de casos aún más complicados) se supone que integramos sobre todas las posibilidades para los momentos permitidos para las líneas internas, y también suma-

[26.12] ¿Por qué no? Explique cómo la conservación del 4-momento en cada vértice determina el 4-momento del fotón virtual. *Sugerencia*: ¡Todos los electrones tienen la misma masa en reposo!

[26.13] ¿Cuál es esta libertad?

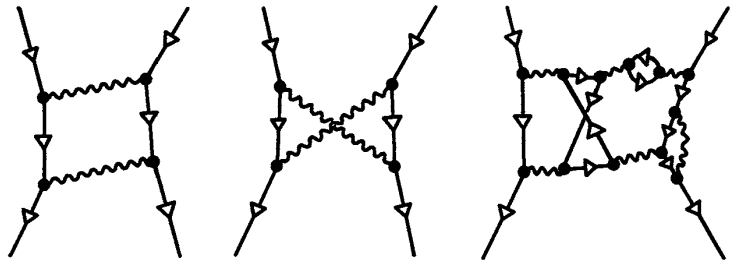


Fig. 26.7. Procesos de orden superior que dan correcciones a la dispersión Möller. (a) y (b) muestran intercambios de dos fotones entre dos electrones, mientras que (c) es un proceso de orden mucho más alto que incluye creaciones y aniquilaciones internas de pares. Cada diagrama semejante de Feynman representa una integral, y hay que sumar las contribuciones de todos ellos.

mos todas las contribuciones diferentes para todas las posibles «topologías de diagramas de Feynman» que son compatibles con las líneas externas dadas con sus momentos dados. (Una «topología» se refiere simplemente a una de las varias formas diferentes en las que puede conectarse un diagrama de Feynman, sin considerar los valores concretos del 4-momento que se han asignado a las líneas en el diagrama.)

Este proceso va a darnos la amplitud total para el conjunto particular de momentos «entrantes» y «salientes» que se han especificado como «dados». El conjunto de amplitudes para los diversos estados iniciales y finales posibles constituye un tipo de matriz (aunque de dimensión infinita) cuyas «filas» y «columnas» corresponden a una base para estados finales e iniciales, respectivamente. Esta matriz se conoce como la *matriz de dispersión*, o más habitualmente como la *matriz S*. El cálculo de la matriz S se considera un objetivo principal de la QFT.[24]

Este procedimiento supone una enorme mejora, en términos de cálculo, sobre la «suma sobre historias» original, porque ya hemos hecho efectivamente las integrales de camino de dimensión infinita (y en apariencia desesperadamente divergentes) que corresponden a cada línea en el diagrama por separado. Cada elección de topología de diagramas de Feynman representa una integral ordinaria en dimensión finita (como aquellas consideradas en §12.6), y esto es un gran avance respecto a las integrales en dimensión infinita incontroladamente di-

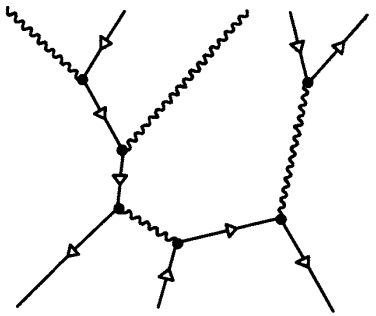


Fig. 26.8. Un diagrama de árbol no contiene lazos. Los momentos internos están determinados, en consecuencia, por los momentos externos, de modo que no se incluye ninguna integración. Los diagramas de árbol reproducen la teoría clásica.

vergentes a las que nos llevaría la interpretación directa de una integral de camino. Además, estas integrales en dimensión finita pueden tratarse mediante los potentes métodos de la integral de contorno compleja (como se ha discutido en §7.2). El parámetro de Feynman ε, que aparece en el propagador (véase el último párrafo de §26.7), es en realidad tan solo una receta para guiar la integración de contorno por el lado adecuado de las singularidades que aparecen en las expresiones.

Pese a todo, aún nos falta mucho para estar «fuera de peligro», porque la integral de dimensión finita que nos ha quedado, para cada topología de diagrama de Feynman, va a ser ella misma divergente cuando quiera que haya *lazos cerrados* en el diagrama de Feynman. Esto podría ser una «mala noticia»; en efecto, solo con lazos cerrados es cuando empezamos a entender la realización de cualquier integración. En todos los demás casos (i.e., los denominados «diagramas de árbol», que no tienen lazos cerrados; véase la Fig. 26.8), los momentos internos están simplemente determinados por los valores externos. ¡Los diagramas de árbol reproducen solo la teoría *clásica*!

26.9. Renormalización

Así pues, parece que pese a todos nuestros esfuerzos (o, más bien, los de Feynman), aún estamos atascados en una expresión divergente para la amplitud total de cualquier proceso genuinamente cuántico. El lector fatigado puede estar preguntándose legítimamente en este momento qué tiene todo esto de bueno para nosotros. De hecho, desde un punto de vista matemático estricto, hemos llegado oficialmente «a ninguna parte», en el sentido de que todas nuestras expresiones son aún «matemáticamente carentes de significado» (como lo era el $1 + 2^2 + 2^4 + + 2^6 + \ldots = -1/3$ de Euler). Pero los buenos físicos no abandonan tan fácilmente. Y tenían razón en no hacerlo. Sus esfuerzos se vieron finalmente recompensados[25] cuando se vio que en el caso de la QED (electrodinámica cuántica: la teoría de las interacciones de electrones, positrones y fotones) todas las partes divergentes de los diagramas de Feynman podían reunirse en varios «paquetes», de modo que podía considerarse que los infinitos proporcionaban simplemente factores de «reescalamiento» que pueden ignorarse, de acuerdo con un proceso conocido como *renormalización* (ya apuntado antes en §26.5).

Estos infinitos concretos aparecen porque los diagramas de Feynman dan integrales que divergen cuando los valores del momento se hacen ilimitadamente grandes, o, de forma equivalente, cuando las distancias se hacen ilimitadamente pequeñas. (Recordemos la relación de incertidumbre de Heisenberg $\Delta_p \Delta_x \geq \frac{1}{2}\hbar$; véase §21.11.) Los infinitos se conocen como divergencias *ultravioletas*. Aunque no son las únicas divergencias en QFT, se consideran las más graves. Hay también divergencias *infrarrojas* que podemos considerar que proceden de distancias indefinidamente grandes (i.e., de momentos indefinidamente pequeños). Estas se consideran normalmente «curables» por diversos medios, a menudo restringiendo el tipo de pregunta que se considera físicamente razonable para hacer a un sistema.

Para hacernos una idea de lo que aquí está implícito, examinemos el significado físico del que es el más claro ejemplo de renormalización. Este ocurre con el valor de la carga eléctrica que posee el electrón. Imaginemos un electrón como una carga puntual situada en algún punto E del espacio. Existe un efecto conocido como *polarización del vacío* que puede entenderse de la forma siguiente. Supongamos que en algún punto próximo a E pudiera darse la creación de un par (virtual) de partículas: un electrón y un positrón, que al cabo de un período de tiempo muy corto se aniquilan mutuamente. (Consideremos que este período de tiempo es corto, de modo que la energía requerida para la producción del par está dentro de las incertidumbres

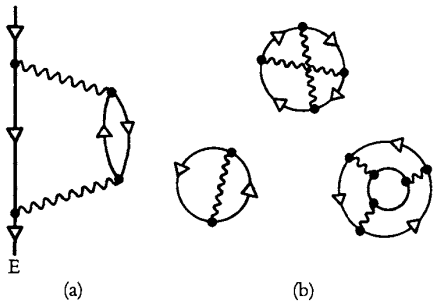


Fig. 26.9. (a) Diagrama de Feynman implicado en la renormalización de carga. Este representa una creación de un par positrón-electrón y la posterior aniquilación en el campo de un electrón de fondo (véase la Fig. 26.10). (b) Diagramas de Feynman totalmente desconectados. Se considera que estos no tienen efectos directamente observables.

de la relación energía-tiempo de Heisenberg $\Delta E\, \Delta t \geq \frac{1}{2}\hbar$; [§21.11].)

El diagrama de Feynman para este proceso se indica en la Fig. 26.9a. La presencia de la línea del fotón (virtual) en el comienzo (y también en el final) de este proceso indica que la creación (y posterior aniquilación) ocurre en el campo eléctrico ambiente del electrón en E. (También podríamos contemplar «lazos» de Feynman completamente desconectados; véase la Fig. 26.9b, donde los procesos de creación y aniquilación tienen lugar sin la presencia del campo ambiente del electrón en E; pero se considera que tales procesos «totalmente desconectados» no tienen efectos físicamente observables.) El efecto de este campo ambiente es que el electrón creado es ligeramente repelido por el electrón en E, mientras que el positrón creado es ligeramente atraído por él, de modo que hay una ligera separación física entre estas cargas durante su momentánea existencia. Esto está sucediendo continuamente en torno al electrón en E, y el efecto neto, conocido como «polarización del vacío», consiste en reducir[26] el valor aparente de la carga de dicho electrón, tal como se mide por su efecto sobre otras cargas; véase la Fig. 26.10.[26.14] El vacío sirve para «apantallar» la carga del electrón, y hace que parezca tener un valor más pequeño —llamado valor de la carga *vestida*— que el valor «real» de la carga *desnuda* del electrón. Es el valor vestido el que se mediría directamente en experimentos físicos.

[26.14] ¿Puede ver por qué debería ser así?

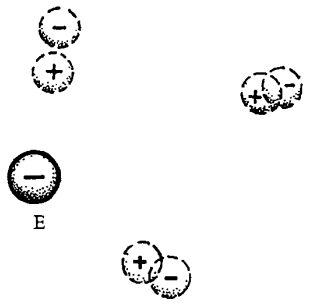


Fig. 26.10. Polarización del vacío: la base física de la renormalización de carga. El electrón E induce una ligera separación de carga en pares virtuales electrón-positrón creados momentáneamente a partir del vacío. Esto reduce algo la carga efectiva de E respecto a su valor desnudo, desgraciadamente en un factor infinito, según el cálculo directo.

Todo esto parece muy razonable. Pero el problema es que el factor numérico calculado por el que hay que escalar el valor desnudo, con respecto al valor vestido, ¡resulta ser infinito! Este infinito puede ser identificado con claridad como uno de los infinitos en el cálculo electrodinámico cuántico (básicamente diagramas como el de la Fig. 26.9a y desarrollos del mismo). Se puede adoptar el punto de vista de que, de acuerdo con alguna teoría futura, las integrales divergentes deberían ser reemplazadas por algo finito, quizá porque hay un «corte» que interviene a distancias muy pequeñas, i.e., a momentos muy grandes (§21.11), y el factor de renormalización correcto debería ser algún número finito bastante grande, antes que ∞. (De hecho, en términos de las «unidades naturales» a las que llegaremos más adelante —en §27.10—, la carga vestida medida del electrón es de aproximadamente 0,0854, y es tentador imaginar que el valor de la carga desnuda debería ser 1, por ejemplo. Esto correspondería a un factor de escala de 11,7062, o aproximadamente $\sqrt{137}$, en lugar de ∞.) Otro punto de vista es considerar que la carga desnuda no es más que una conveniencia conceptual, y adoptar la postura de que la noción de «carga desnuda» es realmente «carente de significado», porque es «inobservable».

Cualquiera que sea la posición filosófica que se adopte sobre esta cuestión, la renormalización es un aspecto esencial de la QFT moderna. De hecho, tal como están las cosas, no hay ninguna forma aceptada de obtener respuestas finitas sin un procedimiento semejante de «reescalamiento infinito» aplicado no necesariamente solo a la carga, o a la masa, sino también a otras magnitudes. Las teorías en las que funciona este tipo de procedimientos se denominan *renormalizables*.

En una QFT renormalizable es posible reunir todas las partes divergentes de los diagramas de Feynman en un número finito de «paquetes»[27] que pueden ser «escalados» mediante la renormalización, estimando que cualesquiera expresiones divergentes remanentes se *cancelarán* mutuamente de acuerdo con ciertos principios generales (tales como los principios de simetría que desempeñan un papel importante en el modelo estándar). La QED es una teoría renormalizable, y así lo es el modelo estándar en conjunto. La mayoría de las QFT, por el contrario, son *no* renormalizables. Un punto de vista común entre los físicos de partículas es considerar la renormalizabilidad como un principio de selección para las teorías propuestas. Por consiguiente, cualquier teoría no renormalizable sería automáticamente rechazada como inadecuada para la naturaleza. De hecho, este principio ha proporcionado una poderosa guía hacia la elección concreta de teoría que ha llegado a ser el modelo estándar de la física de partículas en el siglo XX que hemos encontrado en el capítulo 25. Así pues, desde este punto de vista, la predominancia de infinitos en las QFT no es algo «malo» en absoluto, sino que es una característica que puede volverse poderosamente en nuestro favor.[28] Muy pocas teorías superan el test de la renormalizabilidad, y solo aquellas que sí lo superan tienen una oportunidad de ser consideradas aceptables para la física.

Pese a todo, no todos los físicos suscriben en rigor esta postura. Incluso el premio Nobel Gerard 't Hooft, que proporcionó el ingrediente clave para demostrar la renormalizabilidad del modelo estándar, ha expresado ciertas reservas sobre la estricta adhesión a la renormalizabilidad. (En 1971, mientras aún era un estudiante de doctorado en la Universidad de Utrech, 't Hooft conmocionó a la comunidad física al demostrar la renormalizabilidad de las teorías donde hay una simetría «espontáneamente rota», que se convirtió en una característica esencial de la teoría electrodébil.) En cierta ocasión me expresó su punto de vista de que la importancia de la renormalizabilidad de una teoría depende del tamaño de la constante de acoplamiento en la interacción que se considera. Mencionó concretamente la gravedad, que es extraordinariamente débil comparada con las fuerzas de la física de partículas, pese a lo cual su teoría cuántica resulta ser no renormalizable según los enfoques estándar para la cuantización de las ecuaciones de Einstein

de la relatividad general; véanse §19.6 y §31.1. (La atracción gravitatoria entre el electrón y el protón en un átomo de hidrógeno es menor que la fuerza eléctrica en un factor del orden de 10^{-40}, lo que hace la gravedad algo incomparablemente más débil que las «interacciones débiles» de la desintegración radiactiva.) Sus comentarios expresan lo que podría llamarse una visión *pragmática* de la QFT. Ni siquiera las teorías renormalizables están libres de infinitos, un punto que desarrollaré enseguida. Su idea era cuestionar si los infinitos, potencialmente presentes en una teoría, son en realidad físicamente relevantes a energías siquiera remotamente accesibles al experimento. En el caso de una «gravedad cuantizada», las energías extremas relevantes están absurdamente más allá de lo que es factible, y muchas otras incertidumbres de la teoría física entrarían en escena mucho antes de que la no renormalizabilidad de la gravedad empezara a dejar su huella.

En el otro extremo de la escala, argumentaba él, tenemos las interacciones *fuertes*, con una constante de acoplamiento tan grande que es dudoso que sea completamente útil una descripción solo en términos de diagramas de Feynman. La renormalizabilidad por sí sola es insuficiente para asegurar que la cromodinámica cuántica pueda suministrar respuestas finitas. En este caso se saca provecho de lo que se denomina la *libertad asintótica* de la fuerza fuerte. Para momentos muy grandes —que en teoría cuántica equivalen a distancias minúsculas—, la fuerza fuerte tiene la notable propiedad de que efectivamente desaparece. Esto contrasta con la familiar fuerza electrostática, o la gravitatoria, entre partículas, donde la ley de la inversa del cuadrado nos dice que la fuerza aumenta cuando la distancia se hace más pequeña. La fuerza fuerte se parece más a una goma elástica, donde la intensidad de la fuerza aumenta en proporción a la distancia, y cae a cero cuando la distancia se hace cero.[29] Esta ley de fuerzas es responsable del hecho —conocido como *confinamiento* (véanse §§25.7,8)— de que los quarks no pueden ser extraídos individualmente de un hadrón. A diferencia de una goma elástica ordinaria, la fuerza fuerte no puede «romperse», aunque si se tira de ella lo suficiente otras entidades tales como los antiquarks o los pares de quarks pueden extraerse también del vacío, que es lo que ocurre con los «chorros» que pueden ocurrir en aceleradores de partículas. Esta notable propiedad de la libertad asintótica es lo que salva a

la teoría de las interacciones fuertes de ser inútil desde el punto de vista del cálculo, a pesar de su renormalizabilidad. La constante de acoplamiento fuerte vale aproximadamente 10, lo que puede contrastarse con la constante de acoplamiento electromagnético —la denominada *constante de estructura fina*—, que vale aproximadamente 1/137; y la fuerza débil, aunque no es numéricamente comparable de forma directa, es muchísimo más débil (véase también §31.1).

26.10. Diagramas de Feynman a partir de lagrangianos

En mis descripciones de los diagramas de Feynman, renormalización, etc., he dado un salto muy grande hacia delante y no he explicado cómo se obtienen estos diagramas para cualquier teoría de campos concreta. Tampoco he relacionado la descripción de diagramas de Feynman con el formalismo general de la QFT con el que se ha iniciado este capítulo. Permítanme ahora rectificar esta omisión y clarificar el estatus de los diagramas de Feynman dentro del marco general de la QFT.

El punto de partida sería un lagrangiano apropiado para la teoría en consideración. Los diagramas de Feynman representan entonces un *desarrollo perturbativo* de la teoría cuántica asociada a dicho lagrangiano. Un desarrollo perturbativo es en esencia tan solo un desarrollo en serie de potencias de algún parámetro (o familia de parámetros) que normalmente consideramos pequeño. Este tipo de desarrollo es algo similar a lo que se ha expuesto en §4.3, donde una función $f(x)$ se desarrolla como una serie de potencias de x. El análogo de x para los diagramas de Feynman sería normalmente alguna constante de acoplamiento. En el caso de la QED, por ejemplo, este parámetro sería la carga eléctrica e. Para cada vértice de un diagrama de Feynman habría un factor de e, de modo que los términos de la serie serían diagramas con un número creciente de vértices, donde los diagramas con n vértices proporcionarían, en conjunto, el coeficiente de e^n. Para teorías con más de una constante de acoplamiento, obtendríamos una serie de potencias en más de una variable. Un ejemplo sería una versión de QED en la que las líneas electrónicas del enfoque estándar se reemplazan por zigzags, de acuerdo con las Figs. 25.2 y 25.3b. Las dos

«constantes de acoplamiento» serían entonces la carga eléctrica y la masa M del electrón.

He comentado que las teorías renormalizables no son necesariamente finitas. Incluso esa teoría renormalizable arquetípica, la QED, no es realmente una teoría finita, ni siquiera después de la renormalización. ¿Cómo puede ser esto? La renormalización se refiere a la eliminación de infinitos de colecciones finitas de diagramas de Feynman. No nos dice que la *suma* de todas esas cantidades finitas resultantes sea realmente convergente. Lo que nos da la QED es una serie de potencias como $f_0 + f_1 e + f_2 e^2 + f_3 e^3 + \ldots$, donde cada uno de los coeficientes $f_1, f_2, f_3, \ldots$ es una cantidad finita, obtenida a partir de los procedimientos «renormalizados» aceptados para calcular integrales de diagramas de Feynman, en cada orden sucesivo, 0, 1, 2, 3, ... (De hecho, en cualquier caso concreto solo aparecerán potencias pares o potencias impares.)[26.15] La renormalizabilidad no nos dice que la suma de la serie entera sea finita. De hecho, no es finita, sino que tiene una divergencia logarítmica (como la serie $1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \ldots$ para $-\log(1-x)$ en $x = 1$) que, para la QED, no empieza a manifestarse hasta que llegamos a términos de orden 137, más o menos, lo que está mucho más allá de lo que normalmente se considera relevante.

En el caso de una teoría cuántica de campos general, para calcular con exactitud qué diagrama se da en cada orden necesitamos apelar a la expresión de integral de camino original, incluso si esta expresión representa algo que sería gravemente divergente si tratáramos de sumarla directamente. El procedimiento consiste en tratar la integral de camino como una cantidad completamente *formal*, a la que se aplican procedimientos simples pero formales de derivada funcional (§20.5). Se obtienen diagramas de Feynman con un número cada vez mayor de vértices cuando se realizan derivadas funcionales de orden cada vez mayor. No pretendo entrar aquí en esta materia con más detalle, excepto para decir que la generación de los diagramas de Feynman es inequívoca,[30] de acuerdo con este procedimiento formal. El lagrangiano será, por supuesto, un lagrangiano de *campo*, del tipo general que se

[26.15] ¿Puede ver por qué?

ha examinado en §20.5. Para un lagrangiano semejante, el «camino» involucrado no será una curva 1-dimensional ordinaria, digamos en algún espacio de configuración de dimensión infinita. Para una imagen relativísticamente invariante, la «historia» debe ser toda una configuración de campo 4-dimensional en una región espaciotemporal específica. La integral de la densidad del lagrangiano sobre dicha región sería la acción S, y entonces $e^{iS/\hbar}$ proporciona la amplitud (densidad) que hay que asignar a dicha configuración concreta.

26.11. Los diagramas de Feynman y la elección del vacío

En el caso de una teoría con una simetría bajo cierto grupo, como la simetría U(2) de la teoría electrodébil o la simetría SU(3) de la cromodinámica cuántica, o ambas, esta simetría se tomaría normalmente como una simetría manifiesta del lagrangiano. La presencia de dicha simetría sería importante para la renormalizabilidad de la QFT. En términos generales, la simetría se utiliza para asegurar que ciertos términos divergentes se cancelen mutuamente, y la cancelación se debe (o así se estima) a que si fuera a existir una expresión divergente superviviente, tal expresión no podría compartir la simetría postulada de la teoría.

Al menos, esa es la idea general. Sin embargo, en el caso de la teoría electrodébil existe otra sutileza, porque la teoría resultante no posee, después de todo, la simetría U(2) originalmente postulada.[31] La falta de simetría U(2) se considera el resultado de ruptura de simetría (§25.5), pero para entender cómo hay que acomodar esto necesitamos volver a nuestro formalismo general de la teoría cuántica de campos. La idea básica es que la ruptura de la simetría se explica por una elección del *estado vacío* U(2)-asimétrico. En consecuencia, el estado «$|0\rangle$» fantasma que se supone que está en el extremo derecho de todos los operadores creación y aniquilación, y que en general ha sido ignorado hasta ahora en nuestras consideraciones de los diagramas de Feynman, debe empezar a salir de sus sombras.

En primer lugar, tendremos que ver aproximadamente cómo relacionar los elementos del álgebra $\mathcal{A}$ de la QFT con los diagramas de Feynman. Un punto clave es que los propagadores de Feynman, que

representan las líneas de un diagrama de Feynman, son básicamente los valores de los *conmutadores* o *anticonmutadores* que hemos visto en §26.2 (i.e., los «$\langle\psi|\phi\rangle$» en estas expresiones). En la práctica, estos se expresan normalmente en términos de espacio de momentos, aunque en la definición de los propagadores de Feynman precisos hay algunas sutilezas que surgen de las cuestiones de energía positiva/negativa (que quizá pueden entenderse mejor desde la perspectiva hiperfuncional de §9.7). Aquí no nos preocuparemos por estas sutilezas.

Supongamos ahora que estamos interesados en una situación que empieza con cierto conjunto de partículas entrantes y de la que finalmente emerge un conjunto de partículas salientes. Empezamos con el estado vacío $|0\rangle$ y luego aplicamos los diversos operadores de creación que se necesitan para producir el estado requerido para las partículas entrantes. Este procedimiento da el estado inicial $|\psi_{\text{in}}\rangle$. Análogamente, podemos adoptar el mismo procedimiento, pero ahora utilizando los operadores creación para las partículas salientes, actuando de nuevo sobre $|0\rangle$, para producir el estado final $|\psi_{\text{fin}}\rangle$. Queremos calcular la amplitud $\langle\psi_{\text{fin}}|\psi_{\text{in}}\rangle$, a partir de la que podemos obtener la probabilidad de ir de «in» a «fin» utilizando simplemente la fórmula estándar, que se ha dado en §22.5, que es precisamente $|\langle\psi_{\text{fin}}|\psi_{\text{in}}\rangle|^2$, si los estados están normalizados.

Ahora bien, la expresión $\langle\psi_{\text{fin}}|\psi_{\text{in}}\rangle$ incluye operadores aniquilación a la izquierda (porque la conjugación hermítica implicada al pasar de $|\psi_{\text{fin}}\rangle$ a $\langle\psi_{\text{fin}}|$ cambia todos los operadores de creación en operadores de aniquilación). Todos estos están a la izquierda de los operadores de creación en $|\psi_{\text{in}}\rangle$, de modo que podemos imaginar que «empujamos» todos estos operadores de aniquilación a través de los operadores de creación que hay a su derecha, hasta que llegan a dar con el $|0\rangle$ en el extremo derecho. Cuando esto sucede, se ha «matado» el $|0\rangle$ (véase §26.2), de modo que la expresión obtenida es cero. Pero cada vez que empujamos un operador de aniquilación a través de un operador de creación debemos tener en cuenta el conmutador (y los requisitos de frecuencia positiva/negativa) antes mencionado, lo que nos da una línea del diagrama de Feynman, como he indicado. Cada vez que hacemos esto aparece otra de estas líneas. Finalmente, todo lo que obtenemos es $\langle 0|0\rangle$ multiplicado por el conjunto de propagadores de

Feynman que representan las líneas de un diagrama de Feynman, y $\langle 0|0\rangle = 1$, para un estado vacío normalizado, de modo que obtenemos precisamente el propio diagrama de Feynman.

Hasta ahora, nuestro diagrama de Feynman es completamente trivial, sin ningún vértice en absoluto, pero esto se debe a que no he incluido ninguna interacción en el álgebra de operadores $\mathcal{A}$. Para hacerlo tendríamos que examinar el lagrangiano específico que es relevante para nuestro problema concreto y utilizarlo para generar la $\mathcal{A}$ correcta. Básicamente, estos procedimientos reflejarían los mencionados en §26.10 para generar diagramas de Feynman, con sus términos de vértice apropiados.

Hasta aquí no hemos ganado mucho, pero una ventaja de introducir nuestros diagramas de Feynman en el marco general de la QFT es que ahora podemos reemplazar el estado vacío $|0\rangle$ por un vacío alternativo $|\Theta\rangle$, que puede no ser equivalente a él (como sucede con el estado $|\Sigma\rangle$ mar de Dirac que hemos considerado en §26.4). Esto tiene la virtud, con respecto a la teoría electrodébil y otras teorías que dependen de forma crucial de una simetría fundamentalmente rota, de que mientras el lagrangiano —y en consecuencia los diagramas de Feynman de la teoría— está sujeto a una simetría exacta (el grupo U(2), en el caso de la teoría electrodébil), los estados reales del sistema están sujetos solo a una simetría más baja (el grupo gauge U(1) del electromagnetismo, en el caso de la teoría electrodébil), porque el estado vacío $|\Theta\rangle$ posee solo esta simetría más baja. Por este método, la renormalizabilidad de la teoría que confiere la completa simetría intacta no es perturbada, pese al hecho de que la teoría en conjunto exhibe solo un grupo de simetría «rota» más pequeño.

Este es evidentemente un artificio maravilloso para producir teorías físicas que pueden beneficiarse de una simetría exacta aunque la situación observacional sea una situación en la que la simetría está lejos de ser satisfecha. Es el tipo de cosas que ha ofrecido una gran tentación para los físicos en sus tanteos en busca de esquemas mejores y más profundos. De hecho, todas las ideas modernas para ir más allá del modelo estándar tratan de sacar partido de este tipo de «ruptura de simetría». Pese a todo, y por muy populares que sean todos estos intentos —tales como los que abordaré en §§28.1-5—, deben considerarse todavía muy

especulativos. Tendremos que mantener una postura crítica y escéptica respecto a propuestas de esta naturaleza para no extraviarnos con demasiada facilidad.

Como preludio antes de abordar algunas de estas propuestas, tendremos que familiarizarnos un poco con el *big bang* en el capítulo siguiente. Luego, en el capítulo 28, trataremos de entender algunas de las cuestiones alarmantes que pueden acompañar a la idea de ruptura espontánea de simetría en el contexto concreto del universo primitivo. Por último, tendremos que bregar aún más en los usos necesarios de esta idea ubicua, cuando, en el capítulo 31, lleguemos a examinar la supersimetría, las ideas originales de la teoría de cuerdas, y luego algunas de sus extraordinarias descendientes.

Notas

Sección 26.1

26.1. Véanse Aitchison y Hey (2004), o Zee (2003).

Sección 26.2

26.2. Aquí voy a ser un poco «no estándar» en mis descripciones al admitir que la «función de onda» ψ sea simplemente un campo *general* de frecuencia positiva, no necesariamente normalizado. En correspondencia, la falta de normalización se aplica también al operador creación $\Psi^\dagger$ (y al operador destrucción Ψ). En muchas descripciones convencionales, ψ se consideraría un estado de momento.

26.3. En buena parte de la literatura estándar se utiliza el símbolo a para un operador aniquilación, y se utiliza $a^\dagger$ (el conjugado hermítico de a) para el correspondiente operador creación, adoptándose normalmente una descripción en el espacio de momentos; véanse Shankar (1994) y Zee (2003).

26.4. Algunos lectores familiarizados con la literatura estándar pueden sentirse confundidos por esto, porque con frecuencia se da el caso de que los operadores creación y aniquilación que se utilizan están restringidos a ser aquellos para los diferentes estados de momento que forman una *base ortogonal*. En tal caso, los operadores aniquilación *sí* eliminan estados específicos.

26.5. Véanse Zee (2003) y Peskin y Schröder (1995).
26.6. Véase Zee (2003) para una incisiva demostración de este requisito.
26.7. Hay también algunas cuestiones topológicas intrigantes que interconectan el intercambio de partículas con una rotación de 2π, pero todavía no están claras las plenas consecuencias con respecto a la QFT. Véanse Finkelstein y Rubinstein (1968), Feynman (1987) y Berry y Robbins (1997).

Sección 26.3

26.8. Véanse Landsman (1998) para una referencia técnica más bien exigente y también Ashtekar y Magnon (1980).
26.9. Escrito quizá en términos de un potencial.
26.10. Las cuestiones generales de normalización de estados se tratan, por ejemplo, en Ryder (1996). La cuantización del campo electromagnético recibe un tratamiento algo más convencional en Shankar (1994).

Sección 26.4

25.11. Véase Shrock (2003) para algunas de las noticias más recientes sobre neutrinos: ¡un área muy «caliente» en física actualmente!

Sección 26.6

26.12. El espacio de Fock para el caso sencillo de un campo bosónico, en donde la partícula es su propia antipartícula, puede escribirse $\mathbb{C} \oplus \mathcal{H} \oplus \{\mathcal{H} \odot \mathcal{H}\} \oplus \{\mathcal{H} \odot \mathcal{H} \odot \mathcal{H}\} \oplus \{\mathcal{H} \odot \mathcal{H} \odot \mathcal{H} \odot \mathcal{H}\} \oplus \ldots$, donde la operación suma directa se denota por $\oplus$ y donde el símbolo $\odot$ denota el producto tensorial *simetrizado*. Casos más complicados en los que hay espín y carga, etc., pueden tratarse de forma correspondiente. Véase Shankar (1994) para la idea general; también puede ser útil Davydov (1976).
26.13. Véanse Hannabuss (1997) y Shankar (1994) para una discusión de los estados coherentes, que pueden entrar en muchas variedades (fermiónica, espín, etc.).
26.14. Véanse Wald (1994) y Birrell y Davies (1984).
26.15. Véanse Dirac (1933) y Schwinger (1958).
26.16. Véase Feynman (1948, 1949). Feynman y Hibbs (1965) constituye una excelente visión general de la idea. El enfoque competidor de Schwinger respecto a la electrodinámica cuántica (véase, por ejemplo, Schwinger, 1951) era en muchos aspectos más riguroso, pero la mayoría de los

que trabajan en el tema utilizan hoy la imagen más intuitiva de integrales de camino y diagramas de Feynman.

Sección 26.7

26.17. Como ha señalado Feynman, una precisión de este orden determinaría la distancia entre Los Ángeles y Nueva York ¡con un error menor que el grosor de un cabello humano!
26.18. Algunas otras ideas dignas de mención, tal como la denominada «euclideanización», se discutirán en §28.9.
26.19. Este es un ejemplo de lo que se denomina una función de Green (por el inglés George Green, 1793-1841, hijo de un molinero, que se convirtió en un extraordinario matemático autodidacta). El propagador de Feynman es una función de Green particular $K(p, q)$, definida por los requisitos de frecuencia positiva de la teoría cuántica mencionados en §24.3.
26.20. Véase, por ejemplo, el clásico Bjorken y Drell (1965).

Sección 26.8

26.21. Existen cantidades conocidas como «constantes de acoplamiento móviles» que tienen dependencia funcional de la energía en reposo del sistema total de partículas entrantes en un vértice de Feynman. Estas tienen importancia en muchas teorías modernas de física de partículas.
26.22. Así, si $P_a^{(1)}$, $P_a^{(2)}$, ... son los momentos entrantes, y $Q_a^{(1)}$, $Q_a^{(2)}$, ... los momentos salientes, en un vértice, entonces incluimos el término $\delta(P_a^{(1)} + P_a^{(2)} + \ldots - Q_a^{(1)} - Q_a^{(2)} - \ldots)$.
26.23. Véase la nota 25.5.
26.24. El importante concepto de una «matriz S» (debido, básicamente, a Heisenberg y al muy original físico estadounidense J. A. Wheeler) no está ligado a la noción de un diagrama de Feynman, y puede ser evaluado por otros medios. Véase Eden *et al.* (2002).

Sección 26.9

26.25. Véanse Zee (2003) o Ryder (1996) para más (escabrosos) detalles.
26.26. Puesto que la carga del electrón es negativa, «reducir» significa aquí «hacer el módulo más pequeño».
26.27. Existen algunos procedimientos matemáticos elegantes, adaptados para la sistematización de este método, que sacan partido de la noción de «coproducto», relacionada con las ideas de la geometría no conmutativa discutidas brevemente en §33.1; véase Connes y Kreimer (1998).

26.28. Un conjunto importante de técnicas es el suministrado por la idea de «grupo de renormalización». Zee (2003) y Ryder (1996), así como Peskin y Schröder (1995), tratan estas ideas, y su relación con la mecánica estadística se discute en el enciclopédico Zinn-Justin (1996).

26.29. Las fuerzas gravitatorias siguen siendo manifiestas (incluso más allá de las escalas galácticas), pese a su disminución con la distancia de acuerdo con la ley de la inversa del cuadrado. Por otra parte, el lector podría preocuparse de por qué la fuerza fuerte es apenas manifiesta a distancias mayores que las distancias nucleares incluso si realmente *aumenta* con la distancia. La razón es que mientras que la gravitación se acumula, al ser siempre atractiva, la fuerza fuerte es un compuesto de componentes atractivas y repulsivas que necesariamente se cancelan por completo entre núcleos separados (pues los núcleos individuales son necesariamente «singletes de color»).

Sección 26.10

26.30. Zee (2003) y Zinn-Justin (1996) enseñan el algoritmo; para un toque bastante divertido e intuitivo, se recomienda también Mattuck (1976).

Sección 26.11

26.31. Véase la nota 25.12.

27

El big bang y su legado termodinámico

27.1. Simetría temporal en la evolución dinámica

¿Qué tipo de leyes configuran el universo con todos sus contenidos? La respuesta que ofrecen prácticamente todas las teorías físicas satisfactorias, desde la época de Galileo en adelante, vendrá dada en forma de una *dinámica*, es decir, una especificación de cómo evolucionará un sistema físico con el tiempo, dado el estado físico del sistema en un instante concreto. Estas teorías no nos dicen cómo es el mundo; lo que dicen en su lugar es: «si el mundo era de tal y cual manera en un instante, entonces será de esta otra manera en un instante posterior». Una teoría semejante no nos dirá cómo *está* configurado el mundo a menos que le digamos cómo *estaba* configurado.

Ha habido excepciones importantes a esta forma de plantear el tema, tales como la maravillosa conclusión de Kepler (en 1609), de que las órbitas de los planetas en torno al Sol tienen ciertas formas geométricas —elipses con el Sol en uno de sus focos— que son descritas a velocidades que satisfacen reglas específicas. Esa era una afirmación acerca de cómo *es* el universo, en lugar de cómo podría evolucionar su estado de un instante a otro de acuerdo con una ley dinámica. Pero desde nuestro punto de vista actual, los movimientos geométricos de Kepler son meras consecuencias de la dinámica gravitatoria del siglo XVII, como fue demostrado por primera vez por Newton y publicado en sus grandes *Principia* (1687), y las leyes de Kepler no se consideran directamente fundamentales para las formas de la naturaleza. De hecho, podría argumentarse que Kepler —y la ciencia en general— tuvo la inmensa fortuna de que la ley de fuerzas específica que gobierna la

gravedad de Newton, la ley de la inversa del cuadrado (§17.3), tiene la propiedad de que todas las órbitas de cuerpos pequeños sometidos a una fuerza central son realmente formas matemáticas simples y elegantes (y, de hecho, formas que habían sido estudiadas en profundidad por los antiguos griegos unos dieciocho siglos antes). Esta es una propiedad muy excepcional, difícilmente compartida por cualquier otra ley de fuerzas centrales simple. En general, nuestra perspectiva moderna mantiene que es de las *leyes dinámicas* de las que hay que esperar que tengan una forma matemática elegante, y es una cuestión de buena suerte para nosotros el que, como consecuencia de dichas leyes, encontremos formas matemáticas simples.

Según la forma habitual de considerar la actuación de estas leyes dinámicas, es la elección de condiciones iniciales la que determina qué realización concreta de la dinámica va a darse. Normalmente se piensa en términos de sistemas que evolucionan hacia el futuro a partir de datos especificados en el pasado, y la evolución concreta que tiene lugar está determinada por ecuaciones diferenciales. (Estas serán ecuaciones en derivadas parciales —ecuaciones de campo— cuando hay campos o funciones de onda que evolucionan dinámicamente; véanse §10.2, §§19.2,6, §21.3, el ejercicio [19.2] y la nota 21.1.) Por el contrario, no se suele considerar la evolución de estas mismas ecuaciones hacia el *pasado*, ¡pese al hecho de que las ecuaciones dinámicas de la mecánica clásica y la cuántica son simétricas respecto a una inversión en la dirección del tiempo! Por lo que respecta a las matemáticas, no hay ninguna diferencia en especificar condiciones finales, en un tiempo remoto, futuro y evolucionar hacia atrás en el tiempo. Desde el punto de vista matemático, las condiciones finales son tan buenas como las iniciales para determinar la evolución de un sistema.

Son necesarios algunos comentarios acerca de este determinismo dinámico simétrico respecto al tiempo. En primer lugar, el lector puede estar tranquilo porque no está sustancialmente invalidado por el marco de la teoría de la relatividad especial ni de la general. Los datos que definen el estado del sistema están especificados en un «tiempo» inicial, que es una 3-superficie inicial de género espacio, y estos datos evolucionan de acuerdo con las ecuaciones dinámicas para determinar el estado físico del sistema en el futuro, y también en el pasado, de dicha 3-superficie. No obstante, la relatividad *general* plantea algunas cuestiones nuevas, porque la estructura misma del espaciotiempo en el que tiene lugar la evolución es parte del estado físico a determinar. (Esto tiene implicaciones concretas en el contexto de los agujeros negros, que tendremos que abordar más adelante; véanse §§27.8,9, §28.8 y §§30.4,9.)

En el caso de la mecánica cuántica, el determinismo se refiere solo a la parte **U** de dicha teoría, donde el estado cuántico está gobernado por la ecuación de Schrödinger (o equivalente). Bajo inversión temporal —la **T** mencionada en §25.4—, el operador derivada temporal $i\hbar\partial/\partial t$ de la ecuación de Schrödinger (§21.3) debe reemplazarse por $-i\hbar\partial/\partial t$ (puesto que $t \mapsto -t$). Con tal de que el hamiltoniano sea un hamiltoniano ordinario, que se transforma en sí mismo bajo la acción de **T**, vemos que la evolución de Schrödinger también se transforma en sí misma, siempre que acompañemos la inversión temporal $t \mapsto -t$ por una inversión del signo de la unidad imaginaria: $i \mapsto -i$. De hecho, así es como consideramos la acción de **T** en mecánica cuántica. (Podemos notar que una función $f(t)$ de frecuencia positiva se transforma en una función de frecuencia positiva bajo los reemplazamientos combinados $t \mapsto -t$, $i \mapsto -i$, de modo que todo está bien por lo que a esto se refiere.)[27.1] Sin embargo, el comportamiento de la reducción **R** del estado cuántico bajo la acción de **T** es otra cuestión, y ofrecerá un punto importante para nuestras deliberaciones en el capítulo 30 (§30.3).

27.2. Ingredientes submicroscópicos

Hay, no obstante, otras cuestiones que podrían preocupar al lector iniciado, incluso con respecto a la dinámica clásica. En la mecánica clásica, la simetría de inversión temporal es ciertamente verdadera en la dinámica submicroscópica de las partículas individuales y sus campos acompañantes. Pero en la práctica tenemos poco conocimiento del comportamiento de los ingredientes individuales de un sistema. Nor-

[27.1] ¿Por qué? Explique también por qué el momento espacial es tratado de forma consistente por este reemplazamiento.

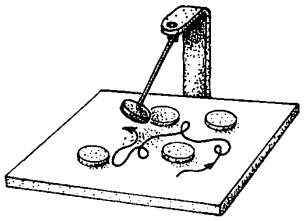


Fig. 27.1. Movimiento caótico. Un «juguete de ejecutivo» que consiste en un péndulo magnético que oscila por encima de un conjunto de imanes fijos. La trayectoria real seguida por el péndulo es extraordinariamente sensible a su posición y velocidad iniciales.

malmente se estima que un conocimiento de la posición y el momento detallados de cada partícula es imposible e innecesario, y que el comportamiento global del sistema está suficientemente bien descrito por algunos promedios adecuados de los parámetros físicos de las partículas individuales. Estos serían cosas como las distribuciones de masa, momento y energía, la localización y la velocidad del centro de masas, la temperatura y la presión en diferentes lugares, las propiedades elásticas, el momento de inercia, la forma global detallada y su orientación en el espacio, etc. Una cuestión importante, por consiguiente, es si un buen conocimiento inicial de tales parámetros medios «globales» bastará o no, en la práctica, para determinar el comportamiento dinámico del sistema hasta un grado adecuado.

Este no es siempre el caso. Los sistemas conocidos como *caóticos* tienen la propiedad de que el comportamiento final depende críticamente de sus condiciones iniciales exactas. Un ejemplo familiar lo constituye un conocido «juguete de ejecutivo» en el que un péndulo magnético oscila sobre un conjunto de imanes colocados en cierta disposición en la base del juguete. Véase la Fig. 27.1. El comportamiento dinámico está bien gobernado, de forma determinista, por las leyes de Newton y las leyes de la magnetostática, junto con el amortiguamiento debido al rozamiento del aire. Pero el lugar de reposo final del péndulo depende de forma tan crítica del estado inicial que es efectivamente impredecible, aunque un conocimiento perfectamente detallado de dicho estado inicial, con todas las partículas y campos constituyentes, determinaría esta evolución de forma unívoca.[1] Se conocen otros muchos ejemplos de tales «sistemas caóticos». Buena parte de la imprecisión en la predicción del tiempo meteorológico se atribuye normalmente a la naturaleza caótica de los sistemas dinámicos involucrados.

Incluso el muy ordenado (y predecible) movimiento gravitatorio newtoniano de los cuerpos del sistema solar constituye (probablemente) un sistema caótico, técnicamente hablando, aunque las escalas de tiempo que son relevantes para semejante «caos» son enormemente mayores que las de la observación astronómica.

¿Qué pasa con la evolución hacia el pasado, en lugar de hacia el futuro? Un comentario justo sería que tal «impredecibilidad caótica» es normalmente mucho peor para la «retrodicción» que está implicada en la evolución dirigida al pasado que para la «predicción» de la evolución normal dirigida al futuro. Esto tiene que ver con la *segunda ley de la termodinámica*, que en su forma más simple afirma básicamente:

El calor fluye desde un cuerpo más caliente a otro más frío.

De acuerdo con esta ley, si ponemos en contacto un cuerpo caliente y uno frío utilizando algún material conductor del calor, el cuerpo caliente se enfriará y el cuerpo frío se calentará hasta que se estabilicen a la misma temperatura.[2] Esta es la expectativa de la *predicción*, y esta evolución tiene un carácter determinista. Si, por el contrario, consideramos este proceso en la dirección inversa del tiempo, observamos que dos cuerpos de la misma temperatura evolucionan hacia temperaturas desiguales, y sería una imposibilidad práctica decidir qué cuerpo se hará más caliente y cuál se hará más frío, cuánto, y cuándo. Este procedimiento de *retrodicción* dinámica, para este sistema, es en la práctica una perspectiva claramente sin esperanzas.

De hecho, esta dificultad se aplicaría a la retrodicción de casi cualquier sistema macroscópico, con grandes números de partículas constituyentes, que se comporte de acuerdo con la segunda ley. Por razones de este tipo, la física se suele interesar en la predicción antes que en la retrodicción.[3] Otro aspecto de ello es que la segunda ley se considera un ingrediente esencial para el poder predictivo de la física, ya que elimina aquellos problemas que acabamos de encontrar en la retrodicción.

De todas formas, muchos físicos adoptarán el punto de vista de que esta ley no es «fundamental» en el mismo sentido en que lo son, por ejemplo, la ley de conservación de la energía, el principio de superposición lineal en mecánica cuántica, y quizá el modelo estándar

de la física de partículas. Argumentarán que la segunda ley es un ingrediente casi «obvio» y necesario en cualquier teoría física razonable. Muchos adoptarán el punto de vista de que es algo vaga e imprecisa, y que de ninguna manera puede compararse con la precisión extraordinaria que encontramos en las leyes dinámicas que controlan la física fundamental. Quiero argumentar de forma muy diferente, y demostrar la precisión casi «alucinante» que hay tras este principio estadístico aparentemente vago al que habitualmente nos referimos como la «segunda ley».

27.3. Entropía

Examinemos de forma algo más exacta lo que en realidad afirma la segunda ley. Como preliminar, debería informar al lector de lo que es la primera ley de la termodinámica. La primera ley es simplemente el enunciado de que la energía total de un sistema aislado se conserva. El lector podría muy bien quejarse de que esto apenas es nada nuevo (§18.6, §20.4 y §21.3). Pero cuando esta ley fue propuesta (inicialmente por Sadi Carnot a comienzos de la década de 1820, aunque no publicada por él),[4] todavía no se había establecido claramente que el calor es solo una forma de energía —ni siquiera estaba del todo clara la propia nocion ordinaria y macroscópica de energía—. La primera ley hace explícito que la energía total no se pierde cuando, por ejemplo, un cuerpo pierde su energía cinética (§18.6) cuando se frena debido a la resistencia del aire; simplemente se utiliza para calentar el aire y el cuerpo. Esta energía *térmica* se entiende como (primariamente) energía cinética de los movimientos de las moléculas del aire y las vibraciones de las partículas que componen el cuerpo. Más aún, la *temperatura* es simplemente una medida de la energía por grado de libertad, de modo que las nociones termodinámicas de calor y temperatura son básicamente las mismas que las nociones dinámicas previamente entendidas, pero aplicadas en el nivel de los constituyentes individuales de los materiales y tratadas de forma estadística. La primera ley tiene el tipo de precisión que nos es familiar: el valor de algo, a saber, la energía total, permanece exactamente constante pese al hecho de que pueden tener

lugar todo tipo de procesos complicados. La energía total después del proceso es igual a la energía antes del proceso.

Mientras que la primera ley es una igualdad, la segunda ley es una desigualdad. Nos dice que una magnitud diferente, conocida como la *entropía*, tiene un valor mayor (o al menos no menor) después de que se haya tenido lugar un proceso que el valor que tenía antes. Hablando muy vagamente, la entropía es una medida de la «aleatoriedad» en el sistema. Nuestro cuerpo en movimiento a través del aire empieza con su energía de una forma organizada (su energía cinética de movimiento), pero cuando se frena por la resistencia del aire esta energía se reparte entre los movimientos aleatorios de las partículas del aire y las partículas individuales del cuerpo. La «aleatoriedad ha aumentado»; más específicamente, la entropía ha aumentado.

La noción de entropía fue introducida por Clausius en 1865, pero fue el destacado físico austríaco Ludwig Boltzmann quien, en 1877, hizo clara la definición de entropía (o al menos tan clara como parece posible hacerla). Para entender la idea de Boltzmann (para un sistema clásico), necesitamos la noción del *espacio de fases* (§12.1, §§14.1,8 y §§20.1,2,4) que, recordemos, para un sistema clásico de n partículas (sin características distintivas), es un espacio $\mathcal{P}$ de $6n$ dimensiones, cada uno de cuyos puntos representa la familia completa de posiciones y momentos de todas las n partículas. Para hacer precisa la noción de entropía necesitamos una idea de lo que se denomina *granulado grueso*.[5] Para ello dividimos el espacio de fases $\mathcal{P}$ en un número de subregiones que llamaré «cajas». Véase la Fig. 27.2. La idea consiste en considerar que los conjuntos de puntos de $\mathcal{P}$ que representan estados del sistema que son indistinguibles unos de otros mediante observaciones macroscópicas están agrupados en la misma caja, pero puntos de $\mathcal{P}$ que pertenecen a cajas diferentes se consideran macroscópicamente distinguibles. La entropía de Boltzmann S para el estado del sistema representado por cierto punto x de $\mathcal{P}$ es

$$S = k \log V,$$

donde V es el volumen de la caja $\mathcal{V}$ que contiene a x (y *log* es un logaritmo *natural*; véase §5.3), y donde k es la *constante de Boltzmann*,[6] que tiene el valor

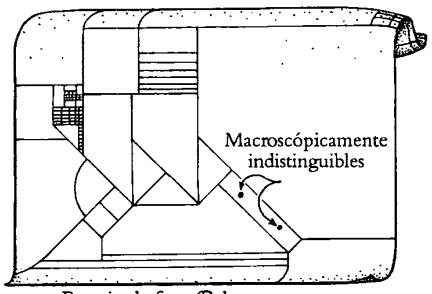


Fig. 27.2. Entropía de Boltzmann. Esta implica la división del espacio de fases $\mathcal{P}$ en subregiones («cajas»), lo que se denomina un «granulado grueso» de $\mathcal{P}$. Los puntos de una caja dada representan estados físicos que son macroscópicamente indistinguibles. La definición de Boltzmann de la entropía de un estado x en una caja $\mathcal{V}$ de volumen V es $S = k \log V$, donde k es la constante de Boltzmann.

$$k = 1{,}38 \times 10^{-23}\ \mathrm{J\ K^{-1}}$$

(donde J significa julios y $\mathrm{K^{-1}}$ significa «por grado kelvin»).

He dicho que la definición de Boltzmann hace «clara» la noción de entropía. Pero para que la fórmula anterior para S represente algo físicamente preciso sería necesario tener una receta precisa para el granulado grueso que se supone que representa nuestra familia de «cajas». Sin duda hay algo «arbitrario» en la división concreta en cajas que pudiera seleccionarse. La definición parece depender de la cercanía con que decidamos examinar un sistema. Dos estados que son «macroscópicamente indistinguibles» para un experimentador podrían ser distinguibles para otro. Además, es también muy arbitrario dónde hay que trazar exactamente la frontera entre dos cajas, puesto que a dos puntos vecinos en $\mathcal{P}$, uno a cada lado de la frontera, podría asignárseles entropías completamente diferentes pese a que son prácticamente idénticos. Sigue habiendo algo muy subjetivo en esta definición de S, pese a que es un claro avance respecto a nociones anteriores de aplicabilidad más limitada, y una indudable mejora sobre la idea de que es solo una medida de la «aleatoriedad» en un sistema.

Mi propia posición respecto al estatus físico de la entropía es que no la veo como una noción «absoluta» en la teoría física actual, aunque resulta muy útil. Sin embargo, existe la posibilidad de que pueda adquirir un estatus más fundamental en el futuro. Para ello habrá que tener siempre en cuenta la física cuántica, y, en cualquier caso, es la mecánica cuántica la que proporciona una *medida* absoluta para cualquier

región concreta $\mathcal{V}$ del espacio de fases, contenida en $\mathcal{P}$, donde pueden escogerse las unidades de modo que $\hbar = 1$ (como sucede con las unidades de Planck; véase §27.10).[27.2] Como quiera que sea, es notable cuán pequeño es el efecto que tiene la arbitrariedad del granulado grueso en las consideraciones termodinámicas. Parece que la razón para ello es que, en la mayoría de las consideraciones de interés, uno está tratando con razones absolutamente inimaginables entre los tamaños de los volúmenes de las cajas relevantes en el espacio de fases y no supone mucha diferencia dónde se trazan las fronteras, siempre que el granulado grueso refleje «razonablemente» la idea intuitiva de cuándo va a considerarse que los sistemas son macroscópicamente distinguibles. Puesto que la entropía se define como un logaritmo del volumen de la caja, se necesitaría un retrazado «increíble» de las fronteras para obtener cualquier cambio significativo en S.[27.3] En mi opinión, más que ser «fundamental», la entropía tiene el estatus de una «conveniencia» en la teoría actual, aunque hay indicios de que en un contexto más profundo, donde lleguen a ser importantes consideraciones de gravitación cuántica (especialmente en relación con la entropía de los agujeros negros), puede haber un estatus más fundamental para una noción de este tipo. Llegaremos a esta cuestión más adelante en este capítulo (§27.10) y en §§30.4-8, §31.15 y §32.6.

27.4. El carácter robusto del concepto de entropía

Una ilustración sencilla podría clarificar algo más el papel de la fórmula de la entropía de Boltzmann. Consideremos un recipiente cerrado en el que una región $\mathcal{R}$ está marcada como *especial*; por ejemplo, una protuberancia en forma de bulbo de una décima parte del volumen del recipiente que está comunicada con el resto del recipiente a través de una pequeña abertura; véase la Fig. 27.3. Supongamos que en el recipiente hay un gas que consiste en m moléculas. Estamos buscando la

[27.2] Muestre cómo se puede asignar una medida absoluta a un volumen del espacio de fases si se escogen unidades de modo que $\hbar = 1$.
[27.3] ¿Qué relación tiene el logaritmo en la fórmula de Boltzmann con las discrepancias «increíbles» en los volúmenes de las cajas?

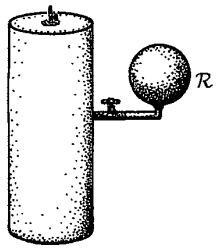


Fig. 27.3. Recipiente cerrado, una parte del cual es una región $\mathcal{R}$ con forma de bulbo cuyo volumen es 1/10 del recipiente entero. ¿Cuánto aumenta la entropía cuando a un gas contenido inicialmente en $\mathcal{R}$ se le permite fluir al volumen entero?

entropía S que hay que asignar a la situación en la que todo el gas se encuentra en $\mathcal{R}$, en comparación con la entropía que hay que asignar a la situación en la que el gas está distribuido aleatoriamente en el recipiente.

Tendremos $S = k \log \mathcal{V}_{\mathcal{R}}$, por la fórmula de Boltzmann, donde $\mathcal{V}_{\mathcal{R}}$ es el volumen de la región $\mathcal{V}_{\mathcal{R}}$ del espacio de fases que representa la situación con todas las moléculas en $\mathcal{R}$. Supongamos, por simplicidad, lo que se denomina *estadística de Boltzmann*, frente a la «estadística de Bose-Einstein» para bosones y la «estadística de Fermi-Dirac» para fermiones, tal como se ha descrito en §23.7; es decir, suponemos que todas las moléculas del gas son *distinguibles* unas de otras (al menos en principio).[27.4] Suponiendo que el gas es aire ordinario a presión atmosférica, con un recipiente de un litro de volumen total, tenemos $m = 10^{22}$ moléculas aproximadamente. La razón entre el volumen de la región $\mathcal{V}_{\mathcal{R}}$ del espacio de fases y el volumen del espacio de fases total $\mathcal{P}$ es $10^{-m}(= \frac{1}{10^m})$,[27.5] que es

$$10^{-100000000000000000000},$$

y así vemos algo de las «increíbles» razones de volúmenes que aparecen en consideraciones de este tipo. La cifra anterior representa la probabilidad ridículamente minúscula de encontrar —solo por puro azar— que todas las moléculas del gas estén realmente en $\mathcal{R}$. La entropía de esa situación extraordinariamente improbable es mucho más pequeña

[27.4] Explique qué diferencias habría en el recuento en los tres casos.
[27.5] ¿Por qué?

que la entropía de la situación en la que el gas está distribuido aleatoriamente, siendo la diferencia aproximadamente

$$-k \log (10^{-100000000000000000000}) = 2{,}3 \times 10^{22} k$$
$$= 0{,}32 \text{ J K}^{-1}$$

por la fórmula de Boltzmann,[27.6] donde he utilizado el hecho de que el logaritmo natural de 10 es aproximadamente 2,3. Así pues, si suponemos que el gas se mantiene inicialmente en $\mathcal{R}$, por ejemplo mediante una válvula cerrada que lo aísla del resto del recipiente, y luego abrimos la válvula para liberar el gas en el resto del recipiente, encontramos un apreciable incremento de entropía de aproximadamente $2{,}3 \times 10^{22} k$, que en unidades ordinarias es solo de un tercio de un julio por grado kelvin.

Al lector podría preocuparle que no fuera factible tener un recipiente en el que inicialmente no hay absolutamente ninguna molécula de gas en su parte no especial. Por ello, relajemos un poco nuestra definición de la región $\mathcal{V}_{\mathcal{R}}$, definiéndola de modo que $\mathcal{R}$ solo tiene que contener al menos el 99,9 % de las moléculas del gas. Así pues, $\mathcal{V}_{\mathcal{R}}$ exige ahora que no más de una milésima de dichas moléculas estén fuera de $\mathcal{R}$. Está perfectamente dentro de la tecnología actual tener un vacío de esta perfección relativa en la región no especial. Sucede que el resultado apenas se ve afectado, y el aumento de entropía, al abrir la válvula, *sigue siendo* del orden general[27.7] de $2{,}3 \times 10^{22} k$. Esta es una ilustración sorprendente del hecho de que aunque hay subjetividad en el dibujo de las cajas de granulado grueso (por ejemplo $\mathcal{V}_{\mathcal{R}}$), esto no causa problemas serios mientras las cajas estén «razonablemente» dibujadas.

El logaritmo en la fórmula de Boltzmann tiene una finalidad importante, además de hacer que parezcan tratables estos números enormemente grandes. Dicha finalidad es que la definición de entropía resultante sea *aditiva* para sistemas independientes. Así, si las entropías asignadas a dos sistemas independientes son S_1 y S_2, entonces la entro-

[27.6] Demuestre por qué este resultado no se ve afectado de forma significativa si introducimos consideraciones de fermión/bosón o si tenemos en cuenta la probable disminución de momento de las moléculas del gas cuando dejan de estar limitadas a $\mathcal{R}$.
[27.7] Trate de ver por qué el incremento de entropía ha caído tan poco, de $2{,}30 \times 10^{22} k$ a $2{,}29 \times 10^{22} k$, aproximadamente, haciendo varias estimaciones groseras para las cantidades matemáticas involucradas. (Si lo desea, utilice la fórmula de Stirling $n! \approx (n/e)^n (2\pi n)^{-1/2}$.)

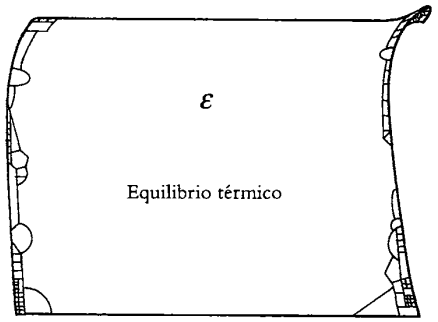


Fig. 27.4. La caja particular $\mathcal{E}$, que representa el equilibrio térmico, tiene un volumen E, que es normalmente casi igual al volumen P del espacio de fases entero $\mathcal{P}$, y por lo tanto supera con mucho los volúmenes de todas las demás cajas juntas.

pía asignada al sistema total, consistente en los dos sistemas juntos, sería $S_1 + S_2$. Estoy suponiendo que el espacio de fases del sistema total es $\mathcal{P} = \mathcal{P}_1 \times \mathcal{P}_2$, donde $\mathcal{P}_1$ y $\mathcal{P}_2$ son los espacios de fases para los dos sistemas individuales, y que las cajas de granulado grueso del sistema total son los productos de las cajas de granulado grueso de $\mathcal{P}_1$ y $\mathcal{P}_2$, que es una hipótesis muy natural[27.8] para sistemas independientes S_1, S_2. (Véanse §15.2, el ejercicio [15.1] y la Fig. 15.3a para la definición de $\times$, aplicada a los espacios.) Puesto que los volúmenes de las cajas se multiplican, las entropías correspondientes se suman (por la propiedad estándar del logaritmo; véase §5.2).

En ejemplos normales de sistemas físicos —y ciertamente en el caso muy estudiado de un gas ordinario en un recipiente ordinario— existe una caja concreta $\mathcal{E}$ del granulado grueso cuyo volumen E supera con mucho el volumen de cualquiera de las otras cajas. Esta representa el estado de *equilibrio térmico*. De hecho, normalmente E será prácticamente igual al volumen P del espacio de fases entero, y por eso E superará fácilmente los volúmenes de todas las otras cajas puestas juntas. Véase la Fig. 27.4. Para un gas ordinario, que consideraremos formado por bolas idénticas con simetría esférica en equilibrio térmico, la distribución de velocidades adopta una forma particular conocida como *distribución de Maxwell* (descubierta por el mismo James Clerk Maxwell de quien he hablado antes, en relación con el electromagnetismo). Tiene la forma

[27.8] ¿Por qué es natural esta hipótesis de grano grueso?

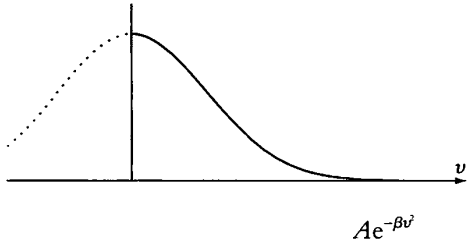


Fig. 27.5. La distribución maxwelliana de velocidades, para un gas en equilibrio, tiene la forma $Ae^{-\beta v^2}$, donde A y β son constantes, con β relacionada con la temperatura del gas, y siendo v la velocidad de la partícula. La parte de puntos de la curva la extiende a valores negativos de v, lo que revela la familiar forma de «curva de campana» de la distribución gaussiana de la estadística.

(que quienes sepan de ello reconocerán como una distribución *gaussiana*, a veces llamada «curva de campana»), donde v es el módulo de la 3-velocidad de la partícula de gas en cuestión, β es una constante relacionada con la temperatura y A es una constante tal que la integral de la probabilidad sobre el espacio de todas las velocidades posibles es 1; véase la Fig. 27.5. El equilibrio térmico, que tiene con mucho la máxima entropía posible para el sistema, es el estado en que uno esperaría que se asiente un sistema si se le deja aislado un tiempo suficientemente largo, de acuerdo con la segunda ley.

La distribución de Maxwell recién descrita se refiere a un gas constituido por cuerpos clásicos idénticos sin grados de libertad internos. Las cosas pueden hacerse mucho más complicadas cuando hay muchos tipos de constituyentes de tamaños diferentes y varios grados de libertad internos (tales como el espín o las vibraciones entre sus partes constituyentes). Para tales sistemas en equilibrio térmico existe un principio general, conocido como *equipartición de la energía*, según el cual la energía del sistema se distribuye por igual (con dispersión estadística) entre todos los diferentes grados de libertad del sistema.

Otra forma de llegar a la distribución de Maxwell consiste en apartarse del equilibrio térmico exacto y preguntar cómo cabe esperar que se mueva un gas en su aproximación al equilibrio (de acuerdo con la

segunda ley). En tales circunstancias se utiliza una ecuación conocida como *ecuación de Boltzmann* para describir la evolución. El lector percibirá que aquí hay una disciplina muy vasta, de relevancia considerable para la comprensión teórica del comportamiento de los cuerpos macroscópicos, cuando hay demasiadas partículas constituyentes para poder seguir sus trayectorias individuales. Esta disciplina se conoce como *mecánica estadística*.

27.5. Derivación de la segunda ley... ¿o no?

Tratemos ahora de entender lo que hay tras la segunda ley. Imaginemos que tenemos un sistema físico representado por un punto x en un espacio de fases de granulado grueso apropiado $\mathcal{P}$. Supongamos (Fig. 27.6a) que x empieza AHORA en alguna pequeña caja de grano grueso

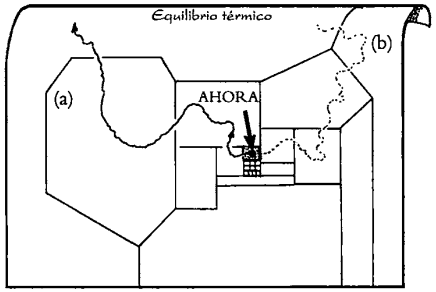


Fig. 27.6. La segunda ley en acción. La evolución de un sistema físico está representada por una curva en el espacio de fases. (a) Si sabemos que en el instante AHORA nuestro sistema está representado por un punto x en una caja $\mathcal{V}$ de volumen muy pequeño V y tratamos de ver cuál sería su probable comportamiento futuro, concluimos, debido a las enormes discrepancias en los volúmenes de las cajas, que en ausencia de cualquier sesgo grande en su movimiento entrará casi con certeza en cajas cada vez más grandes, de acuerdo con la segunda ley. (b) Pero supongamos que aplicamos este argumento hacia el *pasado*, preguntando por la manera más probable en que la curva encontró su camino a $\mathcal{V}$ en primer lugar. El mismo argumento parece llevar a la conclusión aparentemente absurda de que la manera más probable era que x encontró simplemente su camino a $\mathcal{V}$ desde cajas que se hacían cada vez mayores a medida que retrocedemos en el pasado, en flagrante contradicción con la segunda ley.

$\mathcal{V}$ de volumen V. El punto x se moverá por $\mathcal{P}$ de cierta forma, de acuerdo con las ecuaciones dinámicas apropiadas a la situación física en consideración. Teniendo en cuenta la enorme discrepancia de los tamaños de las diferentes cajas de grano grueso, y dando por supuesto que no hay ningún sesgo especial en el movimiento dinámico de x en relación con las posiciones de la caja, esperamos que en la aplastante mayoría de los casos x se paseará por cajas de volumen cada vez mayor. En otras palabras, la entropía del sistema se hará cada vez mayor a medida que avanza el tiempo. Una vez que x entra en una caja con cierta medida de entropía, se hace muy poco probable que pueda volver en un período de tiempo razonable a una caja de entropía significativamente menor que aquella de la que procedía. Alcanzar una entropía significativamente menor supondría encontrar un volumen absurdamente más minúsculo, y las probabilidades en contra de ello son inmensas. Consideremos el ejemplo que acabamos de plantear y la absolutamente inimaginable reducción en volumen del espacio de fases que acompañaría a una reducción muy modesta en la entropía, debido al logaritmo en la fórmula de Boltzmann y al pequeño tamaño de la constante de Boltzmann. Una vez que el gas ha encontrado su vía de salida de $\mathcal{R}$, es ridículamente poco probable que encuentre su vía de regreso de nuevo a $\mathcal{R}$ (al menos dentro de una escala de tiempo que no sea «ridículamente larga»).[27.9]

Este argumento contiene la razón esencial para esperar que sea válida la segunda ley. Notemos que el argumento no parece depender en absoluto de las particularidades de la dinámica, salvo que exigimos que no haya ningún sesgo que tenga como efecto que el punto x busque deliberadamente cajas más pequeñas. ¿Es esto realmente todo lo que hay en la segunda ley? Todo parece demasiado fácil, y la naturaleza aparentemente universal de este tipo de argumento es quizá la razón de que muchos físicos adopten el punto de vista de que no hay nada fundamentalmente enigmático en la segunda ley, y que cualquier teoría física razonable debe satisfacerla. Aquí es pertinente una maravillosa cita del destacado astrofísico sir Arthur Eddington:

[27.9] Trate de estimar cuán largo tendría que ser, tanto en el caso en que *todo* el gas vuelve a $\mathcal{R}$ como en el caso en que lo hace un 99,9 % del mismo. ¿Necesita conocer realmente con qué rapidez se están moviendo las moléculas de gas?

> Si alguien le dice que su teoría favorita del universo está en desacuerdo con las ecuaciones de Maxwell, entonces tanto peor para las ecuaciones de Maxwell. Si se encuentra que es contradicha por la observación... bien, estos experimentadores meten la pata a veces. Pero si resulta que su teoría está en contra de la segunda ley de la termodinámica no puedo darle ninguna esperanza, no hay nada que hacer excepto sumirse en la más profunda humillación.[7]

Pero unos instantes de reflexión nos dicen que hay algo singular en la conclusión del argumento que acabo de presentar, o quizá algo significativamente ausente de estas consideraciones elementales. Parece que hemos deducido una ley con *asimetría* temporal cuando la física subyacente puede considerarse *simétrica* en el tiempo. ¿Cómo ha sucedido esto? Podríamos imaginar que aplicamos el mismo argumento en la dirección del tiempo *pasado* (Fig. 27.6b). Parece que si en el instante AHORA situamos nuestro punto x del espacio de fases en la misma caja pequeña que hemos escogido antes, y examinamos la evolución en dirección al pasado anterior a AHORA, entonces concluimos que es abrumadoramente probable que x haya entrado en esta caja procedente de cajas ¡que se hacen cada vez mayores a medida que nos vamos hacia el pasado cada vez más lejano! Pero esto nos diría que la *inversa* de la segunda ley era válida en el pasado, con entropía creciente en la dirección del *pasado*, pese a nuestras expectativas, basadas en este argumento, de que la versión familiar de la segunda ley debería aplicarse en el futuro. Esta conclusión está en abierto conflicto con las observaciones del modo en que realmente se comportó nuestro universo en el pasado. Véase la Fig. 27.7.

¿Qué ha ido mal? Para tratar de entenderlo, apliquemos estos argumentos al comportamiento de nuestro gas en su recipiente, donde empezamos (en el instante t_0) con el gas enteramente en $\mathcal{R}$, de modo que x yace en $\mathcal{V}_{\mathcal{R}}$. Parece que obtenemos el comportamiento futuro correcto para el gas, en el que, una vez que la válvula está abierta, el gas fluye desde $\mathcal{R}$ al recipiente entero, y la entropía aumenta enormemente cuando x pasa rápidamente a la región $\mathcal{E}$ que representa el equilibrio térmico. Pero ¿qué pasa con el comportamiento pasado? Tenemos que preguntarnos qué es lo que sucedió inmediatamente antes del instante

t_0. ¿Cuál es la forma más probable de que el gas haya encontrado realmente su camino hacia $\mathcal{R}$? Si imaginamos que la válvula se abrió inmediatamente antes de t_0, entonces lo que parece que hemos obtenido como «evolución más probable» es que el gas empezó disperso por todo el recipiente, efectivamente en equilibrio térmico, en algún instante anterior a t_0, y luego se concentró espontáneamente cada vez más en la región $\mathcal{R}$, hasta encontrarse enteramente en $\mathcal{R}$ en el instante t_0.

Por absurdo que pueda parecer, esta es realmente la respuesta correcta al problema planteado de esta forma concreta, sin interferencia externa. En la práctica, nunca encontraríamos todo el gas en $\mathcal{R}$. El argumento nos dice solo cómo tendría que comportarse un gas que se mueve aleatoriamente *si* llegáramos a encontrar todo el gas *espontáneamente* en $\mathcal{R}$, lo que no hacemos. Aquí no hay ninguna paradoja. Pero esto elude el problema que quiero que considere el lector, que es: ¿cómo podría surgir en la práctica una situación semejante en la que todo el gas estuviese en $\mathcal{R}$? No hay problema en que esto suceda en nuestro universo real (donde adoptamos la definición relajada de la región $\mathcal{V}_{\mathcal{R}}$ que permite que una milésima parte del gas esté fuera de $\mathcal{R}$). Podríamos imaginar que un experimentador bombeó inicialmente una cantidad de gas diez veces mayor que la requerida dentro del recipiente entero, a continuación cerró la válvula, y por último conectó una bomba de vacío a la parte principal del recipiente para retirar prácticamente todo ese 90 % del gas. A lo largo de este proceso, la entropía habría estado aumentando continuamente, de acuerdo con la segunda ley.

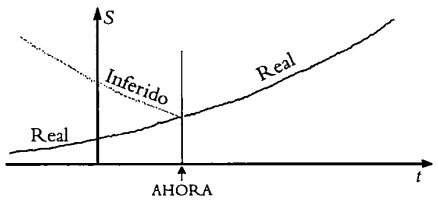


Fig. 27.7. Las conclusiones de la Fig. 27.6, expresadas en una representación de la entropía S frente al tiempo t. El razonamiento nos lleva correctamente a esperar una segunda ley para el comportamiento hacia nuestro futuro AHORA, pero nos da la respuesta aparentemente absurda de que la inversa de la segunda ley era válida en el pasado, en fuerte contradicción con la experiencia real.

Por supuesto, para discutir esto en términos de espacio de fases necesitamos un espacio de fases más grande en el que esté también incorporado el experimentador (y probablemente también una buena parte del universo, que quizá se extienda hasta el Sol o más allá). La entropía del cuerpo del experimentador se mantiene muy baja mediante los actos de comer y respirar. Supongamos, por simplicidad, que la bomba se opera manualmente, de lo contrario tenemos que preocuparnos por el origen de la baja entropía en la fuente de combustible (lo que plantea cuestiones que no son esenciales para nuestros objetivos actuales). Parte de la pequeña cantidad de entropía del experimentador se transfiere al recipiente con su gas, y se emplea en llevar el gas a $\mathcal{R}$. La baja entropía en las comidas del experimentador, y en última instancia en el aire, proviene del Sol exterior. Pronto volveré a este papel específico del Sol.

Así pues, hemos sido capaces de obtener la situación requerida, en la que prácticamente todo el gas del recipiente está en la región $\mathcal{R}$, sin violar la segunda ley, que tiene su forma físicamente apropiada: «la entropía aumenta con el tiempo». ¿Qué ha sucedido con nuestra dificultad con la deducción aparente de una segunda ley invertida en el tiempo para el comportamiento hacia el pasado? ¿Se ha resuelto? ¡No, ciertamente no lo ha hecho! El cuerpo del experimentador debería actuar (y, de hecho, lo hace) de acuerdo con la segunda ley real, como hace el Sol y cualquier cosa descrita por nuestro espacio de fases ampliado. Pero si tratamos de aplicar nuestro argumento del espacio de fases —ahora a este espacio de fases ampliado—, entonces parece que seguimos obteniendo el absurdo físico de que la entropía debe crecer de nuevo hacia el pasado, antes de cualquier instante en que examinamos nuestro sistema entero.

27.6. ¿Es el universo en su conjunto un «sistema aislado»?

Algunos teóricos tratan de hacer una distinción entre sistemas «aislados» y «abiertos», argumentando que mientras que la entropía aumenta (hasta que se alcanza el equilibrio) en un sistema aislado, existe siempre la posibilidad de un aporte del mundo externo que pueda servir

para reducir ocasionalmente la entropía, tal como una intervención por parte de un experimentador o un aporte de baja entropía procedente del Sol, etc. En mi opinión, cualquier explicación que se intente para la asimetría temporal en la segunda ley según estas líneas solo puede tener un estatus provisional, porque estas influencias exteriores pueden ser incorporadas en su totalidad dentro del sistema. Esto implica que el «sistema» bajo consideración tiene que ser realmente el universo en su conjunto. A veces la gente pone objeciones a esto, pero no veo ninguna justificación para tal objeción. Muy bien podría ser que el universo tenga una extensión infinita, pero esto no es obstáculo para que sea considerado como un todo (véase el capítulo 16). En cualquier caso, el universo podría ser espacialmente finito (una clara posibilidad a la que llegaremos inmediatamente) y parecería extraño confiar en un argumento para la segunda ley cuya validez dependa de que el universo sea en realidad espacialmente infinito. Como veremos más adelante, la distinción infinito/finito no tiene especial relevancia para la cuestión del origen de la segunda ley. La discusión de la entropía puede aplicarse, de hecho, al universo entero $\mathcal{U}$, donde un espacio de fases $\mathcal{P}_{\mathcal{U}}$ (cuyo volumen podría ser infinito) describe una totalidad de universos posibles, que incorpora todas sus evoluciones de acuerdo con las ecuaciones dinámicas de la dinámica clásica (apropiada).

Sin embargo, hay que hacer frente a algunos puntos difíciles. Para tratar el universo como un todo, tenemos que entrar en el ámbito de la cosmología, lo que no puede hacerse adecuadamente sin acudir a la relatividad general. Para hacer una discusión que esté completamente de acuerdo con el principio de covariancia general de la relatividad general (§19.6), sería necesario utilizar una descripción en la que no se singularice ninguna coordenada temporal con respecto a la cual se supone que «evoluciona» el universo. Una imagen que evoluciona en el tiempo es la forma en que hemos estado considerando un sistema físico, cuando lo representamos como un punto x que se mueve en un espacio de fases $\mathcal{P}$. Cada localización de x representa una descripción espacial (momento incluido) del sistema en un instante. Pero adoptar un punto de vista más relativista complicaría innecesariamente esta descripción, y no creo que para los puntos que quiero establecer aquí sea útil tratar de adoptar un punto de vista estrictamente relativista. De he-

cho, como veremos de inmediato, los modelos cosmológicos estándar poseen una coordenada temporal definida de forma natural, y esto da una buena aproximación a un «parámetro temporal» t con respecto al cual puede describirse la evolución del universo. Se considera que cada punto de $\mathcal{P}_{\mathcal{U}}$ describe no solo el contenido material del universo en el instante t, sino también la distribución (y el momento) de los campos continuos. El campo gravitatorio es uno de estos, de modo que la *geometría espacial* del universo (junto con su ritmo de cambio —dado por los datos iniciales apropiados para el campo gravitatorio—)[8] estará también codificada en la localización de x dentro de $\mathcal{P}_{\mathcal{U}}$.

De hecho, $\mathcal{P}_{\mathcal{U}}$ será de dimensión infinita, pero esto sucede con independencia de que el universo $\mathcal{U}$ sea infinito en extensión, y es una característica que se da también con todos los demás campos, tales como el campo electromagnético. Esto causa algunos problemas técnicos para la definición de la entropía, puesto que cada región $\mathcal{V}$ del espacio de fases requerido tendrá volumen infinito. Es habitual tratar este problema tomando prestadas ideas de la teoría cuántica (de campos), lo que permite obtener una respuesta finita para los volúmenes del espacio de fases que se refieren a sistemas que están apropiadamente acotados en energía y dimensión espacial. Los detalles no son importantes para nosotros. Aunque no hay ninguna forma completamente satisfactoria de tratar estas cuestiones en el caso de la gravedad —debido a la carencia de una teoría satisfactoria de la gravedad cuántica—, voy a considerarlas como tecnicismos que no afectan a la discusión general de las cuestiones planteadas por la segunda ley.

En este punto debería mencionar un equívoco que a menudo causa gran confusión con respecto a la segunda ley en un escenario cosmológico. Existe la idea común de que el incremento de entropía en la segunda ley es de algún modo solo una consecuencia necesaria de la expansión del universo. (Llegaremos a esta expansión en §27.11.) Esta opinión parece estar basada en la idea errónea de que hay relativamente pocos grados de libertad disponibles para el universo cuando es «pequeño», lo que pone un tipo de «techo» bajo para los valores posibles de la entropía, y hay más grados de libertad disponibles cuando el universo se hace más grande, lo que da un «techo» más alto y permite con ello entropías más altas. A medida que el universo se expande, este máximo admisible aumentaría, de modo que la entropía real del universo también podría aumentar.

Hay muchas maneras de ver que este punto de vista no puede ser correcto. Implica, por ejemplo, que en aquellos modelos de universo donde hay una fase de colapso la entropía necesariamente empieza a disminuir, en violación de la segunda ley. Hay quienes no se sentirían incómodos con esto,[9] pero este punto de vista tropieza con dificultades fundamentales, especialmente en presencia de agujeros negros.[10]

Consideraremos los agujeros negros enseguida (en §27.8), pero en realidad no necesitamos saber de ellos para ver por qué el punto de vista mencionado antes —que exige un «techo» para los valores de la entropía que depende del tamaño del universo— es erróneo. Esta no puede ser la explicación correcta para el incremento de entropía; en efecto, los grados de libertad disponibles para el universo están descritos por el espacio de fases total $\mathcal{P}_{\mathcal{U}}$. La dinámica de la relatividad general (que incluye el grado de libertad que describe el tamaño del universo) está tan descrita por el movimiento de nuestro punto x en el espacio de fases $\mathcal{P}_{\mathcal{U}}$ como lo están todos los demás procesos físicos involucrados. Este espacio de fases está precisamente «allí», y no «crece con el tiempo» en ningún sentido, pues el tiempo no es parte de $\mathcal{P}_{\mathcal{U}}$. No hay tal «techo», porque todos los estados que son dinámicamente accesibles al universo (o familia de universos) en consideración deben estar representados en $\mathcal{P}_{\mathcal{U}}$. Puede necesitarse algún tiempo para que x llegue a alguna caja de grano grueso desde alguna otra caja más pequeña dada, pero la noción de un «techo de entropía» no es apropiada (véase también §27.13).

Volvamos al argumento dado antes para demostrar la segunda ley. Utilizaremos el espacio de fases $\mathcal{P}_{\mathcal{U}}$ apropiado para el universo entero, de modo que la evolución del universo como un todo está descrita por el punto x que se mueve a lo largo de una curva ξ en $\mathcal{P}_{\mathcal{U}}$. La curva ξ está parametrizada por la coordenada temporal t, y a partir de la segunda ley podemos esperar que ξ entre en cajas de grano grueso de tamaño cada vez mayor a medida que t aumenta. Suponemos que a $\mathcal{P}_{\mathcal{U}}$ se le ha aplicado algún granulado grueso «razonable», pero si deseamos obtener valores finitos para las entropías que encuentra x, necesitaremos que los volúmenes de dichas cajas sean finitos. Parecería que para

conseguir esto, para un granulado grueso físicamente apropiado, hay que considerar el universo finito, con una cota finita para su energía disponible. De hecho, como veremos pronto, uno de los tres modelos cosmológicos estándar es de esta naturaleza, de modo que podemos imaginar que el argumento se está aplicando en una situación semejante. Pero no hay ningún requisito definido para esto, si no nos preocupa tener un valor numérico infinito para la entropía en cualquier otro instante. (Aún podemos dar un sentido matemático a la noción de que algunas cajas son «inmensamente mayores» que otras cajas, incluso si los volúmenes reales, y por consiguiente las entropías, son infinitas.)

27.7. El papel del big bang

¿Cómo vamos a imaginar que debe situarse nuestra curva parametrizada ξ, que representa una posible historia del universo, en el espacio de fases $\mathcal{P}_U$? Si ξ fuera arrojada simplemente de forma aleatoria en $\mathcal{P}_U$, entonces esperaríamos que con una abrumadora probabilidad yacería enteramente (o casi enteramente) en la caja más enorme de equilibrio térmico, $\mathcal{E}$, y no habría ninguna medida consistentemente discernible de «aumento de entropía» a lo largo de su longitud. Véase la Fig. 27.8a. Una situación semejante es absolutamente incompatible con el universo tal como realmente lo conocemos, en el que impera una segunda ley. También lo es la situación que se indica en la Fig. 27.8b,c, donde se obliga a que el punto x en ξ esté, en un instante concreto $t_0(> 0)$ que representa AHORA, en una región $\mathcal{V}$ de tamaño razonable, aunque no especialmente grande (que representa un universo con un valor de la entropía que resulta ser el que ahora observamos); por lo demás, la curva ξ se ha escogido de forma aleatoria. Esto corresponde a un universo cuya entropía aumenta en el futuro de AHORA, pero también aumenta en el pasado de AHORA —¡en violación de la segunda ley!—. Lo que realmente encontramos para un universo con nuestra familiar segunda ley es algo como la Fig. 27.8,b,d: aquí ξ tiene un extremo —el extremo pasado (digamos $t = 0$)— en una región extraordinariamente minúscula $\mathcal{B}$ dentro de $\mathcal{P}_U$ (que, por consiguiente, tiene una entropía extraordinariamente minúscula), a partir del cual serpentea (siguiendo las

leyes dinámicas), encontrando volúmenes de tamaño muchísimo mayor a medida que t aumenta; y para nuestro valor de t particular t_0, que representa AHORA, x resulta encontrarse en el aún bastante pequeño volumen $\mathcal{V}$ correspondiente al universo que observamos. Esto es simplemente lo que afirma la segunda ley, y obtenemos (d), (b), en contraste con (c), (b).

Permítanme tratar de reformular lo que se ha dicho antes. Supongamos que consideramos las cosas desde el punto de vista de un instante particular $t_0(> 0)$ que llamamos «AHORA», y encontramos que x, en el instante t_0, está en alguna región $\mathcal{V}$ de tamaño razonable. Entonces, mirando por donde se mueve ξ para valores mayores de t, vemos que entra en cajas de tamaños cada vez mayores a medida que t aumenta. Esto es compatible con la segunda ley y con la hipótesis anterior de que x no muestra «ningún sesgo concreto» en relación con las posiciones de las cajas. Pero visto desde una perspectiva de tiempo invertido, «partiendo» de t, con x en $\mathcal{V}$, da la impresión de que el punto x está guiado a propósito, hacia atrás en el tiempo, hacia la región absurdamente minúscula del espacio de fases que he etiquetado como $\mathcal{B}$.

Con respecto a esta dirección de tiempo invertido, el comportamiento de x parece increíblemente «sesgado», buscando cajas que se hacen sucesivamente más pequeñas en un grado extraordinario a medida que el tiempo se mueve hacia el pasado. ¿Tenemos que entender esto como una perversa búsqueda «deliberada» de cajas cada vez más pequeñas, simplemente por magia diabólica? No, lo que sucede simplemente es que $\mathcal{B}$ está rodeado por cajas que se hacen sucesivamente más pequeñas (véase la Fig. 27.8), de modo que si ξ va a llegar a $\mathcal{B}$, cuando t se acerca a 0 *tiene* que encontrar cajas cada vez menores en este camino. ¡El enigma reside simplemente en el hecho de que un extremo de ξ tiene que estar en $\mathcal{B}$! Esto es lo que debemos entender si queremos comprender la fuente de la segunda ley. La región $\mathcal{B}$ representa el big bang origen del universo, ¡y pronto veremos qué tamaño ridículamente minúsculo tiene en realidad esta región!

Debemos tratar de comprender lo que está implicado aquí. ¿Qué tiene $\mathcal{B}$ de especial? ¿Podemos asignar una medida numérica a este grado de especialidad? En cualquier caso, ¿cuáles son las razones observacionales para creer en el big bang?

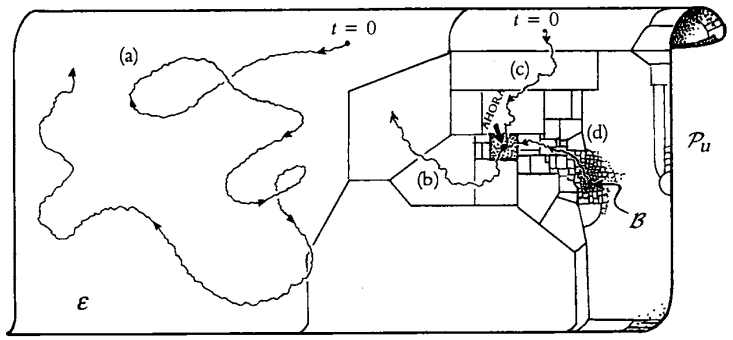


Fig. 27.8. Diferentes evoluciones posibles del universo, descritas por una curva parametrizada ξ en el espacio de fases $\mathcal{P}_u$ de los posibles estados de universo (de masa total fija, digamos, o de cualquier otra cantidad conservada). (a) Si la curva ξ se arroja aleatoriamente en $\mathcal{P}_u$, pasa casi toda su vida en $\mathcal{E}$, y, aparte de fluctuaciones menores, el universo apenas difiere del «equilibrio térmico» (y podría parecerse a la Fig. 27.20d si es cerrado). (b) Si especificamos meramente que la curva empieza AHORA en una caja muy pequeña $\mathcal{V}$ (que se muestra rayada), considerando que los puntos de $\mathcal{V}$ se parecen al universo en que ahora vivimos, pero donde, por lo demás, ξ se arroja aleatoriamente, entonces encontramos una evolución futura consistente con lo que seguimos viendo, con el aumento de entropía de la segunda ley. (c) Si aplicamos la misma consideración a la situación en la que la curva entera ξ está obligada meramente a atravesar $\mathcal{V}$ en un tiempo concreto $t_N > 0$ (AHORA), entonces encontramos un futuro razonable para el universo, pero, como sucede con la Fig. 27.6b, encontramos una fuerte violación de la segunda ley en el pasado. (d) Esto se remedia si especificamos, además, que el extremo inicial ($t = 0$) de ξ yace en la región absurdamente minúscula $\mathcal{B}$, en la que el universo empieza con el big bang extraordinariamente especial que aparentemente ocurrió en nuestro universo real.

Las razones para creer en un origen explosivo del universo procedían inicialmente de un estudio teórico de la ecuación de Einstein en un contexto cosmológico, llevado a cabo por Alexandr Friedmann en 1922 (véase §27.11 más adelante). Luego, en 1929, Edwin Hubble hizo el extraordinario descubrimiento de que las galaxias lejanas se están alejando realmente de nosotros[11] de una forma que parecía implicar que la materia en el universo era el resultado de una tremenda explosión. Según los cálculos modernos, la explosión —ahora denominada el big bang— tuvo lugar hace unos $1{,}4 \times 10^{10}$ años. Las conclusiones de Hubble se basaban en el hecho de que la luz procedente

de objetos que se alejan a gran velocidad está *desplazada hacia el rojo* (es decir, que las líneas espectrales están desplazadas hacia «el extremo rojo del espectro», i.e., hacia longitudes de onda más largas) debido al efecto Doppler.[27.10] Encontró que este desplazamiento hacia el rojo era sistemáticamente mayor cuanto más lejana parecía estar la galaxia, lo que indica una velocidad de recesión que es proporcional a su distancia a nosotros, compatible con la imagen de la «explosión».

Pero la prueba directa más impresionante de apoyo observacional para el big bang es la presencia universal de una radiación que llena el espacio y que tiene una temperatura de unos 2,7 K (i.e., 2,7 °C por encima del cero absoluto).[12] Aunque esta puede parecer una temperatura extraordinariamente baja para un suceso tan violento, se cree que esta radiación es realmente el «destello» del propio big bang, enormemente atenuada («desplazada hacia el rojo») y enfriada, debido a la vasta expansión del universo. La radiación de 2,7 K desempeña un papel de extraordinaria importancia en la cosmología moderna. Normalmente se la conoce como el «fondo (cósmico) de microondas», o a veces como la «radiación de cuerpo negro de fondo», o la radiación «reliquia cósmica». Es extraordinariamente uniforme (aproximadamente hasta una parte en 10^5), lo que indica que el propio universo primitivo era extraordinariamente uniforme inmediatamente después del big bang, y está muy bien descrita por los modelos cosmológicos que consideraremos en §27.11.

Tratemos ahora de hacernos una idea física de la *naturaleza* de la ligadura de entropía extraordinariamente baja en el big bang que restringe $\mathcal{B}$ a un volumen tan minúsculo.[13] Veremos que lo que fue tan especial en el big bang era en realidad su gran uniformidad, como acabamos de mencionar. Debemos tratar de comprender por qué esto corresponde a una entropía muy baja, y cómo nos proporciona una segunda ley que es relevante para nosotros, aquí en la Tierra, en la forma familiar que conocemos.

Consideremos primero, una vez más, el papel del Sol como una

[27.10] Deduzca el desplazamiento de frecuencia Doppler según la relatividad especial para una fuente que se aleja a velocidad v (a) utilizando una imagen ondulatoria de la luz, y (b) utilizando productos escalares de 4-vectores y $E = h\nu$.

fuente de baja entropía. Hay extendida una falsa idea de que nuestra supervivencia depende de la *energía* que el Sol suministra. Esto es equívoco, pues para que dicha energía sea de alguna utilidad para nosotros debe proporcionarse en una forma de baja entropía. Si el cielo entero hubiera estado uniformemente iluminado, por ejemplo, con alguna temperatura uniforme —ya fuera la del Sol o cualquier otra cosa—, entonces no habría forma de utilizar esta energía (cualquiera que sea el tipo de criatura que pudiéramos imaginar que había evolucionado para tratar de aprovecharla). Un suministro de energía en equilibrio térmico es inútil. Sin embargo, nosotros tenemos la fortuna de que el Sol es un punto caliente en un *fondo frío*. Durante el día, la energía llega a la Tierra desde el Sol, pero durante el transcurso del día y la noche vuelve de nuevo al espacio. El balance neto de energía se reduce simplemente (en promedio) a que devolvemos toda la energía que recibimos.[14]

Sin embargo, la energía que obtenemos del Sol está en forma de fotones individuales de *alta* energía (básicamente fotones amarillos de alta frecuencia debido a la alta temperatura del Sol), mientras que la mayor parte de esta energía vuelve al espacio en forma de fotones de *baja* energía (infrarrojos, baja frecuencia). (Esta relación para la energía del fotón se sigue de la fórmula de Planck $E = h\nu$ y sus ideas sobre la radiación de cuerpo negro; véase §21.4.) Debido a su mayor energía (mayor temperatura), hay muchos menos fotones procedentes del Sol que fotones devueltos al espacio, porque la energía *total* transportada por ellos es la misma cuando entra que cuando sale. El menor número de fotones del Sol significa menos grados de libertad y, por consiguiente, una menor región del espacio de fases y con ello una menor entropía que en los fotones devueltos al espacio. Las plantas hacen uso de esta energía de baja entropía en su fotosíntesis, reduciendo con ello su propia entropía.Y nosotros nos aprovechamos de las plantas para reducir la nuestra, comiéndolas, o comiendo algo que las come, y respirando el oxígeno que las plantas liberan; véase la Fig. 27.9.

Pero ¿por qué es el Sol un punto caliente en un cielo frío? Aunque la historia detallada es complicada, en última instancia se reduce al hecho de que el Sol —y todas las demás estrellas— se han condensado gravitacionalmente a partir de un gas previamente uniforme (com-

Fig. 27.9. La Tierra devuelve la misma cantidad de energía que recibe del Sol, pero la que recibe del Sol está en una forma de entropía mucho menor, debido al hecho de que la luz amarilla del Sol tiene frecuencia más alta que la infrarroja que devuelve la Tierra. Por consiguiente, por la $E = h\nu$ de Planck, los fotones del Sol llevan más energía por fotón que los que devuelve la Tierra, de modo que la energía del Sol es transportada por menos fotones que la devuelta por la Tierra. Menos fotones significa menos grados de libertad y, por lo tanto, una menor región en el espacio de fases y una entropía más baja que en los fotones devueltos al espacio. Las plantas hacen uso de esta energía de baja entropía en la fotosíntesis, con la que reducen su propia entropía, y nosotros nos aprovechamos de las plantas para reducir la nuestra, comiéndolas o comiendo algo que las come, y respirando el oxígeno que liberan las plantas. Esto procede, en última instancia, de un desequilibrio en la temperatura en el cielo que fue resultado de la acumulación gravitatoria que dio lugar al Sol.

puesto principalmente de hidrógeno). Cualesquiera que sean las otras influencias presentes (fundamentalmente fuerzas nucleares), ¡el Sol ni siquiera podría existir sin gravedad! La «pequeñez» en la entropía del Sol (a distancia considerable del equilibrio térmico) procede de una enorme reserva de baja entropía que está potencialmente disponible en la *uniformidad* del gas a partir del cual el Sol se ha condensado gravitatoriamente.

La gravitación es algo confusa, en relación con la entropía, a causa de su naturaleza universalmente atractiva. Estamos acostumbrados a pensar en la entropía en términos de un gas ordinario; en este caso, tener el gas concentrado en pequeñas regiones representa *baja* entropía (como en el caso de nuestro recipiente en la Fig. 27.3), y el estado de alta entropía de equilibrio térmico corresponde al gas uniformemente

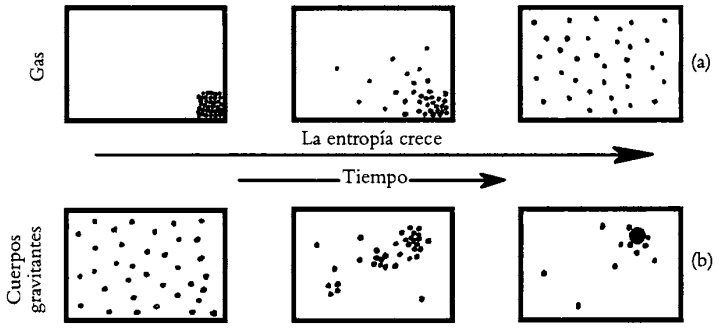


Fig. 27.10. Aumento de entropía con aumento del tiempo, de izquierda a derecha. (a) Para el gas en una caja, inicialmente agregado todo en una esquina, la entropía aumenta a medida que el gas se difunde por toda la caja, hasta alcanzar el estado uniforme de equilibrio térmico. (b) Con gravedad, las cosas tienden a ser de otra forma. Un sistema inicial de cuerpos gravitantes uniformemente dispersos representa una entropía relativamente baja, y las acumulaciones tienden a ocurrir a medida que aumenta la entropía. Por último, hay un enorme aumento de entropía cuando se forma un agujero negro que engulle la mayor parte del material.

disperso. Pero con la gravedad las cosas tienden a ser de otra manera. Un sistema uniformemente disperso de cuerpos gravitantes tendría una entropía relativamente *baja* (a menos que las velocidades de los cuerpos sean enormemente altas y/o los cuerpos sean muy pequeños y/o muy dispersos, de modo que las contribuciones gravitatorias se hacen insignificantes), mientras que la alta entropía se consigue cuando se amontonan los cuerpos gravitantes (Fig. 27.10).

¿Qué hay del estado de máxima entropía? Mientras que para un gas el estado de máxima entropía de equilibrio térmico corresponde al gas uniformemente disperso en la región en cuestión, en el caso de grandes cuerpos *gravitantes* la entropía máxima se consigue cuando toda la masa se concentra en un lugar, en la forma de una entidad conocida como un *agujero negro*. Tendremos que entender algo de estos objetos extraños y maravillosos para poder avanzar y obtener una buena estimación de la entropía potencialmente disponible en el universo en su conjunto. Esto nos permitirá estimar los volúmenes requeridos de $\mathcal{B}$ y $\mathcal{P}_U$.

27.8. Agujeros negros

¿Qué es un agujero negro? En términos generales, es una región del espaciotiempo que ha resultado de un colapso gravitatorio de material, donde la atracción gravitatoria se ha hecho tan fuerte que ni siquiera la luz puede escapar. Para obtener una imagen intuitiva de por qué puede darse una situación semejante, consideremos la noción newtoniana de *velocidad de escape*. Si se lanza una piedra hacia arriba desde el suelo a cierta velocidad v, la piedra volverá a caer al suelo una vez que haya alcanzado determinada altura; esta altura es aquella para la que la energía cinética de la piedra ha sido enteramente utilizada para superar la diferencia de energía potencial con respecto del nivel del suelo (§17.3 y §18.6). La altura desde el suelo depende enteramente de la velocidad de lanzamiento, si se ignoran los efectos de la resistencia del aire.[27.11] Sin embargo, para una velocidad que supere el valor $(2GM/R)^{1/2}$, conocido como *velocidad de escape*, la piedra escaparía por completo del campo gravitatorio de la Tierra. (Aquí M y R son la masa y el radio de la Tierra, respectivamente, y G es la constante gravitatoria de Newton.) Supongamos ahora que en lugar de la Tierra tenemos un cuerpo mucho más masivo y concentrado. Entonces la velocidad de escape será mayor (puesto que M/R aumenta si M aumenta y si R disminuye), y podríamos imaginar que la masa y la concentración fueran tan enormes que la velocidad de escape en la superficie superara incluso la velocidad de la luz.

Podemos creer que cuando esto sucediera, en la teoría newtoniana, el cuerpo aparecería completamente oscuro visto desde grandes distancias, porque ninguna luz podría escapar del mismo; esta fue, de hecho, la conclusión a la que llegó el astrónomo y clérigo inglés John Michell en 1784. Más tarde, en 1799, el gran físico y matemático francés Pierre-Simon Laplace llegó a la misma conclusión.[15] Sin embargo, la situación no me parece tan clara debido a que la velocidad de la luz no tiene ningún estatus absoluto en la teoría newtoniana, y uno puede hacer un buen argumento en el sentido de que, para un cuerpo seme-

[27.11] Demuestre que dicha altura es $v^2R(2gR - v^2)^{-1}$, donde R es el radio de la Tierra y g la aceleración debida a la gravedad en la superficie de la Tierra.

jante, la velocidad de la luz en su superficie debería ser considerablemente mayor que la medida en el espacio libre, y que la luz podría seguir escapando al infinito por muy masivo y concentrado que pudiera ser dicho cuerpo.[16,[27.12]] Así pues, la «estrella oscura» de Michell, aunque es un clarividente precursor de la idea del agujero negro, no proporciona, en mi opinión, un argumento convincente a favor de objetos gravitantes «invisibles» en la teoría newtoniana.

Esta cuestión es mucho más pertinente en el contexto de la teoría de la relatividad, puesto que aquí la velocidad de la luz es fundamental y de hecho representa la velocidad límite para toda señalización (§17.8). No obstante, puesto que estamos interesados en un fenómeno gravitatorio, exigimos un espaciotiempo de relatividad *general* en lugar de solo el espacio de Minkowski. En la relatividad general existe la expectativa de que ocurrirán situaciones en las que la velocidad de escape supere la velocidad de la luz, dando como resultado lo que ahora llamamos un *agujero negro*.

Se espera que resulte un agujero negro cuando un gran cuerpo masivo alcanza una etapa en la que las fuerzas de presión interna son insuficientes para mantener el cuerpo contra el implacable tirón hacia dentro de su propia influencia gravitatoria. De hecho, debe esperarse semejante colapso gravitacional cuando una gran estrella, de una masa total varias veces superior a la del Sol (digamos $10M_\odot$, donde $1M_\odot$ es una masa solar) agota todas sus fuentes internas de energía disponibles, de modo que se enfría y no puede mantener una presión suficiente para evitar el colapso. Cuando esto sucede, el colapso puede hacerse imparable a medida que los efectos gravitatorios aumentan implacablemente.

La imagen detallada puede hacerse muy complicada, especialmente por el hecho de que en condiciones de presión muy alta cobran importancia cuestiones sofisticadas relativas al comportamiento de la materia. De especial relevancia es la *presión de degeneración* electrónica o neutrónica. Esta tiene que ver con el principio de Pauli que, como se ha señalado en §23.7, impide que dos o más fermiones idénticos estén

[27.12] ¿Puede ver por qué? *Sugerencia*: Piense en términos de una teoría de luz como partículas, y de luz externa que cae a la superficie del cuerpo. ¿Qué sucede si la luz cae sobre un espejo horizontal en la superficie del cuerpo?

en el mismo estado cuántico. Una *enana blanca*, que podría tener algo del orden de una masa solar concentrado en aproximadamente el tamaño de la Tierra, se mantiene por la presión de degeneración electrónica; una *estrella de neutrones* de la misma masa sería un cuerpo de unos 10 kilómetros de diámetro, mantenido principalmente por la presión de degeneración neutrónica. (Una pelota de tenis llena del material de una estrella de neutrones ¡pesaría tanto como Deimos, una luna de Marte!) Sin embargo, debido a los requisitos de la relatividad, resulta que la presión de degeneración no puede sostener una estrella si la masa es mayor de aproximadamente $2M_\odot$. El resultado clave fue obtenido por Subrahmanyan Chandrasekhar, en 1931, cuando estableció este límite de aproximadamente $1{,}4M_\odot$ para las enanas blancas. Refinamientos posteriores obtuvieron un límite ligeramente mayor para las estrellas de neutrones.[17] El resultado de todo esto es que no hay ninguna configuración en reposo para un objeto frío de más de aproximadamente $2{,}8M_\odot$ (y probablemente no más de $1{,}6M_\odot$). Un objeto semejante colapsaría hacia dentro, y seguiría colapsando hasta alcanzar una dimensión en la que las consideraciones de Michell empiezan a hacerse relevantes. ¿Qué sucede entonces?

Volvamos a nuestra gran estrella de, digamos, $10M_\odot$, supuesta inicialmente a una temperatura suficientemente alta para que las presiones térmicas puedan mantenerla. Sin embargo, cuando la estrella se enfríe pasará por una etapa en la que su núcleo comprimido superará el límite de Chandrasekhar y colapsará. La caída hacia dentro de las regiones exteriores podría desencadenar una violenta explosión, que se conoce como una *supernova*. Tales estrellas en explosión han sido observadas con frecuencia, sobre todo en otras galaxias, y durante algunos días el brillo de la supernova puede sobrepasar al de la galaxia entera en la que reside. Pero si no se expulsa material suficiente en una explosión semejante —y, para una estrella de inicialmente $10M_\odot$, es poco probable que se pierda tanto—, las predicciones dicen que la estrella colapsará sin parar hasta que alcance la escala donde se aplican las consideraciones de Michell. Examinemos la Fig. 27.11, que es un diagrama espaciotemporal que muestra el colapso a un agujero negro. (Por supuesto, ha habido que prescindir de una dimensión espacial.) Vemos que la materia sigue colapsando hacia dentro, a través de dicha super-

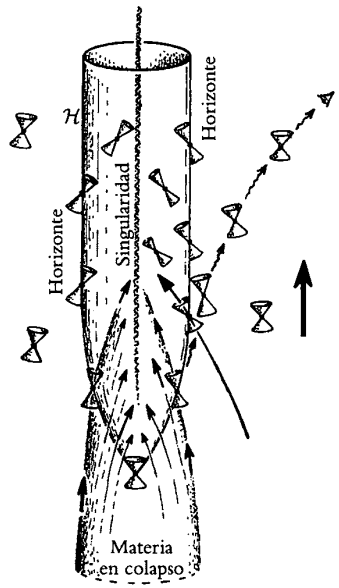


Fig. 27.11. Diagrama espaciotemporal del colapso a un agujero negro. (Se ha suprimido una dimensión espacial.) La materia colapsa hacia dentro, a través de la 3-superficie que se convierte en el horizonte de sucesos (absoluto). Ninguna materia o información puede escapar del agujero una vez que se ha formado. Los conos nulos son tangentes al horizonte y permiten que la materia o las señales pasen hacia dentro pero no hacia fuera. Un observador externo no puede ver el interior del agujero, sino solo la materia —enormemente oscurecida y desplazada hacia el rojo— inmediatamente antes de que entre en el agujero.

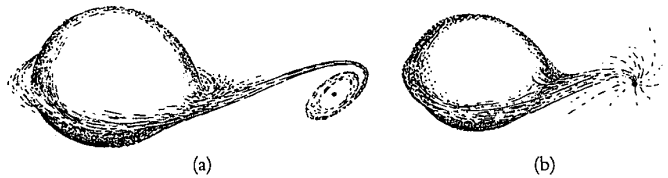
ficie —denominada un *horizonte de sucesos* (absoluto)— donde la velocidad de escape alcanza la velocidad de la luz. A partir de entonces, ninguna información adicional procedente de la propia estrella puede llegar a un observador exterior y se forma un agujero negro.

La imagen de la Fig. 27.11 está basada en la famosa solución de Schwarzschild de la ecuación de Einstein, descubierta por Karl Schwarzschild en 1916,[18] poco después de que Einstein publicara su teoría, y solo unos meses antes de que Schwarzschild muriese de una rara enfermedad contraída en el frente oriental durante la Primera Guerra Mundial. Esta solución describe el campo gravitatorio estático que rodea un cuerpo con simetría esférica, ya se esté contrayendo o no. El horizonte ocurre a la distancia radial $r = 2MG/c^2$ (exactamente el valor crítico de Michell).[27.13]

El horizonte de sucesos no está hecho de ninguna sustancia material. Es meramente una (hiper)superficie particular en el espaciotiempo que separa aquellos lugares de los que pueden escapar señales al infinito externo de aquellos otros lugares tales que todas las señales que partieran de ellos serían atrapadas inevitablemente por el agujero negro. Un infortunado observador que atravesara el horizonte de sucesos, de fuera a dentro, no advertiría nada localmente peculiar en el momento de cruzar el horizonte. Además, el propio agujero negro no es un cuerpo ponderable; lo consideramos como una región gravitante del espaciotiempo de cuyo interior no puede escapar ninguna señal. ¿Y qué pasa con el destino de la pobre estrella? Volveremos a este enigma en §27.9.

En primer lugar, consideremos la situación observacional. ¿Hay evidencia de la existencia real de agujeros negros? De hecho, la hay. En la década de 1970 se conocían varios ejemplos de sistemas de «estrellas dobles», en los que solo un miembro del par era luminoso con luz visible. La existencia, la masa y el movimiento del otro miembro se inferían a partir de los detalles finos del movimiento de su compañera visible. Además, a partir de la emisión de señales de rayos X procedentes de su vecindad, se deducía que la compañera invisible era un objeto compacto, con una masa demasiado grande para que el objeto fuera de uno de los dos tipos —una enana blanca o una estrella de neutrones— que los principios físicos aceptados permitían para una estrella com-

[27.13] La forma métrica original de Schwarzschild era $ds^2 = (1 - 2M/r)dt^2 - (1 - 2M/r)^{-1}dr^2 - r^2(d\theta^2 + \text{sen}^2\theta\, d\phi^2)$, donde se han escogido unidades tales que $G = c = 1$ y donde θ y ϕ son coordenadas polares esféricas estándar (§22.11). Explique cómo queda determinada la coordenada radial r por un requisito sobre el área de las esferas de r y t constantes. Esta forma métrica no se extiende suavemente a la región $r \leq 2M$; para esta puede utilizarse la forma de Eddington-Finkelstein de la métrica $ds^2 = (1 - 2M/r)dv^2 - 2dv\,dr - r^2(d\theta^2 + \text{sen}^2\theta d\phi^2)$. Encuentre un cambio de coordenadas que relacione a ambas explícitamente. Explique por qué las curvas nulas en cada plano (v, r) deben ser las geodésicas radiales nulas, y utilice este hecho para obtener sus ecuaciones y representarlas. (Dibuje las líneas de r constante como verticales y las líneas de v constante con pendiente de 45° hacia la derecha.) Identifique el horizonte de sucesos y la singularidad (cf. §27.9).



(a) (b)

Fig. 27.12. Sistemas de estrellas dobles, un miembro de los cuales es un agujero negro (minúsculo). (a) La materia atraída de la estrella más grande por el agujero negro forma un disco de acreción a su alrededor, que poco a poco cae en espiral y se calienta hasta que se emiten rayos X antes de que el material entre realmente en el agujero. (b) En algunos casos no hay disco de acreción y el material cae «en línea recta». Si el objeto compacto atrayente tuviera una superficie ponderable, el material en caída se calentaría, pero no se ve ningún brillo, lo que confirma la presencia de un agujero negro.

pacta. La emisión de rayos X era compatible con que el objeto invisible fuera un agujero negro rodeado de lo que se conoce como un «disco de acreción» de gas y polvo, que se acerca poco a poco en espiral hacia el agujero, calentándose enormemente a medida que se aproxima al centro. Por último, se emitirían rayos X antes de que el material entrara realmente en el agujero (véase la Fig. 27.12a). El más conocido (y en esa época el más convincente observacionalmente) de estos candidatos a agujero negro era la fuente de rayos X Cygnus X-1, donde el miembro compacto y oscuro del par tiene una masa de aproximadamente $7M_\odot$, lo que ciertamente lo descarta como enana blanca o estrella de neutrones, de acuerdo con la teoría aceptada.

Este tipo de evidencia era siempre más bien indirecta y menos que completamente satisfactoria, porque se basaba en la *teoría* para decirnos que tales objetos compactos masivos no pueden existir como cuerpos ponderables. Sin embargo, ahora hay una evidencia bastante impresionante a favor de los agujeros negros. Los discos de acreción no son las únicas configuraciones que adopta el material que cae en agujeros negros. En ciertos casos el material cae «directo hacia dentro», y este comportamiento parece haber sido observado en la actualidad (Fig. 27.12b). Si el objeto compacto atrayente tuviese una superficie material de algún tipo, entonces el material en caída calentaría dicha superficie, y su brillo se haría visible al cabo de cierto tiempo. Pero no se ve

tal brillo. Así pues, ahora hay evidencia directa de que una entidad tan compacta no tiene ninguna superficie, y puede inferirse de forma muy convincente que la entidad es realmente un agujero negro.[19]

Todo esto se refiere a agujeros negros «estelares», cuya masa sería solo de algunas veces la masa del Sol. Hay también una evidencia impresionante a favor de agujeros negros mucho más grandes. Parece que la mayoría —y quizá todas— de las galaxias tienen agujeros negros de gran tamaño en sus centros. En particular, parece haber un agujero negro masivo de $3 \times 10^6 M_\odot$ en el centro de nuestra Vía Láctea, y se han rastreado en detalle los movimientos reales de las estrellas que describen órbitas a su alrededor, que son plenamente compatibles con esta imagen de agujero negro.

27.9. Horizontes de sucesos y singularidades espaciotemporales

En la Fig. 27.11 he dibujado algunos de los conos nulos futuros, de modo que las propiedades de causalidad del espaciotiempo deberían hacerse razonablemente evidentes. La característica más esencial es la existencia del horizonte de sucesos del agujero negro que, en el espaciotiempo, es una 3-superficie $\mathcal{H}$. Como se ha establecido en §27.8, esta tiene la propiedad de que ninguna señal que se origine en la región interior a $\mathcal{H}$ puede escapar a la región exterior. Esto puede verse como un efecto de la inclinación de los conos hacia dentro, de modo que se encuentran siendo tangentes a $\mathcal{H}$. Cualquier línea de universo que cruza del interior al exterior de $\mathcal{H}$ tendría que violar la causalidad que definen los conos (§17.7). He representado el caso donde el colapso gravitatorio tiene simetría esférica completa, que fue la situación original estudiada por J. Robert Oppenheimer y Hartland Snyder en 1939, y que utiliza la geometría de Schwarzschild para describir la región externa a la materia en caída.

Aunque el horizonte $\mathcal{H}$ tiene propiedades extrañas, la geometría local allí no es significativamente diferente de la de cualquier otra parte. Como se ha señalado, un observador en una nave espacial no notaría que suceda nada especial cuando se cruza el horizonte desde el ex-

terior al interior. Pese a todo, una vez que se haya emprendido este viaje peligroso no habrá retorno. La inclinación de los conos nulos es tal que no hay escape, y el observador encontraría rápidamente efectos de marea crecientes (curvatura del espaciotiempo; véanse §17.5 y §19.6) que divergen a infinito en la singularidad espaciotemporal en el centro ($r = 0$). Estas características no son específicas del caso de simetría esférica, sino que son completamente generales. Existen, de hecho, teoremas de aplicación muy general que nos dicen que las singularidades no pueden ser evitadas en ningún colapso gravitatorio que supere cierto «punto de no retorno».[20] Algunas de las cuestiones relevantes serán expuestas con algo más de detalle en §28.8.

Para un agujero negro de unas pocas masas solares, las fuerzas de marea serían fácilmente suficientes para matar a una persona mucho antes de llegar siquiera al horizonte, y no digamos si lo cruzara, pero para los agujeros negros grandes de $10^6 M_\odot$, o más, que se cree que residen en los centros galácticos, no habría ningún problema particular derivado de los efectos de marea cuando se cruza el horizonte (que en este caso tendría algunos millones de kilómetros de diámetro). De hecho, en el caso de nuestra propia galaxia, la curvatura en el horizonte de su agujero negro es solo de unas veinte veces la curvatura espaciotemporal en la superficie de la Tierra —¡que ni siquiera advertimos!—. Pese a todo, el arrastre implacable del observador hacia la singularidad central haría posteriormente que los efectos de marea aumentaran con gran rapidez hasta el infinito, ¡destruyendo totalmente al observador en menos de un minuto! La destrucción por efectos de marea rápidamente crecientes es, de hecho, lo que espera a cualquier material físico cuando se hunde hacia el centro del agujero negro. Recordemos nuestro interés sobre el destino del material de nuestra estrella colapsante de $10M_\odot$. Incluso las partículas individuales de las que está compuesta encontrarían pronto fuerzas de marea tan intensas que las desgarrarían —¿para dar qué?, ¡nadie lo sabe!

Al menos, lo que *sí* sabemos es que, mientras pueda mantenerse una imagen einsteiniana de un espaciotiempo clásico que actúa de acuerdo con las ecuaciones de Einstein (con densidades de energía no negativas y algunas otras hipótesis suaves y «razonables), se encontrará una singularidad espaciotemporal dentro del agujero.[21] Esperamos que la ecuación de Einstein nos dirá que esta singularidad no puede ser evitada por la materia que hay en el agujero y que las «fuerzas de marea» (i.e., curvatura de Weyl; véase §19.7) divergirán a infinito, muy posiblemente de una forma salvajemente oscilatoria, en el caso general.[22] De hecho, parece inevitable entrar en el dominio de la *gravedad cuántica* (o cualquiera que sea el término apropiado), de modo que estas predicciones de la teoría clásica tendrán que ser modificadas de acuerdo con ello. Aún no sabemos cómo debe ser la teoría de la «gravedad cuántica» correcta, pero estas consideraciones de agujeros negros nos proporcionan un elemento importante; y este elemento debería guiarnos en la dirección apropiada en nuestra búsqueda de la «gravedad cuántica» correcta. Estas cuestiones serán importantes para nosotros en capítulos posteriores, en particular en los capítulos 30, 31 y 32.

Hay una creencia generalizada de que las singularidades espaciotemporales del colapso gravitatorio estarán necesariamente siempre dentro de un horizonte de sucesos, de modo que cualesquiera que resulten ser los extraordinarios efectos físicos en tal singularidad estarán ocultos a la vista de cualquier observador externo. Sin embargo, esta no es una propiedad matemáticamente establecida de la relatividad general. La hipótesis de que las singularidades estarán siempre ocultas se conoce como *censura cósmica*,[23] y se examinará con más detalle en §28.8.

Por otra parte, no tenemos que ir tan lejos como la singularidad para encontrar efectos extraordinarios resultantes del colapso gravitatorio. Existen algunos procesos muy violentos y visibles en el universo. Por ejemplo, se cree que los excepcionalmente luminosos *cuásares* están impulsados por agujeros negros en rotación en los centros galácticos: la rotación del agujero negro es la central energética, aunque el material real expulsado (aparentemente a lo largo del eje de rotación) procede del exterior del agujero (véase §30.7). La energía emitida por algunos cuásares, aunque procedente de una región minúscula (aproximadamente del tamaño del sistema solar), ¡puede sobrepasar en brillo a una galaxia entera en un factor de 10^2 o 10^3 o más! Pueden verse a distancias enormemente grandes, y son importantes herramientas observacionales para la cosmología. Existen también fuentes de potentes rayos γ (fotones extremadamente energéticos) que también se cree que implican agujeros negros, quizá pares de agujeros negros en colisión.[24]

27.10. Entropía de agujero negro

Volvamos a la consideración de las regiones externas «más seguras» de los agujeros negros estacionarios («muertos») aislados. Veremos cuán extraordinariamente grande es la entropía que se puede asignar a un objeto semejante. En primer lugar, deberíamos tomar nota del hecho de que hay teoremas matemáticos[25] que ofrecen una prueba convincente de que los agujeros negros generales, que inicialmente pueden poseer irregularidades complicadas debidas a un colapso asimétrico —quizá en una catástrofe irreversible en caída en espiral incontrolada—, se asentarán rápidamente en cualquier caso (hasta que adquieran una forma geométrica elegante y notablemente simple). Esto está descrito por la *métrica de Kerr*,[26] que está caracterizada por solo dos parámetros (números reales) físico/geométricos etiquetados como m y a.[27] Aquí m describe la *masa* total del agujero negro, y $a \times m$ es el *momento angular* total (en unidades donde $G = c = 1$). Como ha escrito el premio Nobel Subrahmanyan Chandrasekhar (cuyo famoso resultado de 1931, como se ha dicho en §27.8, puso a la astrofísica en el camino hacia el agujero negro):

> Los agujeros negros de la naturaleza son los más perfectos objetos macroscópicos que hay en el universo: los únicos elementos en su construcción son nuestros conceptos de espacio y tiempo. Y puesto que la teoría de la relatividad general proporciona solo una única familia de soluciones para sus descripciones, son también los objetos más simples.[28]

La naturaleza implacable de los agujeros negros, que absorben todo tipo de material —que podría tener una inmensa cantidad de estructura detallada— para convertirlo en una simple configuración descriptible por solo diez parámetros (siendo estos a, m, la dirección del eje de giro, la posición del centro de masas y su 3-velocidad), es una manifestación poderosa de la segunda ley. Estos diez parámetros son todo lo que se necesita para una caracterización macroscópica adecuada del estado final.[29] Aunque un agujero negro no se parece a la materia ordinaria en equilibrio térmico, comparte con ella la propiedad clave de que números enormes de estados microscópicamente distintos llevan a

algo que puede describirse con muy pocos parámetros. Por esta razón, la caja de grano grueso en el espacio de fases correspondiente es realmente inimaginable, y los agujeros negros, en consecuencia, tienen entropías enormes.

De hecho, la entropía de los agujeros negros tiene una notable interpretación geométrica: ¡es proporcional al área del horizonte del agujero! De acuerdo con la famosa *fórmula de Bekenstein-Hawking*, puede atribuirse a un agujero negro una entropía bien definida, que es

$$S_{\mathrm{BH}} = \frac{kc^3 A}{4G\hbar},$$

donde A es el área de la superficie del horizonte del agujero negro, y donde se puede tomar BH como representación de Bekenstein-Hawking, o *black-hole* (agujero negro), ¡a voluntad! Nótese la aparición de la constante de Planck, así como de la constante gravitatoria, lo que indica que esta entropía es un efecto de «gravedad cuántica». De hecho, este es el primer lugar donde hemos encontrado que la constante fundamental de la mecánica cuántica (la constante de Planck, escrita en la forma de Dirac $\hbar$) y la de la relatividad general (la constante gravitatoria de Newton G) aparecen juntas en la misma fórmula.

Para cuestiones de física fundamental, en las que están involucradas la mecánica cuántica y la relatividad general, es a menudo conveniente adoptar unidades en las que ambas constantes se toman como la *unidad*. Ya hemos visto en §17.8 y §§19.2,6,7 (y en otros lugares como el capítulo 24), que a menudo es extraordinariamente conveniente adoptar unidades en las que la velocidad de la luz c se toma como unidad. Sin pérdida de consistencia, podemos extender este convenio de modo que $\hbar$ y G también son la unidad. Esto tiene la notable consecuencia de que todas las unidades de espacio, tiempo, masa y carga eléctrica quedan ahora completamente *fijadas*, proporcionando las que se conocen como unidades de *Planck* (o unidades *naturales* o unidades *absolutas*). Además, también podemos tomar la constante de Boltzmann k como la unidad (véase §27.3)

$$G = c = \hbar = k = 1,$$

y entonces la unidad de temperatura se convierte también en algo absoluto.

Estas están lejos de las unidades prácticas para uso cotidiano, como puede verse cuando tratamos de poner nuestras unidades convencionales en términos de unidades de Planck:

$$\text{gramo} = 4{,}4 \times 10^{4},$$
$$\text{metro} = 6{,}3 \times 10^{34},$$
$$\text{segundo} = 1{,}9 \times 10^{43},$$
$$\text{grado Kelvin} = 4 \times 10^{-33}.$$

En estas unidades, la carga del protón (o *menos* la del electrón) es aproximadamente $e = \frac{1}{\sqrt{417}}$, y con más exactitud[30]

$$e = 0{,}085\ 42\ 45\ldots$$

También podemos expresar estas relaciones a la inversa, y encontramos que

$$\text{masa de Planck} = 2{,}1 \times 10^{-5}\ \text{g},$$
$$\text{longitud de Planck} = 1{,}6 \times 10^{-35}\ \text{m},$$
$$\text{tiempo de Planck} = 5{,}3 \times 10^{-44}\ \text{s},$$
$$\text{temperatura de Planck} = 2{,}5 \times 10^{32}\ \text{K},$$
$$\text{carga de Planck} = 11{,}7 \text{ cargas del protón}.$$

Veremos más sobre las unidades de Planck en §31.1

Volviendo a la fórmula de Bekenstein-Hawking para la entropía del agujero negro, vemos ahora que en unidades de Planck la entropía S_{BH} de un agujero negro de área superficial A es simplemente

$$S_{\text{BH}} = \frac{1}{4}A.$$

En el caso de la solución de Kerr, encontramos explícitamente

$$A = \frac{8\pi G^2}{c^4} m(m + \sqrt{m^2 - a^2})$$

$$S_{\text{BH}} = \frac{2\pi Gk}{c\hbar} m(m + \sqrt{m^2 - a^2})$$

(en unidades generales). Llegaremos a algunas de las razones que hay detrás de la extraordinaria fórmula de Bekenstein-Hawking en §30.4.

Para hacernos una idea de los extraordinarios valores de la entropía que pueden alcanzarse en un agujero negro, consideremos primero lo

que en la década de 1960 se había pensado que proporciona la mayor de todas las contribuciones a la entropía del universo, a saber, la entropía en la radiación de microondas de 2,7 K: los restos del «destello» del big bang. Esta entropía es de aproximadamente 10^8 o 10^9 por barión, en unidades naturales. (Hablando en términos muy generales, este es el número de fotones por barión que quedó del big bang.) Comparemos esta cifra evidentemente enorme con la entropía debida a los agujeros negros en el universo. Los astrónomos no tienen una idea completamente clara de cuántos agujeros negros hay, ni qué tamaño pudieran tener, pero hay buena evidencia de un agujero negro en el centro de nuestra propia Vía Láctea de aproximadamente $3 \times 10^6 M_\odot$, que podría ser razonablemente típico. Algunas galaxias tienen agujeros negros mucho más grandes, y estos deberían compensar fácilmente los grandes números de otras galaxias que pudieran tenerlos más pequeños, puesto que son los agujeros negros grandes los que dominan los valores de la entropía por encima de todo.[27.14] Como estimación aproximada (quizá muy conservadora), tomemos nuestra galaxia como típica, y encontramos una entropía por barión de aproximadamente 10^{21}, que supera abrumadoramente la cifra de 10^8 o 10^9 de la radiación de cuerpo negro. Además, cualquiera que sea ahora la cifra, crecerá implacable e inimaginablemente en el futuro.

27.11. Cosmología

Antes de tratar de encontrar una estimación para la cifra colosal de la entropía que es potencialmente accesible a nuestro universo —de modo que podamos hacernos una idea de cuán «especial» es realmente nuestro universo ahora, y de cuán «particularmente especial» debe de haber sido nuestro universo en el momento del big bang—, tendremos que saber algo sobre cosmología. Trataremos de utilizar la evidencia cosmológica para estimar el tamaño de la caja $\mathcal{B}$ del espacio de fases que representa el big bang, y la compararemos con el tamaño del espacio de fases entero $\mathcal{P}_\mathcal{U}$, y también con el volumen del espacio de fases de

[27.14] ¿Puede ver por qué?

la caja de grano grueso $\mathcal{N}$ que representa el universo tal como es ahora.

Permítanme empezar describiendo brevemente los que se conocen como los *modelos estándar* de la cosmología, de los que hay (en esencia) tres. Como hemos visto en §27.7, la discusión se remonta al ruso Aleksandr Friedmann, que en 1922 encontró por primera vez las soluciones cosmológicas adecuadas de las ecuaciones de Einstein, con una fuente material que puede utilizarse para aproximar una distribución de galaxias uniforme a gran escala (a veces denominado un «fluido sin presión» o «polvo»). Los modelos cosmológicos de la clase general que estudió Friedmann (a veces con un tipo de fuente de materia diferente del «polvo» de Friedmann) se conocen ahora habitualmente como modelos de Friedmann-Lemaître-Robertson-Walker (FLRW), debido a las contribuciones, clarificaciones y generalizaciones posteriores de estos otros.

Básicamente, un modelo FLRW se caracteriza por el hecho de que es *espacialmente homogéneo* e *isótropo*. En general, «isótropo» significa que el universo parece igual en todas las direcciones, de modo que tiene un grupo de simetría de rotación O(3). Del mismo modo, «espacialmente homogéneo» significa que el universo parece igual en cada punto del espacio y en cualquier instante; por consiguiente, hay un grupo de simetrías que es *transitivo* (§18.2) en cada miembro de una familia de 3-superficies de género espacio, siendo estas las 3-superficies $\mathcal{T}_t$ de «espacio» a «tiempo» t constante (lo que da un grupo de simetría 6-dimensional en total).[27.15] Este par de hipótesis está en buen acuerdo con las observaciones de la distribución de materia a muy gran escala, y con la naturaleza del fondo de microondas. Se encuentra directamente que la isotropía espacial es una aproximación muy buena (por observación de fuentes muy lejanas, y principalmente a partir de la radiación de 2,7 K). Además, si el universo *no* fuera homogéneo, solo podría parecer isótropo visto desde lugares muy especiales,[27.16] de modo que tendríamos que estar en una posición muy privilegiada para que el

[27.15] ¿Por qué 6-dimensional?

[27.16] Dé un argumento general para demostrar por qué un (3-)espacio conexo no puede ser isótropo respecto a dos puntos distintos sin ser también homogéneo.

universo nos *pareciera* isótropo, a menos que fuera también homogéneo. Por supuesto, la isotropía observacional no es exacta, puesto que vemos galaxias, cúmulos de galaxias y supercúmulos de galaxias solo en ciertas direcciones. Hay distribuciones desiguales de materia, no siempre visibles, a escalas inconcebibles, tales como lo que se conoce como «el Gran Atractor», que parece estar tirando no solo de nuestra propia galaxia sino de varios cúmulos de galaxias vecinos. Pero parece que las desviaciones de la uniformidad espacial se hacen proporcionalmente menores cuanto más lejos miramos. La mejor información que tenemos de las regiones más lejanas del universo que nos son accesibles procede de la radiación de fondo de cuerpo negro de 2,7 K. Los datos de COBE, BOOMERANG, WMAP, etc., nos dicen que aunque en dicha escala hay desviaciones de temperatura muy ligeras, en un nivel minúsculo de solo unas pocas partes en 10^5, la isotropía está bien apoyada.[31]

Parece así que las cosmologías homogénea e isótropa —los modelos FLRW— son excelentes aproximaciones a la estructura del universo real, al menos dentro de los límites del universo observable que se extiende hasta una distancia que incluye unas 10^{11} galaxias que contienen unos 10^{80} bariones. (Pronto veremos lo que significa esta noción de «universo observable».) La homogeneidad y la isotropía espacial implican[32] que las secciones espaciales 3-dimensionales $\mathcal{T}_t$ a «tiempo constante» llenan el espaciotiempo entero $\mathcal{M}$ (sin cortarse unas a otras), compartiendo cada 3-geometría $\mathcal{T}_t$ el grupo de simetría de homogeneidad/isotropía de $\mathcal{M}$; véase la Fig. 27.13. Las (esencialmente) tres diferentes posibilidades para la 3-geometría dependen de si la curvatura espacial (constante) es positiva ($K > 0$), cero ($K = 0$), o negativa ($K < 0$). En la literatura cosmológica es habitual normalizar el radio de curvatura, en los casos $K \neq 0$, y referirse a $K > 0$ y $K < 0$ simplemente como $K = 1$ y $K = -1$, respectivamente. Sin embargo, no voy a hacerlo aquí, y preferiré las descripciones respectivas $K > 0$ y $K < 0$.

En la Fig. 27.13a,b,c, he tratado de representar la evolución temporal del universo, de acuerdo con el análisis original de Friedmann de las ecuaciones de Einstein, para las diferentes elecciones alternativas de la curvatura espacial. En cada caso, el universo empieza en una *singularidad* —el denominado big bang— donde la curvatura espacial

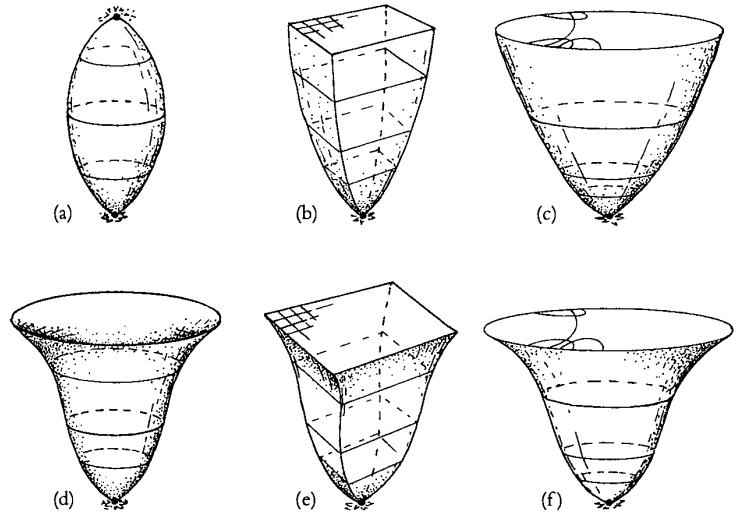


Fig. 27.13. Modelos cosmológicos espacialmente homogéneos e isótropos de Friedmann-Lemaître-Robertson-Walker (FLRW). El tiempo se representa hacia arriba y cada modelo empieza con un big bang. Cada uno está lleno con una familia de uniparamétrica de 3-superficies de género espacio homogéneas y no intersectantes, $\mathcal{T}_t$, que dan «espacio» en el tiempo t. En los modelos de Friedmann la materia se trata como un fluido sin presión («polvo»). Se ilustran los tres casos: (a) $K > 0$, donde las $\mathcal{T}_t$ son 3-esferas S^3 (indicadas en los diagramas como un círculo límitante S^1), donde el modelo colapsa finalmente en un big crunch; o (b) las $\mathcal{T}_t$ son 3-espacios euclídeos $\mathbb{E}^3$, representados como el 2-plano en la parte superior; o (c) las $\mathcal{T}_t$ son 3-espacios hiperbólicos (indicados por una representación conforme en la parte superior). En (d), (e) y (f), se incorpora una constante cosmológica positiva Λ en (a), (b), (c), respectivamente, con expansión exponencial final, donde en el caso (d) se supone que Λ es suficientemente grande para impedir la fase de colapso.

se hace infinita, y luego se expande rápidamente hacia fuera. El comportamiento final depende críticamente del valor de K. Si $K > 0$ (Fig. 27.13a), la expansión se invierte finalmente, y el universo vuelve a una singularidad, a menudo conocida como el *big crunch*, que es una precisa inversa temporal del big bang inicial en el modelo de Friedmann exacto. Si $K = 0$ (Fig. 27.13b), la expansión consigue resistir lo justo y no tiene lugar una fase de colapso. Si $K < 0$ (Fig. 27.13c), no hay perspectiva de colapso, pues la expansión alcanza finalmente a un ritmo casi constante. (Aquí existe una analogía con una piedra lanzada hacia arriba desde el suelo, como se ha expuesto en §27.8. Si la velocidad inicial de la piedra es menor que la velocidad de escape, finalmente caerá de nuevo al suelo, como el universo de Friedmann para $K > 0$; si es igual a la velocidad de escape, ya no vuelve a caer, como $K = 0$; si es mayor que la velocidad de escape, continúa subiendo y se acerca a una velocidad límite que no se frena, como $K < 0$.)

El trabajo original de Friedmann no incluía una constante cosmológica Λ, pero en prácticamente todas las discusiones sistemáticas posteriores en cosmología[33] se había admitido la sugerencia de Einstein en 1917 de un término cosmológico Λg_{ab}, pese a la preferencia de Einstein (después de 1929; véase §19.7) por $\Lambda = 0$. Esto ha resultado ser afortunado, pues evidencias observacionales recientes de diferente tipo han empezado a apuntar claramente en dirección de que realmente hay una constante cosmológica positiva ($\Lambda > 0$) en el comportamiento de nuestro universo. Examinaré más estas cuestiones en §28.10, pero por el momento el lector puede remitirse a la Fig. 27.13d,e,f, que presenta los análogos de la Fig. 27.13a,b,c, pero donde se ha incorporado una Λ positiva (suficientemente grande) en las ecuaciones de Friedmann. Según el balance actual de observación y opinión entre cosmólogos, uno de estos modelos parecería ser una justa descripción de la historia de nuestro universo real, al menos desde el momento del *desacoplamiento*, cuando la edad del universo era de unos meros $\sim 3 \times 10^5$ años, que es aproximadamente 1/50.000 de su edad actual de unos $1{,}5 \times 10^{10}$ años. Este desacoplamiento es el instante al que efectivamente «miramos en retrospectiva» cuando observamos el fondo de microondas.

Antes del desacoplamiento, el universo habría estado básicamente «dominado por la radiación», y después del desacoplamiento, «dominado por la materia». No esperamos que el modelo del «polvo» de Friedmann sea apropiado en la fase dominada por la radiación, y más apropiado podría ser el modelo lleno de radiación de Tolman (1934). Esto no supone una gran diferencia en nuestras imágenes. Acorta la vida media del universo entre el big bang y el desacoplamiento en un factor de aproximadamente 3/4 con respecto a lo que habría sido la pre-

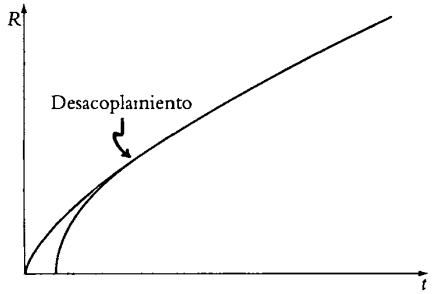


Fig. 27.14. Antes del «desacoplamiento», que ocurrió cuando el universo tenía unos 300.000 años (solo 1/50.000, aproximadamente, de su edad actual) —la época a la que «miramos en retrospectiva» con la radiación de microondas—, el universo estaba «dominado por la radiación» y la aproximación de «polvo» de Friedman no era válida. En su lugar, tenemos la algo más rápida expansión de Tolman, indicada por la curva interna.

dicción de Friedmann,[27.17] como se indica en la Fig. 27.14. Los defensores de la cosmología inflacionaria sugieren un cambio mucho mayor en la evolución, a saber, una expansión exponencial que incrementa la escala del universo en un factor de quizá 10^{60}. ¡Pero esto habría terminado en la época en que el universo solo tenía unos 10^{-32} segundos, de modo que no supone ninguna diferencia para la apariencia de la Fig. 27.13 o la Fig. 27.14! Pese a todo, si la imagen inflacionaria es correcta, las implicaciones en otros aspectos serían enormes. Consideraré la cosmología inflacionaria en §§28.4,5. En cualquier caso, creo que es razonable no incluir la inflación en lo que va a llamarse «el modelo estándar de la cosmología», y aquí no lo haré.[34]

Pero ¿cuál de los tres modelos de la Fig. 27.13d,e,f es probable que sea el apropiado para el universo real? Examinaré esta cuestión en §28.10. Por el momento consideremos que cualquiera de ellos *podría* ser básicamente correcto. Examinemos un poco más cada una de estas diferentes geometrías espaciales.

El caso $K > 0$ se representa normalmente como la 3-esfera. Debería mencionarse, sin embargo, que existe también el *espacio proyectivo* $\mathbb{RP}^3$ obtenido al identificar los puntos antipodales de S^3 (véanse §2.7 y

[27.17] Vea si puede obtener este factor 3/4, suponiendo que el comportamiento del modelo del polvo de Friedmann es de la forma $t = AR^{3/2}$ para valores pequeños del tiempo t, y el de la radiación de Tolman es $t = BR^2$, donde $R = R(t)$ es una medida del «radio» del universo, y A y B son constantes. *Sugerencia*: ¿Deben empalmar las tangentes a las curvas?

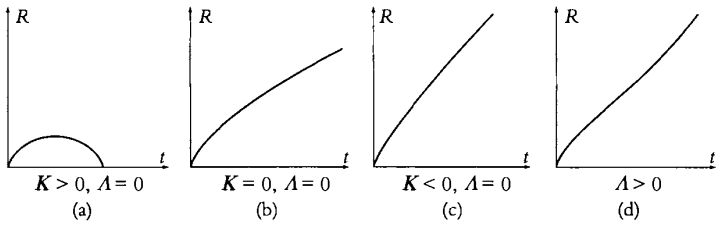


Fig. 27.15. Gráficas de $R = R(t)$ para modelos de Friedmann, primero con $\Lambda = 0$: (a) $K > 0$, (b) $K = 0$, (c) $K < 0$, y luego (d) con $\Lambda > 0$. (El caso (d) se representa para $K = 0$, pero los otros casos son muy similares siempre que Λ sea suficientemente grande en relación con la curvatura espacial.)

§§15.4-6); es difícil imaginar que los dos fueran en la práctica observacionalmente distinguibles. Existen otras identificaciones entre puntos separados de S^3, que dan lo que se denominan *espacios lente*, pero ninguno de estos es globalmente isótropo.[35] El caso $K = 0$ (isótropo) es el 3-espacio euclídeo ordinario, y $K < 0$ da la 3-geometría hiperbólica que hemos estudiado en §§2.4-7 y §18.4. Véanse las Figs. 2.22a, b y c, respectivamente, para las elegantes e ingeniosas representaciones de M. C. Escher de (las versiones 2-dimensionales de) las respectivas geometrías espaciales para $K > 0$, $K = 0$ y $K < 0$. El caso $K > 0$ habitual se denomina un universo *cerrado*, lo que significa *cerrado espacialmente* (i.e., contiene una hipersuperficie de género espacio compacta).[36] Con frecuencia los cosmólogos se refieren a $K < 0$ como el caso «abierto», aunque técnicamente el caso $K = 0$ es también abierto espacialmente. En consecuencia, no utilizaré aquí esta terminología algo confusa. Si abandonamos la isotropía global, entonces, como sucede con los espacios lente $K > 0$ antes mencionados, hay modelos de universos cerrados (no isótropos) también para $K = 0$ y $K < 0$.[37]

El 4-espacio completo $\mathcal{M}$ viene descrito en términos de una evolución temporal para la 3-geometría espacial, como hemos visto, donde existe una escala global que cambia con el tiempo. En la imagen estándar, el universo se expande inicialmente de forma muy rápida a partir de un big bang, pero es una imagen incorrecta pensar en un «punto central» en el que ocurrió la explosión y del que todo se aleja. Una imagen más apropiada, en el caso de dos dimensiones espaciales,

es la superficie de un globo cuando se está inflando. Cada punto de la superficie se aleja poco a poco de todos los demás, a medida que pasa el tiempo, y no hay «punto central» en el modelo de universo. En esta analogía, la superficie representa el universo entero. Así pues, el interior del globo no cuenta como parte del universo en expansión; ni lo hace ningún otro punto que no yace en la superficie.

Utilicemos la notación $d\Sigma^2$ para denotar la forma métrica de una de estas tres 3-geometrías, donde en los casos $K \neq 0$ normalizamos la métrica para que sea la de la 3-esfera *unidad* o el espacio hiperbólico *unidad* (i.e., tomamos $K = 1$ o $K = -1$ respectivamente).[27.18] La 4-métrica del espaciotiempo puede entonces expresarse en la forma

$$ds^2 = dt^2 - R^2 d\Sigma^2,$$

donde t es un parámetro de «tiempo cósmico», cuyos valores constantes determinan la $\mathcal{T}_t$ individual, y donde

$$R = R(t)$$

es una función del parámetro temporal t que da el «tamaño» del universo espacial «en el instante t». Así pues, la métrica de cada $\mathcal{T}_t$ está dada por $R^2 d\Sigma^2$. En la Fig. 27.15a,b,c he representado la gráfica de $R = R(t)$ para $K = 1, 0, -1$, respectivamente, en el caso original de Friedmann con «polvo» (fluido sin presión)[27.19] y con $\Lambda = 0$, y en la Fig. 27.15d he mostrado lo que sucede con una Λ positiva, siendo similares las curvas para los tres valores de K (con tal de que, en el caso $K > 0$, Λ sea suficientemente grande para superar el colapso posterior, como realmente sugieren las observaciones). El ritmo de expansión final es entonces exponencial.

[27.18] Vea si puede demostrar que $d\Sigma^2 = dr^2 + \text{sen}^2\varphi(d\varphi^2 + \text{sen}^2\theta\, d\theta^2)$ describe la métrica de una 3-esfera unidad, y utilice los procedimientos de §18.1 para deducir que $d\Sigma^2 = dr^2 + \text{sen}^2\chi(d\chi^2 + \text{sen}^2\theta\, d\theta^2)$ describe el espacio hiperbólico unidad. *Sugerencia*: Escriba primero la métrica para una 3-esfera de radio arbitrario.

[27.19] La solución del «polvo» de Friedmann para $K > 0, \Lambda = 0$ puede expresarse en la forma $R = C(1 - \cos\xi)$, $t = C(\xi - \text{sen}\,\xi)$, donde C es una constante y ξ es un parámetro conveniente. Demuestre que esta es la ecuación de una *cicloide* —la curva trazada por un punto de la circunferencia de un círculo que rueda sobre una recta horizontal—. ¿Puede ver cómo pasar del caso $K > 0$ al caso $K < 0$ utilizando un «truco» similar al utilizado en §18.1 y el ejercicio [27.16], y a $K = 0$ tomando un límite adecuado (que implica un reescalamiento de coordenadas)?

27.12. Diagramas conformes

Para comprender lo que se entiende por el término «universo observable», es útil emplear lo que se conoce como un *diagrama conforme*[38] en el que una representación (con frecuencia 2-dimensional) del espaciotiempo se presenta de modo que las direcciones nulas están dibujadas a 45° respecto de la vertical, y donde también se representa el infinito como (parte de) la frontera del diagrama. La letra script $\mathscr{I}$ se utiliza normalmente —y se pronuncia «scri»— para esta noción de «infinito», donde «$\mathscr{I}^+$» se utiliza para el infinito *futuro* (o futuro nulo), «alcanzado» en última instancia por rayos de luz salientes, y «$\mathscr{I}^-$» para el infinito *pasado*, para rayos de luz entrantes. Normalmente resultan ser 3-superficies *nulas* en la teoría de Einstein estándar con $\Lambda = 0$ (sin término cosmológico), y 3-superficies *de género espacio* si $\Lambda > 0$.[39]

Los diagramas conformes muestran la estructura *causal* del espaciotiempo, donde lo que nos interesa es la familia de conos nulos, antes que la métrica espaciotemporal completa. Esta es la versión lorentziana de la geometría conforme que hemos visto en §2.4, §8.2 y §§18.4,5 (definida como una clase de equivalencia de métricas, siendo $\boldsymbol{g}$ equivalente a $\Omega^2\boldsymbol{g}$, donde Ω es una función escalar en el espaciotiempo, de modo que Ω modifica la escala de distancia de un lugar a otro). En §2.2 hemos visto cómo puede representarse el plano hiperbólico entero de forma conforme en una región finita del plano euclídeo (Figs. 2.11, 2.12 y 2.13). La idea de un *diagrama espaciotemporal conforme* es básicamente la misma, pero ahora es la métrica lorentziana (no definida positiva) la que se está representando de forma conforme. La nueva característica clave es que en geometría lorentziana los propios conos nulos definen la geometría conforme.

En dos dimensiones, el cono nulo consiste en un par de direcciones nulas, y esto determina la 2-métrica salvo un factor local conforme. Una situación en la que una representación 2-dimensional semejante es especialmente valiosa es aquella en la que existe simetría esférica en el 4-espacio entero. Entonces podemos considerar que este 4-espaciotiempo es un 2-espaciotiempo que está «rotado», de modo que cada punto del 2-espacio representa un S^2 entero en el 4-espacio. Para tales espaciotiempos, los diagramas conformes pueden hacerse completa-

mente precisos, y para estos haré uso de la noción de un diagrama conforme *estricto*. Los diagramas conformes que no son estrictos se denominarán *esquemáticos*. Los *puntos* de un diagrama conforme estricto representan esferas S^2 (métricas) enteras. (En el caso de un espaciotiempo lorentziano n-dimensional del tipo que se podría considerar en teoría de cuerdas, etc. —véanse §§31.4,7—, estas serían $(n-2)$-esferas S^{n-2}.) Lugares excepcionales, donde puntos del diagrama representan simples puntos espaciotemporales, se dan en aquellas partes de la frontera del diagrama que describen un *eje de simetría*. Dichas partes se indican por líneas de puntos, de modo que hay que considerar que el diagrama se rota alrededor de dicha línea de puntos.[27.20] Las partes de la frontera que representan el infinito están indicadas por líneas continuas, y aquellas partes que representan singularidades están indicadas por líneas quebradas. Véase la Fig. 27.16a. Hay también ciertas esquinas donde se unen diferentes líneas de la frontera de un diagrama conforme. Las que están indicadas por pequeños círculos vacíos ∘ deben considerarse como una representación de 2-esferas enteras (como la frontera de un 3-espacio hiperbólico; véase §2.4), mientras que las indicadas por puntos llenos • se consideran mejor como representación de puntos (esferas de radio cero). La Fig. 27.16b es el diagrama conforme estricto para el espacio de Minkowski, y la Fig. 27.16c, para el colapso gravitatorio de un agujero negro de Schwarzschild (el colapso con simetría esférica descrito en la Fig. 27.11). En la Fig. 27.17 he representado los respectivos modelos cosmológicos de la Fig. 27.13.[27.21]

Los diagramas conformes son útiles porque hacen particularmente manifiestas las propiedades de causalidad de los espaciotiempos. Nótese, por ejemplo, que en el colapso esféricamente simétrico a un agujero negro mostrado en la Fig. 27.16c, el horizonte del agujero negro

[27.20] Vea si puede obtener explícitamente el 4-espacio minkowskiano de la Fig. 27.16b, tomando la mitad derecha del 2-espacio minkowskiano entero (métrica $ds^2 = dt^2 - dr^2$, con $r \geq 0$), y rotando alrededor del eje vertical de este modo. Exprese la métrica del 4-espacio utilizando funciones apropiadas de t, r y ángulos polares esféricos θ, ϕ (véase el ejercicio [27.18]). (Para fines de visualización, trate de obtener primero un 3-espacio minkowskiano, en el que la rotación es ahora de un tipo más familiar.)

[27.21] Trate de ver cómo encajan los diagramas de las Figs. 27.11 y 27.16b. Encuentre factores conformes que multiplican las métricas para cada uno de los ejemplos de las Figs. 27.16 y 27.17.

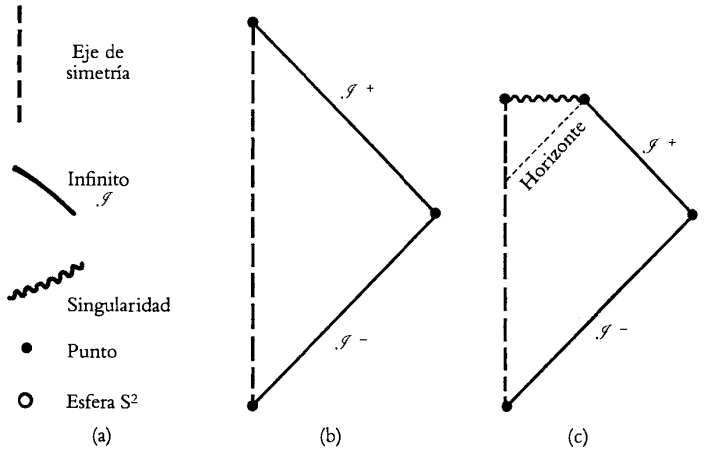


Fig. 27.16. Los diagramas conformes son representaciones planas de espaciotiempos, dibujadas habitualmente de modo que las líneas nulas en el espaciotiempo que yacen en el propio plano están orientadas a 45° de la vertical, y donde el «infinito» suele representarse como una frontera finita para la figura, donde el factor conforme entre la métrica física y la del diagrama va a *cero* en la frontera. (a) En un diagrama conforme *estricto* (en contraposición a *esquemático*), cada punto en el interior del diagrama representa una 2-esfera exacta; pero sobre un eje de simetría (mostrado con una línea de trazos), esta 2-esfera se contrae a un punto, como lo hace en una esquina marcada con •; pero en una esquina marcada con ∘, el punto frontera sigue siendo conformemente una 2-esfera. El infinito está indicado con una línea sólida frontera (a menudo denotada $\mathscr{I}$—pronunciado «scri»—); las singularidades se indican como fronteras con líneas onduladas. (b) Diagrama conforme estricto para el espacio de Minkowski $\mathbb{M}$. (c) Diagrama conforme estricto para la Fig. 27.11, que demuestra el colapso esféricamente simétrico a un agujero negro.

está a 45°. Cualquier línea de universo de una partícula material no puede inclinarse más de 45° respecto a la vertical, de modo que no puede escapar de la región interior que hay tras el horizonte una vez que la ha cruzado. Además, una vez en el interior de dicha región está obligada a caer en la singularidad (Fig. 27.18a). La singularidad parece ser una frontera futura de género espacio para la parte interior del espaciotiempo, un hecho que es algo contraintuitivo desde la perspectiva más convencional de la Fig. 27.11. La situación mostrada por el big bang desempeña un papel similar al inverso temporal de este, actuando

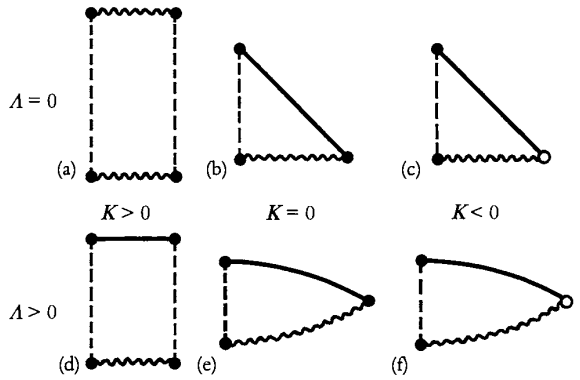


Fig. 27.17. Diagramas conformes estrictos para los respectivos modelos de Friedmann de la Fig. 27.13: (a) $K > 0, \Lambda = 0$; (b) $K = 0, \Lambda = 0$; (c) $K < 0, \Lambda = 0$; (d) $K > 0, \Lambda > 0$ (y Λ suficientemente grande); (e) $K = 0, \Lambda > 0$; (f) $K < 0, \Lambda > 0$.

como una frontera pasada de género espacio para el espaciotiempo (Fig. 27.18b). Esto es de nuevo algo contraintuitivo, puesto que tendemos a pensar en el big bang como un *punto* (singular).[40]

El carácter de género espacio de esta frontera inicial nos lleva a la noción de un horizonte de partículas, que es un aspecto importante del big bang. Consideremos la Fig. 27.18b, donde un observador está en un punto p próximo a la frontera big bang. La región del universo que puede transmitir información al observador es aquella sobre o dentro del cono de luz pasado de p, y notamos que esta intersecta solo a una porción P de la hipersuperficie inicial del big bang.[41] Partículas creadas en el big bang en la región exterior a P están ocultas a la observación en p. Estas regiones están más allá del *horizonte de partículas de* p. Decimos que yacen fuera del *universo observable* de p, siendo esta parte observable del universo la que yace sobre o dentro del cono de luz pasado de p.[27.22]

[27.22] Remitiéndose a los diagramas conformes que se dan aquí, demuestre que para $K = 0$ o $K < 0$, donde $\Lambda = 0$, el universo observable de una partícula que se origina en p aumenta para incluir el universo entero, en el límite de tiempo futuro de la partícula, mientras que esto no es cierto para el caso $K > 0$, o para cualquier K (en los casos mostrados en la Fig. 27.17) si $\Lambda > 0$ (donde se da un «horizonte de sucesos cosmológico»).

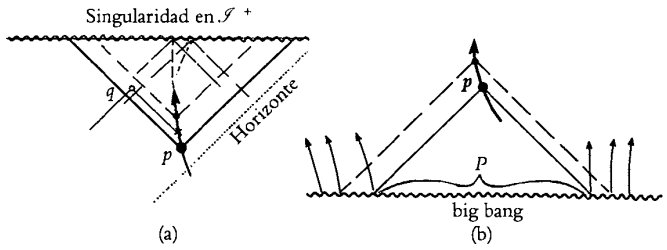


Fig. 27.18. Horizontes. (a) El horizonte de sucesos ocurre cuando una frontera futura —ya sea una singularidad o el infinito— es de género espacio en un diagrama conforme esquemático. A medida que un observador p se aproxima a la frontera, siempre queda alguna porción del espaciotiempo (cuya frontera está definida como un horizonte de sucesos) que p no puede ver, aunque de qué parte se trata exactamente, depende de cómo se mueve p. (Por ejemplo, el suceso q es visto finalmente si p toma la ruta izquierda, pero no la ruta derecha.) En el caso de un agujero negro, el «horizonte de sucesos» más familiar de una naturaleza más absoluta (de puntos, en la figura) que es común a todos los observadores externos. (b) Los horizontes de partículas ocurren en todas las cosmologías estándar, y aparecen debido a que la singularidad pasada es de género espacio. El observador en p ve solo la porción limitada P del big bang (y de las partículas producidas allí), aunque esta porción crece con el tiempo.

27.13. Nuestro extraordinariamente especial big bang

Volvamos ahora al extraordinario «carácter especial» del big bang. El hecho de que debe haber tenido una entropía absurdamente baja es ya evidente por la mera existencia de una segunda ley de la termodinámica. Pero una entropía baja puede tomar muchas formas diferentes. Queremos entender en qué forma *concreta* era especial nuestro universo inicialmente.

Una propiedad especialmente chocante —y en apariencia contradictoria— del big bang procede de una excelente evidencia observacional a favor de que el universo muy primitivo estaba en un *estado térmico*. Parte de esta evidencia es la excepcional proximidad a la curva teórica de «cuerpo negro» *de Planck* (véanse §21.4 y la Fig. 21.3b) que muestra la radiación de fondo de microondas de 2,7 K que representa el «destello» real del big bang, todavía hoy en evidencia, aunque enormemente enfriada por el «desplazamiento hacia el rojo» debido a la ex-

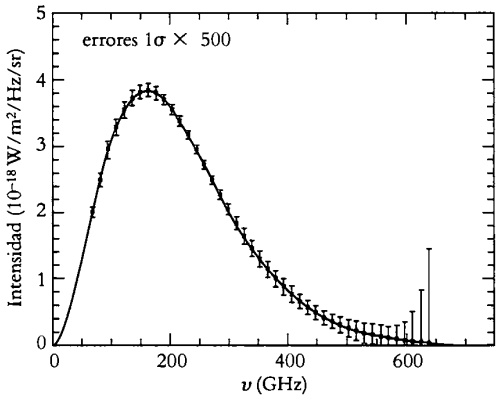


Fig. 27.19. El fondo de microondas tiene una intensidad en términos de frecuencia que es extremadamente precisa de acuerdo con la curva de cuerpo negro de Planck (Fig. 21.3b). (Nótese que las «barras de error» mostradas están exageradas en un factor de 500.)

pansión del universo (Fig. 27.19). Otra evidencia procede del acuerdo notablemente detallado entre lo que nos dicen la teoría y la observación acerca de los procesos nucleares en el universo primitivo. Estos cálculos teóricos dependen crucialmente de suponer el *equilibrio térmico* de la materia en el universo primitivo —tomado conjuntamente con la rápida expansión del universo.

En mi opinión, este aparente equilibrio térmico en el universo primitivo ha confundido bastante a algunos cosmólogos y les ha hecho pensar que el big bang era de alguna forma un estado «aleatorio» (i.e., térmico) de alta entropía, pese al hecho de que debido a la segunda ley debe de haber sido realmente un estado muy organizado (i.e., de baja entropía). Al parecer, ha predominado la idea de que la resolución de esta paradoja debe residir en el hecho de que inmediatamente después del big bang el universo era «pequeño», de modo que le eran accesibles relativamente pocos grados de libertad, lo que pone un «techo» bajo a las entropías posibles. Sin embargo, este punto de vista es falaz, como se ha señalado en §27.6. La resolución correcta de la aparente paradoja está en el hecho de que los *grados de libertad gravitatorios* no han sido «termalizados» junto con todos aquellos grados de libertad materiales

y electromagnéticos que definen los parámetros implícitos en el «estado térmico» del universo instantes después del big bang. De hecho, estos grados de libertad gravitatorios —que proporcionan una enorme reserva de entropía— ¡ni siquiera se suelen tener en cuenta!

Recordemos (Fig. 27.10) que mientras que, en *ausencia* de gravedad, la entropía máxima está representada por un «estado térmico» ordinario, tenemos algo completamente diferente para la entropía máxima cuando empiezan a dominar los efectos gravitatorios, a saber, en el caso de un agujero negro. Con gravitación, la acumulación de material puede representar una entropía mucho más alta que los movimientos térmicos ordinarios, especialmente cuando esta acumulación nos lleva a agujeros negros. Esto se hace particularmente manifiesto si consideramos el caso de un universo cerrado. Imaginemos que el universo está próximo a un modelo FLRW (lo que es compatible con la observación), y tomemos por el momento $K > 0$ y $\Lambda = 0$. La presencia de algunas irregularidades[42] en el material original puede conducir a condensaciones gravitacionales, y supongamos que estas son suficientes para producir galaxias que contienen agujeros negros sustanciales (digamos $10^6 M_\odot$), lo que nos da una entropía por barión de aproximadamente 10^{21}. Si consideramos que nuestro universo cerrado tiene aproximadamente 10^{80} bariones (aproximadamente el contenido bariónico del universo observable), esto da una entropía total de 10^{101}, mucho mayor que el 10^{88} que habría en la radiación y la materia en el instante del desacoplamiento, del oren de 300.000 años después del big bang. Los agujeros negros galácticos crecerían poco a poco, pero el incremento principal ocurriría durante la fase de colapso del universo, cuando las galaxias vuelven a juntarse y sus agujeros negros coalescen. El big crunch final no es del tipo ordenado mostrado en la Fig. 27.16a, que es el inverso temporal del big bang simétrico neto FLRW; es más parecido al temible revoltijo de la coalescencia de singularidades de agujero negro mostrada en la Fig. 27.20a. Podemos estimar la entropía de este revoltijo de un big crunch utilizando la fórmula de la entropía de Bekenstein-Hawking un poco antes de su estado final, cuando todavía podemos considerar que el revoltijo está compuesto de agujeros negros reales, acercándose a la aglomeración de agujero negro final en la que están implicados la totalidad de los 10^{80} bariones. El valor de S_{BH} para

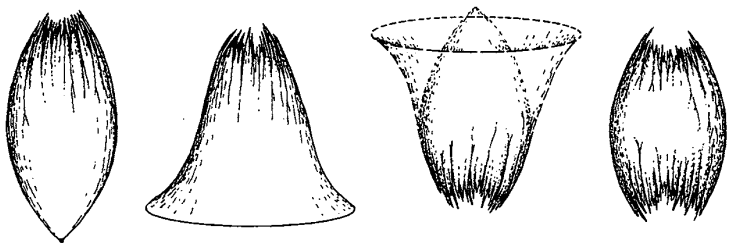


Fig. 27.20. (a) Si en el caso $K > 0$, $\Lambda = 0$ de la Fig. 27.13a permitimos irregularidades del tipo que vemos en nuestro universo real, entonces, en lugar del big crunch «limpio» del modelo de Friedmann exacto, obtenemos un temible amasijo de singularidades de agujero negro coaguladas de entropía enormemente mayor ($S \approx 10^{123}$). (b) Esto no depende de $\Lambda = 0$, pues análogamente podríamos considerar perturbaciones correspondientes de la inversa temporal de la Fig. 27.13d ($K > 0$, $\Lambda > 0$) y de nuevo obtenemos una amasijo similar de entropía enormemente alta ($S \approx 10^{123}$) de agujeros negros coagulados. (c) Un big bang genérico se parecería al de la inversa temporal de semejante colapso genérico (ilustrado para $K > 0$ y $\Lambda = 0$ o $\Lambda > 0$). (d) La situación más «probable» (como la curva de la Fig. 27.8a) —ilustrada en el caso $K > 0$, $\Lambda = 0$ por claridad— no guarda ninguna similitud con el universo real en sus etapas primitivas.

este número de bariones, que es 10^{123}, no debería estar demasiado lejos de la respuesta para la entropía que debe asignarse a este revuelto big crunch.

En este punto, el lector puede objetar razonablemente que, incluso si es cierto que $K > 0$, las observaciones actuales parecen apuntar fuertemente en contra de la hipótesis $\Lambda = 0$ que he hecho, y que (junto con los límites observados sobre la curvatura espacial) el valor positivo observado de Λ parece ser suficientemente grande para impedir la ocurrencia de la fase de colapso que he considerado, aunque con la expectativa de una expansión exponencial final en su lugar. Sin embargo, si se expresa adecuadamente, la discusión precedente sigue siendo aplicable, y se encuentra la misma medida del valor de la entropía ($\sim 10^{123}$) disponible para un universo cerrado de 10^{80} bariones independientemente de $\Lambda > 0$. Para el inverso temporal del universo descrito por la Fig. 27.13d, tan solución de las ecuaciones dinámicas es esta como la descrita por la propia Fig. 27.13d (mientras consideremos leyes dinámicas que deberían ser reversibles en el tiempo). Si consideramos perturbaciones de este universo, podemos encontrar modelos en los que agujeros negros ya formados se juntan y producen un tipo de «revoltijo» de agujeros negros que coagulan similar al que teníamos antes (véase la Fig. 27.20b). Una vez más, llegamos a un valor de la entropía que, por el mismo razonamiento que antes, es del orden de 10^{123}. (Este tipo de razonamiento tendrá importancia de nuevo cuando consideremos la cosmología inflacionaria en §28.5.)

Nos vemos llevados así a una estimación razonable para el volumen total de $\mathcal{P}_U$ (que es en esencia el mismo que el volumen E de la caja $\mathcal{E}$ de entropía máxima de la Fig. 27.4), a saber, la exponencial de este valor de entropía:

$$E = e^{10^{123}} \approx 10^{10^{123}}$$

muy aproximadamente.[27.23] (Esto procede de la $S = k \log V$ de Boltzmann, en unidades naturales.) Ahora bien, ¿cómo se compara esto con lo que sabemos del volumen N de la caja $\mathcal{N}$ para la entropía hoy y con el tamaño B de la caja $\mathcal{B}$ para la entropía en el big bang (suponiendo, por ahora, que vivimos en un universo de 10^{80} bariones)? Tomando la estimación de agujeros negros dada arriba para la entropía actual y el valor de 10^8 para la entropía por barión en la radiación de 2,7 K, obtenemos

$$B: N: E = 10^{10^{88}}: 10^{10^{101}}: 10^{10^{123}}.$$

Se sigue que cada uno de los B y N es solo

una parte en $10^{10^{123}}$

del volumen total E. Además, el volumen B es solo

una parte en $10^{10^{101}}$

del volumen N del espacio de fases del universo actual.

[27.23] ¿Por qué son estas cifras —dentro de la precisión expresada por el número «123»— prácticamente iguales? ¿Por qué el valor real de B no aparece en las conclusiones siguientes?

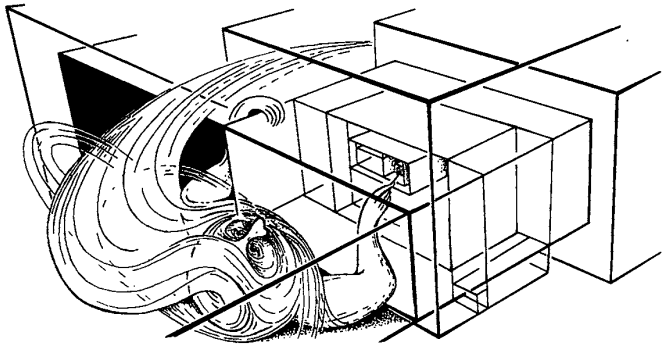


Fig. 27.21. Creación del universo: ¡una descripción extravagante! La aguja del Creador tiene que encontrar una caja minúscula, solo de 1 parte en $10^{10^{123}}$ del volumen del espacio de fases entero para crear un universo con un big bang tan especial como el que realmente encontramos.

Para apreciar el problema planteado por el volumen absurdamente minúsculo del espacio de fases de $\mathcal{B}$, podemos imaginar al Creador tratando de utilizar una aguja para localizar este minúsculo punto en el espacio $\mathcal{P}_U$, y poner en marcha el universo de modo que se parezca al que conocemos hoy. ¡En la Fig. 27.21 he dibujado una representación imaginaria de este tremendo suceso! Si el Creador se equivocase en lo más mínimo al señalar este punto y hundiera la aguja de forma efectivamente aleatoria en la región de máxima entropía $\mathcal{E}$, entonces el resultado sería un universo inhabitable como el de la Fig. 27.20d en el caso $\Lambda = 0, K > 0$, pero por lo demás bastante parecido al caso de eterna expansión de 27.20c, en el que no hay ninguna segunda ley que defina una direccionalidad temporal estadística (como en la Fig. 27.8a). (Las cosas no se explican mucho mejor si imaginamos que nuestro Creador pretende meramente construir un universo en el que hay seres sintientes como nosotros. Esto plantea la cuestión del «principio antrópico»; discutiré estas materias en §§28.6,7 y §34.7.)

Por el contrario, podría ser muy bien que el universo sea espacialmente *infinito*, quizá como los modelos FLRW con $K = 0$ o $K < 0$. Esto no invalida el argumento anterior. Podemos imaginar que lo aplicamos solo al universo observable (en el instante actual) en lugar de al universo entero. Suponiendo que el universo actualmente observable contiene aproximadamente 10^{80} bariones, es difícil ver que las condiciones anteriores resultasen seriamente afectadas. Por el contrario, si aplicamos el argumento al universo como un todo (tomando todavía FLRW como una buena aproximación), obtenemos simplemente un requisito de precisión infinita por parte del Creador, en lugar de una mera precisión absurdamente grande. No veo cómo resuelve esto en modo alguno el dilema que presenta el «ajuste» extraordinariamente preciso que estaba inherente en el big bang —un correlato esencial de la segunda ley.

¿Qué mensaje extraemos de estas consideraciones? Hemos aprendido no solo que el big bang origen del universo fue extraordinariamente especial, sino también algo importante sobre la naturaleza de este carácter especial. En lo que concernía a la materia (incluido el electromagnetismo), la descripción de «equilibrio térmico», en el contexto de un universo en expansión, parece haber sido muy apropiada. Esta es la imagen acertada de «big bang caliente», que es un ingrediente importante del modelo estándar de la cosmología. Al cabo de aproximadamente 10^{-11} s, la temperatura cayó a 10^9 K. Esta caída en temperatura habría estado de acuerdo con el ritmo de expansión de Tolman-Friedmann, y muchos detalles observacionales (por ejemplo, razones hidrógeno/deuterio/helio) son compatibles con los procesos nucleares que tendrían lugar a esas temperaturas posteriores.

Pero para la gravitación las cosas eran completamente diferentes en cuanto que los grados de libertad gravitatorios no fueron «termalizados» en absoluto. La misma uniformidad (i.e., naturaleza FLRW) de la geometría espaciotemporal inicial era lo que resultaba especial en el big bang. El hecho de que un estado singular inicial para el universo «no necesita haber sido así» se ilustra en la Fig. 27.20c —o en el *inverso temporal* del big crunch físicamente apropiado de la Fig. 27.20a. La gravedad parece tener un estatus muy especial, diferente del de cualquier otro campo. En lugar de compartir la termalización que en el universo primitivo se aplica a todos los demás campos, la gravedad permaneció aparte, con sus grados de libertad a la espera, de modo que la segunda ley entraría en juego a medida que estos grados de libertad empiezan a asumirse. Esto no solo nos da una segunda ley, sino que nos da una de la forma concreta que observamos en la naturaleza. ¡La gravedad parece haber sido diferente!

Pero ¿por qué era diferente? Entramos en áreas más especulativas cuando intentamos dar respuestas a una cuestión de este tipo. En el capítulo 28 veremos algunas de las formas que han ensayado los físicos para tratar de entender este enigma y otros relacionados concernientes al origen del universo. En mi opinión, ninguno de estos está muy cerca de abordar el enigma del párrafo precedente. Creo que tendremos que volver a un examen de los mismos fundamentos de la mecánica cuántica, pues tengo la firme convicción de que estas cuestiones están conectadas muy profundamente. Haremos esto en el capítulo 29. En el capítulo 30 trataré de presentar una buena medida de mi propia perspectiva sobre estas cuestiones fundamentales.

Notas

Sección 27.2

27.1. Esto es cierto si consideramos que la dinámica es totalmente clásica. Técnicamente, un «sistema caótico» es un sistema clásico en el que un minúsculo cambio en el estado inicial puede dar como resultado un comportamiento posterior que crece exponencialmente con el tiempo en lugar de, por ejemplo, linealmente. Esta «impredecibilidad» es, por supuesto, una cuestión de grado y no la cuestión de principio con respecto al determinismo que a veces se le atribuye.

27.2. Esto supone que los calores específicos son positivos, que es lo normal. Pero con los agujeros negros esta hipótesis suele ser falsa; véase §31.15.

27.3. No obstante, existe la curiosa «paradoja» de que en la vida ordinaria ¡las cosas son normalmente al revés! Con frecuencia hacemos «retrodicciones» precisas recordando simplemente lo que sucedió en el pasado, mientras que no tenemos un correspondiente acceso al futuro. Además, las investigaciones arqueológicas pueden extender estos «recuerdos» a tiempos muy anteriores a la existiencia de seres humanos. Sin embargo, esta retrodicción no implica la evolución de las ecuaciones dinámicas, en ningún sentido obvio, y su conexión detallada con la segunda ley sigue estando oscura para mí. (Véase Penrose, 1979.)

Sección 27.3

27.4. Véase Pais (1986).

27.5. Véanse Gibbs (1960), Ehrenfest y Ehrenfest (1959) y Pais (1982).

27.6. En realidad, el propio Boltzmann nunca utilizó esta constante, puesto que no estaba interesado en las unidades reales que pudieran utilizarse en la práctica; véase Cercignani (1999). La fórmula $S = k \log V$, que incluye esta constante, parece haber sido escrita explícitamente por primera vez por Planck; véase Pais (1982).

Sección 27.5

27.7. De Eddington (1929a).

Sección 27.6

27.8. Véanse Hawking y Ellis (1973), Misner *et al.* (1973), Wald (1984) y Hartle (2002).

27.9. Véase Gold (1962); véase Tipler (1997) para estas ideas llevadas a una conclusión bastante fantasiosa.

27.10. Véase Penrose (1979a).

Sección 27.7

27.11. Algunos años antes que Hubble, en 1917, el astrónomo estadounidense Vesto Slipher ya había encontrado algún indicio de que el universo se está expandiendo. Véase Slipher (1917). Aunque rara vez se le reconocen estas observaciones, ¡también tiene la distinción de haber descubierto Plutón!

27.12. Esta radiación fue predicha teóricamente por primera vez por George Gamow en 1946, sobre la base de la imagen del big bang, y más explícitamente por Alpher, Bethe y Gamow en 1948; luego lo fue también independientemente por Robert Dicke en 1964. Fue descubierta observacionalmente (por accidente) por Arno Penzias y Robert Wilson, en 1965, e interpretada enseguida por Dicke y sus colegas. Véanse Alpher *et al.* (1948), Dicke *et al.* (1965), y por supuesto Penzias y Wilson (1965), ¡posiblemente el artículo con el título más modesto de todos los tiempos!

27.13. Para más discusiones, véase Penrose (1979, 1989).

27.14. De hecho, la Tierra devuelve en total una energía ligeramente *mayor* que la que recibe. Ignorando la cuestión del quemado por parte del hombre de «combustibles fósiles», que finalmente devuelve alguna

energía recibida del Sol y almacenada en la Tierra hace muchos millones de años (y, en el extremo opuesto de la escala, ignorando el consecuente calentamiento global que resulta del «efecto invernadero» por el que la Tierra atrapa un poco más de energía solar que antes), existe el calentamiento del interior de la Tierra por la desintegración radiactiva que se pierde poco a poco en el espacio a través de la atmósfera. Véase §34.10.

Sección 27.8

27.15. Véanse Michell (1784) y Tipler *et al.* (1980).
27.16. Véase Penrose (1978).
27.17. Véase Van Kerkwijk (2000) para la situación en este tema.
27.18. Véanse Schwarzschild (1916), o la presentación moderna en Wald (1984).
27.19. Véase Narayan (2003) para una reciente evidencia.

Sección 27.9

27.20. La ocurrencia de lo que se conoce como una «superficie atrapada» es una caracterización útil de semejante «punto de no retorno». Una superficie atrapada es una 2-superficie de género espacio compacta $\mathcal{S}$ con la propiedad de que las dos familias de normales nulas a $\mathcal{S}$ convergen en el futuro. (En términos más «coloquiales», esto significa que, si un destello de luz se origina en $\mathcal{S}$, entonces las áreas de las partes entrantes y salientes del destello empezarán a disminuir.) Esperamos encontrar superficies atrapadas dentro del horizonte $\mathcal{H}$ de un agujero negro. La fuerza del criterio de la superficie atrapada reside en que no depende de ninguna hipótesis de simetría, y es «estable» frente a pequeñas perturbaciones de la geometría. Una vez que se ha formado una superficie atrapada, entonces las singularidades son inevitables (suponiendo ciertas condiciones muy débiles y razonables respecto a la causalidad y positividad de la energía en la teoría de Einstein). Resultados similares son válidos para la singularidad big bang cosmológica. Véanse Penrose (1965b) y Hawking y Penrose (1970).
27.21. Véanse Penrose (1965b) y Hawking y Penrose (1970). Wald (1984) revisa estos teoremas en un marco pedagógico.
27.22. Véanse Penrose (1969a, 1998b) y Belinskii *et al.* (1970).
27.23. Véase Penrose (1969a, 1998b).
27.24. Véanse Reeves *et al.* (2002) para una visión más actualizada sobre estas materias, y Chen y Wang (1999); también Hansen y Murali (1998) para la teoría de colisiones.

Sección 27.10

27.25. Véanse Israel (1967), Carter (1970), Hawking (1972) y Robinson (1975).
27.26. Véanse Kerr (1963) y Newman *et al.* (1965) en el caso cargado. Wald (1984) tiene una presentación pedagógica.
27.27. Como las elipses de Kepler mencionadas al principio de este capítulo, la métrica de Kerr aporta otra de aquellas situaciones excepcionales donde hemos sido bendecidos con la buena fortuna de que, a partir de las leyes dinámicas, aparecen realmente configuraciones geométricas relativamente simples.
27.28. Véase Chandrasekhar (1983), p. 1.
27.29. De hecho (como veremos en §31.15; véase también la nota 27.26), hay un parámetro adicional que describe la carga eléctrica total (que es una cantidad conservada; véase §19.3). Pero para agujeros negros astrofísicos realistas, puede ser ignorada en la geometría de agujeros negros, ya que es minúscula en comparación con m y a, debido a la fuerte tendencia del agujero negro a neutralizarse eléctricamente.
27.30. Por supuesto, habría que tener cuidado en no confundir esta «e» con la base de los logaritmos naturales $e = 2{,}7182818285\ldots$; véase §5.3.

Sección 27.11

27.31. Véanse Smoot *et al.* (1991) para la evidencia de COBE y Spergel *et al.* (2003) para WMAP.
27.32. Liddle (1999) es una soberbia introducción a la cosmología. Wald (1984) cubre el tema con un nivel más sofisticado.
27.33. Véanse Bondi (1961), Rindler (2001) y Dodelson (2003).
27.34. El término «modelo de concordancia» ha surgido para describir la situación para la que $K = 0$ y $\Lambda > 0$, donde la inflación está también incorporada. Véanse Blanchard *et al.* (2003) y Bahcall *et al.* (1999). Véase §28.10 para mi valoración del estatus actual de esto.
27.35. Una posibilidad más bien peculiar es que los antiguos griegos tuvieran razón (Fig. 1.1) y el universo sea realmente un dodecaedro (o más bien una versión encolada de uno). Véase Luminet (2003).
27.36. El término *hipersuperficie* se refiere a una subvariedad $(n-1)$-dimensional de una n-variedad. Aquí $\mathcal{T}_i$ es una 3-superficie de género espacio.
27.37. Véanse Killing (1893) y Wolf (1974).

Sección 27.12

27.38. A veces se conocen como «diagramas de Penrose» o «diagramas de Carter-Penrose». Los utilicé por primera vez en mi conferencia de Var-

sovia (1962); la noción sistemática de un diagrama conforme estricto fue introducida por Carter (1966).

27.39. Véanse Penrose (1964, 1965a), Carter (1966), Penrose (1963, 1964, 1965a) y Penrose y Rindler (1986), capítulo 9.

27.40. Ciertos modelos hipotéticos en los que el big bang (o más bien el big crunch) es conformemente (i.e., causalmente) un punto —conocido como el «punto Ω»— encuentran apoyo en algunos teóricos; véase Tipler (1997). No obstante, no conozco ninguna discusión, compatible con los argumentos del capítulo 27, que haga tales modelos físicamente plausibles.

27.41. Véase la nota 27.36 para el término hipersuperficie. En este caso vemos que el big bang, en su representación conforme, es *3-dimensional*. (Podemos contrastarlo con algunas otras representaciones, véase Rindler, 2001.)

Sección 27.13

27.42. A menudo se considera que el fenómeno responsable, en última instancia, de tales irregularidades son las «fluctuaciones cuánticas» en la densidad inicial de materia en el big bang. (Esto se discutirá en §30.14.)

28

Teorías especulativas del universo primitivo

28.1. Ruptura espontánea de simetría en el universo primitivo

Hasta este punto del libro, nuestras consideraciones se han mantenido dentro del alcance de la teoría física firmemente establecida, donde impresionantes datos observacionales han proporcionado un poderoso apoyo a las a veces algo extrañas ideas teóricas que han entrado en juego. Algunos de mis argumentos han sido presentados de una forma ligeramente diferente de como se suelen encontrar en la literatura, pero no creo que exista nada controvertido en ello. En este capítulo empezaré a abordar algunas ideas más especulativas que están relacionadas con las cuestiones que plantea la naturaleza especial del big bang.

En particular, consideraré las ideas de la cosmología inflacionaria, además de otras que se relacionan con la ruptura espontánea de simetría en el universo primitivo (véase §25.5). Algunos lectores versados en ciertas ideas de uso común en cosmología quizá encuentren enigmático que sitúe tan firmemente la cosmología inflacionaria en el campo «especulativo». De hecho, las exposiciones divulgativas suelen dar la impresión de que es un hecho establecido que en las etapas muy primitivas del universo hubo un período de expansión exponencial en el curso del cual el universo aumentó de tamaño en un factor de 10^{30}, o quizá incluso de 10^{60} o más. Otros lectores entendidos quizá estén incluso más alarmados por el hecho de que considere el fenómeno general de la ruptura espontánea de simetría en el universo primitivo como una idea especulativa. En cualquier caso, las nociones que deseo abordar en este capítulo no tienen de momento ningún apoyo signifi-

cativo e inequívoco (si es que tienen alguno) por parte de la observación, y se puede plantear perfectamente la pregunta de si estas ideas tienen o no auténtica relevancia para la naturaleza.

Empecemos con la idea general de ruptura espontánea de simetría. Recordemos la potencia de esta idea para producir QFT renormalizables, en las que la renormalizabilidad saca provecho de una simetría («oculta») mayor que la que se manifiesta directamente en el comportamiento observado. Esta ausencia de una simetría completa en lo que se observa se atribuye a la elección por parte del sistema de un «estado vacío» que no comparte la simetría completa de la teoría dinámica. En particular, esto constituyó un ingrediente clave en la parte electrodébil del modelo estándar de la física de partículas. Más aún, este tipo de ideas, que implican diferentes «vacíos» posibles, son también un ingrediente esencial de la inflación, y estas nociones de ruptura espontánea de simetría y «falsos vacíos» son también invocadas habitualmente por los teóricos que buscan esquemas cada vez más unificados. Sin embargo, debería dejar claro que la ruptura espontánea de simetría propiamente dicha no es una idea especulativa. Tiene una relevancia indudable para muchos fenómenos físicos genuinos (la superconductividad es un excelente ejemplo). Se aplica por supuesto a muchos fenómenos bien establecidos, con frecuencia de una forma elegante y satisfactoria. De ninguna manera voy a arrojar dudas sobre la idea en sí misma. Mi problema con ella es que temo que sea una idea con tanto atractivo que pueda incitar a los físicos a emplearla a veces con demasiada generalidad, y a veces en circunstancias inadecuadas.

La idea de ruptura espontánea de simetría se suele introducir de una forma gráfica haciendo referencia al fenómeno del *ferromagnetismo*. Imaginemos una bola esférica y maciza de hierro. Podemos considerar sus átomos como pequeños imanes que, debido a las fuerzas involucradas, tienen tendencia a alinearse paralelamente a sus vecinos, norte con norte y sur con sur. Cuando la temperatura es suficientemente alta, por encima de un valor crítico que es de aproximadamente 770 °C (1.043 K), la energía de agitación térmica de los átomos superará esta tendencia al alineamiento magnético y el material no mostrará ninguna propensión a convertirse en un imán a gran escala, pues habrá una disposición efectivamente aleatoria en las orientaciones de sus pequeños

imanes atómicos. Pero a una temperatura por debajo de 770 °C (el denominado «punto de Curie»), la alineación de los átomos resultará energéticamente favorable, y en una situación ideal el hierro quedaría completamente magnetizado.[1]

Imaginemos ahora que nuestra bola de hierro está inicialmente calentada por encima de 770 °C (pero no a una temperatura tan alta que pueda fundirse), de modo que inicialmente es una bola desmagnetizada. Luego se enfría poco a poco hasta bajar de los 770 °C críticos. ¿Qué sucede entonces? La tendencia natural para la esfera es encontrar un estado de mínima energía, y la energía de las vibraciones internas de sus átomos se transfiere al entorno más frío. Debido a las interacciones entre átomos vecinos, la energía mínima se alcanza cuando todos los átomos están alineados, de modo que la bola se magnetiza con una dirección definida para su polaridad norte/sur. Pero no hay ninguna dirección favorecida respecto a las demás. Existe lo que se denomina una *degeneración* en los estados de mínima energía (compárese con §22.6). Al no haber una dirección favorecida en su estado caliente y desmagnetizado inicial, la dirección de magnetización final se produce *al azar*. Este es un ejemplo de *ruptura espontánea de simetría*: el estado inicial esféricamente simétrico se asienta en un estado con una simetría menor, a saber, tan solo la simetría de rotación alrededor del eje magnético norte/sur resultante. Un estado SO(3)-simétrico (la bola caliente desmagnetizada original) evoluciona hacia un estado SO(2)-simétrico (la bola fría magnetizada; véanse §§13.1,2,3,8,10 para el significado de estos símbolos).

La imagen utilizada para describir una situación de este tipo es la de un potencial tipo «sombrero mexicano» que se muestra en la Fig. 28.1. El «sombrero» representa la familia de estados permitidos del sistema (cuando la temperatura ambiente se ha enfriado hasta cero), y la «altura» representa la energía del sistema. Observamos que hay un *estado de equilibrio* (i.e., que tiene un plano tangente horizontal) representado por el punto más alto de la copa del sombrero que posee la simetría completa del grupo original, y que en la imagen viene representada como una rotación alrededor del eje vertical. (Esta simetría de rotación SO(2) se considera análoga a la simetría SO(3) completa de la bola de hierro, pero hemos tenido que perder una dimensión espacial para ha-

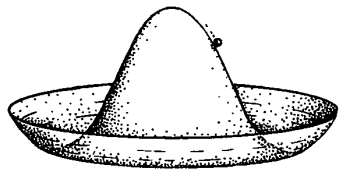


Fig. 28.1. Ruptura espontánea de simetría, con un potencial tipo «sombrero mexicano» para los estados permisibles en un sistema donde la altura mide la energía. El estado de un sistema está representado por una canica, restringida a la superficie del sombrero. Cuando la temperatura ambiente es suficientemente alta (por encima del punto de Curie), el estado de equilibrio del sistema está representado por la canica en reposo en la cima, y tiene simetría rotacional completa (SO(2)), en esta imagen simplificada). Pero cuando la temperatura se enfría, la canica rueda hacia abajo, hasta alcanzar finalmente un punto arbitrario de equilibrio en el ala, rompiendo la simetría rotacional completa.

cer visualizable la imagen. Este punto más alto de la copa del sombrero representa la total falta de magnetización para la bola en conjunto.) Pero este equilibrio —que representa el estado no magnetizado— es inestable y no representa el mínimo de las energías disponibles. Estos mínimos son los estados representados por las partes horizontales —el equivalente a un círculo completo— justamente dentro del ala del sombrero (los puntos diferentes en el ala representan diferentes direcciones de magnetización total en la bola de hierro).

Podemos imaginar que el estado está inicialmente en la copa, representado por una «canica» colocada inicialmente en dicho punto para representar el estado físico, dejada allí por el estado anterior de alta temperatura. Pero la falta de estabilidad significa que la canica caerá rodando desde dicho punto (suponiendo la existencia de ciertas influencias perturbadoras aleatorias) y finalmente encontrará un punto de reposo en el ala. Cada punto en el ala en el que pudiera asentarse la canica representa una dirección de magnetización diferente que la bola podría adquirir finalmente. Esta localización de la canica representa el estado físico final. Pero debido a la degeneración rotacional, no hay ningún lugar favorecido donde la canica llegue al reposo. Todos estos puntos de equilibrio en el ala están en pie de igualdad. Se supone que la elección que hace la canica es aleatoria, y una vez

que se ha hecho dicha elección, la simetría se ha roto, en una dirección escogida al azar.

Un fenómeno de esta naturaleza, en el que una reducción en la temperatura ambiente induce un abrupto cambio global en la naturaleza del estado de equilibrio estable del material, se denomina *transición de fase*. En nuestro ejemplo de la bola de hierro, la transición de fase ocurre cuando la bola pasa del estado desmagnetizado (cuando la temperatura está por encima de 770 °C) al estado uniformemente magnetizado (temperatura por debajo de 770 °C). Más familiares son los fenómenos de congelación (en el que el estado pasa de líquido a sólido cuando desciende la temperatura) y, en un proceso inverso, la ebullición (en la que el estado pasa de líquido a gas cuando aumenta la temperatura). Una transición de fase, cuando desciende la temperatura, suele estar acompañada por una reducción de simetría, pero esto no es esencial.

En procesos QFT, una transición de fase se describiría frecuentemente en términos de una nueva elección de estado vacío (como el $|\Theta\rangle$ de §26.11), donde se imagina que el estado pasa por «efecto túnel»[2] de un vacío a otro. Sin embargo, esta descripción debe tomarse como una aproximación, puesto que no hay ningún proceso mecanocuántico (unitario) por el que un estado pueda evolucionar desde un *sector* a otro (aquí un «sector» se refiere a los estados que pueden construirse a partir de una elección concreta del estado vacío $|\Theta\rangle$, y los estados en diferentes sectores pertenecen a diferentes espacios de Hilbert; véanse §§26.5,11). La aproximación, que implica considerar que un sistema es infinito, cuando en la práctica es finito, es evidentemente una buena aproximación en situaciones prácticas. Por ejemplo, el fenómeno bien establecido de la superconductividad (en el que la resistencia eléctrica se reduce a cero cuando la temperatura es suficientemente baja) se trata de esta forma, y la superconductividad es una transición de fase que acompaña a la reducción de simetría que rompe la simetría U(1) ordinaria del electromagnetismo.

En el ejemplo específico ilustrado en la Fig. 28.1, la simetría se rompe a partir del grupo de rotaciones axiales SO(2) para dar el grupo trivial («SO(1)»), que contiene solo un elemento (de modo que toda la simetría se pierde finalmente, en este ejemplo, pues la posición

de reposo de la canica rompe por completo la simetría).[3] Pero versiones de este «sombrero» en dimensiones más altas ilustran la ruptura espontánea de simetría desde SO(p) a SO($p-1$), con $p > 2$.[28.1] (Nuestra bola de hierro ilustra el caso $p = 3$.) También podemos utilizar la imagen del «sombrero mexicano» para ilustrar la ruptura desde U(2) a U(1) que ocurre en el modelo estándar de la física de partículas,[28.2] donde se considera que la simetría electrodébil U(2) (véase §25.5) se rompe en la simetría U(1) del electromagnetismo a una temperatura de aproximadamente 10^{16} K, algo que habría ocurrido 10^{-12} s después del big bang. En las teorías GUT más generales (véase §25.8), están involucrados otros grupos tales como SU(5), y podemos concebir diferentes etapas de ruptura de simetría que ocurren a diferentes temperaturas. Así pues, a cierta temperatura mucho más alta que 10^{16} K (i.e., en un instante significativamente anterior a los 10^{-12} s inmediatamente después del big bang), SU(5) podría romperse primero en algo que contiene de manera adecuada[4] tanto el SU(3) para las interacciones fuertes como el $SU(2) \times U(1)/Z_2$ (i.e., U(2)) que se necesita para la teoría electrodébil.

28.2. Defectos topológicos cósmicos

No obstante, deberíamos tener en cuenta que es poco probable que esta ruptura de simetría tenga lugar «de una vez», y que muy bien pueden aparecer *dominios* en los que la simetría está rota en «direcciones» diferentes. Consideremos de nuevo nuestra bola de hierro; cabría esperar que la elección aleatoria de la dirección de magnetización fuera diferente en diferentes lugares dentro de la bola. Podríamos imaginar que, si el enfriamiento es suficientemente lento, entonces estas no uniformidades podrían «alisarse», para dar solo un imán uniforme.[5] Pero,

[28.1] Demuestre que el «sombrero» con forma $E = (x_1^2 + \ldots + x_p^2 - 1)^2$ muestra esta ruptura de simetría.

[28.2] Demuéstrelo, con U(2) actuando en $\mathbb{C}^2$, con coordenadas complejas (w, z), estando dado el «sombrero» por $E = (|w|^2 + |z|^2 - 1)^2$. ¿Puede ver la geometría de esta reducción de simetría en la configuración de los paralelos de Clifford en S^3, descrita en §15.4 e ilustrada en las Figs. 15.8 y 33.15?

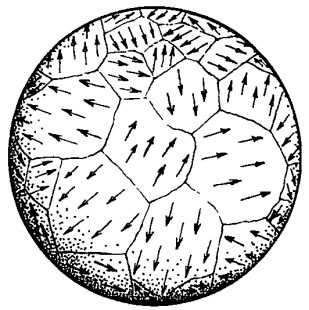


Fig. 28.2. Idealmente, cuando un material ferromagnético se enfría lentamente desde su punto de Curie, las direcciones de magnetización de sus átomos se alinearán todas en la misma dirección (arbitraria). Pero, en la práctica (o con un enfriamiento demasiado rápido), obtenemos un «mosaico» de tales direcciones de magnetización.

alternativamente, con un enfriamiento más rápido, podríamos encontrarnos con un «mosaico» de direcciones parecido al ilustrado en la Fig. 28.2. El tamaño de las celdas resultantes y las pautas que presentan podría depender, entre otras cosas, de la velocidad a la que tiene lugar el enfriamiento. Se plantea así la cuestión de con qué facilidad tiene lugar la «comunicación» entre diferentes regiones, y con qué facilidad podría ser «reorientada» la dirección de magnetización en una región de la bola bajo la influencia de regiones vecinas.

Más serios e interesantes son los defectos topológicos que no pueden ser eliminados en absoluto mediante un movimiento continuo alrededor de las direcciones de magnetización en el interior de la bola. Un defecto semejante es un «monopolo magnético de Dirac» (un polo magnético norte o sur aislado). Un monopolo semejante no puede producirse en el espacio ordinario con un conjunto de imanes y corrientes.[28.3] Sin embargo, se puede conseguir un monopolo efectivo si permitimos que la carga magnética sea «evacuada» a lo largo de un «cable de Dirac», como en la Fig. 28.3. Si se admiten cargas magnéticas en la teoría de Maxwell (§19.2), entonces el «cable» aparece solamente en el potencial A (§19.4), y puede eliminarse por completo por la adopción del apropiado punto de vista «fibrado» (§15.4). Un tipo similar de monopolo existiría también en adecuadas teorías gauge no abelianas.

Estas complicaciones en la imagen de la reducción espontánea de simetría, parcialmente ilustrada antes en el ejemplo «de andar por casa» de una bola de hierro, tienen también relevancia en el nivel más esoté-

[28.3] Demuéstrelo apelando a las expresiones integrales del capítulo 19.

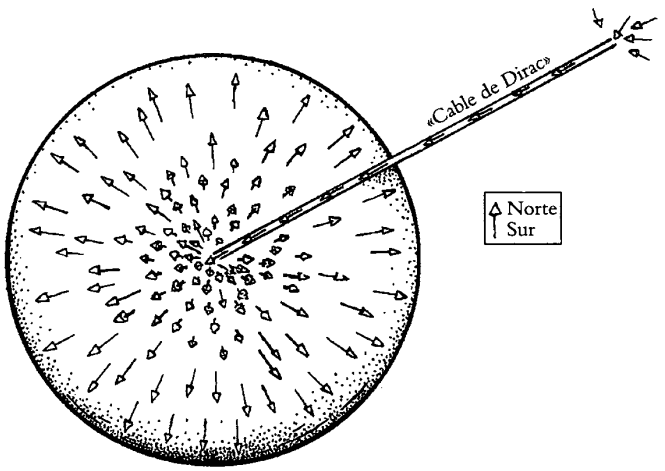


Fig. 28.3. Un monopolo magnético podría aparecer si de alguna forma «evacuamos» el exceso de «polo sur» en el centro de la esfera a lo largo de un «cable magnético». Con fuentes magnéticas permitidas en la teoría de Maxwell, dicho polo podría insertarse en el centro, y el «cable» (Dirac) solo tiene que ocurrir como una falla en el potencial A. Esta falla puede eliminarse con el apropiado punto de vista «fibrado» (tales monopolos también se dan en adecuadas teorías gauge no abelianas).

rico de las teorías físicas básicas (tales como la teoría electrodébil o las GUT) que dependen fundamentalmente de la idea de ruptura espontánea de simetría. Cabe esperar que haya defectos topológicos a gran escala (cosmológica) si tal ruptura espontánea de simetría tuvo lugar en el universo primitivo. En general (para un espacio 3-dimensional), existen tres tipos básicos de defecto topológico, dependiendo de la dimensión de las regiones en las que residen esencialmente. Estos son los denominados *monopolos* (cósmicos, que son espacialmente 0-dimensionales), *cuerdas cósmicas* (espacialmente 1-dimensionales) y *paredes de dominios* (espacialmente 2-dimensionales). La dimensión depende de cuestiones topológicas que tienen que ver con los grupos involucrados. El punto importante en relación con los defectos topológicos es que ningún movimiento continuo de la «dirección» de la ruptura de simetría puede eliminarlos (donde consideramos que, en el defecto propiamente dicho, no hay ninguna dirección bien definida de ruptura de

simetría, aunque una variación continua de esta dirección tiene lugar en otra parte). Debemos tener en cuenta que esta noción de «dirección» no se refiere a una dirección en el espacio ordinario, sino a una noción más abstracta de «dirección» que ocurre dentro del modelo físico en consideración (por ejemplo, en la teoría electrodébil, donde nos dice qué grado de mezcla electrón/neutrino se está considerando). Desde el punto de vista geométrico, deberíamos pensar en términos de un fibrado vectorial sobre el espaciotiempo (véase el capítulo 15, si se necesita recordar esta noción). Las consideraciones topológicas se siguen aplicando, y los defectos topológicos presentarían problemas serios que no pueden ser «tomados a broma» si se va a tomar seriamente la ruptura de simetría como parte de una teoría física básica.

De hecho, se ha considerado seriamente que cuerdas cósmicas de escalas enormes (incluso mayores que galácticas) son los agentes esencialmente responsables de inducir las inhomogeneidades en el gas de fondo que llevan a la formación de galaxias.[6] Podemos pensar que el campo gravitatorio de una cuerda cósmica semejante está construido mediante un procedimiento de «cortar y pegar» aplicado a un espaciotiempo de Minkowski. En términos espaciales (véase la Fig. 28.4), representamos un «sector» eliminado del 3-espacio, sector que está acotado por un par de semiplanos cuyo ángulo diédrico α está centrado en la propia cuerda. Para construir la geometría de la cuerda cósmica, las dos superficies planas se «pegan» de nuevo. (En los modelos sugeridos, α es aproximadamente 10^{-6}.)

El lector puede tener la sensación, en parte justificada, de que estas son medidas extremas para la producción de una entidad tan «tópica» como una galaxia ordinaria. Pero sigue habiendo algunos enig-

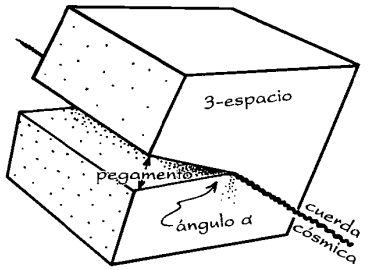


Fig. 28.4. El campo gravitatorio de una cuerda cósmica puede construirse mediante un procedimiento de «cortar y pegar» aplicado al 4-espacio de Minkowski. En el 3-espacio se elimina un sector, limitado por dos semiplanos que se cortan a un ángulo α a lo largo de la cuerda. Luego se «pegan» las superficies semiplanas.

mas teóricos acerca de la formación de galaxias, de modo que ideas tan exóticas no deberían ser descartadas, pese a su naturaleza aparentemente escandalosa. De hecho, uno de los modelos más plausibles de la formación de galaxias —que tiene un impresionante apoyo observacional reciente— propone que son «sembradas» por agujeros negros supermasivos que ahora parecen residir en sus centros.[7] ¡Pero hoy los agujeros negros deben considerarse como física convencional antes que exótica!

La mayoría de estos defectos topológicos sugeridos remiten a teorías (tales como las diversas GUT) que no tienen un apoyo significativo o inequívoco procedente de la observación. Por el contrario, la teoría electrodébil está muy bien apoyada observacionalmente, de modo que debemos prestar atención a lo que implica esta teoría con respecto a procesos en el universo primitivo. Los monopolos cósmicos, que resultan de la ruptura de simetría de la teoría electrodébil, son una posibilidad topológica, pero no una necesidad. Podrían surgir en la ruptura espontánea desde U(2) a U(1), pero solo si lo que se denominan «monopolos gauge» estaban ya presentes en la fase U(2)-simétrica intacta de la teoría, que se supone que ha ocurrido *antes* de 10^{-12} s. Tales monopolos podrían surgir de una ruptura anterior de una simetría GUT mayor, pero estas ideas no son en modo alguno una parte necesaria de la teoría electrodébil.[8]

Tales *monopolos gauge* son los análogos, dentro de alguna teoría Yang-Mills (gauge no abeliana), a los «monopolos magnéticos» que propuso Dirac (en 1931) en el contexto de la teoría (gauge abeliana) del electromagnetismo. Mediante un ingenioso argumento, Dirac demostró que tan solo con que existiera en la naturaleza un único monopolo magnético (un polo magnético norte o sur aislado), todas las cargas *eléctricas* tendrían que tener valores que son múltiplos enteros de cierto valor concreto, valor que es inversamente proporcional a la intensidad magnética del monopolo. De hecho, observaciones actuales sugieren con fuerza que todas las cargas eléctricas *son* múltiplos enteros de un valor concreto (digamos el de la carga del antiquark d, que es un tercio de la del protón; véanse §3.5 y §25.6). Algunos tomarían esto como evidencia circunstancial de la existencia real de monopolos magnéticos. De todas formas, para que tales monopolos no estén en conflicto con la observación, tendrían que ser extraordinariamente poco comunes.[9] (De lo contrario tendrían el efecto de «cortocircuitar» campos magnéticos cósmicos, cuando lo cierto es que la existencia de tales campos es un hecho observado en grandes regiones del universo.) Análogamente, los monopolos de Yang-Mills provocarían serios conflictos observacionales si dichos monopolos estuviesen presentes de forma significativa en el universo actual. ¡Esta cuestión ha tenido importantes consecuencias para el desarrollo de la disciplina de la cosmología, como veremos enseguida!

28.3. Problemas para la ruptura de simetría en el universo primitivo

Antes de llegar a esto es oportuno que consideremos de nuevo la ruptura de simetría en la teoría electrodébil, que se estima que tuvo lugar aproximadamente 10^{-12} s después del big bang. ¿Debemos aceptar que este es un fenómeno real, o podría ser meramente un artificio de la forma concreta en que se presenta normalmente la teoría? Por lo que puedo entender, la mayoría de los teóricos electrodébiles considerarán ciertamente que este proceso es real. Por consiguiente, el lector queda advertido de que mi propuesta de cuestionar aquí su realidad es una posición poco convencional. De todas formas, sigamos adelante y consideremos algunas de las dificultades intrínsecas en la idea de ruptura de simetría.

Supongamos que, contrariamente a mi opinión (menos que convencional) sobre esta materia, hubo un momento en la historia primitiva del universo —anterior a aproximadamente 10^{-12} s a partir del big bang— en que imperaba una simetría U(2) exacta en la que todos los leptones y los quarks carecían de masa, donde electrones y neutrinos «zig» estaban en pie de igualdad, y donde también los bosones W y Z y el fotón podían ser adecuadamente «rotados» en combinaciones de unos en otros de acuerdo con una simetría U(2) (véase §25.5). Luego, en un tiempo de aproximadamente 10^{-12} s, la temperatura en todo el universo cayó justo por debajo del valor crítico. En ese instante se hizo al azar una elección concreta de (W^-, W^+, Z^0, γ) entre toda la va-

riedad $\mathcal{G}$ U(2)-simétrica de conjuntos posibles de bosones gauge. No esperamos que esto suceda de forma exactamente uniforme en todo el espacio, y simultáneamente en el universo entero. Prevemos que, como sucede con los dominios de magnetización en la bola de hierro ilustrada en la Fig. 28.2, en algunas regiones se hará una elección concreta y en otros lugares se harán elecciones diferentes.

En este punto deberíamos abordar la cuestión de lo que se *entiende* realmente por las palabras «mismo» y «diferente» en este contexto. El espacio $\mathcal{G}$ de bosones gauge posibles es, en cada punto espaciotemporal, completamente U(2)-simétrico antes de que tenga lugar la reducción de simetría. Como es inherente a la noción de fibrado, no hay ninguna forma particular preferida a otras de hacer una identificación entre la $\mathcal{G}$ en un punto y la $\mathcal{G}$ en otro punto diferente por completo. Así pues, no parece que haya una regla *a priori* que nos diga a qué elemento de $\mathcal{G}$ en un punto hay que llamar «el mismo» elemento que un elemento de $\mathcal{G}$ en otro punto. Parece que esto nos da la libertad de mantener el punto de vista según el cual simplemente *definimos* la noción de «el mismo» como la que proporciona la elección concreta que ofrece la ruptura espontánea de simetría. Según tal punto de vista, el (W^-, W^+, Z^0, γ) concreto que se «congela» en un punto se identificaría con el correspondiente (W^-, W^+, Z^0, γ) en cualquier otro punto, de modo que parece que no seríamos testigos del tipo de «inconsistencia» entre las rupturas de simetría en diferentes puntos que se da con los dominios de magnetización de hierro ilustrados en la Fig. 28.2.

Pero un punto de vista semejante se enfrenta a la idea general que hay tras la teoría gauge, según la cual no solo son los $\mathcal{G}$-espacios las fibras de un fibrado $\mathcal{B}_g$, con espacio base en el espaciotiempo $\mathcal{M}$, sino que también la teoría gauge concreta —en este caso la teoría electrodébil intacta— se define en términos de una *conexión* sobre este fibrado (véanse §§15.7,8). Esta conexión define la identificación (paralelismo) localmente significativa entre los espacios $\mathcal{G}$ cuando nos movemos a lo largo de cualquier curva dada en $\mathcal{M}$.[10] En general, esta identificación no es globalmente consistente cuando recorremos lazos cerrados (a causa de la curvatura en la conexión, que expresa la presencia de un campo gauge no trivial; véase §15.8). En cualquier caso, la aleatoriedad implicada en la ruptura de simetría en puntos diferentes tendrá como

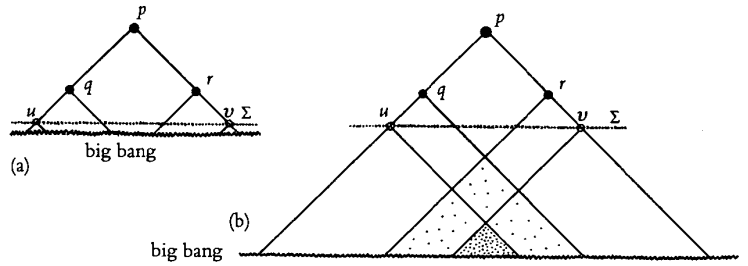


Fig. 28.5. Diagramas conformes esquemáticos que ilustran (in)dependencia causal en el universo primitivo. (a) Un observador en p ve cuásares en direcciones opuestas, en q y r. Si la línea de puntos representa la 3-superficie Σ de tiempo 10^{-11} s, a lo largo de la cual se considera rota la simetría electrodébil exacta anterior U(2) (que relaciona el fotón γ con los bosones W y Z), entonces la particular elección «congelada» de γ en p difiere casi con certeza de la de r, al ser disjuntas las intersecciones de los pasados de p y r con Σ; pero las elecciones γ respectivas no pueden comunicar su identidad/diferencia hasta que se alcanza p. De forma similar, si Σ representa ahora desacoplamiento, en el instante 10^{13} s, las temperaturas en u y v no pueden haberse igualado por termalización, pues sus pasados completos son disjuntos. (b) La «resolución» por inflación del «problema del horizonte» en el último caso consiste en retrasar el big bang de modo que los pasados de q y r se cortan ahora antes de alcanzar la 3-superficie big bang. Sin embargo, el primer problema sigue sin resolver, puesto que las intersecciones de sus pasados ocurren antes de la «congelación» en 10^{-11} s.

consecuencia que el paralelismo local entre los $\mathcal{G}$-espacios no será generalmente consistente con las elecciones que se hagan en la ruptura espontánea de simetría; de modo que la imagen de la Fig. 28.2 no es una analogía tan irrazonable. Podemos imaginar que, como sucede con una bola de hierro enfriada con suficiente lentitud, las inconsistencias se «alisarán» si se deja un tiempo suficiente, con tal de que no haya defectos topológicos (como se indica en las Figs. 28.3 y 28.4). La cuestión que quiero plantear aquí es si puede haber *alguna vez* «tiempo suficiente» en el caso de la ruptura espontánea de simetría electrodébil.

La dificultad tiene que ver con los horizontes de partículas que encontramos en §27.12, Fig. 27.18b. Consideremos el diagrama conforme esquemático de la Fig. 28.5. Un observador situado en el punto p ve cuásares (cf. §27.9) en dos direcciones opuestas, en puntos espaciotemporales respectivos q y r. Según el modelo FLRW estándar, si el desplazamiento hacia el rojo[11] (véase §27.7) de los cuásares es suficiente-

mente grande, entonces los conos de luz pasados de p y q no se intersectarán, de modo que no puede haber ningún tipo de comunicación entre ellos. Puesto que están incomunicados entre sí, no habrán tenido tiempo de «alisar» su ruptura de simetría para que sea consistente en la forma indicada antes. Enseguida consideraremos el «escenario inflacionario» que hace retroceder la línea del big bang, en el diagrama conforme, para poner a q y r en «comunicación» después de todo. Pero eso no nos servirá aquí, porque la 3-superficie Σ, donde va a tener lugar la ruptura de simetría electrodébil, desempeña efectivamente el papel del big bang en nuestras actuales consideraciones de causalidad, suponiendo que la ruptura espontánea de simetría ocurre aleatoriamente en la 3-superficie Σ, sin ninguna influencia causal común efectiva.

Ahora las líneas qp y rp son líneas nulas, de modo que solo el fotón puede viajar de q a p o de r a p, y no los bosones W y Z, pues el fotón es el único miembro sin masa de la familia de bosones gauge. Así pues, a lo largo de estas dos líneas nulas debemos tener una noción consistente de lo que es un fotón. Es muy probable que la noción de «fotón» en q sea incompatible (en el sentido antes indicado) con la noción de «fotón» en r, porque se suponía que cada uno de ellos había sido seleccionado aleatoriamente sin influencia causal común, y sin tiempo para que hubiera comunicación entre ellos.[12] ¿Pueden «alisarse» a tiempo los «diferentes» tipos de fotón para que el observador en p se ahorre una desconcertante confusión W-Z-γ cuando los reciba? No veo cómo puede hacerse esto posible sin alejarse significativamente de las conexiones nulas (i.e., «de género luz») directas desde q a p y desde r a p. Esto podría llevar a un grave conflicto con el hecho de que los objetos lejanos se ven con claridad a través de telescopios ópticos. Me parece que aquí hay peligro de tener una grave incompatibilidad con la observación, aunque yo no lo he visto planteado en la literatura especializada.

Pero algunos lectores tal vez se quejarán (quizá en voz baja) porque parece que he ignorado el muy impresionante apoyo observacional que hay en favor de la teoría electrodébil. ¡Por supuesto, no voy a abandonar todo eso solo por una confusión que pueda tener respecto a los fenómenos propagados desde distancias cosmológicas! Ni mucho menos. De ningún modo estoy sugiriendo que debamos abandonar las ideas esencialmente bellas de la teoría electrodébil, pero respecto a la

ruptura de su simetría U(2), prefiero mantener una actitud ligeramente diferente de la que se presenta normalmente. Tal como lo veo, todavía no ha salido a la luz el verdadero esquema de la naturaleza para la física de partículas. Un esquema semejante debería ser matemáticamente consistente y no tener el hábito desagradable que tienen nuestras QFT actuales de escupir la respuesta «∞» a preguntas físicas tan razonablemente planteadas. Pero de momento no podemos ver por qué esta teoría «correcta» (aún desconocida) va a dar respuestas finitas. Así que hemos recurrido a «trucos» diversos, que nos han abierto camino mediante una combinación de fortuna histórica y excepcional ingenio humano, y que nos permiten dar respuestas finitas que encajan con la observación. En nuestra fase actual de conocimiento, necesitamos una teoría de interacciones débiles y electromagnéticas que sea renormalizable, y la idea de una simetría gauge no abeliana rota no solo ha proporcionado un camino hacia tal teoría renormalizable, sino que las restricciones para hacerlo nos han guiado hasta una familia de verdades profundas sobre la forma en que estas interacciones encajan como parte de una imagen más amplia. Pero no veo por qué una simetría espontáneamente rota tiene que ser el verdadero camino de la naturaleza en física de partículas. De hecho, hay otras rutas para ver por qué las exigencias de renormalización proporcionan las relaciones necesarias entre los parámetros de la teoría electrodébil.[13]

Esto plantea una cuestión importante (a la que volveré en §34.8): la noción de *simetría*, tan dominante en muchas ideas para sondear los secretos de la naturaleza, ¿tiene realmente el papel fundamental que con frecuencia se le supone? No veo por qué tiene que ser así. No me resulta necesario que basar la física de partículas en algún gran grupo de simetría (lo que es parte de la filosofía GUT) sea realmente una imagen «simple», por lo que concierne a una teoría física fundamental. Para mí, los grupos de simetrías geométricas grandes son cosas complicadas, y no simples. Podría muy bien darse el caso de que existan asimetrías fundamentales intrínsecas en las leyes de la naturaleza, y que las simetrías que vemos sean a menudo características meramente aproximadas que no llegan a los niveles más profundos. Volveré más adelante a esta cuestión (en §34.8).

28.4. Cosmología inflacionaria

Volvamos a la cuestión de los monopolos cósmicos, cuya proliferación es una característica de ciertas GUT. El problema con estos monopolos era la ausencia de cualquier indicio de su existencia real. Peor aún, existen fuertes límites observacionales a la abundancia cósmica de tales monopolos, muy por debajo del nivel predicho por las GUT. Sin embargo, en 1981 Alan Guth presentó la propuesta «escandalosa» (también sugerida previamente y de forma independiente por Alexei Starobinski y Katsuoko Sato) de que si el universo se hubiera expandido en un factor de, digamos, 10^{30} o quizá incluso 10^{60} o más en algún período posterior a la producción de los monopolos (aunque anterior al momento de la ruptura de simetría electrodébil a los 10^{-12} s), entonces los monopolos no deseados serían ahora tan escasos que fácilmente podrían escapar a la detección, como se exigía a partir de la observación.

Pronto se advirtió que este «período inflacionario» de extraordinaria expansión exponencial también podría tener otros efectos que tenían que ver con la uniformidad del universo. Como se ha destacado en el capítulo 27, el universo es extraordinariamente uniforme y casi espacialmente plano a gran escala, lo que presentaba un enigma para los cosmólogos. Por ejemplo, la temperatura observada en el universo primitivo es prácticamente la misma (al menos hasta una parte en 10^5) en diferentes direcciones. Esto podría considerarse como el resultado de una «termalización» en el universo muy primitivo, pero solo si las diferentes partes del universo en cuestión estaban «en comunicación» entre sí. (Recordemos que la segunda ley de la termodinámica hace que se igualen las temperaturas de un gas en lugares diferentes, como parte del proceso de llegada al equilibrio térmico; véase §27.2.) Pese a todo, un examen de la Fig. 28.5a nos dice que la igualdad de las temperaturas en puntos u y v lejanos, observados ambos desde nuestra localización actual p en el espaciotiempo, no pueden ser resultado de termalización en los modelos cosmológicos convencionales, debido a que los puntos u y v (donde tomamos Σ en el instante de «desacoplamiento», cuando se formó la radiación cósmica de cuerpo negro) estaban demasiado lejos uno de otro para haber estado en comunicación causal en el modelo estándar.

Esta imposibilidad de la comunicación causal que se requeriría para la termalización, en el modelo estándar, se conoce como el *problema del horizonte*. A este respecto, el efecto del período de inflación se muestra en el diagrama conforme de la Fig. 28.5b. La 3-superficie de género espacio que representa el big bang se ha desplazado ahora a una localización muy anterior, de modo que los pasados de u y v *sí* se cortan antes de alcanzar la 3-superficie que describe el big bang, y hay oportunidad para que la termalización tenga efecto, por lo que podemos imaginar que la igualdad de temperaturas en u y v puede darse de esta forma.

Otra ventaja de este período inflacionario propuesto era que podía proporcionar una explicación de la extraordinaria uniformidad de la distribución de materia y la geometría espaciotemporal, que se conoce como el *problema de la suavidad*. La idea consiste en que, con inflación, el estado inicial del universo podría haber sido muy irregular en detalle, pero la enorme expansión del universo durante la etapa inflacionaria habría servido para «alisar» estas irregularidades, y por ello cabría prever un universo aproximadamente FLRW. El punto de vista inflacionario concibe que incluso un estado inicial «genérico» parecería una variedad suave a pequeña escala, y vemos que esta minúscula porción suave se expandió a escalas cosmológicas —hasta parecer espacialmente plana— en el curso de la fase inflacionaria; véase la Fig. 28.6 (y compárese con la Fig. 12.6). Volveré enseguida a mi propia valoración de esta extraordinaria idea. Por el momento, vale la pena señalar que en esta imagen el universo no solo es uniforme, sino que también tiene curvatura espacial nula ($K = 0$). Como veremos, este es un factor importante para el desarrollo histórico del tema. Pero sea o no espacialmente plano el universo observable en promedio, lo cierto es que está muy cerca de serlo, y esto había presentado un enigma para muchos cosmólogos, conocido como el *problema de la planitud*.

No es inmediato a simple vista que la fase de expansión inflacionaria tenga que ver con el retroceso de la 3-superficie del big bang en los diagramas conformes, como en la Fig. 28.5. Por consiguiente, será instructivo examinar el modelo cosmológico concreto sobre el que se basa esta «fase inflacionaria». Esta es la versión de «estado estacionario» del espacio de *De Sitter*. La forma más rápida de describir el espacio de

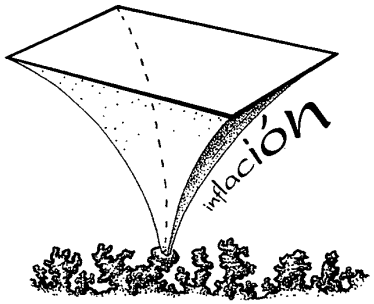


Fig. 28.6. Una de las motivaciones subyacentes de la inflación es que una escala de expansión exponencial de quizá 10^{50} (digamos entre 10^{-35} s y 10^{-32} s) podría servir para «alisar» un estado inicial genérico, y proporcionar así un universo post-inflación esencialmente uniforme y espacialmente plano.

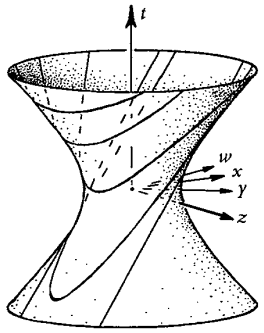


Fig. 28.7. El espaciotiempo de De Sitter (representado como un hiperboloide con dos dimensiones espaciales suprimidas) es una «4-esfera» lorentziana (de radio imaginario que da una signatura métrica intrínseca $+---$) en el 5-espacio de Minkowski $\mathbb{M}^5$ (cuya métrica es $ds^2 = dt^2 - dw^2 - dx^2 - dy^2 - dz^2$). Para obtener el modelo de estado estacionario, «cortamos» el hiperboloide por la mitad a lo largo de $t = w$; el tiempo constante está dado por $t - w$ positiva constante.

De Sitter es decir que es una «4-esfera» lorentziana (signatura $+---$) en un 5-espacio de Minkowski (signatura $+----$). Esta descripción está de acuerdo con las ideas de «salto de signatura» geométrica de §18.4, pero es geométricamente más evidente si representamos el espacio de De Sitter como el *hiperboloide* de la Fig. 28.7. En este punto, vale la pena mencionar otro modelo, llamado espacio *anti-De Sitter*, que es una «4-esfera» lorentziana en un 5-espacio pseudominkowskiano de signatura $++---$ (Fig. 28.8).[28.4] Nótese que el espacio anti-De Sitter no es un espaciotiempo muy razonable, desde un punto de vista físico, debido a que posee *curvas cerradas de género tiempo* que violan la

[28.4] Escriba explícitamente las ecuaciones para los 4-espacios de De Sitter y anti-De Sitter en el 5-espacio de fondo, utilizando las coordenadas t, w, x, y, z indicadas en las Figs. 28.8 y 28.9. Encuentre coordenadas dentro de la «mitad» del espacio de De Sitter, de modo que su métrica intrínseca tome la forma de «estado estacionario» que se da más adelante en esta sección.

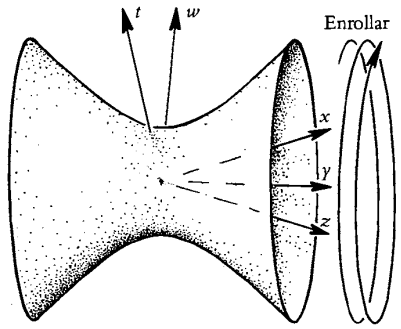


Fig. 28.8. El espaciotiempo anti-De Sitter (representado como un hiperboloide con dos dimensiones espaciales suprimidas) es una «4-esfera» lorentziana (de radio positivo, que da una signatura métrica intrínseca $+---$), en el 5-espacio pseudominkowskiano (con métrica $ds^2 = dt^2 + dw^2 - dx^2 - dy^2 - dz^2$). Tal como está definido, tenemos curvas cerradas de género tiempo, pero estas pueden eliminarse «enrollando» infinitamente en el plano (t, w).

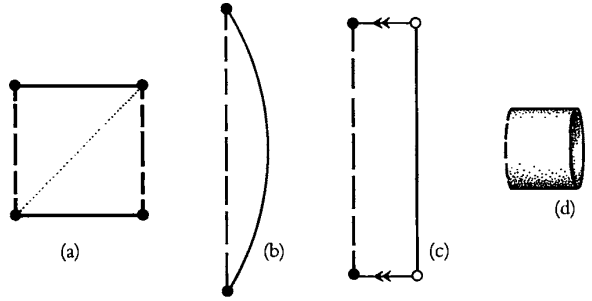


Fig. 28.9. Diagramas conformes estrictos (con los convenios de la Fig. 27.16a) de: (a) espacio de De Sitter, donde la región por encima de la línea de puntos interna da el modelo de estado estacionario; (b) espacio anti-De Sitter (versión completamente desenrollada, sin violación de causalidad); y (c) espacio anti-De Sitter en la forma «hiperboloidea» original que viola la causalidad, donde deben identificarse los bordes superior e inferior. (d) Lo mismo que (c), pero con la identificación realizada, de modo que el diagrama aparece como un cilindro.

causalidad (por ejemplo, el círculo en el plano que generan los ejes t y w); véanse §17.9 y la Fig. 17.18. A veces el término «espacio anti-De Sitter» se refiere a una versión «desenrollada» en la que cada círculo en un plano (x, y, z)-constante se ha desenrollado en una recta, y el espacio entero se hace simplemente conexo (§12.1). He dibujado un diagrama conforme estricto para el espacio de De Sitter en la Fig. 28.9a, y también para la porción del mismo que representa el modelo de es-

tado estacionario (la línea frontera *de puntos* indica el corte); para el espacio anti-De Sitter que viola la causalidad en la Fig. 28.9c (donde deben identificarse las partes inferior y superior del diagrama) y en la Fig. 28.9d, y para el espacio anti-De Sitter causal (desenrollado) en la Fig. 28.9b.

Para obtener explícitamente el universo en estado estacionario, «cortamos» por la mitad el espacio de De Sitter, a lo largo del 4-plano $t = w$ del 5-espacio de Minkowski que se muestra en la Fig. 28.7, y retenemos solo la mitad «superior».[14] Curiosamente, aunque hay una «incompleción» en este modelo debida al corte (líneas de trazos en la Fig. 28.10b), esta incompleción no se considera normalmente como un defecto, porque ninguna partícula real entra en el espaciotiempo desde la mitad inferior «borrada». La métrica para la mitad superior puede reexpresarse de la forma

$$ds^2 = d\tau^2 - e^{A\tau}(dx^2 + dy^2 + dz^2)$$

(siendo A una constante), que es un caso particular de las métricas FLRW que se han dado en §27.11, con secciones espaciales planas $K = 0$ y una expansión exponencial (el factor $e^{A\tau}$).[28.5] (Esta métrica fue de particular interés durante las décadas de 1950 y 1960, cuando Hermann Bondi, Thomas Gold y Fred Hoyle la defendieron firmemente como un modelo para el universo real: el modelo de «estado estacionario» de considerable atractivo estético. Perdió el favor en los años sesenta, después de que se hiciese manifiesto que el modelo estaba en conflicto con las observaciones, especialmente con las medidas del fondo de microondas y los recuentos de galaxias lejanas.)

El tensor de Ricci R_{ab} para el espacio (anti)De Sitter es *proporcional* a la métrica g_{ab}.[28.6] (Véase §19.6 para la definición de este tensor, y también para la ecuación de campo de Einstein, etc.) Recordemos la forma original de la ecuación de campo de Einstein $R_{ab} - \frac{1}{2}Rg_{ab} = -8\pi G T_{ab}$, que afirma que el tensor energía-momento para la materia es $-(8\pi G)^{-1}$ multiplicado por el inverso de la traza del tensor de Ricci. Así pues, para los modelos de De Sitter y anti-De Sitter, el «tensor de materia» T_{ab} debe ser proporcional al tensor métrico. De hecho, ninguna materia ordinaria puede tener esta propiedad (por ejemplo, porque su energía-momento no definiría ningún sistema en reposo). El punto de vista habitual es considerar que los espacios (anti)De Sitter representan vacíos sin materia, donde hay que tomar la ecuación de Einstein en la forma en que incluye una *constante cosmológica* Λ, de modo que las ecuaciones de campo nos dan ahora

$$R_{ab} = \Lambda g_{ab}.$$

Aquí $\Lambda = A^2$, donde A es la constante que fija la escala del factor de crecimiento exponencial en la métrica de estado estacionario anterior. En la cosmología inflacionaria se considera que el «material» inflacionario es un «falso vacío», sobre el que diré algo más enseguida.

Para construir un modelo de universo inflacionario, tomamos una porción del universo en estado estacionario, entre dos 3-superficies de τ constante, y la pegamos a dos partes de un modelo FLRW estándar $K = 0$. Este procedimiento se ilustra en la Fig. 28.10. En la Fig. 28.10a se corta el espacio de De Sitter entero para producir el modelo de estado estacionario. En la Fig. 28.10b se ha seleccionado una porción del modelo de estado estacionario que se hincha enormemente. En la Fig.

[28.5] Encuentre formas métricas para los espacios de De Sitter y anti-De Sitter del tipo FLRW $ds^2 = dt^2 - (R(t))^2 d\Sigma^2$, donde $d\Sigma^2$ da la 3-métrica hiperbólica de acuerdo con la segunda expresión en el ejercicio [27.18]. ¿Qué porción del espacio (anti)De Sitter completo cubre esto?

[28.6] ¿Puede ver por qué esto debe ser así sin hacer ningún cálculo?

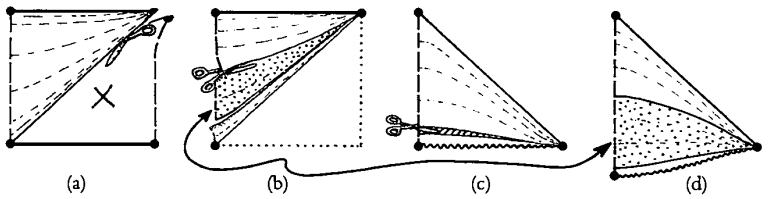


Fig. 28.10. Kit para construir un modelo de universo inflacionario. (a) Corte del espacio de De Sitter para dar un modelo de estado estacionario. (b) Porción que se infla enormemente, del modelo de estado estacionario seleccionado, entre dos líneas de tiempo constante. (c) Pequeño intervalo de tiempo constante eliminado del modelo $K = 0$ FLRW. (d) Porción de b insertada en c para obtener un modelo de universo inflacionario. Esto retrasa el big bang, como en la Fig. 28.5b.

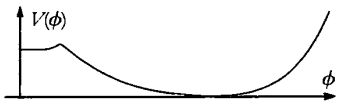


Fig. 28.11. La densidad de energía efectiva del universo muy primitivo, según el modelo inflacionario, estaría dominada por el potencial efectivo $V(\phi)$ para el campo cuántico escalar «inflatón» ϕ. La gráfica demuestra una forma habitualmente supuesta de $V(\phi)$, donde se considera que la inflación ocurre cuando el estado (la «canica» de la Fig. 28.1) «rueda» pendiente abajo a la izquierda (donde se supone que hay un «falso vacío»). La inflación cesa cuando se alcanza el fondo.

28.10c se corta un trozo del modelo FLRW $K = 0$, de modo que pueda recibir la porción inflacionaria para completar el modelo en la Fig. 28.10d. La porción de estado estacionario invertida «empuja hacia atrás» el big bang (desde el punto de vista conforme, i.e., causal), de modo que el horizonte de partículas está enormemente expandido; véase la Fig. 28.5b.

Para conseguir este período inflacionario, es necesario introducir un nuevo campo escalar φ en el bestiario de las partículas/campos físicos conocidos (y conjeturados). Por lo que conozco, no se considera que este campo φ esté directamente relacionado con ninguno de los otros campos conocidos de la física, sino que se introduce solamente para obtener una fase inflacionaria en el universo primitivo. A veces se conoce como un campo de «Higgs», pero no parece ser el campo «ordinario» relacionado con la teoría electrodébil (véase §25.5). Algunos modelos requieren más de una fase inflacionaria independiente, en cuyo caso tendría que haber un campo escalar independiente para cada fase. El proceso de inflación se describe en términos de una imagen que guarda cierta relación con la del «sombrero mexicano» de la Fig. 28.1, pero sin la simetría inicial. A menudo se utiliza un diagrama similar al de la Fig. 28.11, donde el eje vertical representa la «energía efectiva». La idea consiste en que antes del período de inflación, el estado —nuestra «canica», como en la Fig. 28.1— está representado en la cima de la protuberancia, pero luego rueda hacia abajo poco a poco. La inflación tiene lugar en el curso de esta caída y cesa cuando la «canica» llega al fondo. Durante la fase de inflación tenemos una región de «falso vacío» que representa una transición de fase me-

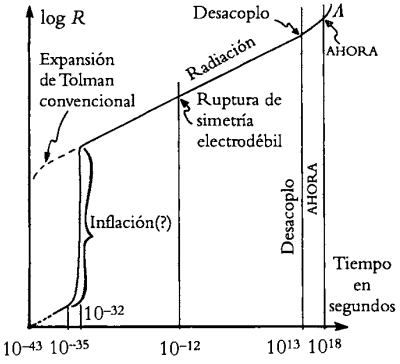


Fig. 28.12. «Historia del universo» descrita habitualmente como una representación logarítmica, que incluye una fase inflacionaria. Aquí se representa $\log R(t)$ frente a $\log t$.

canocuántica a un vacío diferente de aquel con el que hoy estamos familiarizados.

Como se ha mencionado en §27.11, ahora hay evidencias a favor de una Λ positiva en nuestra época actual, pero de un valor extraordinariamente pequeño en términos ordinarios, correspondiente a una densidad que es de solo aproximadamente 10^{-30} veces la del agua. Por el contrario, el falso vacío habría tenido una Λ efectiva correspondiente a una densidad que supera a la del agua en aproximadamente 10^{80}. Esto dominaría por completo el tensor energía-momento de cualquier materia ordinaria, y por esta razón el modelo de De Sitter puede utilizarse para esta fase.

En la Fig. 28.12 he indicado el tipo de imagen de la historia del universo muy primitivo que frecuentemente se nos presenta, y que ahora ha llegado a ser casi «estándar». Nótese que las escalas de tiempo y distancia son «logarítmicas» (como la regla de cálculo de la Fig. 5.6) y están marcadas con las diferentes potencias de 10, en unidades de un segundo (en vertical) y un centímetro (en horizontal). El «radio» denota la historia de la «$R(t)$» de §27.11 (que no debe confundirse con la curvatura escalar «R» de §19.6). En mi opinión, esta imagen debe considerarse muy especulativa hasta aproximadamente 1/10 (y con certeza 10^{-30}) de segundo, ¡aunque a menudo se presenta como un hecho prácticamente establecido!

28.5. ¿Son válidos los motivos para la inflación?

¿Qué razón hay para creer que una imagen semejante del universo está probablemente cerca de la verdad? A pesar de su evidente popularidad, quiero presentar mis propias razones para arrojar dudas considerables ¡sobre toda la idea! Una vez más debo dar al lector mi advertencia de rigor. La cosmología inflacionaria se ha convertido en una parte principal del corpus del moderno pensamiento cosmológico. El lector se dará cuenta de que incluso entre aquellos que todavía no están convencidos de la necesidad de la inflación, hay pocos que sean tan negativos como yo en la crítica que sigue. Si siente la necesidad de «compensar» mi exposición con otra más favorable a la idea inflacionaria, consulte el muy accesible libro de Alan Guth *El universo inflacionario*.[15] Por lo que a mí respecta, debo presentar las cosas tal como las veo, y puesto que creo que hay razones poderosas para dudar de la propia base de la cosmología inflacionaria, no me contendré al presentar estas razones al lector.

Pero antes de hacer mi juicio crítico, debería dejar claro que mis comentarios *no* nos dicen que la cosmología inflacionaria sea falsa. Simplemente ofrecen serias razones para dudar de muchos de los motivos iniciales que hay tras la idea inflacionaria. Recordemos que, después de todo, muchas ideas científicas importantes en el pasado se han basado (en parte) en motivos que no resisten un examen a la luz del conocimiento posterior. Uno de los más importantes fue la significativa dependencia de Einstein del *principio de Mach* como guía para su descubrimiento final de la relatividad general. El principio de Mach afirma que la física debería definirse enteramente en términos de la relación mutua entre cuerpos, y que la propia noción de un espacio de fondo debería ser abandonada.[16] No obstante, un análisis posterior de la teoría de Einstein demostró que el principio de Mach no está incorporado en la relatividad general,[17] con independencia de cuál pueda ser la importancia motivacional de la idea de Mach.[18] Otro ejemplo fue el descubrimiento por parte de Dirac de la ecuación de onda del electrón, para lo que se basó fundamentalmente en lo que él veía como necesidad de una ecuación de primer orden (véanse §§24.5,6). La comprensión posterior de la QFT demostró que este requisito no es necesario (§26.6).

Del mismo modo, si las predicciones observacionales de la cosmología inflacionaria se confirman de forma convincente, entonces cualquier inadecuación de las motivaciones iniciales no será tan importante, y la teoría podrá mantenerse por sí misma sin la «escalera» original que condujo a Guth y a otros a este esquema concreto. De hecho, los inflacionistas han hecho algunas predicciones precisas que en años recientes se han mostrado de acuerdo con varias observaciones nuevas.

Creo que, a diferencia de muchas de las otras ciencias, es recomendable ser cauteloso en materias de cosmología, sobre todo en relación con el origen del universo. La gente suele tener fuertes respuestas emocionales a las preguntas sobre el origen del universo, que a veces están implícita o explícitamente relacionadas con tendencias religiosas. Esto no deja de ser natural, pues se trata en realidad de la creación del mundo en que vivimos. Como se ha señalado en §27.13, debido a la segunda ley existe un extraordinario grado de precisión en la forma en que empezó el universo, en el big bang, y esto presenta un enigma profundo. Preguntemos: ¿es la solución a este enigma de la precisión del big bang algo a lo que se pueda responder con una teoría científica futura, incluso si está aún más allá de nuestro conocimiento científico actual? (esta es esencialente mi postura optimista; véanse §§30.10-13) ¿O debemos resignarnos a que esto sea un tipo de «acto divino»? La visión de los inflacionistas es diferente, a saber, que este rompecabezas queda «resuelto» esencialmente por su teoría, y esta creencia ofrece una poderosa fuerza impulsora tras la postura inflacionaria. Sin embargo, ¡nunca he visto que el profundo enigma planteado por la segunda ley sea seriamente planteado por los inflacionistas!

En su lugar, los inflacionistas tienden a señalar tres problemas concretos en el modelo estándar de la cosmología, y los tres son cuestiones relacionadas con la precisión inicial del universo primitivo. Han sido abordados específicamente en §28.4, y se conocen como el problema del horizonte, el problema de la suavidad y el problema de la planitud. En el modelo estándar, estas cuestiones se tratan mediante el «ajuste fino» del estado del big bang inicial, algo que se considera «feo» por parte de los inflacionistas. Ellos afirman que la necesidad de tal ajuste fino del estado inicial queda eliminada en la imagen inflacionaria, que estéticamente se considera una imagen física más agradable. La conclusión de

la planitud espacial global que resulta de la inflación se considera también un aspecto positivo, desde un punto de vista estético.[19]

Creo que habría que ser muy cauteloso en relación con estos argumentos con base estética. Por supuesto, hay algunos elementos fundamentales para la imagen inflacionaria cuyo estatus estético es algo cuestionable, tales como la introducción de un campo escalar (o quizá varios campos escalares independientes, si se contempla más de un período de inflación) no relacionado con otros campos conocidos de la física y con propiedades muy específicas diseñadas solo al efecto de hacer que la inflación funcione. Además, la preferencia estética por $K = 0$ es muy discutible. ¡Conozco a muchos matemáticos (entre los que me incluyo) que consideran mucho más bello el caso hiperbólico ($K < 0$)! Y otros prefieren lo «acogedor» de un universo espacialmente finito ($K > 0$). La cuestión general del papel de la belleza como guía en la física teórica básica será examinada más adelante en este libro (véase §34.9), como lo serán otras cuestiones relacionadas específicamente con la inflación (§34.4) y con el papel de las modas científicas (§34.3). La inflación está ciertamente muy de moda entre los cosmólogos actuales, y es importante ver hasta qué punto se justifica este estatus.

Como se ha dicho antes, mis objeciones básicas a esta idea de la inflación cósmica tienen que ver principalmente con las motivaciones que subyacen en ella. Consideremos primero el problema del horizonte y cómo lo trata la cosmología inflacionaria en la que, por ejemplo, las temperaturas de fondo, casi iguales en cualquier dirección, se ven como resultado de la *termalización*. La inflación se introduce para eliminar los horizontes de partículas que de otro modo impedirían dicha termalización.

Sin embargo, hay algo fundamentalmente erróneo en tratar de explicar la uniformidad del universo primitivo como resultado de un proceso de termalización (§28.4), ya se trate de una uniformidad en la temperatura de fondo, en la densidad de materia o en la geometría espaciotemporal en general. De hecho, es fundamentalmente erróneo tratar de explicar por qué el universo es especial en *cualquier* aspecto concreto apelando a un proceso de termalización. Pues si la termalización está realmente haciendo algo (tal como que las temperaturas en regiones diferentes sean más iguales que las que eran antes), entonces

representa un definido incremento de la entropía (§27.2). Así pues, el universo tendría que haber sido incluso más especial antes de la termalización que después. Esto solo sirve para incrementar cualquier dificultad que pudiéramos haber tenido previamente al tratar de entender la extraordinariamente especial naturaleza inicial del universo (§27.13). Hay, por supuesto, profundos interrogantes con respecto al estado restringido del universo primitivo. Pero estas ligaduras son fundamentales para la propia existencia de la segunda ley de la termodinámica, como se ha señalado en el capítulo 27. ¡No cabe esperar que seamos capaces de explicar estas ligaduras apelando simplemente a *manifestaciones* de la segunda ley (siendo la termalización un ejemplo)!

Para desarrollar este punto, consideremos la cuestión de la igualdad de temperaturas tal como se ven en direcciones diferentes desde nuestro punto de vista particular en el universo. Supongamos que descubrimos que las temperaturas en dos regiones lejanas han sido iguales en algún instante cósmico temprano t_1, y supongamos que encontramos desconcertante este «hecho especial». Consideremos dos posibilidades. Podríamos imaginar (a) que en una era aún anterior —tiempo t_0— las temperaturas eran realmente desiguales y solo llegaron a igualarse una vez que tuvo lugar un proceso de termalización entre los instantes t_0 y t_1. Alternativamente, podríamos imaginar (b) que en el instante anterior t_0 las dos temperaturas eran realmente iguales entre sí, y no tuvo lugar ninguna termalización. En el caso (a) encontramos que ha habido un incremento de entropía entre t_0 y t_1, de modo que encontramos un grado aún mayor de especialidad en t_0 que la que había en t_1, por lo que deberíamos estar aún más intrigados por la naturaleza especial del universo en el instante t_0 de lo que lo estábamos por su carácter especial en el instante t_1. ¡El problema ha empeorado! En el caso (b), por el contrario, el problema del carácter especial en t_0 no es peor al menos que el que había en t_1. En ninguno de los dos casos hemos explicado el *enigma de por qué el universo es especial*, en este o en cualquier otro aspecto, y vemos que invocar argumentos de termalización para abordar este problema particular ¡es peor que inútil!

¿Qué pasa con la uniformidad (y planitud) del universo? Aquí el argumento inflacionario principal es diferente. Se afirma que la expansión exponencial de la fase inflacionaria fue la que sirvió para ha-

cer el universo tan uniforme (y espacialmente plano). De nuevo hay un equívoco fundamental. Parece que existe la idea de que si partimos de un estado inicial «genérico», entonces el «efecto de estiramiento» de la expansión exponencial de la fase inflacionaria servirá para alisar las irregularidades de dicho estado inicial. Por supuesto, para saber si tal proceso tiene una oportunidad, tenemos que tener alguna idea de a qué podría parecerse una geometría inicial «genérica». Un presupuesto importante es que un estado semejante tendría que ser, en cierta escala pequeña, suave. Pero los conjuntos fractales, por ejemplo, nunca se alisan por mucho que se estiren. Recordemos el conjunto de Mandelbrot, algunas de cuyas porciones se han mostrado en la Fig. 1.2. Si acaso, el conjunto de Mandelbrot parece hacerse menos suave cuanto más se amplía.

Pero oigo murmurar al lector: ciertamente eso es tan solo una objeción menor; muy bien, quizá haya algunas situaciones patológicas en las que estirar no suaviza las cosas, pero ciertamente en el caso realista más general no esperaríamos tales cosas. Por desgracia, esto no está tan claro; tendríamos que estar preparados para aceptar algo fractal (o peor que fractal) en un estado de partida genérico. Cualquiera que sea esta estructura genérica singular, no es algo de lo que podamos esperar que se alise simplemente debido a una física que permita procesos inflacionarios. ¿Por qué es así? Las razones no tienen nada que ver con tecnicismos detallados, y son simplemente intrínsecas al carácter equívoco de tratar de suponer que el universo podría haber empezado realmente en un estado genérico,[20] lo que no puede haber hecho, debido a la segunda ley; véase §27.7. Si queremos hacernos una idea de a qué podría parecerse dicho estado «genérico», consideremos las etapas finales de un universo cerrado que colapsa, tal como el que se ilustra esquemáticamente en la Fig. 27.20a,b, y luego invirtamos el flujo del tiempo, como en la Fig. 27.20c (o Fig. 27.20d). El gran revoltijo de singularidades de agujero negro que coagulan es el tipo de cosa que, en *forma de tiempo invertido*, deberíamos esperar para un big bang genérico.

¡Por supuesto, no estoy pidiendo que el lector tenga una comprensión instantánea de la geometría complicada y detallada implicada en un big crunch genérico desordenado! Yo mismo tengo una idea muy

pobre y no creo que nadie sepa mucho más sobre ello.[21] Pero no necesitamos saber muchos detalles sobre esta geometría. Para entender la cuestión esencial, consideremos cualquier modelo de universo en colapso, que podemos construir partiendo de un estado expandido inicial muy irregular (compárese con la Fig. 27.20b). Tiene que colapsar a *algo*; de hecho, su colapso dará como resultado algún tipo de *singularidad espaciotemporal* genérica, como podemos inferir razonablemente a partir de teoremas matemáticos precisos.[22] Si ahora invertimos la dirección del tiempo en nuestro modelo —suponiendo leyes dinámicas con simetría temporal—, obtenemos una evolución que parte de una singularidad de apariencia general y luego se convierte en cualquier tipo irregular de universo que queramos elegir. Muy bien podría suceder que no haya inflación en esta evolución, aunque nuestras leyes físicas invertidas en el tiempo permiten la posibilidad de inflación. La cuestión es que tengamos o no realmente inflación, la posibilidad física de un período inflacionario no aporta nada en los intentos por asegurar que la evolución a partir de una singularidad genérica llevará a un universo uniforme (o espacialmente plano).

Tratemos de entender cuál es el problema real. Esto se ha examinado extensamente en el capítulo 27. El universo *era* muy especial en el big bang. Tuvo que serlo para que haya habido una segunda ley de la termodinámica que se extiende hasta el comienzo. Todos los procesos de termalización *dependen* de la segunda ley, así que no explican por qué tenemos una segunda ley ni por qué teníamos un universo muy especial en el comienzo. Más aún, todos los procesos de ruptura espontánea de simetría y todas las transiciones de fase (que son necesarias para la inflación) tienen lugar solo gracias a la segunda ley. Estos procesos no explican la segunda ley: la *utilizan*. Además, todos los cálculos serios en cosmología inflacionaria suponen una geometría espaciotemporal que es FLRW, o próxima a ella, lo que no da ninguna idea respecto a lo que sucedería en el caso genérico. Si queremos saber por qué el universo era tan sumamente especial inicialmente, debemos apelar a argumentos muy diferentes de aquellos de los que depende la cosmología inflacionaria.

28.6. El principio antrópico

Antes de que lleguemos a estos argumentos, tengo que abordar otra cuestión que se suele invocar como parte del punto de vista inflacionario. Se trata del *principio antrópico*, y este principio se utiliza también en muchos otros argumentos para explicar por qué el universo es tal como lo vemos. Hablando vagamente, el argumento antrópico parte del hecho de que el universo que percibimos a nuestro alrededor debe ser de tal naturaleza que produzca y acomode seres que puedan percibirlo. Podríamos utilizar este argumento para explicar por qué el planeta en que vivimos tiene unos rangos de temperaturas, una atmósfera, una abundancia de agua, etc., tan agradables. Si las condiciones no fueran tan agradables en este planeta concreto, no estaríamos aquí, sino que ¡estaríamos en algún otro lugar![23]

Uno de los usos más admirables del argumento antrópico fue el que hicieron Robert Dicke en 1957 y Brandon Carter en 1973,[24] cuando resolvieron un enigma —señalado por Dirac (1937)— concerniente a una relación aparentemente accidental entre la edad del universo, cuando se mide en unidades de Planck (§27.10), y la razón entre las intensidades de la gravedad y el electromagnetismo.[25] Si esta coincidencia reflejara una relación fundamental entre los parámetros de la naturaleza, entonces debería mantenerse constante a lo largo de la historia del universo. Pero puesto que la edad del universo es algo que (¡obviamente!) aumenta con el tiempo, debería reducirse en consecuencia la intensidad de las fuerzas gravitatorias respecto a las eléctricas. De hecho, Dirac hizo realmente esa sugerencia, pero la evidencia actual afirma que una variación semejante en la constante gravitatoria es incompatible con los hechos.[26] Lo que Dicke y Carter demostraron era que existe otra explicación para la coincidencia de Dirac. Examinando el papel exacto que desempeñan las constantes de la naturaleza en la determinación del tiempo de vida de una estrella ordinaria —una estrella agradable para la vida tal como la conocemos—, fueron capaces de demostrar que esta escala de tiempo es de un orden tal que la coincidencia de Dirac debería darse necesariamente para seres que evolucionan (y habitan) en un planeta en órbita alrededor de dicha estrella. Así pues, la coincidencia de Dirac tiene una explicación antrópica. Se

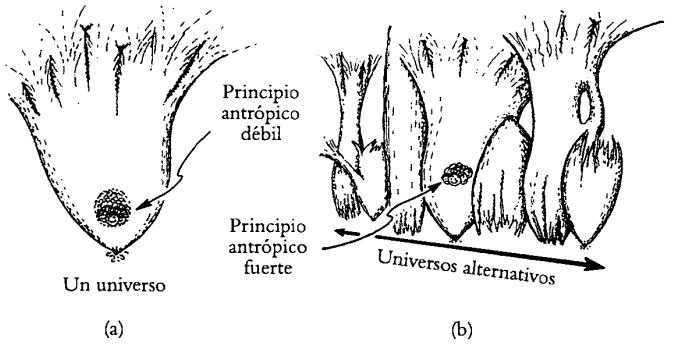


Fig. 28.13. Principio antrópico. (a) Forma débil: los seres sintientes deben encontrarse en una localización espaciotemporal en el universo, en la que haya condiciones apropiadas para la vida sintiente. (b) Forma fuerte: más que considerar solo un universo, concebimos un conjunto de universos posibles, entre los cuales pueden variar las constantes fundamentales de la naturaleza. Los seres sintientes deben encontrarse localizados en un universo donde las constantes de la naturaleza (además de la localización espaciotemporal) sean favorables.

da porque los parámetros implicados en la producción de vida sintiente (en este caso, los que determinan la edad de una estrella) están relacionados con aquellos parámetros que tal vida sintiente ¡*verá* realmente en el mundo exterior!

Debería ser evidente para el lector que los argumentos tomados del principio antrópico están plagados de incertidumbres, aunque no carecen de significado genuino. No tenemos mucha idea, por ejemplo, de cuáles son las condiciones realmente necesarias para la producción de vida sintiente. De todas formas, la situación no es tan mala cuando se utiliza con ejemplos, tales como los que se han expuesto antes, donde estamos tomando las leyes de la física y la estructura espaciotemporal del universo global como dadas y simplemente preguntamos cuestiones como dónde o cuándo pueden darse en el universo tales y cuales condiciones para que conduzcan a la vida sintiente. Esta versión del principio antrópico se conoce, como la llamó Carter, como el principio antrópico *débil* (Fig. 28.13a).

Mucho más problemáticas son las versiones del principio antrópico *fuerte*, según las cuales tratamos de extender el principio antrópico

para determinar las constantes reales de la naturaleza (tales como la razón entre la masa del electrón y la del protón, o el valor de la constante de estructura fina; véanse §26.9 y §31.1). Algunas personas podrían considerar que el principio antrópico fuerte nos lleva a una creencia en un «designio divino», con el que el Creador del universo aseguraba que las constantes físicas fundamentales estuvieran preordenadas para que tuvieran valores específicos que hicieran posible la vida sintiente. Por el contrario, podemos pensar que el principio antrópico fuerte es una extensión del débil en la que ampliamos nuestras preguntas de «dónde» y «cuándo», de modo que se apliquen no solo a un espaciotiempo único, sino al conjunto completo de los espaciotiempos posibles (Fig. 28.13b).[27] Cabría esperar que diferentes miembros del conjunto posean diferentes valores de las constantes físicas básicas. Las preguntas dónde y cuándo también implican ahora una elección de universo dentro del conjunto, de modo que de nuevo debemos encontrarnos en un universo que permite que se produzca la vida sintiente.

Que yo sepa, el primer ejemplo de este tipo fue señalado por Fred Hoyle, cuando dedujo que debía existir un hasta entonces inobservado nivel energético nuclear en el carbono para que fuera posible que las estrellas construyan elementos más pesados que el carbono en el proceso de nucleosíntesis. Este es el proceso mediante el cual se producen los elementos más pesados de los que dependen nuestros cuerpos. (Estos elementos se producen en las estrellas, y finalmente son escupidos en explosiones de supernovas para proporcionar el material para la formación de planetas; véase §27.8.) ¡Sin ello no podríamos existir como seres vivos (del tipo que conocemos)! A instancias de Hoyle, William Fowler y sus colaboradores[28] encontraron posteriormente el nivel energético de Hoyle, confirmando así, en 1953, una impresionante predicción por parte de Hoyle. Resulta muy notable que las constantes de la naturaleza estén tan bien ajustadas para que exista un nivel de energía semejante exactamente en el lugar adecuado para que pudiera producirse la vida tal como la conocemos. Otro ejemplo de buena suerte cósmica es el hecho de que la masa del neutrón es solo ligeramente mayor que la del protón (1.838 y 1.836 masas electrónicas, respectivamente). La existencia de una familia adecuada de núcleos estables, de la que depende la casi totalidad de la química, se basa en este hecho aparentemente fortuito.

Mi postura es sumamente cauta sobre el uso del principio antrópico, y muy en especial del principio fuerte. Tengo la impresión de que el principio antrópico fuerte se utiliza a menudo como una forma de «escurrir el bulto» cuando parecen haberse agotado las consideraciones teóricas genuinas. Con frecuencia he oído a teóricos que recurren a argumentos como: «Los valores de los parámetros desconocidos en mi teoría serán determinados, en última instancia, por el principio antrópico». Por supuesto, podría resultar que en definitiva no exista ninguna forma matemática de fijar ciertos parámetros en la «teoría verdadera», y que estos parámetros deban elegirse simplemente de forma que el universo en el que nos encontramos permita vida sintiente. ¡Pero tengo que confesar que no me gusta mucho esa idea!

Creo que con un universo espacialmente infinito y esencialmente uniforme (por ejemplo, $K \leq 0$ en el modelo estándar) el principio antrópico fuerte no sirve prácticamente de nada para ajustar parámetros físicos, más allá de exigir que las leyes físicas sean tales que permitan que la vida sintiente sea posible (lo que es en sí mismo inutilizable, puesto que no sabemos cuáles son los prerrequisitos para la vida sintiente). Pues si la vida sintiente es posible, entonces esperamos que en un universo espacialmente infinito ocurrirá. Esto sucederá incluso si las condiciones para la vida sintiente son extraordinariamente poco probables para que se den en cualquier región finita del universo. En un universo espacialmente infinito, nuestra expectativa es que debería haber *algún lugar* dentro de sus infinitos confines donde se dé la vida sintiente, aunque solo sea por que el mero azar reúna todos los ingredientes necesarios. Esto ocurriría simplemente por azar, incluso si es extraordinariamente infrecuente.

Ahora bien, si encontramos que las constantes físicas fundamentales resultan ser tales y cuales —quizá determinadas por criterios matemáticos—, entonces podemos plantear una pregunta mejor: ¿cuáles son las *circunstancias más probables* para que se produzca la vida inteligente, dados estos valores de las constantes físicas? En el universo que conocemos, con los valores de las constantes físicas que tenemos, *parece* que la respuesta es: «En algún planeta bastante parecido a la Tierra, próximo

a una estrella bastante parecida al Sol, que ha existido durante quizá 10^9 o 10^{10} años, tiempo suficiente para permitir que tenga lugar una evolución darwiniana apropiada». Pero para un universo con valores diferentes de las constantes, la respuesta podría ser muy diferente.

Para terminar esta sección, debería mencionar un punto de vista afín con respecto a las constantes físicas fundamentales, propuesto originalmente por John A. Wheeler en 1973. Tiene alguna relación con el principio antrópico. Según este punto de vista, el universo pasa por ciclos, donde continuamente ocurren nuevos «big bangs», cada uno de ellos nacido a partir de una fase de colapso previa.

Recordemos el modelo de Friedmann en el caso $K > 0$, $\Lambda = 0$. El universo se expande a partir de la singularidad del big bang inicial y luego se contrae hasta otra singularidad, el big crunch final. Sin embargo, en los primeros días de la cosmología esto se conocía como un modelo «oscilante», porque la curva que representa $R(t)$ frente a t es una cicloide que admite un número infinito de ciclos de expansión y contracción (véanse la Fig. 27.15a y el ejercicio [27.19]). No obstante, ahora se aprecia mejor que antes que no hay forma de «suavizar» la singularidad que une cada «crunch» con el siguiente «bang», dentro de los límites de la relatividad general clásica convencional.[29] Si se ignora este hecho, o se presume que alguna forma de «gravedad cuántica» permitirá que se produzca tal «rebote», entonces cabe especular que el cicloide de Friedmann sea una aproximación plausible a lo que realmente podría suceder. La idea de Wheeler era que la física cuántica extrema que tiene lugar en el punto de retorno singular implica un cambio en las constantes fundamentales de la naturaleza. En consecuencia, el «conjunto» de universos que se contempla en relación con el principio antrópico fuerte se realiza físicamente en la propuesta de Wheeler.

Lee Smolin, en su extraordinario libro de 1997 *The Life of the Cosmos*,[30] sugiere una modificación intrigante de esta idea. En lugar de exigir un universo cerrado cuyo big crunch omniabarcador se convierte en el big bang de la próxima fase de universo, Smolin considera que las singularidades en el interior de los agujeros negros son fuentes de nuevas fases de universos, donde cada singularidad de agujero negro produce individualmente una fase de universo diferente,[31] y donde en cada caso habría un ligero reajuste de las constantes físicas fundamentales. Smolin presenta la idea ingeniosa de que podría haber entonces alguna forma de «selección natural» de universos, donde las constantes fundamentales evolucionan lentamente para obtener fases de universo «mejor adaptadas», y toma la proliferación de agujeros negros como un mejor indicio de un «ajuste» del universo (porque produce muchos «hijos») que cualquier consideración antrópica. Argumenta que hay algún indicio de que las constantes físicas fundamentales que encontramos realmente en nuestro universo son tales que favorecen una proliferación de agujeros negros. Sin embargo, creo que el argumento antrópico tendría también un papel importante en esta discusión, ¡puesto que no podríamos encontrarnos en una fase de universo de «muerte de la vida sintiente», por muchas de ellas que haya!

El lector puede preguntarse cómo la masa-energía de un único agujero negro podría convertirse en la de un universo entero, que muy bien podría ser 10^{22} veces más masivo. De hecho, puesto que se necesita una física desconocida para evitar la singularidad y alterar las constantes fundamentales, «ya no hay apuestas» con respecto a las leyes de conservación estándar de la física convencional. En cualquier caso, puede argumentarse que la ley de conservación de masa-energía es problemática en el contexto de la relatividad general sin la hipótesis de planitud asintótica (véase §19.8).

Tengo un montón de dificultades con las propuestas de Wheeler y Smolin. En primer lugar, está la naturaleza extraordinariamente especulativa de la idea clave de que alguna física actualmente desconocida en la actualidad no solo pueda convertir la singularidad espaciotemporal del colapso en un «rebote», sino también reajustar ligeramente las constantes físicas fundamentales cuando esto sucede. No conozco ninguna justificación a partir de la física conocida que sugiera tal extrapolación. Pero, en mi opinión, todavía es menos plausible geométricamente que las singularidades altamente irregulares que resultan del colapso puedan convertirse por arte de magia en (o adherirse a) el big bang extraordinariamente suave y uniforme que cada nuevo universo necesitaría si va a adquirir una respetable segunda ley del tipo que nos es familiar (véase §27.13).

28.7. La naturaleza especial del big bang: ¿una clave antrópica?

¿Puede invocarse el principio antrópico para explicar la naturaleza tan especial del big bang? ¿Cómo puede incorporarse dicho principio como parte de la imagen inflacionaria, de modo que un estado inicialmente caótico (máxima entropía) pueda conducir, pese a todo, a un universo como este en el que vivimos, en el que impera la segunda ley? Básicamente, el argumento general consiste en decir que la segunda ley es esencial para la vida tal como la conocemos; más aún, las densidades, temperaturas, distribuciones de materia y composiciones, etc., globales, deben ser tales que conduzcan a la vida. Además, el universo debe haber existido durante el tiempo suficiente para que la evolución actúe, y todo lo demás. A veces se utiliza este argumento junto con un argumento inflacionario. Por consiguiente, aunque un estado inicial completamente genérico podría no hincharse para darnos un universo suavizado como el que observamos, lo único que habría que pedir es que alguna pequeña región de la «variedad» espaciotemporal inicial, inmediatamente después del big bang, sea suficientemente suave para que la inflación domine en dicha región y todo el universo observable actual surja como resultado de una inflación de esa minúscula región suave (véase la Fig. 28.14a). El argumento diría aproximadamente: «Para que exista vida sintiente necesitamos un gran universo con escalas de tiempo suficientemente grandes para que tenga lugar la evolución, en condiciones propicias, etc.; esto requiere cierta inflación, que se origina a partir de nuestra minúscula región inicial suave,y una vez que empieza, la inflación continúa para ofrecernos el universo observable maravilloso y enorme que hoy conocemos».

Aunque pueda parecer que esta imagen es de una naturaleza tan maravillosamente romántica que resulta completamente inmune al ataque científico, no creo que sea así.Volvamos al extraordinario grado de precisión (o «ajuste fino») que parece exigido por un big bang como el que parecemos observar. Como se ha argumentado en §27.13, la precisión requerida, en términos de volumen de espacio de fases, es al menos de una parte en $10^{10^{123}}$. El exponente «10^{123}» procede de la entropía de un agujero negro de masa igual a la del universo observable.

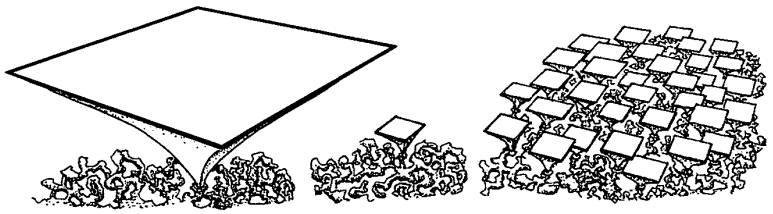


Fig. 28.14. (a) Un estado inicial completamente general para el universo no se infla, pero podemos buscar meramente una pequeña región inicial que es suficientemente suave para inflarse hasta el universo que observamos (coste: $10^{10^{123}}$). (b) Pero ¿cuánto de nuestro enorme universo es realmente necesario para nuestra existencia sintiente? Para el Creador es absurdamente «más barato», para la creación de vida sintiente, producir un universo de una décima parte de la dimensión lineal (coste: $10^{10^{117}}$). (c) Para crear tantos seres sintientes como en (a), el Creador puede producir de forma mucho más barata 10^3 ejemplares independientes de los universos «más pequeños» de (b) (a un coste de «ganga»: $(10^{10^{117}})^{1.000} = 10^{10^{129}}$). Por ello el principio antrópico no explica la aparente extravagancia de la inflación.

Pero ¿realmente necesitamos todo el universo observable para que pueda darse la vida sintiente? Esto parece poco probable. Resulta difícil imaginar que fuera necesario siquiera algo externo a nuestra galaxia. Pese a todo, pudiera ser que la vida inteligente sea muy rara, y podría ser un poco más cómodo tener algo más de espacio que esto. Seamos muy generosos y pidamos que una región del radio de una décima parte de la distancia al límite del universo observable deba parecerse al universo que conocemos, pero sin que nos preocupe lo que suceda fuera de ese radio. El volumen del espacio de fases puede calcularse como antes. Calculamos que la masa en dicha región es 10^{-3} de la que teníamos antes, y que esto nos da una entropía de agujero negro de 10^{-6} de la que teníamos antes.[28.7] Así pues, la precisión necesaria por parte de nuestro «Creador» (véase la Fig. 27.21) para construir esta región más pequeña es ahora de solo:

una parte en $10^{10^{117}}$.

[28.7] ¿Por qué?

Echemos una ojeada a la Fig. 28.14b. Ahora nuestro Creador solo requiere una «minúscula región suave» de la «variedad» inicial, bastante *más pequeña* que antes. Es mucho más probable que el Creador encuentre una región suave de este menor tamaño que la región algo mayor que hemos considerado antes. Suponiendo que la inflación actúa sobre la región pequeña de la misma forma que lo haría sobre la región algo mayor, aunque produciendo un universo inflado menor, en proporción, podemos estimar cuánto más frecuentemente encuentra el Creador las regiones más pequeñas antes que las más grandes. La cifra no es mejor que

$$10^{-10^{117}} \div 10^{-10^{123}} = 10^{10^{123}}$$

(dentro de la precisión expresada por los exponentes más altos).[28.8] Se ve qué extravagancia tan increíble era (en términos de probabilidad) que el Creador se molestase en producir esta parte lejana extra del universo, que en realidad no necesitamos —y que el principio antrópico en realidad no necesita— para nuestra existencia.

Algunos lectores podrían lamentarse de que esta «economía» por parte del Creador haya podido producir un número relativamente menor de seres sintientes. Esté esto en cuestión o no, no es la respuesta a por qué tuvo lugar la «extravagancia». En términos de probabilidades (i.e., tamaño de cajas en el espacio de fases; véase la Fig. 27.2), sería mucho «más barato» —en un factor de aproximadamente 1 frente a $10^{10^{123}}$— tener 10^3 regiones menores de universo inflado (que nos dan el mismo número de seres sintientes que una única más grande) que tener solo una región de universo más grande (Fig. 28.14c).[28.9]

Para ver hasta qué punto es impotente el argumento antrópico en este contexto, consideremos los hechos siguientes. La vida en la Tierra no necesita directamente la radiación de fondo de microondas. De hecho, ¡ni siquiera necesitamos la evolución darwiniana! Habría sido mucho más «barato», en términos de «probabilidades», producir vida sintiente a partir de la unión aleatoria de gas y radiación. (Se puede estimar que el sistema solar entero, junto con sus habitantes, podría

[28.8] Explique estas cifras.
[28.9] Explique las cifras con detalle.

crearse a partir de la colisión aleatoria de partículas y radiación con una probabilidad de una parte en $10^{10^{60}}$ —o probablemente mucho menos—, que es el «chocolate del loro» en comparación con los $10^{10^{123}}$ necesarios para el big bang del universo observable.)[32] No necesitamos que haya un big bang en esta configuración uniforme observada. No necesitamos la segunda ley en tiempos anteriores a que hubiera vida. Sería mucho más «barato» para el Creador no molestarse con eso. Y la inflación no sirve en absoluto. La curva de economía «barata» que debería adoptar el Creador en la Fig. 27.8 solo para producir vida sintiente se parecería mucho más a la curva (c)(b) que a la observada (d)(b), ¡con o sin inflación!

Todo esto refuerza simplemente el argumento de que es erróneo buscar razones del tipo anterior, donde se supone que las condiciones adecuadas del universo han resultado de algún tipo de elección inicial aleatoria. Había algo muy especial en el punto de partida del universo. Me parece que hay dos rutas posibles para abordar esta cuestión. La diferencia entre ambas es una cuestión de actitud científica. Podríamos adoptar la postura de que la elección inicial fue un «acto divino» (parecido al ilustrado imaginativamente en la Fig. 27.21), o podríamos buscar alguna teoría científico-matemática para explicar la naturaleza extraordinariamente especial del big bang. Por supuesto, siento una gran inclinación por tratar de ver hasta dónde podemos llegar con la segunda posibilidad. Nos hemos acostumbrado a utilizar leyes matemáticas —leyes de extraordinaria precisión— que controlan el comportamiento físico del mundo. Parece que una vez más necesitamos algo de precisión excepcional, una ley que determine la naturaleza misma del big bang. Pero el big bang es una singularidad espaciotemporal, y nuestras teorías actuales no son capaces de manejar estos aspectos. No obstante, pensamos que lo que se requiere es una forma apropiada de *gravedad cuántica*,[33] donde las reglas de la relatividad general, de la mecánica cuántica, y quizá también de algunos otros ingredientes físicos desconocidos, se reúnan de la forma adecuada.

28.8. La hipótesis de curvatura de Weyl

Pospondré mis principales consideraciones de la actividad actual en el campo de la gravedad cuántica hasta los capítulos 30-33. Por el momento, concentrémonos solo en tratar de entender cuáles parecen haber sido realmente las ligaduras geométricas del big bang. Después examinaremos la única propuesta seria que conozco, a saber, la de Hartle y Hawking, que intenta explicar esta geometría sobre la base de una teoría de gravedad cuántica seria.

Recordemos de §19.7 que los grados de libertad *gravitatorios* vienen descritos por el tensor de Weyl conforme C_{abcd}. Así pues, en el *espacio vacío* (donde una posible constante cosmológica Λ, pequeña en cualquier caso, será aquí ignorada por el momento) encontramos que la curvatura espaciotemporal es enteramente curvatura de Weyl (anulándose la curvatura de Ricci). La curvatura de Weyl es el tipo de curvatura cuyo efecto sobre la materia es de naturaleza distorsionante o de marea, en lugar de la reductora de volumen de las fuentes materiales. El efecto de la curvatura de Weyl se ilustró en la Fig. 17.9a (y el hecho de que esta imagen fuera originalmente una imagen espaciotemporal newtoniana no le resta validez). Esta imagen debe contrastarse con la Fig. 17.9b, en la que vemos el efecto de reducción de volumen de la materia, i.e., tensor de Ricci. No obstante, existen realmente algunas cuestiones complicadas cuando consideramos (como aquí) los efectos de la curvatura de Weyl y de Ricci sobre geodésicas *de género tiempo* (partículas masivas que se mueven libremente), puesto que el tensor de Ricci también puede tener a veces un efecto distorsionante, además de su efecto de reducción de volumen.

Estas cuestiones complicadas se eliminan si pensamos en la actuación de estos tipos de curvatura sobre geodésicas *nulas* (rayos de luz). Más aún, podemos reinstaurar entonces una constante cosmológica Λ, puesto que un término de la forma Λg_{ab} no concentra rayos de luz.[28.10] Podemos considerar que las geodésicas en la Fig. 17.9 son rayos de luz que pertenecen a cierto cono de luz (a la manera de la Fig. 17.16). De hecho, si pensamos que pertenecen al cono de luz pasado de cierto observador, los efectos distorsionantes pueden entenderse muy gráficamente en términos de lentes colocadas entre una fuente de luz y el observador. El efecto del tensor de Ricci,[34] debido a una distribución de materia, es el de actuar como una lente con focalización positiva, mientras que el efecto del tensor de Weyl, debido a un campo gravitatorio libre, es el de actuar como una lente puramente *astigmática*, con tanta focalización positiva en un plano como focalización negativa en un plano perpendicular (Fig. 28.15). Podemos hacernos una idea muy buena de los efectos (de orden más bajo) de estos dos tipos diferentes de curvatura si imaginamos que simplemente estamos mirando a través de un cuerpo esférico sólido y muy masivo que tiene el índice de refracción del vacío. (Quizá deberíamos pensar en «mirar» a través del Sol con neutrinos —considerados como partículas *sin masa*— que atraviesan el Sol directamente, ¡prestando atención solo a su campo gravitatorio!) Con razonable aproximación, podemos considerar que los rayos que atraviesan el Sol son afectados principalmente por la curvatura de Ricci, de modo que obtenemos una ampliación aparente (lente positiva) del campo estelar detrás del Sol. Por el contrario, más allá del borde del Sol obtenemos en efecto los efectos distorsionantes puramente astigmáticos de la curvatura de Weyl, de modo que una pequeña figura circular en el fondo celeste parecería ser elíptica para el observador. Véase la Fig. 28.16.[28.11] Esta es esencialmente la forma en que el campo gravitatorio del Sol distorsiona la pauta estelar de fondo, como se vio por primera vez en la expedición de Eddington en 1919 (véase §19.8).

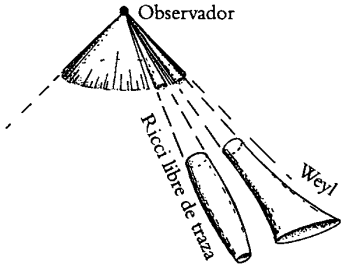


Fig. 28.15. El efecto focalizador del tensor de Ricci libre de traza (debido a una distribución de materia) es una lente positivamente focalizadora, mientras que el del tensor de Weyl (debido a un campo gravitatorio libre) es una lente puramente astigmática, con una focalización mucho más positiva en un plano y focalización negativa en el plano perpendicular.

[28.10] ¿Por qué no? *Sugerencia*: Explique la nota 28.34.

[28.11] Demuestre que las áreas se conservan para un desplazamiento infinitesimal hacia fuera, que varía de forma inversamente proporcional a la distancia.

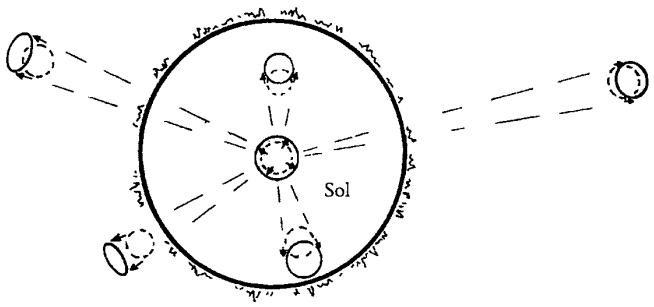


Fig. 28.16. Obtenemos una buena impresión de efectos (de orden más bajo) de los dos diferentes tipos de curvatura espaciotemporal «examinando» el campo estelar a través de un Sol transparente no refractivo (como si fueran neutrinos sin masa). Hasta una aproximación razonable, los rayos que atraviesan el Sol son focalizados solo por la curvatura de Ricci, lo que da como resultado una ampliación (como en una lente positiva), mientras que fuera del borde del Sol obtenemos en esencia distorsiones de Weyl puramente astigmáticas, de modo que una pequeña figura circular en el campo estelar parecería elíptica.

Consideremos ahora un universo que evoluciona de modo que una distribución inicialmente uniforme de materia (con ciertas fluctuaciones de densidad) se amontona poco a poco gravitatoriamente, de modo que finalmente partes del mismo colapsan en agujeros negros. La uniformidad inicial corresponde principalmente a una distribución de curvatura de Ricci, pero a medida que se acumula gravitatoriamente más material obtenemos cantidades crecientes de curvatura de Weyl, que básicamente habita las regiones de distorsión espaciotemporal que rodean a la materia acumulada. La curvatura de Weyl diverge finalmente a infinito cuando se alcanzan las singularidades de agujeros negros. Si consideramos que el material había sido vomitado originalmente del big bang de una forma casi completamente uniforme, entonces empezamos con una curvatura de Weyl que, para todos los efectos, es *nula*. De hecho, un rasgo característico de los modelos FLRW es que la curvatura de Weyl se anula por completo (siendo estos modelos, por consiguiente, conformemente planos; véase §19.7). Para un universo que empiece *muy próximo* a un FLRW, esperamos que la curvatura de Weyl sea extraordinariamente pequeña comparada con la curvatura de Ricci, que realmente diverge en el big bang.

Esta imagen sugiere con fuerza cuál es la diferencia geométrica entre la singularidad big bang inicial —de entropía extraordinariamente baja— y las singularidades de agujero negro genérico de muy alta entropía. La curvatura de Weyl se anula (o es al menos muy pequeña —por ejemplo, simplemente finita— comparada con la que podría haber sido) en la singularidad inicial y no tiene ligaduras, divergiendo sin duda incontroladamente al infinito, en las singularidades finales. Es esta caracterización geométrica la que parece distinguir la Fig. 27.20a de la Fig. 27.20d, por ejemplo, incluso si pudiera ser difícil reconocer la distinción en términos de diagramas conformes.

Esta observación debería considerarse junto con otra característica conjeturada de las singularidades espaciotemporales, conocida como *censura cósmica*. Esta es una afirmación (en la actualidad no demostrada) según la cual, a grandes rasgos, el resultado de un colapso gravitatorio imparable será un agujero negro, en lugar de algo peor conocido como una *singularidad desnuda*. Una singularidad desnuda sería una singularidad espaciotemporal resultante de un colapso gravitatorio que es visible para observadores externos, de modo que no está «vestida» por un horizonte de sucesos. Hay varias maneras técnicas ligeramente diferentes de especificar lo que se entiende por el término «singularidad desnuda», y aquí no pretendo entrar en las distinciones.[35] Para nuestros propósitos, bastaría con decir que una singularidad desnuda es de «género tiempo», en el sentido de que pueden entrar y salir señales de la singularidad, tal como se indica en la Fig. 28.17a. La censura cósmica prohibiría tales objetos (excepto tal vez en ciertas situaciones muy raras o «especiales» que no ocurrirían en un colapso gravitatorio realista).

La censura cósmica es básicamente una conjetura matemática —todavía no demostrada ni refutada— concerniente a las soluciones generales de la ecuación de Einstein. Si aceptamos esta conjetura, entonces las singularidades espaciotemporales físicas tienen que ser de «género espacio» (o quizá «nulas») pero nunca de «género tiempo». Hay dos tipos de singularidades de género espacio (o nulas), a saber, «iniciales» o «finales», dependiendo de si curvas de género tiempo pueden escapar de la singularidad hacia el futuro o entrar desde el pasado; véanse las Figs. 28.17b,c. La conjetura física que yo llamo *la hipótesis de curvatura de Weyl* afirma que (en un sentido apropiado) la curvatura de

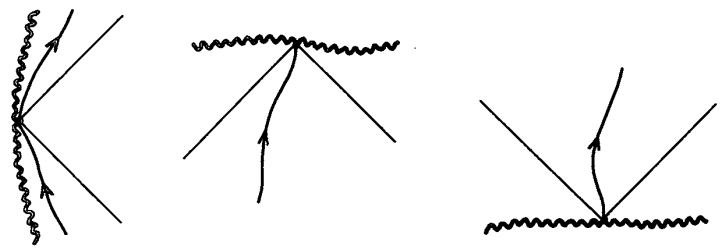


Fig. 28.17. (a) Las señales causales pueden entrar o salir de una «singularidad desnuda». Si estas están excluidas —por censura cósmica—, nos quedamos básicamente con (b) «singularidades futuras» (que resultan del colapso gravitatorio) de las que las señales causales pueden entrar pero no salir, y (c) «singularidades pasadas» (en el big bang, o quizá sucesos de creación más localizados) de las que las señales causales pueden salir pero no entrar. La hipótesis de curvatura de Weyl afirma que la curvatura de Weyl está limitada (apropiadamente) a ser cero (o a ser muy pequeña) en las singularidades iniciales (c) del universo físico real.

Weyl está obligada a ser nula (o al menos muy pequeña) en las singularidades *iniciales*, en el universo físico real. La creación de un universo de un modo que satisfaga la hipótesis de curvatura de Weyl representaría una limitación absolutamente enorme sobre la elección del Creador, en el proceso representado en la Fig. 27.21. Como resultado, habría una segunda ley de la termodinámica y tomaría la forma que observamos. Ahora hay buena evidencia de que una forma de «hipótesis de curvatura de Weyl» restringe adecuadamente el big bang de una forma tal que el modelo de universo resultante se parece estrechamente a un modelo FLRW en sus etapas tempranas.[36]

28.9. La propuesta de «ausencia de frontera» de Hartle-Hawking

Simplemente como afirmación, la hipótesis de curvatura de Weyl es quizá aún más parecida a «un acto divino» que a una teoría física. Lo que se requiere es una justificación teórica para una hipótesis de esta naturaleza. ¿A qué tipo de teoría tendríamos que apelar? El punto de

vista habitual con respecto a las singularidades espaciotemporales es que este es el dominio de la *gravedad cuántica.*

La dificultad aquí es que, a pesar de más de cincuenta años de esfuerzos decididos por unir la relatividad general y la mecánica cuántica, no hay nada todavía que se aproxime siquiera a un consenso respecto al enfoque correcto de este tema. Abordaré algunos de los esquemas actualmente más divulgados en los capítulos 31 y 32, pero incluso entre estos hay pocos intentos serios de entender la naturaleza particular del big bang. No obstante, hay una notable excepción, propuesta por James Hartle y Stephen Hawking en 1983, y por lo tanto es oportuno que haga algunos comentarios en relación con su idea principal.

Uno de los ingredientes de la propuesta de Hartle-Hawking es lo que normalmente se conoce como «euclideanización». La idea subyacente está estrechamente relacionada con la de una *rotación de Wick* aplicada a un espacio de Minkowski, por la que la coordenada temporal t se «rota» en $\tau = it$. La métrica espaciotemporal (espacial) $d\ell^2$ se convierte entonces en $d\ell^2 = d\tau^2 + dx^2 + dy^2 + dz^2$ (véase §18.1). La idea original[37] (Gian Carlo Wick) consistía en que se puede construir una teoría cuántica de campos relativista (especial) formulándola primero con el espaciotiempo de Minkowski reemplazado por este 4-espacio $\mathbb{E}^4$ euclídeo, donde ahora se considera la teoría invariante bajo el grupo euclídeo de simetrías de $\mathbb{E}^4$. Suponiendo que las cantidades obtenidas en la versión euclídea de la teoría son analíticas en las coordenadas, la rotación de Wick puede aplicarse, con τ rotado de forma continua de nuevo en t, de modo que ahora obtenemos una teoría correspondiente que es invariante bajo el grupo de Poincaré del 4-espacio de Minkowski. Este procedimiento tiene dos ventajas importantes. En primer lugar, las cantidades que probablemente van a ser divergentes en el espacio de Minkowski pueden resultar *con*vergentes en la versión euclídea de la teoría. (La razón se reduce a que el grupo de rotación euclídeo O(4) es compacto, y así de volumen finito, mientras que el grupo de Lorentz relativista O(3,1) no es compacto y de volumen infinito.) En particular, las integrales de camino (véase §26.6) tienen una oportunidad mucho mejor de una definición matemáticamente significativa en la versión euclídea que en la minkowskiana. La otra

ventaja es que los requisitos de frecuencia positiva (véanse §§9.3,5 y §24.3) pueden garantizarse si se aplica cuidadosamente la rotación de Wick de la forma correcta.

En el esquema de Hartle-Hawking es necesario utilizar una ingeniosa modificación, debida a Hawking, de la idea de Wick, en la que la «rotación» no se aplica a un espacio que es un *fondo* para los caminos en una integral de camino —que es la idea habitual—, sino a los espaciotiempos individuales que constituyen en sí mismos cada camino de la integral de camino.[38] Por consiguiente, a estos «espaciotiempos» se les permite tener métricas riemannianas definidas positivas, en lugar de las lorentzianas que se aplican a un espaciotiempo normal. (Estas métricas riemannianas se conocen a menudo confusamente como «euclídeas», ¡pese al uso estándar de dicho nombre para un espacio euclídeo plano $\mathbb{E}^n$!) Sin embargo, debería quedar claro que hay un «salto imaginativo» implicado en la versión de Hawking de la «euclideanización» que va mucho más allá de la idea original de Wick. Queda por ver si esto proporciona o no una ruta fructífera para la correcta unión de la relatividad general con la mecánica cuántica.[39]

La sorprendente propuesta de Hartle y Hawking era que este enfoque de integrales de camino de Hawking podría describir la teoría cuántica relevante para el propio big bang, y que en lugar de un espaciotiempo singular real habría una superposición cuántica (i.e., «integral de camino») de «espaciotiempos» que podrían tener métricas riemannianas en lugar de lorentzianas. Ellos bautizaron a su idea como la propuesta de «ausencia de frontera», porque en lugar de tener la frontera singular para el espaciotiempo clásico que representa el big bang, habría una familia superpuesta de espacios *no singulares*, dominada por espacios riemannianos que simplemente «cierran» el extremo inferior a la manera indicada en la Fig. 28.18, de modo que la frontera singular desaparece por completo. Momentos «después» del big bang tiene que haber una transición donde la dominancia pasa de la geometría riemanniana a la lorentziana. (Podemos imaginar que esto se logra vía una métrica *compleja* adecuada.) Incluso en la región lorentziana sigue habiendo una superposición de «espaciotiempos» (algunos de los cuales son riemannianos), pero se supone que fuera del big bang domina un espaciotiempo lorentziano, mientras que en la propia región del big

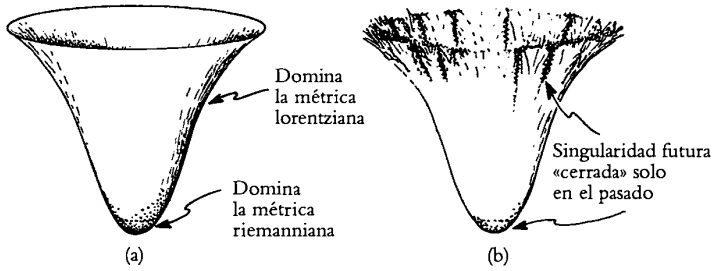


Fig. 28.18. La propuesta de «ausencia de frontera» de Hartle-Hawking sugiere que (a) el big bang puede tratarse según un procedimiento de gravedad cuántica donde las geometrías riemannianas (antes que las lorentzianas) dominan la integral de camino cerca de la singularidad clásica, y proporcionan maneras de cerrar el espaciotiempo de un modo no singular. (b) Con respecto a las singularidades de colapso, el «cierre» parece ser requerido solo en el «extremo lejano» del espaciotiempo, permitiendo con ello las singularidades genéricas de alta entropía que se espera que ocurran en el colapso gravitatorio a agujeros negros (o a big crunch).

bang son las métricas riemannianas con «ausencia de frontera» las que dominan. Este esquema no solo tiene una elegancia genuina, convirtiendo un problema aparentemente intratable en uno que parece ser vagamente manejable, sino que también parece dar algún apoyo directo a un «universo primitivo suave» que podría ser compatible con la hipótesis de curvatura de Weyl.

Hasta aquí, todo muy bien. Pero también tengo algunas dificultades con esta propuesta. En primer lugar, la idea misma de «euclideanización» es problemática en varios aspectos de relevancia para el contexto que encuentra aquí su uso. Incluso en un contexto de espacio plano, el cálculo exacto de una integral de camino es normalmente intratable, y hay que hacer muchas aproximaciones. Lo normal sería discriminar ciertos términos específicos que puede considerarse que dominan la integral y prescindir del resto. Cabría esperar que esto dé una aproximación razonable a la integral de camino «euclídea», pero recordemos que entonces hay que aplicar un proceso de prolongación analítica para obtener la respuesta física apropiada. Este es un procedimiento muy poco fiable, porque es probable que algo que se aproxima a una función holomorfa en una región sea incontrolable en otra región. Para apreciar la esencia de la dificultad, supongamos que tenemos

una función analítica real $f(x)$ que conocemos para valores reales de x, aunque solo aproximadamente, y deseamos inferir sus valores para x puramente imaginarios. Si añadimos a $f(x)$ una función de la forma $\varepsilon \cos(Ax)$, donde ε y A son reales, con ε muy pequeño y A grande, entonces $f(x)$ no resultará muy alterada a lo largo del eje real de x; pero el comportamiento a lo largo del eje imaginario cambiará por completo, ilustrando de este modo la extrema inestabilidad del proceso de prolongación analítica.[28.12] Por lo que puedo ver, el «truco de euclideanización» puede ser muy útil para producir modelos exactos de QFT, pero tengo dificultades con él cuando se utiliza junto con aproximaciones, como en este caso. (No obstante, no está claro para mí con qué fuerza depende la propuesta de Hartle-Hawking de este paso de prolongación analítica.)

También tengo algunas dificultades técnicas con la generalidad de la euclideanización. Por lo que puedo ver, es un truco astuto para producir QFT consistentes (y garantizar una condición de frecuencia positiva; §§9.3,5), pero sería muy optimista esperar que cualquier QFT concreta de interés pueda obtenerse por medio de ello. Las teorías obtenidas vía euclideanización tienen, en efecto, estructuras ocultas, que se originan a partir de su grupo de simetría asociado de la «signatura errónea»; véanse §13.8 y §18.2. No veo por qué una teoría «correcta» tiene que tener este carácter especial.

28.10. Parámetros cosmológicos: ¿estatus observacional?

Está también la cuestión del acuerdo con la observación. Al menos en su forma original, la propuesta de ausencia de frontera de Hartle-Hawking parecería apuntar a un universo cerrado (de hecho, $K > 0$), y durante varios años Hawking ha dado su apoyo a tales modelos. Pero frente a una evidencia cosmológica creciente, que ha aparecido para dar apoyo al caso hiperbólico ($K < 0$), Hawking, en colaboración con Turok, modificó posteriormente sus argumentos para permitir que la propuesta de «ausencia de frontera» acomode también el caso hiperbólico.[40] Existe un interesante paralelismo con las predicciones de la cosmología inflacionaria, pues durante muchos años se había dicho que una consecuencia decisiva de la misma era que el universo debía ser espacialmente plano ($K = 0$). Varios inflacionistas, ante estos datos cosmológicos cada vez más impresionantes, también modificaron después sus argumentos para permitir la posibilidad[41] de $K < 0$.

[28.12] Explique esto utilizando resultados de §5.3. (*Sugerencia*: ¿Qué es $e^{Aix} + e^{-Aix}$?)

¿Cuál es la posición observacional actual? Bien, las cosas han vuelto a cambiar de forma significativa con la extraordinaria evidencia (y procedente de más de una fuente) a favor de una *constante cosmológica Λ positiva*. La consecuencia de esto es que, después de todo, podríamos tener $K = 0$. Y si la evidencia observacional permite $K = 0$, no puede excluir una pequeña curvatura espacial positiva (la $K > 0$ favorita de Hawking) o una pequeña curvatura espacial negativa (mi favorita $K < 0$), ¡de modo que otra vez se han acabado las apuestas!

¿Qué relación guarda este descubrimiento de $\Lambda > 0$ con el valor de K? Debería mencionar primero las razones que había tras la creencia anterior en que la evidencia cosmológica favorecía un valor negativo de K. La cuestión esencial es el *contenido de masa-energía* total del universo, y que si este es demasiado pequeño entonces no será capaz de cerrar el universo con curvatura positiva, o (en los modelos de Friedmann) de tirar de él de nuevo después de su expansión inicial para producir una fase de colapso (véanse las Figs. 27.15a,b,c). Se sabía desde hace tiempo que la densidad de materia «bariónica» (véase §25.6) ordinaria visible en las galaxias es insuficiente para esto, pues solo es un tercio del valor *crítico* que representa la división entre valores positivos y negativos de K, siendo esta densidad crítica la que nos da $K = 0$. Normalmente se introduce la cantidad Ω_b para denotar esa fracción de la densidad de masa-energía crítica que se debe solo a la materia bariónica normal. Así, si $\Omega_b = 1$, la materia bariónica suministraría la densidad crítica, y cualquier otra masa-energía (positiva) importante llevaría a un universo $K > 0$. Sin embargo, como se ha mencionado antes, parece que en lugar de ello tenemos algo parecido a $\Omega_b = 0{,}03$, lo que había dado un poderoso indicio de que $K < 0$.

Sin embargo, esto no tiene en cuenta la fuerte evidencia de que hay mucha más materia en el universo que la materia bariónica que se observa directamente en las estrellas. Durante muchos años se había

hecho claro que la dinámica de las estrellas dentro de las galaxias no tiene sentido, según la teoría estándar,[42] a menos que haya mucho más material en el entorno de una galaxia que el que se ve directamente en forma de estrellas. Un comentario similar se aplica a la dinámica de las galaxias dentro de los cúmulos. En total, parece haber aproximadamente 10 veces más de materia que la que se percibe de forma bariónica ordinaria. Esta es la misteriosa *materia oscura* sobre cuya naturaleza real todavía no hay acuerdo entre los astrónomos, y que puede ser incluso de un material diferente de cualquiera conocido para los físicos de partículas —aunque hay mucha especulación sobre ello en el momento actual.[43] Puesto que la materia oscura parece aportar 10 veces más de masa-energía que la materia bariónica ordinaria, la densidad que proporciona la materia oscura, como una fracción Ω_d de la densidad crítica, viene aproximadamente dada por $\Omega_d = 0{,}3$ (y las incertidumbres son tales que, si queremos, podemos incluir en esta cifra la $\Omega_b = 0{,}03$ bariónica). Esto aún nos deja bastante lejos del valor crítico. Además, varios tipos de observaciones (entre ellos, los efectos de lente gravitatoria, que, recordemos de §19.8, ofrecen una medida directa de la presencia de masa) estaban empezando a mostrar, muy convincentemente, que no puede haber otras concentraciones importantes de masa en el universo. De modo que la conclusión $K < 0$ estaba pareciendo muy firme, y en consecuencia los inflacionistas y los seguidores de Hartle-Hawking empezaron a buscar modos de incorporar $K < 0$ en sus respectivas visiones del mundo.

Entonces llegó la bomba de la constante cosmológica. Recordemos de §19.7 que Einstein había considerado la introducción de Λ como «su mayor error» (quizá principalmente porque contribuyó a su fallo el predecir la expansión del universo). Aunque desde entonces siempre se había considerado como una posibilidad por parte de los cosmólogos, parece que muy pocos de ellos habrían esperado que Λ fuera no nula en nuestro universo real. Una cuestión adicional era el hecho de que cálculos de la «energía de vacío» por los teóricos de campos cuánticos (básicamente un efecto de renormalización como los de §26.9) habían dado una respuesta absurda según la cual debería haber realmente una constante cosmológica *efectiva* ¡que es mayor de la que se ve en un factor de aproximadamente 10^{120} (o al menos 10^{60}, si se hacen hipótesis

diferentes)! Esto llegó a conocerse como el «problema de la constante cosmológica». Podría haber sido plausible que una cancelación desconocida o principio general diera el valor 0 para esta energía del vacío, pero de ninguna manera se preveía encontrar un residuo minúsculo que pudiera tener relevancia para la cosmología en la época actual. (Habría que mencionar que esta energía del vacío debería ser proporcional a la métrica g_{ab}, por invariancia Lorentz local, de modo que se prevé la forma Λg_{ab}, para una Λ constante, que contribuye a la ecuación de Einstein precisamente como Einstein había sugerido en 1919. ¡El único problema es que el valor de Λ resulta completamente erróneo!)

De todas formas, en 1998 dos equipos que observaban supernovas muy lejanas (véase §27.8) —uno dirigido por Saul Perlmutter en California, y el otro, dirigido por Brian Schmidt en Australia y Robert Kirschner en el este de Estados Unidos— llegaron a la extraordinaria conclusión de que la expansión del universo había empezado a acelerarse, lo que es compatible con el giro hacia arriba de la gráfica de la Fig. 27.15d, ¡que es el sello de una constante cosmológica positiva! ¿Qué valor tiene esta Λ que parece observarse? Hay todavía algunas incertidumbres sobre ello (y algunos teóricos han argumentado que el caso de una Λ positiva todavía no se ha establecido de forma convincente[44]), pero la notable conclusión es que la densidad de masa-energía efectiva Ω_Λ que Λ proporciona, como una fracción de la densidad crítica, está dada aproximadamente por $\Omega_\Lambda = 0{,}7$, así que parece que tenemos, para la densidad efectiva total, como fracción de la crítica

$$\Omega \approx \Omega_d + \Omega_\Lambda \approx 0{,}3 + 0{,}7 = 1.$$

En otras palabras, las observaciones parecen ser ahora compatibles con $K = 0$.

Los inflacionistas (al menos los que tenían confianza para no cambiar sus fundamentos) están, por supuesto, exultantes, y ciertamente puede contarse como un éxito predictivo de su teoría el hecho de que, frente a una evidencia aparentemente poderosa a favor de lo contrario, la predicción $K = 0$ parece haber salido ganadora. Sin embargo, las incertidumbres son todavía demasiado grandes para que esta conclusión sea convincente, y es significativo que otro tipo de observaciones recientes tengan también una poderosa relación con esta cuestión. Como

se ha mencionado en §28.5, ha habido varias medidas de las variaciones detalladas de temperatura en el fondo de microondas, empezando con el satélite COBE, lanzado en 1989, y el examen más reciente (en el momento de escribir esta obra) es el realizado por la sonda espacial WMAP.

Estas variaciones de temperatura se analizan normalmente descomponiendo la pauta sobre el cielo en armónicos esféricos, de acuerdo con los procedimientos examinados en §22.11. Recordemos que los diferentes armónicos esféricos están etiquetados por un entero positivo ℓ y un entero m en el rango que va de $-\ell$ a ℓ. (En la mecánica cuántica, a ℓ se le llama normalmente j, y ambos j y m podrían ser semiimpares.) La cantidad m es menos importante, porque depende de una dirección arbitrariamente escogida en el cielo, de modo que la intensidad general para cada valor de ℓ se considera la cantidad de mayor interés. En la Fig. 28.19 he mostrado el resultado de este análisis. Nótese que hay un indicio de que después de que la curva alcanza su máximo, a aproximadamente $\ell = 200$, empieza a oscilar. Estos máximos locales se conocen como «picos acústicos», puesto que reflejan una clara predicción teórica de que en una etapa primitiva del universo empezarían a producirse concentraciones locales de materia que se acumulan y luego rebotan o se deshacen (que es lo que cabría esperar para la materia oscura), dando como resultado un tipo de oscilación sónica. La escala típica en la que ocurriría esta oscilación estaría gobernada por una «escala de horizonte» en el desacoplamiento (véase la Fig. 28.5a, e imaginemos que los puntos u y v se mueven sobre la superficie de desacoplamiento hasta que sus pasados se tocan; esta es la «escala de horizonte»).[45] Es en esta escala donde ocurre el pico principal.

Sin embargo, se plantea la cuestión de qué separación angular en el cielo corresponde a la separación local en el universo en el instante de la recombinación, y es aquí donde la curvatura espacial del universo desempeña un papel importante, pues los picos acústicos se desplazan en una dirección u otra, en ℓ, dependiendo del valor de K (hacia valores menores para K positiva y mayores para K negativa). No obstante, la cuestión no es directamente extrapolable porque el ritmo de expansión del universo también tiene un papel en esto, de modo que se ne-

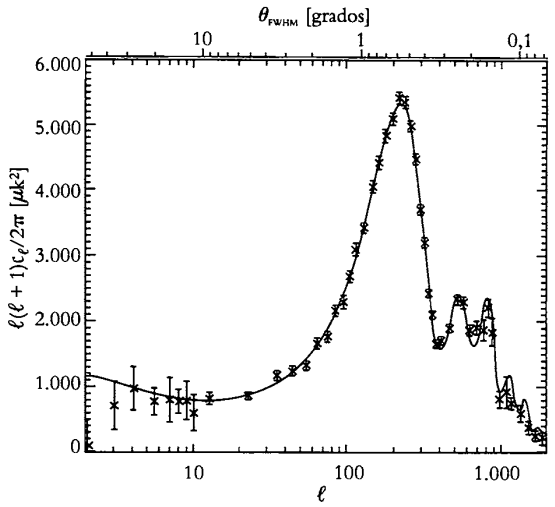


Fig. 28.19. Los «picos acústicos» predichos en el análisis armónico del fondo cósmico de microondas (línea sólida), y los datos observados (cruces, con barras de error). Adviértase la discrepancia muy significativa en el cuadripolo ($\ell = 2$), casi oculto (¿accidentalmente?) por el eje vertical.

cesitan cálculos detallados. El resultado es que este tipo de análisis del fondo cósmico de microondas es, en un sentido general, compatible con $K = 0$, pero sigue habiendo espacio para una valor de K positivo o negativo que sería de importancia observacional.

Los resultados para altos valores de ℓ parecen así compatibles con las expectativas de la inflación (y hay también una invariancia de escala en las fluctuaciones de temperatura observadas que también había sido una predicción de algunos modelos inflacionarios). Pero ¿qué pasa con los bajos valores de ℓ? El valor $\ell = 0$ no es muy ilustrativo, puesto que solo describe la intensidad global. ¿Qué pasa con $\ell = 1$ (el momento dipolar)? Este no nos habla del universo distante, porque el movimiento de la Tierra a través del fondo de microondas lleva a un desplazamiento Doppler asimétrico (véase el ejercicio [27.10]) que da lugar a una distribución de temperatura $\ell = 1$ con una temperatura ligeramente más alta percibida en la dirección del movimiento y una temperatura ligeramente más baja en la dirección contraria. El primer

valor de ℓ con importancia cosmológica es $\ell = 2$ (el «momento cuadripolar»). De hecho, se ve una discrepancia con las predicciones con invariancia de escala de la inflación en este punto, que es confirmada por los armónicos siguientes. La discrepancia no es pequeña, y parece razonablemente clara. La ruptura de invariancia de escala implicada puede interpretarse como algo en las escalas más altas que difiere de la geometría plana $K = 0$, lo que posiblemente indica que $K > 0$ o $K < 0$, puesto que el «radio de curvatura» proporciona tal escala.

Estas consideraciones nos dejan intrigados pero algo inseguros. No obstante, habría que tener en cuenta que la gráfica de la Fig. 28.19 está haciendo uso solo de una minúscula cantidad de la información que está contenida en el mapa de temperatura de WMAP. Para cada valor de ℓ, hay $2\ell + 1$ valores diferentes de m, y es un parámetro real para cada uno de estos. La mayor parte de esta información está siendo ignorada en este análisis, y debe de haber un gran número de datos ocultos que nos están diciendo algo posiblemente de gran importancia sobre el universo primitivo.

Aquí menciono solamente una manera alternativa de analizar tales datos, que se debe básicamente a Vahe Gurzadyan *et al.* (1992, 1994, 1997, 2002, 2003, 2004), y que parece tener consecuencias extraordinarias. En esta aproximación no se utiliza un análisis armónico; en su lugar, se examinan las distorsiones en la forma de regiones distantes de cada temperatura concreta, debidas a la curvatura espacial interpuesta. Si imaginamos que la forma no distorsionada de una de estas regiones es realmente circular, entonces los efectos de curvatura podrían hacer que se vuelva elíptica (recordemos la Fig. 28.15). Por supuesto, en la práctica no conocemos la forma de la región que estamos mirando, pero puede haber efectos estadísticos que hacen que las regiones de temperatura específica se hagan más (o menos) estiradas y delgadas de lo que serían sin ello. Esto es un ejercicio delicado de análisis estadístico, pero la conclusión a la que llegan Gurzadyan y sus colegas es que hay una cantidad importante de elipticidad en los mapas de microondas (originalmente COBE, luego BOOMERANG y posteriormente WMAP). ¿Qué significa esto? El análisis teórico de esta situación nos dice que solo con $K < 0$ podemos esperar este grado de elipticidad, como resultado de un «mezclado geodésico». Estos resultados son nuevos, de modo que hay que esperar para ver si surgen objeciones importantes contra tan extraordinaria conclusión.

Este análisis proporciona también evidencia independiente a favor de una constante cosmológica positiva de aproximadamente el tamaño que está implicado por los datos de supernovas. Así pues, se concluye que la curvatura negativa es pequeña, en el sentido de que $\Omega_d + \Omega_\Lambda$ no puede diferir mucho de la unidad, teniendo quizá un valor de 0,9. Esto resalta un enigma que ha preocupado a muchos cosmólogos. Las cantidades Ω_b, Ω_d y Ω_Λ no son constantes en el tiempo. En las primeras etapas del universo, Ω_b y Ω_d habrían sido mucho mayores y Ω_Λ mucho menor. En las últimas etapas del universo, Ω_b y Ω_d se habrían hecho despreciables, y Ω_Λ dominaría la densidad de masa-energía efectiva. La aparente coincidencia de que los tamaños de Ω_Λ y Ω_d sean del mismo orden parece una coincidencia enigmática.

Curiosamente, el término «constante cosmológica» parece haber pasado de moda casi tan pronto como Λ fue descubierta observacionalmente, pese a ser la terminología estándar desde su introducción teórica por Einstein en 1917. En su lugar, Λ es llamada «energía oscura», o «energía de vacío», o a veces «quintaesencia», quizá porque el frío término «constante cosmológica» no transmite el suficiente aire de misterio, o quizá, un poco más racionalmente, porque la presencia de la palabra «constante» implica más bien que Λ no puede cambiar con el tiempo. Muchos cosmólogos parecen sentirse más a gusto con una Λ variable, posiblemente considerando que la «Λ» actual representa el inicio de una «nueva fase inflacionaria», de la que ellos señalan la similitud con la supuesta muy temprana fase inflacionaria del universo. Recordemos de §28.4 que esta se considera dominada por un «falso vacío» en el que hay una constante cosmológica efectiva que es tan alta que domina por completo toda la (ya enormemente densa) materia ordinaria. Si se permitía que el universo tuviera entonces una «Λ» efectiva tan extraordinariamente diferente del valor que encontramos hoy —dice el argumento—, entonces por supuesto que deberíamos permitir una «Λ variable», y el término «constante cosmológica» es, en consecuencia, inapropiado.

Sin embargo, esta idea, por atractiva que pueda parecer a algunos, tiene sus dificultades con las matemáticas, pues el término «constante

cosmológica» fue introducido por buenas razones. La constancia de Λ es una consecuencia directa de la ecuación de conservación de la energía $\nabla^a T_{ab} = 0$, de §§19.5-7, puesto que sumar a T_{ab} un múltiplo de g_{ab} solo puede dejar inalterada dicha ecuación de conservación si ese múltiplo es una constante.[28.13] Así, cualquier no constancia en «Λ» tendría que ir acompañada de una no conservación compensatoria de la masa-energía de la materia. Ciertamente es mucho más cómodo teóricamente tener Λ constante, como es, de hecho, compatible con la observación.

¿Dónde nos deja esto? Ciertamente en un estado interesante. No veo que la cosmología inflacionaria sea «confirmada» por estas observaciones, e incluso si lo fuera esto no resolvería el problema cosmológico que, en mi opinión, eclipsa a todos los demás, a saber, el big bang extraordinariamente «especial» —al menos en el grado de una parte en $10^{10^{123}}$— que subyace en la segunda ley. Algunos cosmólogos considerarían el «ajuste fino» que está implícito en esto (véase la Fig. 27.21) como inaceptable, y tratan de «explicarlo» en términos de inflación o del principio antrópico (§§28.4,6), aunque, como hemos visto, tales procedimientos fallan por mucho.

En realidad, tengo un problema fundamental con cualquier propuesta (por ejemplo, la inflación o la propuesta de Hartle-Hawking) que intente abordar el problema de las singularidades espaciotemporales dentro de una física aparentemente *simétrica con respecto al tiempo*. No hay asimetría temporal en la física inflacionaria y, por lo que puedo ver, tampoco la hay en la propuesta de Hartle-Hawking, de modo que esta propuesta debería aplicarse también a las singularidades finales del colapso (en agujeros negros, o en el big crunch si existe uno) tanto como en el big bang. Hawking (1982) ha argumentado que es *posible*, aunque de una forma decididamente exótica, que el espacio en la vecindad de una singularidad final sea «cerrado sin frontera», ¡remontando el camino del universo hasta el big bang, y aplicando solo allí la «euclideanización» (Fig. 28.18)! Su argumento es que la propuesta de ausencia de frontera simplemente afirma que hay *alguna* forma de cerrar las cosas sin frontera, y *definimos* el «comienzo» (que determina el sentido temporal del universo) como el extremo en el que ocurre el cierre. Debo decir que tengo grandes dificultades con este argumento, y, de hecho, con *cualquier* argumento donde no hay asimetría temporal explícita en las propias leyes físicas. (En el argumento «exótico» de Hawking, por ejemplo, parecería que sigue habiendo una «frontera» en la singularidad final del colapso, incluso si ha habido un cierre suave y libre de fronteras solo «en el otro extremo» del espaciotiempo. Me parece que solo se ha atendido a una mitad de este problema de eliminación de frontera.)

¿Tenemos que abordar entonces la posibilidad de una física básica realmente *asimétrica respecto al tiempo* tal como estoy afirmando? En el capítulo 30 abordaré esta cuestión de frente; veremos que está relacionada con algo fundamentalmente enigmático que hemos dejado sin resolver en los capítulos referentes a la mecánica cuántica. Por consiguiente, en el próximo capítulo tendremos que volver a este importante interrogante mecanocuántico. En el capítulo 30 presentaré mis propias ideas respecto a la ruta correcta hacia su resolución, y con ello también a la resolución eventual del problema de la asimetría temporal en la singularidad. Pese a todo, una vez más debo hacer al lector mi advertencia reglamentaria: muchos físicos se sentirán bastante incómodos con mi punto de vista.

[28.13] ¿Por qué?

Notas

Sección 28.1

28.1. Véase, por ejemplo, Weinberg (1992), p. 195, donde también utiliza el ejemplo del ferromagnetismo, como parece ser casi universal en las exposiciones divulgativas por parte de los expertos. Pese a todo, debemos tener en cuenta que esta es una idealización considerable para un bloque de hierro *real*, en el que los efectos detallados de las fuerzas pueden ser muy complicados. Aunque para una región suficientemente pequeña dentro del hierro esta tendencia a la magnetización puede ser una buena aproximación, en la práctica tales regiones magnetizadas tienden a orientarse aleatoriamente, de modo que el hierro en conjunto no suele proporcionar un imán efectivo. Además, para que el hierro se vuelva magnetizado de forma significativa, el enfriamiento al cruzar el

punto de Curie debe ser extremadamente lento, y la situación ideal no es fácil de conseguir. Para la discusión teórica presente es oportuno que ignoremos tales complicaciones y aceptemos la idealización que se está describiendo.

28.2. El efecto túnel mecanocuántico ocurre cuando un sistema sufre espontáneamente una transición desde un estado a otro de energía más baja (con la emisión del exceso de energía), donde hay una barrera de energía que impide que esto tenga lugar clásicamente.

28.3. La simetría de reflexión ha sido excluida en este ejemplo debido a la «S» en «SO(2)».

28.4. Parece que este «grupo adecuado» es $SU(3) \times SU(2) \times U(1)/Z_6$.

Sección 28.2

28.5. Véase la nota 28.1.

28.6. Véanse Vilenkin (2000), Gangui (2003) y Sakellariadou (2002).

28.7. Para una teoría de este tipo de apariencia prometedora, véase Silk y Rees (1998). Véase Haehnelt (2003) para una revisión y referencias adicionales.

28.8. Véase Chan y Tsou (1993).

28.9. La Colaboración MACRO ha puesto límites estrictos a la frecuencia de estas partículas. Véase MACRO (2002).

Sección 28.3

28.10. Esta conexión se tomaría inicialmente como una conexión gauge ∇ en el fibrado más pequeño $\mathcal{B}_{\mathcal{L}}$, sobre $\mathcal{M}$, cuyas fibras son los espacios U(2)-simétricos $\mathcal{L}$ de leptones en cada punto. Pero precisamente de la misma forma que, como en §14.3, en el cálculo tensorial ordinario el conocimiento de la actuación de ∇ sobre vectores fija por completo su actuación sobre tensores generales, el conocimiento de la actuación de ∇ sobre $\mathcal{B}_{\mathcal{L}}$ determina por completo su acción sobre los «tensores» definidos a partir de $\mathcal{L}$. Podemos tomar $\mathcal{G}$ como $\mathcal{L}^* \otimes \mathcal{L}$ (un «índice» abajo, uno arriba).

28.11. El «desplazamiento hacia el rojo» z se define de modo que $1 + z$ mide el factor en el que se incrementa la longitud de onda. Liddle (1999) es el texto más accesible; Dodelson (2003) es un tratamiento más avanzado.

28.12. Podría contemplarse un posible papel para una conexión *cuanlazamiento* (véase §23.10) entre q y r. Esto es por supuesto digno de considerar, pero va más allá de las ideas actuales de «ruptura espontánea de simetría». Mi opinión sobre estas cuestiones ha sido influida por conversaciones con George Sparling y Bikash Sinha.

28.13. Véase Lewellyn Smith (1973).

Sección 28.4

28.14. Véase Schrödinger (1956).

Sección 28.5

28.15. Véase Guth (1997). Dodelson (2003) y Liddle y Lyth (2000) son fuentes técnicas. Para un examen crítico y detallado, es recomendable Börner (2003).

28.16. Véanse Barbour (1989, 2001a, 2001b), Barbour, Foster y O'Murchadha (2002), Sciama (1959) y Smolin (2002). Un ejemplo de una aproximación completamente «machiana» es el de *redes de espín*, que se describirá en §32.6.

28.17. Véase Ozsvath y Schücking (1962, 1969).

28.18. Sin embargo, hay perspectivas más recientes sobre estas cuestiones que se argumentan en apoyo de que la teoría de Einstein es «machiana» después de todo. Véanse Barbour (2004), Barbour *et al.* (2002) y Raine (1975).

28.19. Estos desiderata estéticos están específicamente argumentados en la exposición divulgadora de Mario Livio. Véase Livio (2000).

28.20. Hubo una idea precursora de esta en lo que se conocía como «cosmología caótica»; fue propuesta independientemente en los años sesenta por Charles W. Misner y Yakov B. Zeldovich, y en ella se concebía un estado inicial aleatorio —pese al conflicto en apariencia fundamental con la segunda ley— invocando procesos térmicos en un intento de suavizar el universo. Véase Misner (1969).

28.21. La mejor propuesta para una probable estructura caótica en esta singularidad genérica procede del trabajo de 1970 de Belinskii *et al.* (1970).

28.22. Véase la nota 27.21, que da las referencias relevantes.

Sección 28.6

28.23. Creo que oí hablar por primera vez de esta idea antrópica «débil» en las charlas radiofónicas de Fred Hoyle, que se emitían en la BBC en la década de 1950. Llegué a familiarizarme con la forma más fuerte del principio antrópico, que aborda la cuestión del papel «antrópico» de las constantes físicas básicas, en una de las conferencias de Fred Hoyle en

Cambridge sobre «La religión como una ciencia», que se refería a la fabricación de elementos pesados en las estrellas que requiere un específico nivel energético nuclear en el carbono, que se va a describir en breve.

28.24. Véanse Dicke (1961) y Carter (1974).

28.25. De forma aproximada, la raíz cúbica de la edad del universo en unidades de Planck es notablemente próxima a la raíz cuadrada de la razón entre la atracción electrostática y la gravitatoria entre un protón y un electrón.

28.26. Véanse Dirac (1938), Buckley y Peat (1996) y Guenther *et al.* (1998). Una reciente idea de «constante variable» se da en una entretenida exposición de Magueijo (2003).

28.27. Mi uso del término «principio antrópico fuerte» sigue aquí a Carter (1974). Barrow y Tipler (1988) separan esto en varias categorías diferentes.

28.28. Para más información concerniente a esta historia, véanse Barrow y Tipler (1988) y Smolin (1997), p. 111. En cuanto a las implicaciones profundas para los contenidos de nuestro universo, véanse Hoyle *et al.* (1956) y Burbidge *et al.* (1957).

28.29. Véase Hawking y Penrose (1970).

28.30. Véase Smolin (1997).

28.31. En mi ensayo para el premio Adams de 1966 (véase Penrose, 1966, 1968) propuse una idea semejante (aunque sin reajuste de las constantes físicas), ¡aunque no muy en serio! Quizá otros lo hayan hecho también con anterioridad.

Sección 28.7

28.32. Véase Penrose (1989).

28.33. Abhay Ashtekar me ha resaltado un punto de vista alternativo según el cual podría haber algo más, diferente de la «gravedad cuántica», que determina la naturaleza extraordinariamente especial del big bang. Quizá sea así, pero no puedo dejar de sorprenderme por el hecho de que fuera la gravedad lo que era especial en el big bang, y aparentemente solo la gravedad.

Sección 28.8

28.34. En realidad, solo la parte libre de traza del tensor de Ricci $R_{ab} - \frac{1}{4} R g_{ab}$ es aquí relevante, y la constante cosmológica no desempeña ningún papel.

28.35. Véase Penrose (1969a); y véase Penrose (1988) para un examen general de la censura cósmica.

28.36. Véanse Newman (1993), Claudel y Newman (1998), Tod y Anguige (1999a, 1999b) y Anguige (1999). Una versión particularmente atractiva de la hipótesis de curvatura de Weyl es la presentada por K. P. Tod, que simplemente afirma que en cualquier singularidad inicial hay una geometría conforme regular con frontera.

Sección 28.9

28.37. Véase Wick (1956) para la primera utilización de esta técnica, utilizada en Zinn-Justin (1996) con gran y frecuente efecto.

28.38. Véase Hartle y Hawking (1983).

28.39. Un reciente trabajo de Renate Loll y sus colaboradores sugiere que puede haber diferencias profundas entre el uso de métricas riemannianas en la integral de camino, como sucede con la propuesta de Hawking, y las más directamente apropiadas métricas lorentzianas. Véase Ambjorn *et al.* (1999).

Sección 28.10

28.40. Véase Hawking y Turok (1998).

28.41. Véanse Bucher *et al.* (1995) y Linde (1995).

28.42. Mordehai Milgrom (1994) ha hecho la intrigante sugerencia de que no hay materia oscura, sino que la dinámica gravitatoria newtoniana necesita ser alterada de una forma diferente de la de Einstein, tal que para aceleraciones muy bajas el efecto de la gravedad se incrementa de cierta manera específica. Aunque esta idea parece encajar muy bien los hechos, todavía no hay ninguna teoría coherente de esto que tenga sentido teórico general. En mi opinión, estas ideas poco convencionales no deberían descartarse sin más, y valdría la pena el esfuerzo de ver si este esquema puede formar parte de un punto de vista consistente más amplio. (¡Yo no he sido capaz de ver cómo hacerlo!)

28.43. Véase Krauss (2001) para una discusión accesible de la materia oscura (y también de la «energía oscura», i.e., una posible Λ variable).

28.44. Véase Blanchard *et al.* (2003). Para una interpretación más en la línea de la «corriente principal», véanse Perlmutter *et al.* (1998) y Bahcall *et al.* (1999).

28.45. Dodelson (2003) explica cómo hacer esto, así como los análisis relacionados de los datos de CMB.

29

La paradoja de la medida

29.1. Las ontologías convencionales de la mecánica cuántica

No hay duda de que la mecánica cuántica ha sido uno de los logros supremos del siglo XX. Explica muchísimos fenómenos que habían sido profundamente enigmáticos en el siglo XIX, tales como la existencia de las líneas espectrales, la estabilidad de los átomos, la naturaleza de los enlaces químicos, las durezas y colores de los materiales, el ferromagnetismo, las transiciones de fase sólido/líquido/gas y los colores de los cuerpos calientes en equilibrio con sus entornos calientes (radiación de cuerpo negro). Incluso algunas cuestiones enigmáticas de la biología, como la extraordinaria fiabilidad de la herencia, son ahora vistas como resultado de principios mecanocuánticos. Estos fenómenos —así como muchos otros que se han conocido en el siglo XX como los cristales líquidos, la superconductividad y la superfluidez, el comportamiento de los láseres, los condensados de Bose-Einstein, la curiosa no localidad de los efectos EPR y de la teleportación cuántica— son ahora bien entendidos sobre la base del formalismo matemático de la mecánica cuántica. De hecho, este formalismo ha significado una revolución en nuestra imagen del mundo físico real que es mucho mayor incluso que la que supuso la del espaciotiempo curvo de la relatividad general de Einstein.

¿O no es así? Una opinión común entre muchos físicos de hoy es que la mecánica cuántica ¡*no* nos ofrece ninguna imagen de la «realidad»! De acuerdo con esta opinión, el formalismo de la mecánica cuántica es tan solo eso: un formalismo matemático. Este formalismo,

como argumentarán muchos físicos cuánticos, no nos dice nada sobre la *realidad cuántica* del mundo, sino que meramente nos permite calcular probabilidades para realidades alternativas que podrían ocurrir. Esa ontología de los físicos cuánticos —en la medida en que estos estuviesen preocupados por cuestiones de «ontología»— consistiría en la idea (a) que sencillamente no hay ninguna realidad expresada en el formalismo cuántico. En el otro extremo están muchos físicos cuánticos que adoptan el punto de vista (en apariencia) diametralmente opuesto (b) que el estado cuántico que evoluciona unitariamente describe por completo la realidad, con la alarmante implicación de que prácticamente todas las alternativas cuánticas deben seguir siempre coexistiendo (en superposición). Como se ha mencionado en §21.8, la dificultad básica a la que se enfrentan los físicos cuánticos, y que empuja a muchos de ellos a tales opiniones, es el conflicto entre los dos procesos cuánticos **U** y **R**, donde (§22.1) **U** es el proceso determinista de evolución unitaria (descrita por la ecuación de Schrödinger) y **R** es la reducción de estado cuántico que tiene lugar cuando se realiza una «medida». El proceso **U**, cuando se encontró, era algo de un tipo familiar para los físicos: la evolución temporal precisa de una cantidad matemática definida, a saber, el vector de estado $|\psi\rangle$, controlado de forma determinista por una ecuación diferencial (en derivadas parciales), y la evolución temporal de la ecuación de Schrödinger no es diferente de la de las ecuaciones de Maxwell clásicas (véanse §21.3 y el ejercicio [19.2]). Por el contrario, el proceso **R** era algo completamente nuevo para ellos: un salto aleatorio discontinuo de esta misma $|\psi\rangle$, donde solo están determinadas las probabilidades de los diferentes resultados. Si la física del mundo observado hubiera estado descrita simplemente mediante una cantidad $|\psi\rangle$, que actúa solo de acuerdo con **U**, entonces los físicos no hubieran tenido ninguna dificultad seria para aceptar que **U** proporciona un proceso de evolución «físicamente real» para una $|\psi\rangle$ «físicamente real». Pero no es así como se comporta el mundo observado. En lugar de ello, parece que percibimos una curiosa combinación de **U** con la injerencia ocasional del muy diferente proceso **R**. (Recordemos la Fig. 22.1.) Esto hacía mucho más difícil que los físicos creyeran que $|\psi\rangle$ podía ser realmente una descripción de la realidad física después de todo. La enigmática cuestión de cómo puede ocurrir **R**,

cuando se supone que el estado está evolucionando de acuerdo con la evolución **U**, es el problema de la medida —o, como prefiero llamarlo, la *paradoja de la medida*— de la mecánica cuántica (que se ha examinado brevemente en §23.6 y mencionado en §21.8 y §22.1).

El punto de vista (a) es básicamente la ontología de la *interpretación de Copenhague* tal como la expresó específicamente Niels Bohr, que consideraba que $|\psi\rangle$ no representa una realidad en el nivel cuántico, sino que debe tomarse meramente como una descripción del «conocimiento» que tiene el experimentador de un sistema cuántico. El «salto», según **R**, tendría que entenderse entonces simplemente como el hecho de que el experimentador adquiere más conocimiento del sistema, de modo que lo que salta es el *conocimiento*, y no la física del sistema. Según (a), uno no debería pedir que se asigne ninguna «realidad» a los fenómenos de nivel cuántico, pues la única realidad reconocible es la del mundo clásico dentro del cual encuentra su hogar el aparato del experimentador. Como una variante de (a), se podría adoptar la idea de que donde interviene este «mundo clásico» no es en el nivel de alguna pieza de «maquinaria macroscópica» que constituye el aparato de medida del observador, sino en el nivel de la propia *consciencia* del observador. Pronto discutiré estas alternativas con más detalle.

Los defensores de la alternativa (b), por el contrario, *sí* consideran que $|\psi\rangle$ representa la realidad, pero niegan que ocurra **R**. Argumentarán que cuando tiene lugar una medida, todos los resultados alternativos *coexisten* realmente en una gran superposición lineal cuántica de universos alternativos. Esta gran superposición está descrita por una función de onda $|\psi\rangle$ para todo el universo. A veces se le llama el «multiverso»,[1] pero creo que un nombre más apropiado es el *omnium*.[2] Pues aunque este punto de vista se expresa coloquialmente como una creencia en la coexistencia paralela de diferentes mundos alternativos, esto es equívoco. Los mundos alternativos no «existen» realmente por separado en esta visión; solo la enorme *superposición concreta* expresada por $|\psi\rangle$ se considera real.

¿Por qué, según (b), el omnium no es *percibido* como «realidad» por un experimentador? La idea es que los diferentes estados mentales del experimentador también coexisten en la superposición cuántica, entrelazados con los diferentes resultados posibles de la medida que se

está realizando. Por consiguiente, la idea es que existe efectivamente un «mundo diferente» para cada posible resultado diferente de la medida, y existe una «copia» independiente del experimentador en cada uno de estos mundos diferentes, todos los cuales coexisten en una superposición cuántica. Cada copia del experimentador experimenta un resultado diferente del experimento, pero puesto que estas «copias» habitan en mundos diferentes, no hay comunicación entre ellos y cada uno piensa que solo se ha dado un resultado. Los defensores de (b) suelen mantener que es el requisito de que un experimentador tenga un «estado consciente» coherente el que fuerza la impresión de que existe solo «un mundo» en el que *parece* tener lugar **R**. Este punto de vista fue propuesto explícitamente por primera vez por Hugh Everett III en 1957[3] (aunque sospecho que muchos otros habían defendido antes en privado una idea semejante, aunque no siempre con convicción —como hice yo mismo a mediados de la década de 1950—, ¡sin atreverse a exponerla!).

Pese a sus naturalezas diametralmente opuestas, los puntos de vista (a) y (b) tienen algunos puntos significativos en común con respecto a la relación que se establece entre $|\psi\rangle$ y nuestra «realidad» observada —por lo que entiendo el mundo aparentemente real que todos nosotros experimentamos a escala macroscópica—. Se considera que en dicho mundo solo ocurre un resultado de un experimento, y la tarea de la física es precisamente explicar o modelar eso que solemos llamar «realidad». Ni según (a) ni según (b) el vector de estado $|\psi\rangle$ describe dicha realidad. Y en cada caso debemos introducir las percepciones de un experimentador humano para dar sentido a la relación entre el formalismo y este mundo real observado. En el caso (a) es el propio vector de estado $|\psi\rangle$ el que se considera un artificio de las percepciones de dicho experimentador humano, mientras que en el caso (b) es la «realidad ordinaria» la que está delineada de algún modo en términos de las percepciones del experimentador, y el vector de estado $|\psi\rangle$ representa ahora cierto tipo de realidad primordial más profunda (el omnium) que no es directamente percibida. En ambos casos se considera que el «salto» **R** no es físicamente real, pues, en cierto sentido, ¡está «todo en la mente»!

A su debido tiempo, expondré mis propias dificultades con ambas posiciones (a) y (b), pero, antes de hacerlo, debería mencionar una po-

sibilidad adicional para interpretar la mecánica cuántica convencional. Por lo que puedo apreciar, este es el dominante entre los puntos de vista mecanocuánticos —el de la *decoherencia por el entorno* (c)—, aunque es quizá una posición más pragmática que ontológica. La idea de (c) es que en cualquier proceso de medida el sistema cuántico en consideración no puede tomarse aislado de su entorno. Así pues, cuando se realiza una medida cada resultado diferente no constituye un estado cuántico por sí mismo, sino que debe considerarse como parte de un estado entrelazado (§23.3), en el que cada resultado alternativo está entrelazado con un estado diferente del entorno. Ahora bien, el entorno consistirá en muchísimas partículas, en movimiento aleatorio efectivo, y los detalles completos de sus posiciones y movimientos deben considerarse totalmente inobservables en la práctica.[4] Existe un procedimiento matemático bien definido para manejar una situación de este tipo en la que hay una carencia de conocimiento: uno «suma sobre» los estados del entorno desconocidos y obtiene un objeto matemático, conocido como una *matriz densidad*, para describir el estado físico en consideración. Las matrices densidad son importantes para la discusión general del problema de la medida en mecánica cuántica (y también son importantes en muchos otros contextos), pero su estatus ontológico difícilmente se hace claro alguna vez. Explicaré de inmediato (en §29.3) qué es una matriz densidad. Sin embargo, más adelante veremos por qué es importante para la posición (c) que la ontología de la matriz densidad ¡*no* se haga completamente clara! Los que mantienen el punto de vista (c) tienden a considerarse a sí mismos «positivistas» que en ningún caso quieren saber nada de «aburridas» cuestiones de ontología, y afirman creer que no les interesa qué es «real» y qué es «no real». Como ha dicho Stephen Hawking:[5]

> No exijo que una teoría corresponda a la realidad porque no sé lo que es. La realidad no es una cualidad que se pueda comprobar con papel tornasol. Todo lo que me interesa es que la teoría prediga los resultados de medidas.

Mi posición, por el contrario, es que la cuestión de la ontología tiene una importancia crucial en la mecánica cuántica, aunque plantea algunas preguntas que no están ni mucho menos resueltas en el momento actual.

29.2. Ontologías no convencionales para la mecánica cuántica

Antes de entrar en los detalles de todo esto, permítanme considerar otros tres puntos de vista generales con respecto a la mecánica cuántica. No hay que pensar que mi lista sea exhaustiva, ni hay que suponer que estos nuevos puntos de vista son completamente independientes de los que he presentado en la sección anterior. La lista (a), (b), (c), (d), (e), (f) que voy a considerar aquí representa el abanico de puntos de vista que se encuentran con más frecuencia en la literatura actual, pero no hago afirmaciones respecto a la compleción, independencia o especificidad de mi lista. Las tres ontologías adicionales que considero aquí representan cambios reales en el formalismo cuántico habitual; pero en dos de ellas, (d) y (e), no se prevé que haya diferencias experimentales entre el formalismo propuesto y la mecánica cuántica estándar. El punto de vista (d) es el enfoque de «historias consistentes» debido a Griffiths, Omnès y Gell-Mann/Hartle, y (e) es la ontología de la «onda piloto» de De Broglie y Bohm/Hiley.[6] La última posibilidad (f) es que la mecánica cuántica actual sea simplemente una aproximación a algo mejor, y que en esta teoría mejorada, tanto **U** como **R** tengan lugar *objetivamente* como procesos reales; más aún, desde la perspectiva de (f) se prevé que experimentos futuros deberían ser capaces de distinguir dicha teoría de la mecánica cuántica convencional.

En cuanto tengamos las herramientas necesarias, trataré de dar mi valoración de las diferentes alternativas (a), ..., (f). Sin embargo, para que el lector pueda adoptar una actitud objetiva respecto a estas valoraciones, es mejor que «reconozca» claramente mi propia posición en este momento. Soy, de hecho, un firme creyente en que son necesarios ciertos desarrollos en la línea de (f) para que la mecánica cuántica pueda tener un sentido completamente consistente. En el capítulo siguiente presentaré la versión concreta de (f) que me parece la más natural. Hecha esta advertencia, prosigamos; haremos primero una lista

de tales alternativas para ayudar al lector a tenerlas explícitamente en mente.

(a) «Copenhague»,
(b) muchos universos,
(c) decoherencia por el entorno,
(d) historias consistentes,
(e) onda piloto,
(f) nueva teoría con **R** objetiva.

Haré algunos comentarios sobre (d) y (e), puesto que no las he explicado realmente. El esquema (d) de «historias consistentes» ofrece una generalización del marco estándar de la teoría cuántica. Algunos proponentes han dotado a (d) de una ontología un poco parecida a la de los muchos universos (b), aunque todavía más extravagante en un aspecto, pero hasta donde puedo ver semejante ontología extravagante muy bien puede no ser necesaria. Tanto en (b) como en (d) podemos adoptar la postura de que tenemos, como ingredientes básicos, un espacio de Hilbert **H**, un estado de partida $|\psi_0\rangle$ que pertenece a **H** y un hamiltoniano $\mathcal{H}$.[7] En el caso de los muchos universos (b), la posición ontológica consiste en considerar que la realidad (del omnium) está descrita como una familia continua y uniparamétrica de estados (elementos de **H**, y con parámetro t), que parte de $|\psi_0\rangle$ en $t = 0$ y está completamente gobernada, para $t > 0$, por la evolución de Schrödinger determinada por $\mathcal{H}$. No hay aquí **R**, solo **U**. Pero el caso (d) de historias consistentes amplía esto de modo que también incorpora en su «evolución» procesos tipo **R**, incluso si no se considera que estos estén necesariamente asociados con medidas reales.

Para comprender la naturaleza matemática de estos procedimientos, debemos recordar primero (véanse §§22.5,6) cómo se describe matemáticamente una medida mecanocuántica (incluso si en el caso (d) no consideramos estos procedimientos como medidas), en términos de la acción de un operador hermítico (o normal; véase §22.5) **Q**. Si inmediatamente antes de la medida el estado del sistema es $|\psi\rangle$, entonces se considera que inmediatamente después de la medida «salta» al autoestado de **Q** correspondiente al autovalor de **Q** que resulta de la medida.

Pero por lo que concierne a su efecto sobre $|\psi\rangle$, también podemos reemplazar **Q** por un «conjunto completo de proyectores ortogonales» $\boldsymbol{E}_1, \boldsymbol{E}_2, \boldsymbol{E}_3, \ldots, \boldsymbol{E}_r$ (suponiendo que **Q** tiene exactamente r autovalores diferentes, donde por conveniencia consideramos que nuestro espacio de Hilbert **H** es de dimensión finita). Entonces, si la medida da el autovalor q_j, encontramos que $|\psi\rangle$ salta a un estado proporcional a $\boldsymbol{E}_j|\psi\rangle$ (postulado de proyección).

Veamos esto con más detalle. Recordemos de §22.6 que un *proyector* es un operador $\boldsymbol{E}$ cuyo cuadrado es igual a sí mismo y es hermítico, i.e.,

$$\boldsymbol{E}^2 = \boldsymbol{E} = \boldsymbol{E}^*.$$

La afirmación de que los proyectores $\boldsymbol{E}_1, \boldsymbol{E}_2, \boldsymbol{E}_3, \ldots, \boldsymbol{E}_r$ son mutuamente *ortogonales* es

$$\boldsymbol{E}_i\boldsymbol{E}_j = 0, \text{ si } i \neq j$$

y su completitud es que su suma es la identidad $\boldsymbol{I}$ en **H**

$$\boldsymbol{E}_1 + \boldsymbol{E}_2 + \boldsymbol{E}_3 + \ldots + \boldsymbol{E}_r = \boldsymbol{I}.$$

Llamemos *conjunto de proyectores* a un conjunto de $\boldsymbol{E}$ que satisface todas estas condiciones. La conexión entre **Q** y su conjunto de proyectores correspondiente es que, para cada autoestado q_j de **Q**, el espacio de autovectores correspondiente consiste en vectores de la forma $\boldsymbol{E}_j|\phi\rangle$. El papel del proyector $\boldsymbol{E}_j$ es que proyecta en este espacio de autovectores, para el autovalor q_j.[29.1]

El *postulado de proyección* para la operación **R** (véase §22.6), en la medida representada por **Q**, nos dice que si el resultado de la medida es q_j, entonces $|\psi\rangle$ salta a (algo proporcional a) $\boldsymbol{E}_j|\psi\rangle$. Esto ocurre con una probabilidad dada por

$$\langle\psi|\boldsymbol{E}_j|\psi\rangle,$$

[29.1] Explique por qué $\boldsymbol{E}_j|\psi\rangle$ es el resultado (ignorando normalización) de una medida dada por $\mathbf{Q} = q_1\boldsymbol{E}_1 + q_2\boldsymbol{E}_2 + q_3\boldsymbol{E}_3 + \ldots + q_r\boldsymbol{E}_r$ aplicada a $|\psi\rangle$, donde el autovalor es q_j, siendo las cantidades $q_1, q_2, q_3, \ldots, q_r$ números reales diferentes. ¿Puede demostrar que el operador hermítico general de dimensión finita tiene esta forma? (Puede suponer que cualquier matriz hermítica **Q** de dimensión finita puede transformarse en una matriz diagonal mediante una transformación unitaria.) Los $\boldsymbol{E}$ se denominan *idemponentes principales* de **Q**. ¿Qué modificaciones son necesarias para un operador *normal* **Q**?

si suponemos que $|\psi\rangle$ está *normalizado*, i.e., $\langle\psi|\psi\rangle = 1$. Así pues, para describir el efecto sobre el estado cuántico de la medida correspondiente a $\boldsymbol{Q}$ solo necesitamos considerar el conjunto de proyectores definido por $\boldsymbol{Q}$.

Volvamos a la ontología del enfoque (d) de las historias consistentes. La teoría opera con entidades llamadas *historias de grano grueso*,[8] cada una de las cuales se asemeja generalmente a un «omnium» de la aproximación de los muchos universos (b) que evoluciona según Schrödinger, utilizando un hamiltoniano $\mathcal{H}$. Pero en el caso de (d) también admitimos que se inserte un conjunto de proyectores en varios valores de t durante el curso de su evolución.

Todavía no tengo muy claro el estatus ontológico de la inserción de semejante conjunto de proyectores, pero uno se ve animado a adoptar la actitud de que el papel de dicho conjunto de proyectores consiste en ofrecer una especie de «refinamiento» de la historia, antes que representar un cambio fundamental para lo que está sucediendo en el mundo. Ciertamente no cabe asignar a los proyectores el estatus ontológico dado por una medida objetiva. Una analogía más apropiada podría ser que los conjuntos de proyectores proporcionan refinamientos, o alteraciones, para «cajas» de grano grueso, como en el espacio de fases clásico (véase §27.3), y esto explica el término «historia de grano grueso» que se utiliza aquí. En dicha historia de grano grueso, en el momento en que encuentra un conjunto de proyectores (y de forma análoga al procedimiento estándar adoptado en una medida cuántica), el estado actual $|\psi\rangle$ queda reemplazado por (algo proporcional a) $\boldsymbol{E}_j|\psi\rangle$, donde $\boldsymbol{E}_j$ es un miembro del conjunto de proyectores. Podría pensarse que esto es una pérdida de información, pero no hay pérdida si seguimos la pista de la *familia entera* de $\boldsymbol{E}_j|\psi\rangle$ para todos los $\boldsymbol{E}_j|\psi\rangle$ en el conjunto, puesto que $|\psi\rangle$ es simplemente la suma de todos ellos.

De acuerdo con un deseo de que surja algo que se parezca al tipo de mundo clásico que realmente percibimos, algunas familias particulares de historias de grano grueso son distinguidas y calificadas como *consistentes* (o a veces «decoherentes») si se satisface cierta condición, que expresa el hecho de que las probabilidades, calculadas según las reglas cuánticas estándar, satisfacen las reglas clásicas ordinarias para la

probabilidad.[9] Un conjunto consistente de historias de grano grueso se denomina *máximamente refinado* si no se puede insertar otro conjunto de proyectores (no equivalente a ninguno de los que ya han sido incorporados) sin destruir la consistencia. Creo que una historia a partir de un conjunto máximamente refinado ofrece un fuerte candidato para lo que podría considerarse como ontológicamente «real», de acuerdo con el punto de vista (d).

Pese a todo, no he visto este punto de vista presentado de forma explícita; por lo que sé, lo que parece estar más cerca del punto de vista ontológico para las «historias consistentes» es algo más parecido a la *totalidad* de historias en un conjunto máximamente refinado.[10] Esto está quizá más en la línea de lo que hemos visto en el punto de vista (b) de los muchos universos, pero la presencia de muchas posibles, y alternativas, colecciones consistentes de conjuntos de proyectores parece proporcionarnos un conjunto aún mayor de «mundos» alternativos. Sin embargo, recordemos que también en la imagen (b) de los muchos universos puede surgir una confusión ontológica. El omnium ontológicamente «real» (descrito por $|\psi\rangle$) es una *superposición* de numerosos mundos diferentes; la *colección* de todos estos mundos individuales (que no es lo mismo que su superposición concreta $|\psi\rangle$) *no* debe considerarse «real». Frente a este tipo de confusión, el punto de vista (d) de las historias consistentes tiene la ventaja de que las probabilidades cuánticas correctas están dadas por la teoría, lo que no parece ser el caso con (b).

En el caso (e) «bohmiano» (onda piloto), la posición ontológica está, y ello es agradable, mucho más cercana a la Tierra, aunque incluso aquí hay algunas sutilezas considerables, pues hay, en cierto sentido, *dos* niveles de realidad, uno de los cuales es más firme que el otro. Es más sencillo plantear primero el caso de un sistema que consiste en una única partícula sin espín. Entonces este nivel más firme de realidad viene dado por la posición de la partícula. En un experimento de doble rendija (§21.4, Fig. 21.4), y puesto que la posición de la partícula es ontológicamente real, esta atraviesa realmente una rendija o la otra, pero su movimiento está «guiado», en efecto, por ψ, de modo que esta también adquiere un estatus secundario pero que en cualquier caso sigue siendo ontológicamente «real». En esta teoría se suelen tener posturas algo di-

ferentes respecto al *módulo* y el *argumento* de ψ (§5.1); a partir del primero se construye una cantidad conocida como «potencial cuántico», y el segundo se utiliza para definir lo que se denomina la «onda piloto». Sin embargo, este tipo de desdoblamiento no es necesario y su importancia parece hacerse menos clara con sistemas más complicados.

En general, podemos considerar ψ como una función compleja que está definida en el espacio de configuración $\mathcal{C}$, y sirve de «guía» para el comportamiento de un punto P en $\mathcal{C}$. Se admite que la parte más firme de la realidad del sistema es la configuración clásica realmente definida por P, pero se asigna un tipo de realidad (más débil) también a la función compleja ψ, en virtud de su papel como guía del comportamiento de P. Se considera que todas las medidas pueden reducirse, en última instancia, a medidas de «posición», lo que aquí significa medidas de la configuración del sistema. El módulo al cuadrado $|\psi|^2$, en un punto Q de $\mathcal{C}$, define la densidad de probabilidad de encontrar el sistema en la configuración definida por Q, pero la localización de P en $\mathcal{C}$ determina lo que se considera que es la configuración *real* del sistema.

Ahora bien, todo esto parece casi «demasiado fácil», pero hay sutilezas. En particular, se trata de una imagen muy no local, donde ψ es una entidad altamente «holística» (como deber ser, para que esté de acuerdo con la naturaleza holística de las funciones de onda que se ha resaltado en §21.7). No obstante, esto parece inevitable en mecánica cuántica. Algo más serio es el hecho de que hay que imponer condiciones importantes sobre la distribución de probabilidad para el estado inicial $|\psi_0\rangle$, de modo que sea válida la ley de probabilidad cuántica $|\psi|^2$, y continúe siendo válida después de medidas secuenciales. Por otra parte, se podría cuestionar la corrección de la hipótesis de que todas las medidas pueden reducirse, en última instancia, a medidas de posición (en particular, porque las medidas estrictas de posición no son completamente legítimas en mecánica cuántica; véase §21.10), y si la imagen del espacio de configuración es adecuadamente inequívoca cuando se están considerando parámetros no clásicos como el espín. De todas formas, la claridad de la posición ontológica de (e) le da mucho crédito (aunque, como veremos en §29.9, hay también otras cuestiones a las que hacer frente).[11]

Finalmente, existen muchas propuestas diferentes en la línea de (f). No es oportuno que las describa todas aquí en detalle, pero puedo hacer algunos comentarios generales sobre ellas. Muchas de estas propuestas aceptarán (al menos como postura provisional) un estatus ontológicamente *real* para un vector de estado $|\psi\rangle$ que evoluciona. La evolución temporal de $|\psi\rangle$, en una teoría semejante, será algo que se aproxima estrechamente a la alternancia de **U** con **R** que la mecánica cuántica estándar nos dice que adoptemos en la práctica (véase la Fig. 22.1). Pese al hecho de que esas teorías en línea con el punto de vista (f) se consideran «fuera de la corriente principal» del pensamiento mecanocuántico, podría decirse con seriedad que (f) es realmente el punto de vista que más *acepta* la realidad del formalismo de la mecánica cuántica tal como se utiliza hoy en la práctica, ¡puesto que *los dos* procesos de evolución mecanocuántica **U** y **R** se toman con seriedad, ontológicamente, como descripción de la evolución de la realidad! El problema es, no obstante, que **U** y **R** son en realidad matemáticamente *incompatibles*, y esta es la razón por la que (f) exige que haya cambios de la evolución unitaria ordinaria, ¡y es esto lo que separa (f) de la corriente principal!

¿Por qué es **R** matemáticamente incompatible con **U**? Quizá la razón más obvia es que **R** representa un cambio discontinuo en el vector de estado (salvo en la circunstancia excepcional de que el estado previo a la medida sea realmente un autoestado del operador de medida), mientras que **U** actúa siempre de forma continua. Pero incluso si imaginamos que el «salto» inducido por **R** no es absolutamente instantáneo, habría problemas con la unitariedad debido a la falta de determinismo en **R**. Pueden obtenerse resultados diferentes a partir de una misma entrada, algo que nunca sucede con **U**. Más aún, una teoría que hace de **R** un proceso real nunca puede ser unitaria cuando realmente tiene lugar —de acuerdo con **R**— un salto cuántico (no trivial). Pese a esto, existe un notable acuerdo entre los dos procesos **U** y **R**, puesto que la «regla del módulo al cuadrado» que interrumpe **U** para ofrecernos la **R** probabilista, utiliza la propia «unitariedad» de **U** para darnos una ley de conservación de probabilidad para **R** (básicamente el hecho de que los productos escalares $\langle\phi|\psi\rangle$, a partir de los cuales se calculan las probabilidades, se conserven bajo la evolución unitaria;

véanse §§22.4,5). Esta es una parte integral de las maravillas de la mecánica cuántica, y proporciona una razón de peso para que la gente sea reacia a jugar de cualquier manera con los principios de dicha teoría, lo que explica en parte por qué (f) no es particularmente popular entre los físicos cuánticos de hoy.

De todas formas, creo que hay razones poderosas para esperar un cambio. En mi opinión, dicho cambio representaría una revolución fundamental y no puede conseguirse solo «retocando» la mecánica cuántica. Pese a todo, los cambios necesarios deben ser completamente respetuosos con los principios centrales que yacen en el corazón de la física actual. La misma rigidez del formalismo cuántico, tal como se ha indicado en el párrafo precedente, es una razón para ambos requisitos. A modo de comparación, recordemos la rigidez de la física newtoniana. La relatividad y la mecánica cuántica no se obtuvieron de ella haciendo retoques, sino mediante los cambios revolucionarios de perspectiva que, en cualquier caso, mostraban el debido respeto a la estructura geométrica lagrangiana/hamiltoniana/simpléctica muy organizada de la teoría newtoniana. ¿Tienen este carácter respetuosamente revolucionario los cambios de la teoría cuántica que han sido sugeridos hasta ahora por diversas personas,[12] o solo están haciendo retoques? Tengo que decir que en su mayor parte estas ideas deben considerarse como retoques; pero algunas de ellas podrían proporcionar muy bien indicadores de la ruta verdadera hacia una teoría cuántica mejorada.

29.3. La matriz densidad

Pero ¿por qué hay necesidad de «mejorar» la teoría cuántica? La mayoría de los físicos cuánticos no parecen creer que se necesite una teoría semejante, tras haber hecho las paces con las aparentes contradicciones y la oscura ontología de una u otra de las imágenes estándar (o la carencia de una imagen semejante). Antes de que podamos intentar abordar cualquiera de las dificultades que podría haber en cualquiera de las imágenes «estándar» (a), (b) y (c), deberíamos llegar a la idea de una matriz densidad, que no solo es fundamental para el punto de vista (c), sino que desempeña un papel importante en muchas otras considera-

ciones mecanocuánticas. Además, plantea cuestiones intrigantes y profundas respecto a cómo debería representarse la realidad en mecánica cuántica.

Supongamos que tenemos cierto sistema cuántico cuyo estado no nos es completamente conocido. El estado podría ser $|\psi\rangle$, o podría ser $|\phi\rangle$, o podría ser..., o podría ser $|\chi\rangle$. La lista podría ser infinita, pero para nuestros propósitos será suficiente considerar solo una lista finita de posibilidades. A cada una de estas posibilidades hay que asignarle una probabilidad, digamos $p, q, \ldots, s$, respectivamente. Las posibilidades deben ser exhaustivas, de modo que las probabilidades —números reales entre 0 y 1 (inclusive)— deben sumar la unidad:

$$p + q + \ldots + s = 1.$$

Suponemos que cada uno de los $|\psi\rangle$, $|\phi\rangle$, ..., $|\chi\rangle$ está *normalizado*:

$$\||\psi\rangle\| = 1, \quad \||\phi\rangle\| = 1, \quad \ldots, \quad \||\chi\rangle\| = 1.$$

(Recordemos de §22.3 que $\||\psi\rangle\| = \langle\psi|\psi\rangle$, etc.). Entonces definimos la matriz densidad como la magnitud

$$\boldsymbol{D} = p\,|\psi\rangle\langle\psi| + q\,|\phi\rangle\langle\phi| + \ldots + s\,|\chi\rangle\langle\chi|.$$

Recordemos de §22.3 que el vector bra $\langle\psi|$ es el conjugado hermítico del vector ket $|\psi\rangle$. La magnitud $|\psi\rangle\langle\psi|$ es entonces el producto tensorial (o producto exterior) de $|\psi\rangle$ con $\langle\psi|$, etc. En la notación de índices de §23.8, podemos escribir $\langle\psi|$ como $\overline{\psi}_\alpha$, donde ψ^α representa a $|\psi\rangle$. Entonces, $|\psi\rangle\langle\psi|$ podría escribirse como $\overline{\psi}^\alpha\,\psi_\beta$, etc. Por consiguiente, la propia $\boldsymbol{D}$ tendría la estructura de índices $D^\alpha{}_\beta$. La matriz densidad tiene las propiedades algebraicas de que es hermítica, definida no negativa y de traza unidad:

$$\boldsymbol{D}^* = \boldsymbol{D}, \quad \langle\xi|\boldsymbol{D}|\xi\rangle \geq 0 \text{ para todo } |\xi\rangle, \quad \langle\boldsymbol{D}\rangle = 1,$$

donde $\langle\boldsymbol{D}\rangle = \text{traza}\ \boldsymbol{D} = D^\alpha{}_\alpha$ (véase §13.4).[29.2]

La matriz densidad desempeña un papel análogo al que es frecuentemente utilizado en mecánica estadística clásica, donde podríamos no estar especialmente interesados en el estado (clásico) preciso de un sis-

[29.2] Deduzca estas propiedades.

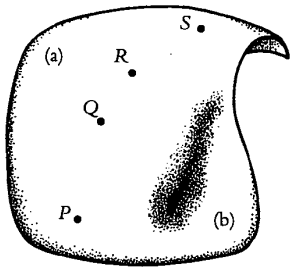


Fig. 29.1. Distribuciones de probabilidad clásicas representadas en el espacio de fases $\mathcal{P}$. (a) Para un conjunto finito de puntos $P, Q, \ldots, S$, en $\mathcal{P}$, se asigna un valor de probabilidad $p, q, \ldots, s$ (números reales entre cero y uno) a cada punto, con $p + q + \ldots + s = 1$. (b) Una distribución continua, con una medida de probabilidad (una densidad no negativa en números reales), cuya integral es 1, asignada en una región de $\mathcal{P}$.

tema, sino que nos contentamos con considerar una distribución de probabilidad de alternativas clásicas. Esto se piensa más fácilmente en términos del espacio de fases $\mathcal{P}$ de las alternativas clásicas posibles. En lugar de representar el sistema como un punto P en $\mathcal{P}$, se consideraría en términos de una distribución de probabilidad sobre $\mathcal{P}$. Si tenemos solo un número finito de alternativas[13] para el sistema, siendo $p, q, \ldots, s$, las diversas probabilidades, entonces representamos esto como un conjunto finito de puntos $P, Q, \ldots, S$, en $\mathcal{P}$, a cada uno de los cuales se le asigna su respectivo valor de probabilidad $p, q, \ldots, s$ (véase la Fig. 29.1). De hecho, podríamos imaginar que hacemos exactamente lo mismo en física cuántica, con el espacio de Hilbert $\mathbf{H}$ del sistema cuántico en el papel del espacio de fases $\mathcal{P}$; entonces tendríamos cierta distribución de probabilidad en $\mathbf{H}$. En relación con la matriz densidad $\boldsymbol{D}$ que acabamos de considerar, esta distribución consistiría solo en un número finito de puntos $P, Q, \ldots, S$, en $\mathbf{H}$, cada uno de los cuales tiene asignado su respectivo valor de probabilidad $p, q, \ldots, s$.

Pero esto *no* es lo que se hace normalmente en mecánica cuántica; en lugar de ello se utiliza la matriz densidad.[14] ¿A qué se debe esto? La razón es que, en mecánica cuántica, una *medida*, que tiene la forma de una pregunta planteada a un sistema cuántico —y restrinjamos la atención a una pregunta SÍ/NO—, se expresa en términos de la acción de un proyector $\boldsymbol{E}$ aplicado al vector de estado (normalizado) $|\xi\rangle$. La probabilidad de la respuesta sí viene dada entonces por[29.3]

$$\text{probabilidad de sí} = \langle \xi | \boldsymbol{E} | \xi \rangle,$$

[29.3] Explique por qué; obtenga también la expresión $\langle \boldsymbol{ED} \rangle$ siguiente.

de lo que se sigue que para la mezcla probabilista de estados alternativos posibles $|\psi\rangle, |\phi\rangle, \ldots, |\chi\rangle$, descritos arriba, con matriz densidad $\boldsymbol{D}$, obtenemos la respuesta

$$\text{probabilidad de sí} = \langle \boldsymbol{ED} \rangle.$$

La importancia de esto es que no necesitamos conocer la información completa de la distribución de probabilidades para los estados alternativos $|\psi\rangle, |\phi\rangle, \ldots, |\chi\rangle$ para poder calcular probabilidades para una pregunta estándar SÍ/NO en mecánica cuántica (o, de hecho, para el valor esperado de cualquier otro observable mecanocuántico);[29.4] toda la información necesaria está almacenada en la matriz densidad, y como veremos muy pronto una matriz densidad dada puede estar compuesta de muchas distribuciones de probabilidad de estados *diferentes*. Hay una economía y elegancia considerable en esta notable entidad matemática (introducida en 1932 por el destacado matemático húngaro-estadounidense John von Neumann). Combina en una expresión lo que parecerían ser dos nociones completamente diferentes de probabilidad. Por una parte, tenemos los números $p, q, \ldots, s$, que son las probabilidades clásicas ordinarias para los estados alternativos $|\psi\rangle, |\phi\rangle, \ldots, |\chi\rangle$, mientras que, por la otra, tenemos las probabilidades cuánticas obtenidas a partir de la regla del módulo al cuadrado de §21.9. La matriz densidad combina las dos y no distingue directamente un tipo del otro.

29.4. Matrices densidad para espín 1/2: la esfera de Bloch

Permítanme ilustrar este punto con un ejemplo sencillo. Supongamos que tenemos una partícula de espín 1/2, cuyo estado de espín sabemos que es o $|\uparrow\rangle$ o $|\downarrow\rangle$, con probabilidad 1/2 para cada alternativa. Si decidimos medir este espín en una dirección arriba/abajo, simplemente obtenemos «arriba» si el estado es $|\uparrow\rangle$, y «abajo» si el estado es $|\downarrow\rangle$. En cada caso la probabilidad es 1/2. Estas son solo probabilidades clásicas directas, y aquí no hay ningún misterio cuántico. Pero supongamos que en lugar de lo anterior medimos el espín en la dirección iz-

[29.4] ¿Puede ver por qué debería ser así?

quierda/derecha. Entonces, si el estado es $|\uparrow\rangle$, las reglas R cuánticas nos dicen que obtenemos una probabilidad 1/2 de que el espín sea «izquierda» y una probabilidad 1/2 de que el espín sea «derecha». Exactamente la misma conclusión se obtiene si el estado es $|\downarrow\rangle$. Así, para la mezcla con probabilidades iguales de $|\uparrow\rangle$ y $|\downarrow\rangle$, seguimos obteniendo probabilidades de 1/2 para cada uno de los resultados «izquierda» y «derecha». Ahora, sin embargo, las probabilidades se obtienen enteramente a partir de la ley mecanocuántica del «módulo al cuadrado». También podríamos decidir medir el espín en cualquier otra dirección. Las probabilidades resultarían ser una vez más 1/2 para cada respuesta, pero esta probabilidad estaría compuesta, en general, por una mezcla de probabilidades clásicas y cuánticas.[29.5]

Alternativamente, podríamos imaginar que rotamos la mezcla de estados en lugar del aparato de medida. Así, una mezcla con probabilidades iguales de $|\leftarrow\rangle$ y $|\rightarrow\rangle$ daría exactamente las mismas respuestas que la mezcla anterior con probabilidades iguales de $|\uparrow\rangle$ y $|\downarrow\rangle$, y lo mismo sucedería con una mezcla con probabilidades iguales de $|\nwarrow\rangle$ y $|\searrow\rangle$ (donde en cada caso consideramos que estos pares de estados son ortogonales y normalizados: $\langle\uparrow|\downarrow\rangle = \langle\leftarrow|\rightarrow\rangle = \langle\nwarrow|\searrow\rangle = 0$, $\langle\uparrow|\uparrow\rangle = \langle\downarrow|\downarrow\rangle = \ldots = \langle\searrow|\searrow\rangle = 1$). Obtenemos, para la matriz densidad $\boldsymbol{D}$, en cada caso

$$\boldsymbol{D} = \tfrac{1}{2}|\uparrow\rangle\langle\uparrow| + \tfrac{1}{2}|\downarrow\rangle\langle\downarrow|,$$

$$\boldsymbol{D} = \tfrac{1}{2}|\leftarrow\rangle\langle\leftarrow| + \tfrac{1}{2}|\rightarrow\rangle\langle\rightarrow|,$$

$$\boldsymbol{D} = \tfrac{1}{2}|\nwarrow\rangle\langle\nwarrow| + \tfrac{1}{2}|\searrow\rangle\langle\searrow|,$$

y la notable propiedad de la matriz densidad es que todas estas $\boldsymbol{D}$ son la *misma*.[29.6] Todas las probabilidades para medidas de espín a las que nos acabamos de referir pueden ser obtenidas mediante el uso de la fórmula $\langle\boldsymbol{ED}\rangle$ anterior; así pues, puesto que las $\boldsymbol{D}$ son la misma, las

[29.5] Calcule esto para un ángulo general de pendiente θ de la dirección de medida, utilizando la expresión $\frac{1}{2}(1 + \cos\theta)$ de §22.9 para la probabilidad.

[29.6] Demuéstrelo por cálculo explícito utilizando los resultados de §§22.8,9 y el ejercicio [22.25].

probabilidades respectivas deben dar el mismo resultado, tal como hemos visto.

Pero ¿cómo vamos a considerar la *ontología* de estas mezclas probabilistas de estados? Si consideramos que el estado cuántico tiene algún tipo de realidad física, entonces estas tres situaciones son decididamente distintas desde un punto de vista ontológico. Decir que existe la misma probabilidad de que el estado se encuentre en una u otra de las alternativas (físicamente reales) $|\uparrow\rangle$, $|\downarrow\rangle$ es completamente diferente de decir que existe la misma probabilidad de que esté en $|\nwarrow\rangle$ o $|\searrow\rangle$. Sin embargo, esta cuestión está extraordinariamente confusa en buena parte de la literatura mecanocuántica. A menudo, los físicos cuánticos parecen estar tomando una posición ontológica completamente diferente de la que se acaba de describir, al considerar la propia matriz densidad como algo que proporciona una mejor descripción de la realidad que los estados individuales. Podrían adoptar el punto de vista de que las tres ontologías aparentemente distintas para $\boldsymbol{D}$ dadas arriba (i.e., las tres diferentes colecciones con probabilidades ponderadas de estados cuánticos alternativos) son físicamente indistinguibles. Por consiguiente, estos físicos —a menudo partidarios del punto de vista (c) de la decoherencia por el entorno— podrían adoptar la postura positivista o pragmática según la cual no tiene sentido distinguir entre dichas alternativas. Estas personas podrían adoptar el punto de vista de que es la matriz densidad lo que mejor describe la realidad cuántica.

De hecho, en muchos contextos se considera que la palabra «estado» se refiere a una matriz densidad en lugar de a la noción más primitiva que, hasta este momento, he estado llamando un «estado cuántico», a saber, una magnitud descriptible por un ket tal como $|\psi\rangle$. Cuando se utiliza la palabra «estado» en el sentido de una matriz densidad, se utiliza el término «estado puro» para una matriz densidad de la forma especial $|\psi\rangle\langle\psi|$, y se utiliza «estado mezcla» para una matriz densidad más general que no puede representarse de este modo. Los «estados puros» en este sentido se refieren a lo que he estado llamando simplemente un «estado». Personalmente, encuentro muy confuso referirse a una matriz densidad (pura o no) como un «estado», y me abstendré de utilizar aquí esta terminología. Para mí, un «estado cuántico» es efectivamente un vector de estado cuántico $|\psi\rangle$, y no una matriz

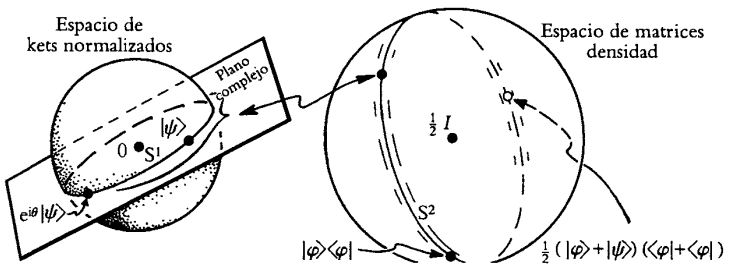


Fig. 29.2. ¿Cómo representamos un estado cuántico puro? (a) Espacio de kets $|\psi\rangle$, normalizados por $\langle\psi|\psi\rangle = 1$. (b) La matriz densidad $|\psi\rangle\langle\psi|$ es «equivalente» a $|\psi\rangle$ salvo la libertad de fase $|\psi\rangle \mapsto e^{i\theta}|\psi\rangle$, y a la familia de kets no nulos proporcionales (factores de proporcionalidad complejos). Pese a todo, la linealidad cuántica básica se oscurece en la descripción de matriz densidad.

densidad. Pese a todo, algunos preferirían distinguir los términos «estado cuántico» y «vector de estado cuántico», siendo el último el ket $|\psi\rangle$ y estando el primero representado como la clase de equivalencia de múltiplos complejos no nulos de $|\psi\rangle$, i.e., el elemento del espacio de Hilbert *proyectivo* $\mathbb{P}\mathbf{H}$ correspondiente al elemento $|\psi\rangle$ de $\mathbf{H}$ (véase §15.6). Si decidimos normalizar $|\psi\rangle$ por $\langle\psi|\psi\rangle = 1$, entonces la única libertad que queda en $|\psi\rangle$ (para un punto dado en $\mathbb{P}\mathbf{H}$) es la libertad de fase $|\psi\rangle \mapsto e^{i\theta}|\psi\rangle$ (con θ real); véase la Fig. 29.2. La noción de una matriz densidad «estado puro» es efectivamente equivalente a esta noción «proyectiva» de un estado cuántico, puesto que $|\psi\rangle\langle\psi|$ es invariante bajo esta libertad de fase. Así pues, podríamos razonablemente adoptar la postura de que una matriz densidad estado puro describe de manera adecuada el estado cuántico.

De todas formas, me siento incómodo al considerar semejante «matriz densidad estado puro» como la representación matemática adecuada de un «estado físico». El factor de fase $e^{i\theta}$ es solo «inobservable» si el estado en consideración representa el objeto entero de interés. Cuando se considera cierto estado como parte de un sistema mayor, es importante seguir la pista de dichas fases. Además, la linealidad compleja fundamental de la estructura básica del espacio de Hilbert de los vectores ket se complica matemáticamente de forma innecesaria si uno tiene que operar siempre con las cantidades $|\psi\rangle\langle\psi|$ en lugar de la matemáticamente más simple $|\psi\rangle$ (o $\langle\psi|$).[29.7] En parte por estas razones, mi propia postura sería no considerar la matriz densidad como realidad, sino solo como un artificio útil. Sin embargo, existen algunos aspectos intrigantes respecto al estatus ontológico confuso de la matriz densidad, como veremos aquí y en §29.5.

Antes de llegar a esto, será útil que nos familiaricemos con la *esfera de Bloch*, que representa el espacio de matrices densidad para un sistema de 2 estados. Esta es la esfera maciza cerrada (o, en la terminología del matemático, la *3-bola* o *3-disco*) B^3 que reside en el 3-espacio euclídeo. Representa las matrices densidad para espín 1/2 (o para cualquier otro sistema de 2 estados); véase §22.9. Podemos escribir la matriz hermítica general (2 × 2) de traza unidad como

$$\frac{1}{2}\begin{pmatrix} 1+a & b+ic \\ b-ic & 1-a \end{pmatrix},$$

donde a, b, c son números reales. Para que esto sea una matriz densidad, debe ser definida no negativa, que es la condición[29.8]

$$a^2 + b^2 + c^2 \leq 1.$$

Esto representa un punto general en la esfera de Bloch B^3, cuya *frontera* S^2 es la 2-esfera $a^2 + b^2 + c^2 = 1$. Aquí, S^2 representa los estados puros en el sistema de 2 estados (por ejemplo, espín 1/2), y este espacio puede identificarse con la esfera de *Riemann* S^2 descrita en §22.9.[15]

Ahora la matriz densidad concreta $\boldsymbol{D} = \left(\frac{1}{2}\boldsymbol{I}\right)$ que acabamos de considerar está representada por el *origen* de la esfera de Bloch, y su interpretación ontológica completamente ambigua es muy obvia a partir de la simetría de la figura (Fig. 29.3). Sin embargo, *cualquier* punto (matriz densidad no pura) $\boldsymbol{L}$ en el interior de B^3 representa una matriz densidad con una interpretación ontológica asimismo ambigua. Para ver esto tracemos simplemente una recta arbitraria (cuerda) que pasa por $\boldsymbol{L}$ y corta la frontera S^2 en dos puntos $\boldsymbol{P}_1$ y $\boldsymbol{P}_2$. Estos representan dos *esta-*

[29.7] Vea si puede caracterizar la familia de matrices densidad de «estado puro» que corresponde a la combinación lineal $w|\psi\rangle - z|\varphi\rangle$ para cierto par de estados dados $|\psi\rangle$ y $|\varphi\rangle$.

[29.8] Demuéstrelo. *Sugerencia*: ¿Cuál es el producto de los autovalores en términos de a, b y c? ¿Qué significa que este producto sea no negativo?

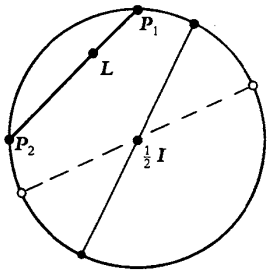


Fig. 29.3. Esfera de Bloch B^3 de matrices densidad para un sistema de 2 estados, centrada en $\frac{1}{2}\boldsymbol{I}$. Cualquier matriz densidad $\boldsymbol{L}$ (no pura) tiene una interpretación ontológica ambigua. Una cuerda arbitraria que pasa por $\boldsymbol{L}$ corta la frontera S^2 en $\boldsymbol{P}_1$ y $\boldsymbol{P}_2$; entonces, $\boldsymbol{L}$ tiene una interpretación como una mezcla probabilística de los estados puros $\boldsymbol{P}_1$ y $\boldsymbol{P}_2$.

dos puros y la matriz densidad $\boldsymbol{L}$ puede ser interpretada entonces como una mezcla probabilista de estos dos.[29.9] La única cosa que hay de especial acerca del origen $\boldsymbol{D}$ de la esfera de Bloch es que todos estos pares de estados puros, en cuyos términos puede representarse $\boldsymbol{D}$, son pares *ortogonales*. Pero no hay nada en la definición de una matriz densidad que requiera que la mezcla probabilista sea entre estados mutuamente ortogonales. En §29.5 veremos cómo pueden aparecer mezclas no ortogonales.

29.5. La matriz densidad en situaciones EPR

Examinemos una situación particularmente clara en la que aparece de una forma natural una colección con probabilidades ponderadas de vectores de estado posibles. Esto se da en el efecto EPR-Bohm (véase §23.4). Supongamos que en alguna parte entre la Tierra y Titán, la luna de Saturno —pero digamos que a distancia doble de la Tierra que de Titán—, se emite un par EPR de partículas de espín 1/2 en un estado combinado de espín 0. Supongamos que mi colega en Titán (nuestro viejo conocido de §§23.4,5) mide el espín de la partícula que llega allí en una dirección arriba/abajo y obtiene cierta respuesta, aproximadamente una hora antes de que yo reciba mi partícula aquí en la Tierra. Supongamos que cuando llega mi partícula no ha habido tiempo para que yo haya recibido ninguna señal procedente de mi colega acerca del

[29.9] Explique por qué es así, demostrando que las dos probabilidades en la mezcla tienen la misma proporción que las dos longitudes en las que $\boldsymbol{L}$ divide la cuerda.

resultado de esa medida anterior. (Titán está aproximadamente a tres horas luz de la Tierra.) Por lo que a mí respecta, mi partícula tiene o bien espín $|\uparrow\rangle$ o bien espín $|\downarrow\rangle$. Será $|\uparrow\rangle$ si mi colega encontró el estado $|\downarrow\rangle$, y mi estado será $|\downarrow\rangle$ si mi colega encontró realmente $|\uparrow\rangle$. Puesto que yo sé que las probabilidades de que mi colega encuentre $|\downarrow\rangle$ o $|\uparrow\rangle$ son iguales, debo adoptar el punto de vista de que el estado de la partícula que yo recibo (una hora después de la medida de mi colega) tiene una probabilidad 1/2 de ser $|\uparrow\rangle$ y una probabilidad 1/2 de ser $|\downarrow\rangle$. Así pues, utilizo la matriz densidad

$$\boldsymbol{D} = \tfrac{1}{2}|\uparrow\rangle\langle\uparrow| + \tfrac{1}{2}|\downarrow\rangle\langle\downarrow|$$

(tomando los dos estados $|\uparrow\rangle$, $|\downarrow\rangle$ ortogonales y normalizados: $\langle\uparrow|\downarrow\rangle = 0$, y $\langle\uparrow|\uparrow\rangle = 1 = \langle\downarrow|\downarrow\rangle$).

Sin embargo, podría darse el caso de que, por un cambio de idea a última hora, mi colega hubiera decidido medir el espín de la partícula que llega a Titán no en una dirección arriba/abajo, sino en una dirección izquierda/derecha. Si mi colega obtuvo el resultado $|\leftarrow\rangle$, entonces debo encontrar $|\rightarrow\rangle$ para mi partícula que llega a la Tierra; si mi colega obtuvo $|\rightarrow\rangle$, entonces yo debo encontrar $|\leftarrow\rangle$. Una vez más, la probabilidad de las dos alternativas de mi colega habría sido 1/2 en cada caso; por eso, aunque yo no conozco todavía cuál de estos resultados obtuvo mi colega, debo concluir que mi partícula podría ser $|\rightarrow\rangle$ o podría ser $|\leftarrow\rangle$, con probabilidad 1/2 en cada caso. Por consiguiente, asigno la matriz densidad

$$\boldsymbol{D} = \tfrac{1}{2}|\rightarrow\rangle\langle\rightarrow| + \tfrac{1}{2}|\leftarrow\rangle\langle\leftarrow|$$

a mi partícula (donde $\langle\leftarrow|\rightarrow\rangle = 0$, $\langle\rightarrow|\rightarrow\rangle = 1 = \langle\leftarrow|\leftarrow\rangle$). Por supuesto, como hemos visto, esta es exactamente la misma $\boldsymbol{D}$ que antes. Así es como debería ser, puesto que la decisión de mi colega respecto a la forma de medir la partícula en Titán no debería afectar a las probabilidades en la Tierra (o de lo contrario, tendríamos un método de enviar señales de Titán a la Tierra más rápidas que la luz).[29.10] Así pues, parecería que para el tipo de situación que estamos considerando la matriz densidad proporciona una excelente descripción matemática de la si-

[29.10] Explique cómo.

tuación física. El estado de espín de la partícula que recibo en la Tierra, siempre que yo no sepa nada de lo que sucede en Titán —ni la dirección que mi colega escogió para medir el espín, ni el resultado de dicha medida—, está muy bien descrito por la matriz densidad anterior $\boldsymbol{D}$.

Por supuesto, esto funciona bien solo si yo no recibo información de Titán. Si conozco el tipo de medida que realiza mi colega, esto afectará a mi idea de la ontología del estado de espín que yo recibo, pero no afectará a las expectativas de las probabilidades de las medidas que yo pudiera realizar aquí en la Tierra.[16] Si sé que la medida de mi colega va a ser derecha/izquierda, yo podría adoptar el punto de vista de que la ontología del estado de espín de mi partícula es o bien derecha o bien izquierda, pero no sé cuál, un punto de vista que yo no podría haber adoptado si no hubiera conocido la dirección de la medida de mi colega. No obstante, este conocimiento ontológico no afectará a mis estimaciones de las probabilidades de los resultados de las medidas de espín que yo realizo en la Tierra, así que podría adoptar la postura alternativa de que la «ontología» no es importante, y quizá incluso científicamente carente de significado, de modo que la matriz densidad *es* todo lo que se necesita científicamente. Por el contrario, si recibo realmente un mensaje de Titán diciéndome los resultados de la medida de mi colega, entonces mis estimaciones de probabilidad muy bien podrían verse afectadas. Más que eso, habrá realmente requisitos de consistencia que limiten los resultados de nuestras medidas conjuntas (por ejemplo: yo no puedo obtener el resultado $\langle\leftarrow|$ si mi colega obtuvo el resultado $\langle\leftarrow|$. *Ahora* está claro que la descripción mediante matriz densidad es completamente inadecuada, y debemos volver a una descripción en términos de un (vector de) estado cuántico real que describa el par entrelazado entero: $|\Omega\rangle = |\uparrow\rangle|\downarrow\rangle - |\downarrow\rangle|\uparrow\rangle$ $(= |\leftarrow\rangle|\rightarrow\rangle - |\rightarrow\rangle|\leftarrow\rangle$, etc.).

La matriz densidad concreta que aparece en el ejemplo anterior es muy especial. En términos de cualquier base ortonormal, tiene la forma

$$\boldsymbol{D} = \begin{pmatrix} \frac{1}{2} & 0 \\ 0 & \frac{1}{2} \end{pmatrix}.$$

Lo que es especial en ella es que todos sus autovalores son iguales (los dos números 1/2 de la diagonal). La consecuencia de esto es que ella tiene la misma forma cualquiera que sea la base (ortonormal) utilizada, puesto que es simplemente un múltiplo de la matriz identidad. Así pues, no hay nada que distinga la base arriba/abajo de la base izquierda/derecha, etc.

Es importante señalar que esto es solo el resultado de la situación particularmente simple que hemos considerado en este ejemplo. Ya hemos visto en §29.4 que no hay nada especial en la matriz densidad $\boldsymbol{D}$ concreta (con valores propios iguales) con respecto a su confusión de ontología. Con una modificación muy ligera del ejemplo podemos obtener cualquier matriz densidad 2×2 que queramos. En lugar del par de partículas EPR de espín 1/2 producido en un estado de espín 0, como en el caso recién considerado, las tomamos inicialmente en un estado de espín 1. Para ver cómo funciona esto en un caso concreto, podemos considerar el ejemplo de Lucien Hardy, estudiado en §23.5. Aquí el estado inicial es $|\leftarrow\nearrow\rangle = |\leftarrow\rangle|\nearrow\rangle + |\nearrow\rangle|\leftarrow\rangle$ (en la descripción de Majorana de §22.10, siendo 4/3 la tangente del ángulo entre $\rightarrow$ y $\nearrow$), y supondré que mi colega decide hacer una medida derecha/izquierda en la partícula que llega a Titán. A partir de los resultados de §23.5, observamos que si mi colega obtiene $|\rightarrow\rangle$, entonces el estado que recibo yo aquí en la Tierra es $|\leftarrow\rangle$, mientras que si mi colega obtiene $|\leftarrow\rangle$, entonces el estado que yo recibo es $|\uparrow\rangle$.[29.11] Así pues, si sé que mi colega realizó una medida derecha/izquierda (y sé que el estado inicial era $|\leftarrow\nearrow\rangle$), entonces concluyo que el estado de espín de la partícula que recibo en la Tierra es una mezcla probabilista de $|\leftarrow\rangle$ y $|\uparrow\rangle$. Nótese que $|\leftarrow\rangle$ y $|\uparrow\rangle$ *no son ortogonales*. La ortogonalidad no es un requisito para la mezcla probabilista de estados que componen una matriz densidad, y lo vemos explícitamente en este ejemplo.

¿Cuál es la matriz densidad que utilizaría yo para mi partícula? Podemos calcularla si sabemos los valores de probabilidad para los dos resultados alternativos, $|\rightarrow\rangle$ y $|\leftarrow\rangle$, que puede obtener mi colega. De hecho, estas probabilidades respectivas resultan ser 1/3 y 2/3, de modo

[29.11] ¿Por qué?

que tengo una probabilidad 1/3 de recibir el estado $|\leftarrow\rangle$ y una probabilidad 2/3 de recibir $|\uparrow\rangle$. Mi matriz densidad es, por consiguiente,

$$\boldsymbol{L} = \tfrac{1}{3}|\leftarrow\rangle\langle\leftarrow| + \tfrac{2}{3}|\uparrow\rangle\langle\uparrow|.$$

Expresada en una base arriba/abajo, esta matriz tiene la forma

$$\boldsymbol{L} = \begin{pmatrix} \frac{5}{6} & -\frac{1}{6} \\ -\frac{1}{6} & \frac{1}{6} \end{pmatrix}$$

(tomando $|\leftarrow\rangle = (|\uparrow\rangle - |\downarrow\rangle)\sqrt{2}$). Ciertamente, esta no tiene valores propios iguales, puesto que sus valores propios son, de hecho, $\frac{1}{2} + \frac{1}{6}\sqrt{5}$ y $\frac{1}{2} - \frac{1}{6}\sqrt{5}$.[29.12] La *ontología* particular «$|\leftarrow\rangle$ con probabilidad 1/3 y $|\uparrow\rangle$ con probabilidad 2/3» para esta matriz densidad no es ni mucho menos única. Por ejemplo, es obvio a partir de la simetría entre ← y ↗, en el estado inicial $|\leftarrow\nearrow\rangle$, que si mi colega decide medir en la dirección de ↗ en lugar de la de izquierda/derecha (dirección de ←), entonces mi propia ontología para la matriz densidad $\boldsymbol{D}$ cambiaría mucho, al incluir $|\nearrow\rangle$ y otro estado perpendicular a este. De hecho, se obtendría una ontología diferente para cada diferente dirección de medida posible que mi colega en Titán pudiera escoger.[29.13]

Podrían obtenerse numerosas ontologías más complicadas, dada cualquier matriz densidad concreta, si permitimos que la mezcla probabilista implique tres o más estados diferentes. Una situación semejante aparecería si el estado inicial tuviera espín $\frac{1}{2}n$, para $n > 2$, que se desintegra en una partícula de espín 1/2 dirigida hacia la Tierra y una de espín $\frac{1}{2}n - \frac{1}{2}$ dirigida a Titán, puesto que la medida de espín de mi

[29.12] Deduzca la forma matricial de $\boldsymbol{L}$, verifique que estos son sus autovalores y encuentre los autovectores. El punto $\boldsymbol{L}$ en la esfera de Bloch de la Fig. 29.3 se ha escogido para coincidir con esto. ¿A qué distancia está del centro?

[29.13] Demuestre que cualquier matriz densidad 2 × 2 preasignada puede obtenerse mediante el procedimiento anterior, donde el estado inicial para el par EPR tiene espín 1. ¿Cómo se relacionan las direcciones de espín de los autovectores de la matriz densidad con la descripción de Majorana de dicho estado inicial?

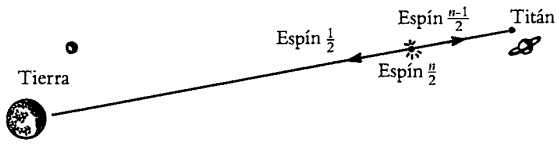


Fig. 29.4. Una matriz densidad puede representar una mezcla probabilística de más estados que la dimensión del espacio. En este ejemplo: en un punto entre la Tierra y Titán, pero más próximo a Titán, un estado inicial conocido de espín $n/2$ (para $n > 2$) se desdobla en una partícula de espín 1/2 dirigida hacia la Tierra y una partícula de espín $\frac{1}{2}(n-1)$ dirigida a Titán. Un colega en Titán mide el valor m de espín de la última, y la probabilidad para cada uno de los n resultados posibles de la medida es un número específico que puede calcularse (en la Tierra), conociendo el estado inicial, de modo que una matriz densidad 2 × 2 específica aparece en la Tierra, compuesta como una mezcla probabilista de n estados. (Esto también se generaliza evidentemente a un espacio de Hilbert de más de dos dimensiones.)

colega admitiría entonces n resultados diferentes, cada uno de ellos con su propia probabilidad (§22.10); véase la Fig. 29.4. Esto también se generaliza evidentemente a situaciones en las que el espacio de Hilbert de estados que utilizo para mi partícula cuando llega a la Tierra es más que 2-dimensional. Todo esto sirve para resaltar que no hay una ontología única de «estados alternativos con probabilidades ponderadas» cualquiera que sea la matriz densidad utilizada.[17] Veremos muy pronto que este hecho es causa de dificultades para la filosofía del punto de vista (c) de decoherencia por el entorno.

Aquí se debería hacer un comentario respecto al cálculo real de una matriz densidad, donde, como antes, parte de la información en un estado entrelazado está oculta (por ejemplo, «en Titán»). Hay un método muy eficaz que se conoce como «sumar sobre los estados desconocidos». Esto se expresa más fácilmente en la notación de índices. Escribamos nuestro estado inicial (el vector ket normalizado $|\psi\rangle$) como $\psi^{\alpha\rho}$, que debe considerarse un estado entrelazado, donde α se refiere a *aquí* (digamos la Tierra) y ρ se refiere a *allí* (digamos Titán); véanse §§23.4,5. El complejo conjugado de este estado (el vector bra $\langle\psi|$) es $\bar{\psi}_{\alpha\rho}$. La normalización del estado es la condición

$$\bar{\psi}_{\alpha\rho}\psi^{\alpha\rho} = 1.$$

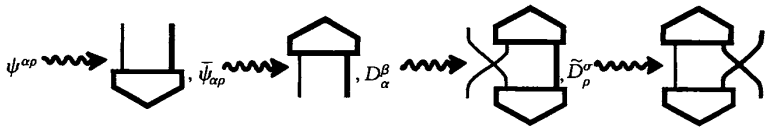


Fig. 29.5. Notación diagramática para las matrices densidad construidas «sumando sobre estados desconocidos». El vector $|\psi\rangle$ normalizado se expresa como $\psi^{\alpha\rho}$, donde «α» se refiere a «aquí» (la Tierra) y «σ» se refiere a «allí» (Titán). El conjugado hermítico (vector bra $\langle\psi|$ es $\bar{\psi}_{\alpha\rho}$ y la normalización es $\bar{\psi}_{\alpha\rho}\psi^{\alpha\rho} = 1$. La matriz densidad utilizada «aquí» es $D_\alpha^\beta = \psi_{\alpha\rho}\psi^{\beta\rho}$, mientras que la utilizada «allí» es $\tilde{D}_\rho^\sigma = \bar{\psi}_{\alpha\rho}\psi^{\alpha\sigma}$.

Entonces, la matriz densidad que yo utilizaría aquí en la Tierra, en ausencia de información de Titán, es la cantidad

$$D_\alpha^\beta = \bar{\psi}_{\alpha\rho}\psi^{\beta\rho}$$

(con una contracción en el índice ρ). En correspondencia, la matriz densidad de mi colega sería $\bar{\psi}_{\alpha\rho}\psi^{\alpha\sigma}$.[29.14] Véase la Fig. 29.5 para la versión diagramática de esto.

29.6. Filosofía FAPP de la decoherencia por el entorno

Las consideraciones anteriores pueden verse como un «preludio» para nuestra investigación del punto de vista (c) de la decoherencia por el entorno, que mantiene que la reducción de estado **R** puede entenderse como resultado de que el sistema cuántico bajo consideración queda inextricablemente entrelazado con su entorno. Para aplicar estas ideas, tomamos el propio sistema como la parte *aquí*, y el entorno como la parte *allí*. Consideramos que el entorno es extraordinariamente complicado y esencialmente «aleatorio», de modo que en la práctica no hay ninguna forma imaginable de extraer la parte *allí* del entorno de la información del estado cuántico total. Por consiguiente, «sumamos sobre los estados desconocidos» en el entorno para obtener

[29.14] Muestre por qué funciona esto. (*Sugerencia*: Escoja una base normal independiente para *aquí* y para *allí* y trabaje en términos de las probabilidades conjuntas para los diversos resultados posibles de medidas en los dos lugares.) Verifique las probabilidades anteriores 1/3 y 2/3 para el caso $|\psi\rangle = |\leftarrow\nearrow\rangle$ que se ha considerado antes.

una descripción mediante la matriz densidad de la parte *aquí* del estado. Se ha trabajado mucho en este tema para demostrar que si el entorno se modela de una forma «razonable», entonces en un período de tiempo muy corto (incluso para un entorno medianamente «ruidoso») la matriz densidad se hace diagonal,

$$\boldsymbol{D} = \begin{pmatrix} p_1 & 0 & \dots & 0 \\ 0 & p_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & p_n \end{pmatrix}$$

hasta un alto grado de aproximación, cuando se expresa en términos de cierta base particular $|1\rangle, |2\rangle, \dots, |n\rangle$, de especial interés.[29.15] Esta se interpreta entonces como una mezcla probabilista

$$\boldsymbol{D} = p_1|1\rangle\langle 1| + p_2|2\rangle\langle 2| + \dots + p_n|n\rangle\langle n|$$

de aquellos estados base particulares que corresponden a los términos diagonales. Se considera que esta mezcla probabilista refleja las alternativas que se dan en el proceso de reducción de estado **R**, y las probabilidades de cada resultado son los respectivos números $p_1, p_2, \dots, p_n$.

Pero, como hemos visto antes, cualquier matriz densidad tiene numerosas interpretaciones ontológicas. Nunca podemos aprender, simplemente a partir de un argumento semejante, que una cualquiera de estas interpretaciones nos proporciona el estado de cosas «real». Además, no podemos deducir entonces que el estado sea *uno* de los $|1\rangle$, $|2\rangle, \dots$, o $|n\rangle$, con probabilidades respectivas $p_1, p_2, \dots, p_n$.

En circunstancias normales, además, se debe considerar la matriz densidad como un tipo de *aproximación* a la verdad cuántica completa, pues no hay ningún principio general que nos proporcione un medio de extraer información detallada del entorno. Quizá una tecnología futura proporcione medios con los que se pueda seguir la pista de las relaciones de fase cuántica detalladas, en circunstancias donde la tecnología actual simplemente «se rinde». ¡Parecería que el recurso a una descripción mediante la matriz densidad es una receta dependiente de

[29.15] De hecho, cualquier matriz densidad es diagonal en *alguna* base. ¿Puede ver por qué, en el caso particular en que todos los autovalores son distintos?

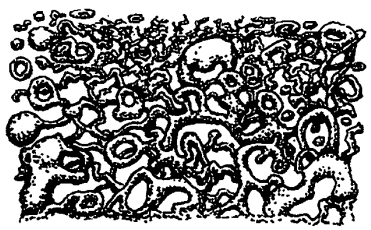


Fig. 29.6. ¿Cuál es la naturaleza del espaciotiempo en la escala de Planck de 10^{-33} cm o 10^{-43} s? Se ha argumentado que fluctuaciones cuánticas en el campo gravitatorio pueden dar como resultado un hervidero de «espuma» con múltiples cambios de topología, y donde las relaciones de fase cuántica detalladas pueden perderse realmente en este nivel.

la tecnología! Con una tecnología mejor, la descripción del vector de estado podría mantenerse más tiempo, ¡y el recurso a una matriz densidad se pospondría hasta que las cosas se hicieran *desesperantemente* embrolladas! Una visión de la realidad física que la considera descrita «realmente» por una matriz densidad parece una visión extraña. Por consiguiente, tales descripciones se conocen a veces como *FAPP*, un acrónimo sugerido por John Bell (el de las famosas desigualdades de Bell; véase §23.3) que significa «para todo propósito práctico». Puede considerarse así que la descripción mediante una matriz densidad es una conveniencia pragmática: algo FAPP, antes que proporcionar una imagen «verdadera» de la realidad física fundamental.

No obstante, podría haber un nivel en el que las relaciones de fase detalladas se pierdan *realmente* debido a algún profundo principio básico primordial. Las ideas en esta dirección suelen apelar a la *gravedad* como algo que tal vez nos lleve a tal principio. A veces podría apelarse a la idea de «fluctuaciones cuánticas en el campo gravitatorio», según la cual la propia estructura del espaciotiempo podría hacerse «espumosa», en lugar de parecerse a una variedad suave (Fig. 29.6) en la «escala de Planck» de 10^{-35} m.[18] (Me referiré a tales ideas en §31.1 y §33.1). Cabría imaginar que a dicha escala las relaciones de fase podrían «perderse en la espuma» inextricablemente. Otra sugerencia, que se debe a Stephen Hawking, es que en presencia de un agujero negro la información acerca del estado cuántico podría ser «engullida» por el agujero y llegar a perderse irrecuperablemente *en principio*. En tales circunstancias, cabría imaginar que un sistema cuántico —que se refiere a una física externa que está entrelazada con una parte que ha caído dentro del agujero— debería estar descrito realmente por una matriz densidad

antes que por un «estado puro».[19] Volveré a estas ideas más adelante, en §§30.4,7,8,14.

29.7. El gato de Schrödinger con la ontología «de Copenhague»

Volvamos al *problema de la medida* mecanocuántica que consiste en cómo podría ocurrir —o parecer que ocurre— **R** cuando se supone que el estado cuántico evoluciona «realmente» de acuerdo con el proceso **U** determinista (§21.8, §§22.1,2 y §23.10). Este problema se presenta frecuentemente de forma muy gráfica como la *paradoja del gato de Schrödinger*. La versión que presento aquí difiere de la versión original de Schrödinger, aunque solo en aspectos no esenciales. Supongamos que hay una fuente de fotones *S* que emite un solo fotón en la dirección de un divisor de haz (espejo semiplateado), donde el estado del fotón se divide en dos partes. En uno de los dos haces emergentes, el fotón encuentra un detector que está acoplado a algún artilugio criminal para matar al pobre gato, mientras que en el otro haz, el fotón escapa y el gato sigue vivo. Véase la Fig. 29.7. (Por supuesto, este es solo un «experimento mental». En un experimento real —tal como el que veremos en §30.13— no hay necesidad de involucrar a una criatura viva. ¡El gato se utiliza solo con un efecto dramático!) Puesto que estas dos alternativas para el fotón deben coexistir en una superposición lineal cuántica, y puesto que la linealidad de la ecuación de Schrödinger (i.e.,

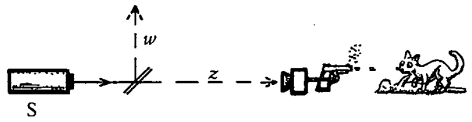


Fig. 29.7. Gato de Schrödinger (modificado del original). Una fuente de fotones S emite un único fotón dirigido hacia un divisor de haz, donde el estado del fotón se divide en una superposición de dos partes. En una de estas, el fotón encuentra un detector que dispara un arma asesina que mata al gato; en el otro, el fotón escapa y el gato vive. La evolución **U** da como resultado una superposición de un gato muerto y un gato vivo.

de **U**) exige que las dos evoluciones temporales subsiguientes deban seguir en una constante superposición ponderada por números complejos, a medida que pasa el tiempo (§22.2) el estado cuántico debe incluir en definitiva una superposición de números complejos tal como la de un gato muerto y un gato vivo, de modo que el gato ¡está muerto y vivo al mismo tiempo!

Por supuesto, una situación semejante es un absurdo para el comportamiento de un objeto del tamaño de un gato en el mundo físico real que experimentamos. ¿Cómo tratan esta paradoja las diferentes interpretaciones «estándar» de la mecánica cuántica? Tomemos el punto de vista (a) de «Copenhague». Por lo que puedo ver, esta interpretación consideraría simplemente que el detector de fotones es un «dispositivo de medida clásico», al que no se le aplican las reglas de la superposición cuántica. El estado del fotón entre su emisión y su detección (o no detección) por el dispositivo está descrito por una función de onda (vector de estado), pero a eso no se le atribuye ninguna «realidad física». La función de onda se utiliza simplemente como una expresión matemática que sirve para calcular probabilidades. Si el divisor de haz es tal que la amplitud del fotón se divide en dos por igual, entonces el cálculo nos dice que hay un 50 % de probabilidades de que el detector registre la llegada del fotón y un 50 % de probabilidades de que no lo haga. Por consiguiente, hay un 50 % de probabilidades de que el gato muera y un 50 % de probabilidades de que siga vivo.

Esta es la respuesta correcta físicamente, donde «físicamente» se refiere al comportamiento del mundo del que realmente tenemos experiencia. Pero esta descripción nos ofrece una imagen de las cosas muy insatisfactoria si queremos seguir los sucesos físicos con mayor detalle. ¿Qué sucede realmente dentro de un detector? ¿Por qué se nos permite tratarlo como un «dispositivo clásico» cuando, después de todo, está construido a partir de los mismos ingredientes cuánticos (protones, electrones, neutrones, fotones virtuales, etc.) que cualquier otra pieza de material físico, grande o pequeña? Puedo entender muy bien que en los primeros días de la mecánica cuántica se necesitase algo parecido a la perspectiva de Niels Bohr sobre el tema para que la teoría pudiera ser utilizada realmente y pudiesen hacerse progresos en física cuántica. Pero, por lo que puedo ver, una perspectiva semejante solo

puede ser temporal, y no resuelve la cuestión de por qué y en qué etapa podría aparecer el «comportamiento clásico» para estructuras grandes y complicadas como los «detectores». Puesto que el punto de vista (a) requiere semejantes «estructuras clásicas» para su interpretación de la mecánica cuántica, solo puede ser una postura «provisional», en la que no se abordan en absoluto las cuestiones más profundas respecto a lo que realmente constituye una medida.

Otra variante de (a) exigiría, en efecto, que el «aparato de medida clásico» sea en última instancia la consciencia del observador. Por consiguiente (si descontamos la consciencia del propio gato), solo cuando un experimentador consciente examina el gato es cuando se ha conseguido la «clasicalidad». Me parece que, una vez que hemos llegado a este nivel, nos vemos llevados a adoptar una posición que está más en la línea de (b) o (f). Si adoptamos el punto de vista de que las reglas **U** de la superposición lineal cuántica siguen siendo válidas hasta el nivel del ser consciente, entonces estamos en el reino de la perspectiva (b) de muchos universos; pero si adoptamos la postura de que **U** falla para seres conscientes, entonces nos vemos llevados a una versión de (f) según la cual entra en juego un nuevo tipo de comportamiento, fuera de las predicciones ordinarias de la mecánica cuántica, cuando intervienen seres que poseen consciencia. Una sugerencia en esta línea fue propuesta realmente por el distinguido físico cuántico Eugene Wigner en 1961.[20]

No obstante, creo que cualquier teoría que exija la presencia de un observador consciente para que pueda efectuarse **R** conduce a una imagen muy sesgada (y yo diría que muy implausible) del universo. Imaginemos un planeta lejano similar a la Tierra pero sin vida consciente, y para el que tampoco hay consciencia en muchos años luz a la redonda. ¿A qué se parece el clima en dicho planeta? Las estructuras climáticas tienen la propiedad de que son «sistemas caóticos», en el sentido de que cualquier estructura particular que se desarrolle dependerá de forma muy crítica de cualquier minúsculo detalle de lo sucedido antes (véase §27.2). De hecho, es probable que en un mes terrestre, pongamos por caso, minúsculos efectos cuánticos lleguen a amplificarse tanto que el clima del planeta entero dependería de ellos. La ausencia de consciencia, según la versión particular de (f)

(o quizá de (a)) en discusión, implicaría que **R** *no ocurre nunca* en dicho planeta, de modo que el clima es en realidad tan solo un revoltijo cuántico superpuesto que no se parece a ningún clima real en el sentido en el que lo conocemos. Pero si una nave espacial que lleve viajeros conscientes, o una sonda con la capacidad de transmisión de una señal a un ser consciente, es capaz de dirigir sus sensores a ese planeta, entonces inmediatamente —y solo entonces— su clima se convierte de repente en un clima ordinario, ¡exactamente como si hubiera sido un clima ordinario todo el tiempo! No hay en esto ninguna contradicción real con la experiencia, pero ¿es esta «realidad de Wigner» una imagen creíble del comportamiento de un universo físico real? Para mí, no lo es; pero puedo entender (apenas) que otros le den más crédito.

29.8. ¿Pueden las ontologías (b) y (c) resolver el «gato»?

¿Qué pasa entonces con el punto de vista (b) de los muchos universos? Aquí la «realidad» de la superposición cuántica de un gato muerto y vivo se *acepta* simplemente (como se aceptaban las estructuras climáticas en superposición cuántica del párrafo anterior); pero esto no nos dice lo que un observador «percibe» realmente al mirar el gato (o el clima). Se considera que el estado de percepción del observador está

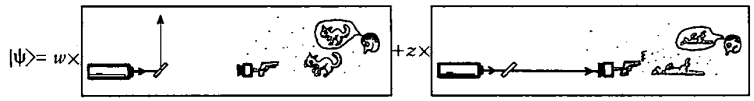


Fig. 29.8. La conclusión de la Fig. 29.7 no se ve afectada por la presencia de entornos diferentes entrelazados con los estados del gato o por las respuestas diferentes de un observador. Así, el estado toma la forma

$$|\Psi\rangle = w\times|\text{gato vivo}\rangle|\text{entorno del gato vivo}\rangle|\text{percibir gato vivo}\rangle + z\times|\text{gato muerto}\rangle|\text{entorno del gato muerto}\rangle|\text{percibir gato muerto}\rangle.$$

Si la evolución **U** tiene que representar realmente la realidad (punto de vista de los muchos universos (b)), entonces debemos adoptar la postura de que la consciencia de un observador puede experimentar solo una u otra alternativa, y «se divide» en experiencias de universo separadas en esta fase.

entrelazado con el estado del gato. El estado de percepción «yo percibo un gato vivo» acompaña al estado «gato vivo», y el estado de percepción «yo percibo un gato muerto» acompaña al estado «gato muerto». Véase la Fig. 29.8. Se supone entonces que un ser perceptivo siempre encontrará que su estado de percepción está en uno de estos dos; por consiguiente, el gato está en el mundo percibido o vivo o muerto. Estas dos posibilidades coexisten en la «realidad» en la superposición entrelazada:

$$|\Psi\rangle = w|\text{gato vivo}\rangle|\text{percibir gato vivo}\rangle + z|\text{gato muerto}\rangle|\text{percibir gato muerto}\rangle.$$

Deseo dejar claro que, tal como está, esto no es ni mucho menos una resolución de la paradoja del gato, pues no hay nada en el formalismo de la mecánica cuántica que exija que un estado de consciencia no pueda estar involucrado en la percepción simultánea de un gato vivo y un gato muerto. En la Fig. 29.9 he ilustrado esta cuestión, donde he considerado la situación simple en la que las dos amplitudes, z y w, para reflexión y transmisión en el divisor de haz, son iguales. Como sucede con el sencillo ejemplo EPR-Bohm con dos partículas de espín 1/2 emitidas en un estado inicial de espín 0, podemos reescribir el estado entrelazado resultante de muchas maneras. En el ejemplo ilustrado en la Fig. 29.9, el estado $|\text{gato vivo}\rangle + |\text{gato muerto}\rangle$ está acompañado por $|\text{percibir gato vivo}\rangle + |\text{percibir gato muerto}\rangle$ y el estado $|\text{gato vivo}\rangle - |\text{gato muerto}\rangle$ está acompañado por $|\text{percibir gato vivo}\rangle - |\text{percibir gato muerto}\rangle$. Esto es exactamente análogo a reescribir el estado $|\Omega\rangle = |\uparrow\rangle|\downarrow\rangle - |\downarrow\rangle|\uparrow\rangle$ como $|\rightarrow\rangle|\leftarrow\rangle - |\leftarrow\rangle|\rightarrow\rangle$, como en §23.4. ¿Por qué no permitimos estos estados de percepción superpues-

$$\sqrt{8}|\psi\rangle = \left(|\quad\rangle + |\quad\rangle\right)\left(|\quad\rangle + |\quad\rangle\right) + \left(|\quad\rangle - |\quad\rangle\right)\left(|\quad\rangle - |\quad\rangle\right)$$

Fig. 29.9. Reformulación de la Fig. 29.8 (en el caso $z = w = \frac{1}{R\sqrt{2}}$, e incorporando el estado del entorno con el del gato) como sigue.

$$\sqrt{8}|\psi\rangle = \{|\text{gato vivo}\rangle + |\text{gato muerto}\rangle\}\{|\text{percibir gato vivo}\rangle + |\text{percibir gato muerto}\rangle\} + \{|\text{gato vivo}\rangle - |\text{gato muerto}\rangle\}\{|\text{percibir gato vivo}\rangle - |\text{percibir gato muerto}\rangle\}.$$

tos? Hasta que sepamos exactamente qué hay en un estado cuántico que le permita ser considerado como una «percepción», y ver en consecuencia que tales superposiciones son «no permitidas», no habremos llegado en realidad a ninguna parte en el intento de explicar por qué el mundo real de nuestras experiencias no puede involucrar superposiciones de gatos vivos y muertos.

A veces la gente pone objeciones a este ejemplo sobre la base de que la igualdad de las amplitudes para las dos alternativas es una situación muy especial, y que en general no hay libertad para reexpresar los estados entrelazados de este modo. Sin embargo, cuando consideramos esta situación con mayor profundidad observamos que el aspecto de «igual amplitud» de este ejemplo particular no tiene realmente ninguna importancia. Es útil tener en cuenta el ejemplo de un par EPR de partículas de espín 1/2, considerado en §29.5. «Igualdad de amplitudes» (en realidad, «igualdad de módulos de amplitudes» $|z| = |w|$) es lo que da lugar a una matriz densidad con valores propios iguales. En §§29.4,5 hemos visto explícitamente que una matriz densidad 2×2 con valores propios desiguales tiene muchas representaciones como una mezcla probabilista de un par de estados, pero el par será en general no ortogonal. De hecho, la ortogonalidad ocurre solo cuando los dos estados son autovectores de la matriz densidad.[29.16] En el caso de «amplitudes iguales» (estrictamente $|z| = |w|$), podemos considerar que los estados $|\text{gato vivo}\rangle$ y $|\text{gato muerto}\rangle$ son ortogonales y, de hecho, los estados acompañantes $|\text{percibir gato vivo}\rangle$ y $|\text{percibir gato muerto}\rangle$ son ortogonales (los «autovectores»). Pero en el caso $|z| \neq |w|$, el par de estados de percepción que acompaña a un par ortogonal de estados de gato superpuestos no será en general ortogonal, y el par de estados de gato que acompaña a un par de estados de percepción ortogonales no será en general ortogonal. No hay nada erróneo en usar cualquiera de estas representaciones del estado total $|\Psi\rangle$. Puesto que **R** no tiene lugar realmente según la posición (b), no hay un estatus especial para las alternativas ortogonales (puesto que nada se «reduce» a ellas en ningún caso).

De hecho, resulta que en el caso general habrá un único par de es-

[29.16] Demuéstrelo.

tados ortogonales que acompañan a un par de estados del gato ortogonales. Esto es algo conocido como la *descomposición de Schmidt* de un estado entrelazado.[21] Sin embargo, esto no sirve de mucho para resolver la paradoja de la medida (pese a la popularidad de la descomposición de Schmidt en relación con la teoría de la información cuántica),[22] porque en general este par de estados del gato «matemáticamente preferidos» (autoestados de la matriz densidad del gato) no serían los deseados $|\text{gato vivo}\rangle$ y $|\text{gato muerto}\rangle$, sino alguna no deseada superposición lineal de ellos. Podemos ver que estos autoestados de la matriz densidad que ocurren en una descomposición de Schmidt no necesitan tener nada que ver con las expectativas de lo que debería ser «ontológicamente real» si miramos de nuevo el ejemplo de Lucien Hardy considerado en §29.5. Encontramos (véase el ejercicio [29.12]) que los autovectores de la matriz densidad (para la partícula que yo recibo aquí en la Tierra) ¡son muy diferentes de las alternativas $|\leftarrow\rangle$ y $|\uparrow\rangle$ que son «alternativas macroscópicamente distinguibles» según las medidas de mi colega en Titán!

Puesto que las matemáticas por sí solas no discriminarán los estados «$|\text{gato vivo}\rangle$» y «$|\text{gato muerto}\rangle$» como estados «preferidos», seguimos necesitando una teoría de la percepción antes de que podamos dar sentido a (b), y carecemos de una teoría semejante.[23] Más aún, una obligación de una teoría semejante sería no solo explicar por qué no ocurren superposiciones de gatos muertos y gatos vivos (o de cualquier otra cosa macroscópica) en el mundo percibido, sino también por qué la maravillosa y extraordinariamente precisa regla del módulo al cuadrado ¡da realmente las respuestas correctas en mecánica cuántica! Una teoría de percepción que pudiera hacer esto necesitaría ser ella misma tan precisa como la teoría cuántica. Los defensores de (b) no han estado nunca próximos a sugerir un esquema semejante.[24]

Volvamos ahora a los intentos de resolver la paradoja del gato mediante decoherencia por el entorno (c). Consideremos que la emisión inicial del fotón es ontológicamente real. (La fuente podría estar preparada para registrar este suceso de una forma macroscópica.) Entonces, una vez que el fotón ha encontrado el divisor de haz, tenemos una superposición ontológicamente real del mismo en los dos haces. La parte transmitida del estado del fotón evoluciona hacia un gato muer-

to, junto con su entorno, y la parte reflejada evoluciona hacia un gato vivo, junto con un entorno diferente. Por el momento, la ontología es aún la superposición de los dos. Las alternativas del entorno, al ser «inobservables», se han de sumar, llevándonos a una matriz densidad 2×2. Ahora la posición ontológica cambia sigilosamente, y la «realidad» llega a describirse por la propia matriz densidad. El argumento de la decoherencia por el entorno afirma que esta matriz llega a hacerse rápidamente casi diagonal, con extraordinaria aproximación, en la base $(|\text{gato vivo}\rangle, |\text{gato muerto}\rangle)$, de modo que hay otro cambio subrepticio en la ontología, y el estado se convierte en una mezcla probabilista de $|\text{gato vivo}\rangle$ y $|\text{gato muerto}\rangle$. ¡Así es como se nos ha «permitido» pasar con este cambio de ontología de la superposición

$$w|\text{gato vivo}\rangle|\text{entorno del gato vivo}\rangle + z|\text{gato muerto}\rangle|\text{entorno del gato muerto}\rangle$$

a las alternativas $|\text{gato vivo}\rangle$ o $|\text{gato muerto}\rangle$! Recordemos que no hay unicidad en la interpretación ontológica de una matriz densidad de una mezcla probabilista de estados (sean o no iguales los autovalores). De hecho, pasar a la mezcla de $|\text{gato vivo}\rangle$ y $|\text{gato muerto}\rangle$ representa un genuino (doble) cambio de ontología desde la superposición original. La posición (c) es, de hecho, FAPP, y no nos da ninguna ontología consistente para la realidad física.

29.9. ¿Qué ontologías no convencionales pueden ayudar?

Debería hacer un breve comentario sobre (d) y (e). Si se adopta la ontología «extravagante» para la aproximación (d) de las historias consistentes, en la que la realidad se representa como la totalidad del conjunto de historias consistentes máximamente refinadas, entonces puede plantearse una crítica que es algo similar a la del caso (b) de los muchos mundos. Como sucede con (b), se necesita una teoría detallada y precisa de los perceptores conscientes para que (d) pueda evocar una imagen que sea consistente con el mundo físico que conocemos. Se han hecho intentos en esta dirección (proporcionados por la noción de un IGUS, «sistema de recogida y utilización de información»), pero por el

momento están muy lejos de ser suficientes.[25] Alternativamente, se podría preferir algo como la ontología más económica sugerida en §29.2, en la que un único conjunto de historias consistentes máximamente refinado se considera un candidato plausible para una ontología del «mundo real». Sin embargo, esta (igual que la más extravagante ontología anterior) depende de que el criterio de «historia consistente» consiga realmente aquello para lo que estaba diseñado, a saber, discriminar historias que se parezcan al tipo de mundo en el que vivimos. No obstante, como demostraron Dowker y Kent en 1996, esta condición de «consistencia» por sí sola está lejos de ser adecuada. Parece que se requieren algunos criterios adicionales.

En mi opinión, una desventaja importante para (d) es que, a pesar de la introducción de procesos tipo **R** (mediante la inserción de conjuntos de proyectores), no parece acercarnos más a una comprensión de lo que es realmente una medida que lo que lo hacen las ontologías más convencionales de (a) o (b). De hecho, en (d) los procedimientos tipo **R** son establecidos explícitamente para no tener nada que ver directamente con medidas físicas reales. Mi problema con esto es que eliminando la asociación entre estos reemplazamientos tipo **R** y las medidas físicas no sacamos ninguna idea de lo que realmente constituye una medida física. ¿Por que, según (d), no vemos realmente cosas como gatos de Schrödinger en un limbo superpuesto entre la vida y la muerte? La teoría no parece ofrecer ninguna mejora sobre la posición de Copenhague estándar (a) para explicar qué sistemas (tales como piezas de aparatos físicos o gatos) deberían comportarse clásicamente, mientras que los neutrones o los fotones no lo hacen. El requisito de «consistencia» para historias de grano grueso (máximamente refinadas) parece estar lejos de lo que se necesita para ofrecer un modelo[26] para la realidad física observada.

Aunque es una característica positiva de (d) el hecho de que hace un serio intento por incorporar procesos tipo **R** en un nivel fundamental, los criterios que se han propuesto hasta ahora no hacen suficiente por estrechar el comportamiento del modelo de modo que pueda aparecer una imagen algo parecida al mundo que conocemos. Esto parece ser cierto tanto en el nivel «tipo clásico» macroscópico (como he comentado antes, en relación con el análisis de Dowker-Kent del

criterio de «historia consistente») como en el «nivel cuántico» en el que uno esperaría ver una evolución unitaria no perturbada. Puesto que la paradoja de la medida está relacionada con el aparente conflicto entre el comportamiento físico en estos dos niveles diferentes, es difícil ver cómo el punto de vista (d) de las historias consistentes esté, pese a todo, en posición de arrojar mucha luz sobre esta paradoja.

¿Qué pasa con (e)? Como se ha comentado en §29.2, el punto de vista (e) de la onda piloto de De Broglie-Bohm parece tener la ontología más clara entre todas aquellas que no alteran realmente las predicciones de la teoría cuántica. Pese a todo, en mi opinión no aborda realmente la paradoja de la medida de una forma más satisfactoria que las otras. Tal como lo veo, (e) puede sacar ventaja conceptual de sus dos niveles de realidad, pues tiene un nivel de «partícula» más firme de la realidad de la configuración del sistema, así como un nivel de realidad secundario «de onda», definido por la función de onda ψ, cuyo papel es guiar el comportamiento del nivel más firme. Pero no tengo claro cómo podemos estar seguros, en cualquier situación experimental real, de a qué nivel deberíamos estar apelando. Mi dificultad reside en que no hay ningún parámetro que defina qué sistemas son «grandes», en el sentido apropiado, de modo que coincidan con imágenes «tipo partícula» o «tipo configuración» más clásicas, y qué sistemas son «pequeños», de modo que cobre importancia el comportamiento «tipo función de onda» (y esta crítica se aplica también a (d)). Sabemos de §23.4, etc., que el comportamiento cuántico puede extenderse sobre distancias de decenas de kilómetros al menos, de modo que no es solo la distancia física la que nos dice cuándo un sistema deja de parecer mecanocuántico y empieza a comportarse como una entidad clásica. Pero, de todas formas, hay un sentido en el que un objeto grande (como un gato) parece no estar de acuerdo con las leyes cuánticas unitarias a pequeña escala. (En §30.11 explicaré mis propias ideas respecto al tipo de «medidas de escala» que serán necesarias.) Pero ya crea uno o no que cualquier medida concreta semejante es apropiada, pienso que *alguna* medida de escala es necesaria para definir en qué momento el comportamiento de tipo clásico empieza a tomar el mando a partir de la actividad cuántica a pequeña escala. En común con las otras ontologías cuánticas en las que no se espera ninguna desviación medible de la

mecánica cuántica estándar, el punto de vista (e) no posee una medida de escala semejante, de modo que no veo cómo puede abordar de manera adecuada la paradoja del gato de Schrödinger.

En relación con esta cuestión, quizá sea apropiado un comentario general respecto a los intentos por «obtener» la ocurrencia aparente de **R** a partir de la dinámica de (digamos) **U**. Podemos ver que la dinámica ordinaria (determinista) por sí sola nunca puede conseguir esto, como es evidente, aunque solo sea por la razón de que no hay probabilidades en una ecuación dinámica como la ecuación de Schrödinger. (Remito al lector a la discusión de §27.1.) Se necesita también algún principio probabilista. Después de todo, **R** es una ley probabilista. Así pues, como se ha comentado en §29.2, es realmente un ingrediente esencial de (e) el que las sucesivas probabilidades de medidas estén correctamente codificadas en la elección de (digamos) el estado inicial.

Esto nos deja con (f). Las dificultades principales con la mayoría de las muchas propuestas diferentes (a menudo heroicas) de una «**R** objetiva» residen en su apariencia poco natural, su carácter esencialmente no relativista, su necesidad de introducir parámetros arbitrarios que no están justificados por la física conocida, sus violaciones de la ley de conservación de la energía y, en algunos casos, su conflicto directo con la observación. No sería oportuno que discuta aquí todas estas propuestas, y sería poco honesto que discriminase algunas de ellas en detrimento de las otras. De hecho, adoptaré el procedimiento de ser uniformemente injusto con todas las propuestas que otros han presentado ¡imponiendo al lector, en el capítulo 30, la propuesta (en algunos aspectos minimalista) que para mí tiene más probabilidad de ser correcta (con disculpas para muchos de mis amigos)! De hecho, ha habido un estímulo y aportaciones muy importantes por parte de varias propuestas que otros han presentado antes, y me referiré a estas (con la gratitud debida), pero solo en relación con la ideas específicas que deseo argumentar.

Notas

Sección 29.1

29.1. Véase Deutsch (2000).

29.2. Debo este término a mi colega clasicista Peter Derow. Véase Penrose (1987a).

29.3. Véanse Everett (1957), Wheeler (1957), DeWitt y Graham (1973) y Deutsch (2000).

29.4. Algunos físicos argumentan que no hay «ningún problema» en la superposición cuántica de estados macroscópicamente diferentes —como el superpuesto gato vivo y muerto de Schrödinger que se ha comentado en §§29.7-9—, porque sencillamente sería «demasiado caro» (o una imposibilidad práctica) diseñar un experimento para detectar interferencias entre los estados vivo y muerto. Esto, de nuevo, es tomar una posición «pragmática» que no aborda en realidad las cuestiones ontológicas que nos interesan aquí. Personalmente, ubicaría a tales físicos, en general, en la categoría (c).

29.5. Véase Hawking y Penrose (1996), p. 121.

Sección 29.2

29.6. Para (d), véase Gell-Mann y Hartle (1995); para (e), véase Bohm y Hiley (1994). La lista (a), (b), (c), (d) es solo representativa, y hay muchos matices diferentes dentro de los puntos de vista que he enumerado. Por ejemplo, algunos han expresado la idea (por ejemplo, Sorkin, 1994) de que la «realidad cuántica» se entiende mejor en términos de las integrales de camino y/o diagramas de Feynman que hemos visto en §§26.6-11. Por lo que puedo deducir, esta familia concreta de ontologías pertenecería a la clase general cubierta por (b) (aunque teniendo algunos elementos importantes en común con (d)), según la cual a una superposición particular que define el «estado cuántico» (o «historia cuántica») se le asignaría estatus de «realidad». Debería mencionar también las ontologías «transaccionales» de Aharonov y Vaidman (2001), Cramer (1988), Costa de Beauregard (1995) y Werbos y Dolmatova (2000), según las cuales una función de onda con propagación de Schrödinger hacia el futuro de la última medida junto con otra función de onda con propagación de Schrödinger hacia el pasado de la próxima medida se reúnen en la descripción de la realidad (véase §30.3). Sin

embargo, no veo que en ninguna de estas alternativas, sin otros ingredientes, se resuelva la paradoja de la medida mejor que en (a), (b), (c), (d) o (e).

29.7. El formalismo (d) permite también que el «estado de partida» pudiera ser una matriz densidad (véase §29.3).

29.8. A veces a esto se le llama simplemente una «historia», pero esto podría confundirse con el uso de dicho término en la «suma sobre historias» de Feynman de §26.6.

29.9. Esta es una condición del tipo siguiente. Supongamos que tenemos una sucesión dada de conjuntos de proyectores (y supongamos por el momento $\mathcal{H} = 0$); entonces construimos la expresión $X = \langle\psi_0|\boldsymbol{E}'\boldsymbol{F}' \ldots \boldsymbol{K}'\boldsymbol{L}'\boldsymbol{D}_\infty\boldsymbol{L}\boldsymbol{K} \ldots \boldsymbol{F}\boldsymbol{E}|\psi_0\rangle$, donde $|\psi_0\rangle$ es el estado inicial y donde el estado final podría tomarse como una matriz densidad $\boldsymbol{D}_\infty$ (véase la nota 29.7). Los proyectores en los pares sucesivos $(\boldsymbol{E}, \boldsymbol{E}')$, $(\boldsymbol{F}, \boldsymbol{F}')$, ..., $(\boldsymbol{K}, \boldsymbol{K}')$, $(\boldsymbol{L}, \boldsymbol{L}')$ pertenecen, respectivamente, a la sucesión dada de conjuntos de proyectores. La condición de consistencia exige que la parte real de X se anule cuando cualquiera de los pares $(\boldsymbol{E}, \boldsymbol{E}')$, $(\boldsymbol{F}, \boldsymbol{F}')$, ..., $(\boldsymbol{K}, \boldsymbol{K}')$, $(\boldsymbol{L}, \boldsymbol{L}')$ es desigual. Este es estrictamente el caso solo cuando se ha ignorado la parte de Schrödinger de la evolución (i.e., tomamos $\mathcal{H} = 0$), pero podemos reinstaurar una evolución de Schrödinger no trivial introduciendo esta evolución apropiadamente entre las aplicaciones de los proyectores. Esta «condición de consistencia» sobre historias de grano grueso puede interpretarse como la condición de «no interferencia» entre las historias que se comparan.

29.10. De hecho, no he localizado un enunciado claro de *ninguna* «(d)-ontología» realmente pretendida en la literatura de las historias consistentes. Lo que estoy presentando aquí es solo mi propia tentativa de entender esta cuestión, basada en extensas discusiones con Jim Hartle y, más en particular, una fructífera correspondencia con Fay Dowker. Es probable que, pese a mis esfuerzos, aún no esté presentando adecuadamente una ontología subyacente defendida por la comunidad «(d)».

29.11. Véanse Bohm y Hiley (1995) y Valentini (2002). Antony Valentini tiene también un libro en curso sobre la teoría de De Broglie-Bohm, ¡que esperamos se publique pronto!

29.12. Véanse Károlyházy (1974), Frenkel (2000), Ghirardi *et al.* (1986), Ghirardi *et al.* (1990), Komar (1964), Pearle (1985), Pearle y Squires (1995), Kibble (1981), Weinberg (1989), Diósi (1984, 1989), Percival (1994,

1995), Gisin (1989, 1990), Penrose (1986a, 1989, 1996a, 2000a) y Leggett (2002), sin ningún orden particular.

Sección 29.3

29.13. Para una distribución de probabilidad continua, necesitamos una función univaluada de valor real f sobre $\mathcal{P}$, cuya integral es 1. El espacio $\mathcal{P}$ tendría una forma de volumen natural —la $2N$-forma $\boldsymbol{\Sigma}$ de §20.4 que intervenía en el teorema de Liouville—, de modo que $f\boldsymbol{\Sigma}$ puede integrarse legítimamente sobre $\mathcal{P}$, y nuestra condición requerida es realmente $\int f\boldsymbol{\Sigma} = 1$.

29.14. Véase Brody y Hughston (1998b). Nielsen y Chuang (2000) ofrecen un buen cubrimiento conceptual de la matriz densidad en la práctica.

Sección 29.4

29.15. Para un sistema de n estados, con $n > 2$, la imagen es más complicada. Solo parte de la frontera del espacio $(n^2 - 1)$-dimensional de matrices densidad es el espacio de estados puros, siendo esta parte un $(n-1)$-*espacio proyectivo complejo* $\mathbb{CP}^{n-1}$ (cf. §21.9 y §22.9).

Sección 29.5

29.16. Quizá el lector se está preguntando cómo podría afectar la noción de *cuanlazamiento*, introducida en §23.10, a estas cuestiones ontológicas. Esta es una pregunta intrigante, y muy bien podría ser que la cuestión global de la «ontología» en un contexto cuántico tenga que verse en definitiva a una nueva luz. Pero por el momento adoptemos simplemente una actitud más de «sentido común» hacia la realidad, en la que no entrarán las cuestiones planteadas por la relatividad.

29.17. Nielsen y Chuang discuten este punto; véase también Hughston *et al.* (1993).

Sección 29.6

29.18. La idea se debe originalmente (como tantas cosas) a Wheeler; véase Ng (2004) para una perspectiva moderna.

29.19. Véanse Hawking (1976b) y Preskill (1992); véase también §30.14.

20.20. No estoy seguro de si este punto de vista representaba la posición real de Wigner con respecto a la medida cuántica, que, después de todo, puede haber evolucionado durante su vida. Debería señalar también que mi posición difiere fundamentalmente de aquellas, como la aquí

mencionada, que afirman que es la consciencia la que reduce el estado. (En relación con esto, mi opinión ha sido a veces malinterpretada por otros comentaristas.) Véanse §§30.9-12.

Sección 29.8

29.21. La descomposición de Schmidt (o *polar*) de un estado entrelazado $|\Psi\rangle$ perteneciente a $\mathbf{H}^2 \times \mathbf{H}^2$, lo expresa (de forma esencialmente unívoca) como $|\Psi\rangle = \lambda|\alpha\rangle|\beta\rangle + \mu|\rho\rangle|\sigma\rangle$, donde $|\alpha\rangle$ y $|\rho\rangle$, que pertenecen al primer $\mathbf{H}^2$, son ortogonales (autoestados normalizados de su matriz densidad), y $|\beta\rangle$ y $|\sigma\rangle$ corresponden análogamente al segundo $\mathbf{H}^2$. Aquí $\bar{\lambda}\lambda$ y $\bar{\mu}\mu$ son autovalores de la matriz densidad. Una expresión similar es válida para $\mathbf{H}^n \times \mathbf{H}^n$, donde la suma en $|\Psi\rangle$ tiene n términos. Véase Nielsen y Chuang (2000).

29.22. Véase Nielsen y Chuang (2000), ¡que trata, después de todo, de la teoría de la información cuántica!

29.23. Véase Page (1995) para una discusión de estas cuestiones.

29.24. Véanse Gell-Mann (1994) y Hartle (2004) para un corte de diez años de tales ideas.

Sección 29.9

29.25. Véase Dowker y Kent (1996).

29.26. Un ejemplo sorprendente que se debe a Adrian Kent muestra con claridad lo insuficiente que es la condición de «consistencia» para ofrecer una imagen físicamente plausible de la «realidad». En este ejemplo, una partícula p puede estar en una de tres cajas A, B, C, descritas por los respectivos estados ortogonales normalizados $|A\rangle$, $|B\rangle$, $|C\rangle$. El hamiltoniano lo tomamos cero, lo que da una evolución unitaria constante. El estado inicial va a ser $|A\rangle + |B\rangle + |C\rangle$ y suponemos que el estado final que se mide es $|A\rangle + |B\rangle - |C\rangle$. (Esto es posible porque $|A\rangle + |B\rangle + |C\rangle$ y $|A\rangle + |B\rangle - |C\rangle$ no son ortogonales.) La inserción del conjunto de proyectores $\{|A\rangle\langle A|, I - |A\rangle\langle A|\}$ entre los dos resulta ser «consistente», y parece que concluimos que p debe estar en la caja A en esta etapa intermedia (básicamente porque $|B\rangle + |C\rangle$ y $|B\rangle - |C\rangle$ *son* ortogonales). El mismo argumento, con B en lugar de A, llega a la conclusión de que ¡p debe estar en B en la etapa intermedia! Este ejemplo parece haberse desarrollado a partir del «problema del rey» de Yakir Aharonov»; véase Albert *et al.* (1985), p. 5.

30

El papel de la gravedad en la reducción del estado cuántico

30.1. ¿Va a quedarse aquí la teoría cuántica actual?

En este capítulo argumentaré que existen razones poderosas y positivas, aparte de las negativas presentadas en el capítulo precedente, para creer que las leyes de la mecánica cuántica actual necesitan un cambio fundamental (aunque presumiblemente sutil). Estas razones proceden de principios físicos aceptados y de hechos observados acerca del universo. Pese a todo, encuentro muy significativo que haya tan pocos físicos cuánticos actuales que estén dispuestos a mantener en serio la idea de un cambio real en las reglas básicas de su disciplina. La mecánica cuántica, a pesar de su notable apoyo experimental sin excepciones y de sus predicciones confirmadas de manera sorprendente, es una disciplina relativamente joven, pues solo tiene unos tres cuartos de siglo (si la datamos a partir del establecimiento de la teoría matemática por Dirac y otros, basada en los esquemas de Heisenberg y Schrödinger, en los años inmediatamente posteriores a 1925). Cuando digo «relativamente», estoy comparando la teoría cuántica con la de Newton, que perduró unos 300 años antes de que necesitase una modificación seria en la forma de la teoría de la relatividad especial y luego la general, y de la mecánica cuántica. Incluso si admitimos que la teoría de Newton sufrió su primera modificación con la introducción de los campos maxwellianos, ¡esto aún le dio un reinado de más de 175 años libre de excepciones!

Además, la teoría de Newton no tenía *una paradoja de la medida*. Mientras que la linealidad del proceso **U** de la teoría cuántica da a dicha teoría una elegancia particular, es esa misma linealidad (o unitariedad) la que nos lleva directamente a la paradoja de la medida (§22.2). ¿No resulta razonable creer que esta linealidad pudiera ser una aproximación a alguna *no* linealidad más precisa (pero sutil)?

Tenemos un precedente claro. La teoría gravitatoria de Newton tiene una elegancia matemática especial porque las fuerzas gravitatorias siempre se suman de una forma completamente lineal. Sin embargo, en la teoría más precisa de Einstein esta linealidad se ve sustituida por un tipo característicamente sutil de no linealidad en la forma en que se combinan los efectos gravitatorios de cuerpos diferentes. Y la teoría de Einstein no carece de elegancia, de un tipo completamente diferente de la de Newton. En la teoría de Einstein también vemos que las modificaciones en la teoría de Newton que fueron necesarias no eran nada parecido al «retoque» al que me he referido en §29.2. En diversos momentos se han sugerido retoques semejantes en la teoría de Newton, tales como reemplazar la potencia 2 en la fórmula de la inversa del cuadrado de Newton GmM/r^2 (véase §17.3) por 2.000 000 16, como sugirió Aspeth Hall en 1894 para acomodar aquellas ligerísimas desviaciones, descubiertas en 1843, respecto a las predicciones newtonianas detalladas del movimiento de Mercurio alrededor del Sol (y la sugerencia de Hall da también buenos ajustes para los otros planetas, como demostró Simon Newcombe).[1] La teoría de Einstein explicó posteriormente dichas desviaciones, sin ningún revuelo, pero la nueva teoría no se obtuvo ni mucho menos con solo retocar la vieja, sino que implicaba un cambio de perspectiva completamente radical. Este es, en mi opinión, el tipo general de cambio en la estructura de la mecánica cuántica al que debemos mirar si queremos obtener la (en mi opinión) necesaria teoría no lineal que reemplace a la teoría cuántica convencional actual.

De hecho, mi perspectiva es que la relatividad general de Einstein proporcionará por sí misma algunas claves necesarias respecto a las modificaciones que se requieren. El siglo xx nos trajo dos revoluciones capitales en el pensamiento físico, y a mi modo de ver la relatividad general ha supuesto una revolución tan impresionante como la teoría cuántica (o la teoría cuántica de campos). Pese a todo, estos dos grandes esquemas están basados en principios difícilmente compatibles. La perspectiva usual con respecto al matrimonio propuesto entre estas

teorías es que una de ellas, a saber, la relatividad general, debe someterse a la voluntad de la otra. Al parecer, es una idea común que las reglas de la teoría cuántica de campos son inmutables y es la teoría de Einstein la que debe modificarse de manera adecuada para encajar en el molde cuántico estándar. Pocos sugerirían que las reglas cuánticas deben admitir una modificación para asegurar un matrimonio armonioso. De hecho, el mismo nombre «gravedad cuántica», que normalmente se asigna a la unión propuesta, conlleva la connotación implícita de que lo que se busca es una teoría *cuántica* (de campos) estándar. Pese a todo, ¡yo diría que hay evidencia observacional a favor de que la idea que tiene la naturaleza de esta unión es muy diferente! Sostengo que su diseño para esta unión debe ser lo que a nuestros ojos sería uno claramente no estándar, y que una de sus características importantes debe ser una reducción de estado objetiva.

30.2. Claves de una asimetría temporal cosmológica

¿Cuál es esta evidencia? Dirijámonos, en primer lugar, a aquellos lugares donde con más claridad se manifiesta la elección por la naturaleza de una unión cuántico/gravedad. Me refiero a las *singularidades* espaciotemporales del big bang y de los agujeros negros (y también del big crunch, si tal cosa va a tener lugar). En el capítulo 27 se ha expuesto la naturaleza extraordinariamente especial del big bang, en abierto contraste con la naturaleza en apariencia «genérica» de las singularidades de colapso. A pesar de las atrevidas sugerencias hechas de acuerdo con la propuesta de Hartle-Hawking (discutidas en §28.9), no veo escapatoria a que una flagrante asimetría temporal sea una característica necesaria de la unión cuántico/gravedad de la naturaleza.

Tal asimetría temporal parecería estar en completo contraste con las implicaciones de cualquier teoría cuántica de campos estándar. Consideremos, por ejemplo, el teorema CPT, señalado en §25.4. (Recordemos que «T» representa inversión temporal, mientras que «P» y «C» representan, respectivamente, reflexión espacial y reemplazamiento de partículas por sus antipartículas.) Si creemos que el teorema CPT se aplica a nuestra deseada unión cuántico/gravedad, entonces estamos en

dificultades. Si aplicamos CPT a cualquier singularidad final «genérica» de colapso gravitatorio, entonces obtenemos que una posibilidad para el big bang (o para parte del big bang) es una singularidad del tipo inicial. Recordemos la enormidad del espacio de fases disponible, como se ha descrito en §27.13 (y se ha ilustrado gráficamente en la Fig. 27.21). Una vez que se hacen *admisibles* tales singularidades iniciales «genéricas», entonces no hay nada que guíe a la aguja del Creador en dicha región $\mathcal{B}$ absurdamente (y, desde la perspectiva «antrópica», en §28.6, innecesariamente) minúscula, que parece haber sido el punto de partida real de nuestro universo. Creo que está claro que el misterio de la naturaleza extraordinariamente especial del big bang no puede resolverse dentro del marco estándar de la teoría cuántica de campos.

Este sería al menos el caso para cualquier teoría para la que la palabra «estándar» implica la validez del teorema CPT (§25.4). Estrictamente hablando, este teorema no es inmediatamente aplicable a una teoría que respete por completo el espaciotiempo curvo base de la relatividad general de Einstein. Una de las premisas del teorema CPT es que el espaciotiempo de fondo es el espacio de Minkowski plano. De todas formas, sospecho que la mayoría de los físicos considerarán esto como un «tecnicismo» sin importancia, adoptando el punto de vista de que la teoría de Einstein se puede reexpresar, si así se desea, en la forma de una «teoría de campos invariante por Poincaré» introduciendo un fondo de Minkowski como conveniencia. Personalmente, tengo grandes reservas sobre este tipo de procedimientos;[2] pero estaría de acuerdo en que parece poco probable que la teoría de la relatividad general de Einstein clásica y completamente simétrica respecto al tiempo se hiciera tan asimétrica cuando se someta a los procedimientos estándar con simetría temporal de la teoría cuántica de campos.

Por otra parte, recordemos que en §25.5 y §§26.5,11 hemos visto situaciones donde una simetría de la teoría clásica se rompe cuando pasamos a la teoría cuántica. ¿Podría ser que esto es lo que sucede cuando la teoría de Einstein se introduce adecuadamente dentro del alcance de las reglas estándar QFT? Supongo que es concebible, pero resulta difícil ver cómo podría esto parecerse mucho al tipo de ruptura de simetría que ocurre, por ejemplo, en la teoría electrodébil, donde se

considera que el «estado vacío» $|\Theta\rangle$ no comparte las simetrías de la dinámica cuántica. Para que esta idea funcione $|\Theta\rangle$ tiene que ser «asimétrico respecto al tiempo». No estoy seguro de cómo se podría dar sentido a una idea de este tipo. Es cierto que el ket $|\Theta\rangle$ tendría que colocarse a la derecha de todos los operadores de campo, a la manera descrita en §26.11, y podría considerarse que representa el estado inicial del universo, que aquí quiere decir el muy concreto estado big bang. Pero en la QFT estándar, el complejo conjugado de $|\Theta\rangle$, a saber el bra $\langle\Theta|$, también intervendría en el formalismo, pues es necesario para la formulación de probabilidades vía expresiones como $\langle\Theta|A|\Theta\rangle$, y desempeñaría un papel completamente simétrico al de $|\Theta\rangle$ pero con el tiempo invertido. Así pues, $\langle\Theta|$ tendría que representar el estado final del universo, y tenemos un estado final de una estructura similar al inicial, en total contradicción con el mensaje entero del capítulo 27.

Hay también otras características que aparecen en el proceso de «cuantización», por las que la teoría cuántica podría no compartir las simetrías de la teoría clásica. Se conocen como *anomalías*. Estas ocurren cuando las reglas de conmutación clásicas que proporcionan la simetría clásica (dada por los paréntesis de Poisson; véase §14.8) no pueden ser completamente realizadas por conmutadores cuánticos, y solo un subgrupo del grupo de simetría clásico completo sobrevive en la teoría cuántica. Parece que las anomalías deben considerarse normalmente como cosas a evitar (y veremos los malabarismos que tienen que realizar a veces los teóricos para eliminarlas cuando consideremos la teoría de cuerdas en el capítulo siguiente). Pese a todo, podríamos imaginar que se adopta un punto de vista diferente y se considera una anomalía como algo «bueno» en aquellas circunstancias en que la simetría mayor es algo que no deseamos. Sin embargo, en nuestro caso presente es una simetría discreta, a saber CPT, además de T, CT y PT —en realidad cualquier cosa que contenga una «T»—, la que hay que violar, y es difícil ver la relevancia de la idea de anomalía usual, que habitualmente (pero no siempre) se refiere solo a las simetrías continuas que pueden realizarse en términos de paréntesis de Poisson.

Como quiera que se mire, es difícil evitar la conclusión de que en aquellas circunstancias extremas en que los efectos cuánticos y los efectos gravitatorios deben unirse —en las singularidades espaciotemporales en el big bang y en el colapso gravitatorio— la gravedad se comporta *de forma diferente* de otros campos. Recordemos la conclusión final en el penúltimo párrafo del capítulo 27 respecto a este punto. Por la razón que sea, la naturaleza ha impuesto una flagrante asimetría temporal en el comportamiento de la gravedad en circunstancias tan extremas.

30.3. Asimetría temporal en la reducción del estado cuántico

¿Tiene esto relación con cualquier otra clave respecto a la posible interrelación entre gravedad y mecánica cuántica? Así lo creo firmemente. Mientras que no percibimos ninguna asimetría temporal en la parte **U** de la teoría cuántica (§27.1), *sí hay* una asimetría temporal esencial en **R**. Podemos ver esto en un sencillo experimento cuántico hipotético. Supongamos que hay una fuente de fotones S que emite fotones de uno en uno y que cada vez que lo hace este suceso queda registrado.[3] Supongamos también que los fotones son de alta energía, posiblemente fotones de rayos X o incluso rayos γ. Los fotones se dirigen a un divisor de haz B (un «espejo semiplateado») colocado a un ángulo de 45° respecto al haz, de modo que si un fotón se transmite, entonces activa un detector D al otro lado, mientras que si el fotón se refleja, entonces queda absorbido en el techo C (véase la Fig. 30.1). Estoy suponiendo amplitudes iguales para estas dos alternativas, de modo que el detector registrará la llegada del fotón solo en la mitad de las ocasiones en que se registra que la fuente lo ha emitido.

Esto es solo una aplicación directa del procedimiento **R**. Hay una amplitud $\frac{1}{\sqrt{2}}$ (ignorando posibles factores de fase) para la historia del fotón SBD y una amplitud $\frac{1}{\sqrt{2}}$ para la historia del fotón SBC. La aplicación de la regla del módulo al cuadrado para **R** da entonces la respuesta (correcta), según la cual cada vez que hay un suceso de emisión en S, hay una probabilidad de un 50 % de un suceso de detección en D

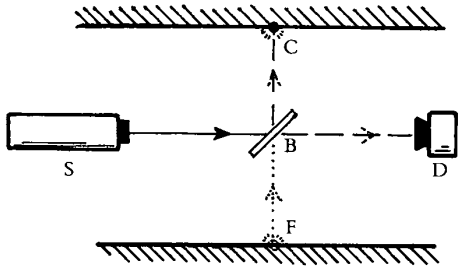


Fig. 30.1. Una fuente S emite aleatoriamente fotones únicos de alta energía (registrándose cada uno de tales sucesos) dirigidos a un divisor de haz B, inclinado a 45° del haz. Si se transmite a través de B, el fotón activa un detector D (ruta SBD); si se refleja, es absorbido en el techo C (ruta SBC). La regla cuántica del módulo al cuadrado predice correctamente las probabilidades 1/2, 1/2. Por otra parte, dado que D registra, el fotón podría haber venido de S (ruta SBD) o del suelo F (ruta FBD). Utilizada en la dirección inversa del tiempo, la regla del módulo al cuadrado retrodice incorrectamente probabilidades 1/2, 1/2, que deberían ser 1,0.

y (por inferencia) una probabilidad del 50 % de que el fotón llegue a C. Esta es sencillamente la respuesta correcta.

Pero imaginemos ahora que leemos este experimento concreto hacia atrás en el tiempo. No estoy proponiendo que tratemos de construir una fuente o un detector «hacia atrás en el tiempo». No, los procesos físicos no deben alterarse en modo alguno. Solo yo propongo expresar mis *preguntas* acerca de ellos de una forma invertida en el tiempo. En lugar de preguntar acerca de las probabilidades *finales*, preguntamos cuáles son las probabilidades *iniciales*, dado que existe un suceso de detección en D. Las amplitudes relevantes se refieren ahora a las dos historias alternativas SBD y FBD, donde F representa un punto en el suelo con la propiedad de que, si se emitiera un fotón desde allí, podría ser reflejado en B para ser recibido en D. Una vez más las amplitudes son $\frac{1}{\sqrt{2}}$ para cada una de estas dos historias (ignorando las fases). Esto debe ser así porque la razón de (los módulos de) las amplitudes de seguir un camino o el otro es solo una propiedad del divisor de haz. Aquí no hay asimetría temporal. Ahora, si aplicáramos la «regla del módulo al cuadrado» para obtener las probabilidades

para estas dos alternativas, encontraríamos una probabilidad del 50 % para la emisión en S y del 50 % (por inferencia) para que el fotón provenga del suelo F, cada vez que haya un suceso de detección en D.

Por supuesto, esto es absurdo. Hay prácticamente una probabilidad cero de que un fotón de rayos X (o de rayos γ) surja del suelo en dirección al divisor de haz. Cada vez que haya un suceso de detección en D, las probabilidades de que haya un suceso de emisión en S son casi del 100 % y las de que el fotón provenga del suelo F son del 0 %. ¡La regla del módulo al cuadrado aplicada en la dirección del pasado nos ha dado una respuesta completamente errónea![4]

Por supuesto, esta regla no fue diseñada para ser aplicada hacia el pasado, pero resulta instructivo ver hasta qué punto sería completamente erróneo hacerlo así. A veces se han planteado objeciones a esta deducción señalando que yo no he tenido en cuenta todo tipo de circunstancias particulares que atañen a mi descripción con el tiempo invertido, tales como el hecho de que la segunda ley solo funciona en un sentido en el tiempo, o el hecho de que la temperatura del suelo es mucho menor que la de la fuente, etc. ¡Pero la maravillosa característica de la ley del módulo al cuadrado mecanocuántica es que nunca tenemos que preocuparnos por circunstancias concretas! El milagro es que las probabilidades cuánticas para predicciones futuras que aparecen en el proceso de medida no parecen depender en absoluto de consideraciones de temperaturas particulares, o geometrías, o cualquier cosa.[5] Si conocemos las amplitudes, entonces podemos calcular las probabilidades futuras. Todo lo que necesitamos conocer son las amplitudes. La situación es muy diferente para las probabilidades de retrodicción. Entonces sí necesitamos conocer toda clase de detalles acerca de las circunstancias. Las amplitudes por sí solas son completamente insuficientes para computar probabilidades pasadas.

Sin embargo, hay situaciones en que las probabilidades cuánticas pueden calcularse de un modo que es completamente simétrico en el tiempo, y quizá sea instructivo echarles un vistazo. Tales situaciones se dan cuando el estado cuántico se mide para ser algo conocido tanto antes como después de una medida cuántica intermedia. Para ser más explícitos, imaginemos una sucesión de tres medidas, la primera de las cuales proyecta el estado en $|\psi\rangle$ y la tercera lo proyecta en $|\phi\rangle$, y entre

estas dos hay una medida sí/NO descrita por el proyector $\boldsymbol{E}$ (§22.6). La probabilidad de sí para la medida intermedia viene dada entonces por[30.1]

$$|\langle\phi|\boldsymbol{E}|\psi\rangle|^2$$

(donde suponemos las normalizaciones $\langle\psi|\psi\rangle = 1 = \langle\phi|\phi\rangle$), que ciertamente tiene simetría temporal. (Para establecer una situación semejante, uno realiza muchas veces las tres medidas sucesivas y escoge para examen solo aquellos casos en los que la primera medida da $|\psi\rangle$ y la tercera da $|\phi\rangle$. La probabilidad anterior se refiere entonces a la fracción de estos casos para los que la medida intermedia dio sí.)[6] Esto ha llevado a algunas personas a concluir que, de entrada, no hay ninguna asimetría temporal en la medida cuántica.[7]

Sin embargo, la mayoría de las medidas cuánticas no son de este tipo. Para el uso normal hacia delante en el tiempo de la regla del módulo al cuadrado no especificamos un $|\phi\rangle$, y para el pretendido uso anterior hacia atrás en el tiempo no especificamos $|\psi\rangle$. Vemos que podemos calcular perfectamente las probabilidades cuánticas, aunque no especifiquemos $|\phi\rangle$, pero no podemos arreglárnoslas sin especificar $|\psi\rangle$. Se podría adoptar el punto de vista de que la razón de que las reglas cuánticas funcionen bien para las probabilidades futuras tiene que ver con que $|\phi\rangle$ sea en algún sentido «aleatorio», lo que tiene que ver con la segunda ley de la termodinámica. Quizá haya algo de verdad en esto, pero encuentro este requisito para $|\phi\rangle$ muy poco claro. ¿Qué significa «aleatorio» en este contexto? De todas formas, parecería ciertamente que hay alguna conexión con la segunda ley en la cuestión de la medida. Podemos tomar nota del hecho de que los dispositivos de medida reales suelen sacar ventaja de ello en algún momento de su actuación. La existencia de cierta relación entre **R** y la segunda ley es parte de mi propia perspectiva sobre el tema. Y puesto que hemos visto que la segunda ley está íntimamente ligada con la buscada unión cuántico/gravedad, debemos esperar también una íntima relación entre **R** y esta unión.

Antes de llegar a esto de forma más explícita, vale la pena señalar que

[30.1] ¿Por qué? ¿Puede obtener esta fórmula?

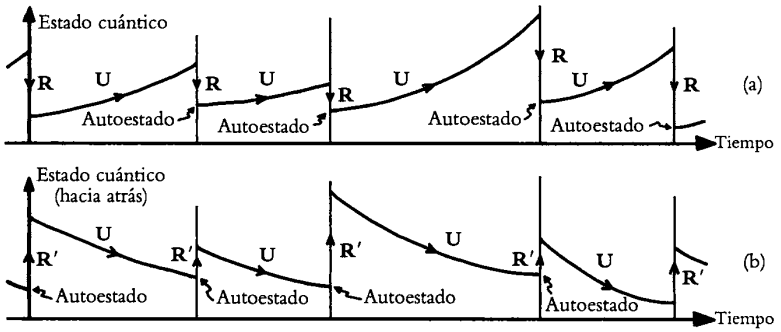


Fig. 30.2. Ilustración esquemática de la alternancia... **U**, **R**, **U**, **R**, **U**, ... de los dos procesos **U** y **R** tal como se usan en la práctica en mecánica cuántica (compárese con la Fig. 22.1), según: (a) la dirección temporal estándar de evolución y donde ocurren autoestados de operador en el extremo pasado de cada tramo de evolución **U**, y (b) el punto de vista del tiempo invertido para la evolución, donde autoestados del operador ocurren en el extremo futuro de cada tramo de evolución **U**. En la interpretación «transaccional» de la mecánica cuántica, hay dos vectores de estado, uno que evoluciona según (a) y el otro que evoluciona según (b).

el otro aspecto de **R**, a saber, el «salto» del estado cuántico —en oposición al cálculo de probabilidades vía la regla del módulo al cuadrado— puede (aparentemente) expresarse tanto según una perspectiva hacia atrás en el tiempo como según una hacia delante en el tiempo. Esto se ilustra de manera esquemática en las Figs. 30.2a, b; en la Fig. 30.2a he mostrado la visión «normal» de la alternancia ..., **U**, **R**, **U**, **R**, **U**, ... (véase la Fig. 22.1) y el estado es un autoestado de la medida *una vez que* esta ha tenido lugar, y en la Fig. 30.2b he representado la visión «invertida en el tiempo» en la que el estado es un autoestado inmediatamente *antes* de la medida. El cálculo de las amplitudes da el mismo resultado cualquiera que sea el punto de vista que se adopte,[30.2] pero aquel con el tiempo invertido tiene un aspecto «teleológico» que algunos encuentran perturbador. Hay también un punto de vista, la interpretación «transaccional» (que se debe, independientemente, a varios teóricos cuánticos),[8] según el cual ambas imágenes se mantienen de

[30.2] Explique por qué esto es básicamente una expresión de la naturaleza «unitaria» de **U**; véase §22.4.

manera simultánea, y en cualquier instante hay dos vectores de estado que evolucionan simultánea y unitariamente para describir el sistema cuántico, uno parecido al de la Fig. 30.2a y el otro como en la Fig. 30.2b. Se dice que esto tiene ventajas con respecto a la interpretación de los fenómenos EPR del capítulo 23. En mi opinión, esta descripción es algo excesiva, y quizá sea mejor adoptar una perspectiva de cuanlazamiento en la que la dirección temporal de «propagación» del estado no es importante, y el cuanlazamiento simplemente proporciona conexiones entre los estados en instantes diferentes (§23.10).

30.4. Temperatura del agujero negro de Hawking

¿Hay formas de conectar **R** con la buscada unión cuántico/gravedad (con asimetría temporal) que sean más directas que el solo hecho de la asimetría temporal en **R**? Creo que las hay, y describiré dos de ellas. La primera de estas tiene que ver con el notable fenómeno de «evaporación de agujeros negros». El argumento es parcialmente sugerente y ciertamente incompleto; más aún, es controvertido en ciertos aspectos fundamentales. Los ingredientes de esta discusión serán el objeto de esta sección, y de las secciones siguientes hasta §30.9 (excepto §§30.5,6, que pueden considerarse una digresión). El segundo argumento es mucho más explícito, pues procede de una tensión fundamental entre los principios básicos de la relatividad general y los de la mecánica cuántica, y conduce a algunas predicciones cuantitativas precisas. Esta línea de razonamiento se expondrá en §§30.10-13. Sin embargo, el primer argumento —que concierne a ciertas implicaciones de la entropía de los agujeros negros— plantea otras cuestiones teóricas que son importantes para nosotros y muy citadas en las discusiones teóricas actuales, y será útil tener cierta idea de ellas.

Recordemos de §27.10 la expresión de Bekenstein-Hawking $S_{\mathrm{BH}} = \frac{1}{4}A$ (en unidades naturales, donde $k = c = G = \hbar = 1$) para la entropía S_{BH} de un agujero negro cuyo horizonte de sucesos tiene un área superficial A. Como parte de su propia discusión, Hawking (1973) demostró que un agujero negro también debe tener una *temperatura*,

que resulta ser proporcional a lo que se denomina la «gravedad superficial» del agujero. Para un agujero estacionario en rotación (geometría de Kerr; véase §27.10), encontramos

$$T_{\mathrm{BH}} = \frac{1}{4\pi m[1 + (1 - a^2/m^2)^{-\frac{1}{2}}]},$$

donde, como en §27.10, m es la masa del agujero negro y am es su momento angular. Esta temperatura puede obtenerse a partir de una fórmula estándar de la termodinámica:

$$T\,\mathrm{d}S = \mathrm{d}E,$$

donde, al variar la energía E, mantenemos constante el momento angular conservado.[30.3] Por consiguiente, el agujero negro emitirá fotones, como si fuera un objeto físico en equilibrio térmico, irradiando energía con el espectro característico de «cuerpo negro» (planckiano), descrito en §21.4 (véase la Fig. 21.3b), para la temperatura T_{BH}. Puede advertirse que, aunque la entropía de Bekenstein-Hawking de un agujero negro es enorme y da lugar a las extraordinarias cifras examinadas en §27.13, la temperatura de Hawking es absurdamente minúscula para los agujeros negros de un tamaño plausible. Para un agujero negro de una masa solar, por ejemplo, la temperatura de Hawking es solo de alrededor de 10^{-7} K, que no es mucho mayor que las temperaturas más bajas conseguidas por el hombre en la Tierra (aproximadamente 10^{-9} K).

Jakob Bekenstein (1972) había obtenido algunos años antes la expresión de la entropía de un agujero negro utilizando un argumento físico (basado en la aplicación de la segunda ley a situaciones en las que partículas cuánticas caen lentamente dentro del agujero), pero no había obtenido un valor claro para el «1/4» que ahora aparece en dicha expresión, ni había obtenido una temperatura del agujero negro. Stephen Hawking proporcionó por primera vez la temperatura y el factor «1/4» en la fórmula de la entropía utilizando ideas de QFT en un fondo espaciotemporal curvo. Aquí el fondo describe un agujero negro que ha sido resultado del colapso de cierto material (digamos una estrella) en el pasado remoto. La situación está descrita por el diagrama

[30.3] Obtenga esta fórmula para T_{BH} suponiendo la expresión para el área del horizonte de un agujero de Kerr dada en §27.10.

conforme de la Fig. 27.16c (que es estricto si el colapso es esféricamente simétrico).

En mi opinión, el extraordinario cálculo de Hawking de la entropía y la temperatura de un agujero negro (junto con el relacionado «efecto Unruh»)[9] es la única conclusión razonablemente fiable que se ha obtenido hasta la fecha de cualquier teoría de la gravedad cuántica. Incluso las conclusiones de Hawking no eran estrictamente de gravedad cuántica, sino que más bien estaban obtenidas a partir de consideraciones de QFT en un espaciotiempo de fondo curvo. En general, surgen problemas graves cuando se intenta formular la teoría cuántica en un fondo curvo, y es sorprendente que Hawking fuera capaz, pese a todo, de llegar a algunas conclusiones firmes.

Uno de los problemas más cruciales consiste en encontrar una noción adecuada de «frecuencia positiva» en un fondo curvo. Como hemos visto en §24.3 y §26.2, esta noción es un ingrediente clave de la visión estándar de las partículas cuánticas y la QFT. El problema de formular esta cuestión en un espaciotiempo curvo general reside en la ausencia de un «parámetro temporal» definido de forma natural en cuyos términos pueda formularse la noción de «frecuencia positiva».

El lector avisado podría muy bien señalar que tampoco hay parámetro temporal definido de forma natural en un espacio de Minkowski plano. Sin embargo, un hecho sorprendente viene en nuestra ayuda y nos dice que para soluciones de las ecuaciones de ondas relativistas (como las estudiadas en los capítulos 19 y 24-26), la frecuencia positiva en una elección del parámetro temporal de Minkowski t es equivalente a la frecuencia positiva en cualquier otro parámetro semejante, para el que la orientación temporal no esté invertida. Para campos sin masa, se puede ir incluso más lejos y obtener la misma condición de frecuencia positiva con el uso de un «parámetro temporal» que se obtiene a partir del parámetro temporal de Minkowski estándar por una transformación conforme que conserva la orientación temporal (lo que tiene relevancia para la teoría de twistores; véanse §§33.3,10).[10]

En un espaciotiempo general no hay un análogo natural a un parámetro semejante, y la noción de frecuencia positiva resultaría, en general, de forma distinta para diferentes elecciones de un parámetro temporal. Aparte del caso de la temperatura de Hawking, los resultados más

plausibles proceden de consideraciones de espaciotiempos *estacionarios*, para los que existe una familia continua de desplazamientos temporales que conservan la geometría espaciotemporal (véase la Fig. 30.3). Tales movimientos espaciotemporales están generados por un *vector de Killing* $\boldsymbol{\kappa}$ *de género tiempo* (véanse §14.7 y §30.6). Las curvas a lo largo de las cuales apuntan los vectores $\boldsymbol{\kappa}$ (curvas integrales de $\boldsymbol{\kappa}$) son las curvas a lo largo de las cuales puede especificarse un «parámetro temporal» t razonablemente natural, de modo que

$$\boldsymbol{\kappa} = \frac{\partial}{\partial t},$$

donde las tres coordenadas restantes x, y, z se toman constantes a lo largo de las curvas. Entonces puede definirse una noción de «frecuencia positiva» con respecto a este parámetro.

Una curiosa situación puede aparecer cuando hay más de un vector de Killing de género tiempo, puesto que entonces puede haber más de una noción de «frecuencia positiva». Esta multiplicidad de vectores de Killing de género tiempo sucede, por supuesto, con el espacio de Minkowski $\mathbb{M}$, pero, por lo que se ha dicho antes, las nociones de frecuencia positiva coinciden cuando pasamos de un sistema inercial minkowskiano a otro. Sin embargo, este no es el caso cuando pasamos a un sistema acelerado. Entonces obtenemos una noción diferente de «frecuencia positiva», y la QFT resultante está en lo que se denomina un *vacío térmico*, según el cual un observador acelerado experimenta una temperatura no nula, aunque absurdamente baja, para cualquier aceleración razonable.

Debería quedar claro que, aunque este es un efecto sorprendente, esta «temperatura de aceleración» es simplemente el tipo ordinario de temperatura, tal como sería medida por un termómetro ordinario (aunque idealizado). En este caso, el termómetro estaría experimentando aceleración uniforme, y se supone que está en un vacío ambiente que tendría temperatura cero si fuera medida por un termómetro no acelerado. (La noción de «vacío térmico» tiene relaciones con las nociones de QFT de «vacíos alternativos» en §§26.5,11 y «falso vacío» en §§28.1,4.) Esto se conoce como el «efecto Unhru», y es compatible con el estado térmico de Hawking de un agujero negro. Un observador que se mantenga estacionario cerca de un agujero negro muy

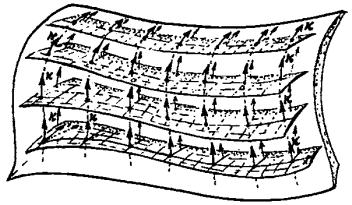


Fig. 30.3. La estacionariedad de un espaciotiempo está expresada como la presencia de un vector de Killing $\boldsymbol{\kappa}$ de género tiempo. Esto genera una familia continua de desplazamientos temporales que conservan la métrica. Si $\boldsymbol{\kappa} = \partial/\partial t$, donde t es el «parámetro tiempo» de un sistema de coordenadas (t, x, y, z), entonces x, y y z deben ser constantes a lo largo de las curvas integrales de $\boldsymbol{\kappa}$. (Véase §14.7.)

grande experimentaría, según el principio de equivalencia (§17.4), una aceleración efectiva, y la temperatura de Unhru de esta aceleración coincide con la temperatura de Hawking, tal como se obtiene con sus propios procedimientos.

La dificultad que presenta en general la falta de una definición natural de «frecuencia positiva» es evitada en algunos enfoques abandonando la noción de «partícula» y concentrándose en el álgebra de los operadores mecanocuánticos.[11] A primera vista, esto puede parecer un sacrificio demasiado grande, y se necesita ingenio en la formulación de muchas preguntas de interés en tal enfoque. En el momento de escribir esto, yo mismo no he sido capaz de evaluar todos los méritos de este tipo de teoría enigmática y de aspecto prometedor. No obstante, sospecho que es superior a los enfoques basados en campos vectoriales de género tiempo específicos. En cualquier caso, no veo por qué la QFT en un fondo dado debería tener necesariamente un sentido físico completo. Es solo una aproximación a un esquema más exacto en el que los grados de libertad en el campo gravitatorio —es decir, en la geometría del propio espaciotiempo— deben formar parte también de la física cuántica.

En su cálculo de la temperatura y la entropía de un agujero negro, Hawking se las arregla para evitar la mayoría de estos problemas exigiendo una noción de frecuencia positiva/negativa que se separa solo en el infinito. Existe una noción en $\mathscr{I}^-$ (infinito nulo pasado) y otra diferente en $\mathscr{I}^+$ (véase §27.12). Esta diferencia lleva a la producción de un «estado térmico» de Hawking por el agujero negro, cuyo resultado es lo que se conoce como *radiación de Hawking*. También es digno de

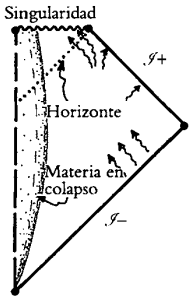


Fig. 30.4. El cálculo de Hawking de la temperatura de un agujero negro, que incluye el colapso de materia a un agujero negro en el pasado distante, requiere solo la noción (estándar) de separación de frecuencias positiva/negativa en $\mathscr{I}^+$ y $\mathscr{I}^-$. El vacío del agujero negro se convierte en un estado térmico (una matriz densidad) porque la información inicial en $\mathscr{I}^+$ se divide entre la información en $\mathscr{I}^+$ y la información en la singularidad final, perdiéndose la última.

mención que este efecto debe su existencia al hecho de que parte de la información definida en $\mathscr{I}^-$ se *pierde* en la singularidad, y no toda lo hace en $\mathscr{I}^+$ (véase la Fig. 30.4). Veremos la importancia de este hecho en §30.8.

30.5. Temperatura del agujero negro a partir de la periodicidad compleja

En este punto es instructivo considerar una ingeniosa derivación posterior de la temperatura de Hawking, obtenida por Gibbons y Perry en 1976, aunque suponga una ligera digresión respecto a las líneas principales de razonamiento de este capítulo. (Las retomaré en §30.8.) El argumento de Gibbons-Perry plantea algunas cuestiones interesantes respecto al papel de las ideas matemáticas elegantes en la derivación de fenómenos físicos genuinos. Lo que ellos advirtieron era que si la solución de la ecuación de Einstein (a saber, la solución de Schwarzschild o de Kerr; véanse §§27.8,10) que representa al agujero negro asentado en su estado final es «complexificada» (i.e., extendida desde valores reales a valores complejos de las coordenadas; véase §18.1), entonces una condición de regularidad básica sobre las cantidades definidas en este espacio complexificado implica que dichas cantidades adquieren necesariamente una periodicidad (véanse §9.1 y la Fig. 9.1a) en el tiempo complexificado, con un período puramente imaginario $2\pi i T_{BH}$. Consideraciones de termodinámica estadística nos dicen que semejante periodicidad compleja corresponde a la temperatura exacta T_{BH}, dada en

§30.4. Esto ofrece una ruta sorprendentemente directa a la temperatura de Hawking para un agujero negro.

Pero ¿qué vamos a hacer de un procedimiento semejante como derivación física? De hecho, es un argumento notablemente elegante, y puede utilizarse directamente para obtener la temperatura de agujero negro de Hawking en varias situaciones diferentes en las que no podría aplicarse de inmediato su discusión original. Por otra parte, tengo dificultad en considerar que este argumento proporciona genuinamente una justificación física real de la temperatura de Hawking. Es un buen ejemplo de un bello fragmento de matemáticas, que resulta dar la respuesta correcta (si se juzga su «corrección» por su acuerdo con la respuesta obtenida a partir de criterios más aceptables físicamente: en este caso, el argumento original de Hawking que se ha mencionado antes), pese al hecho de que algunas de las hipótesis «físicas» que intervienen en las matemáticas pueden juzgarse de validez dudosa.

Consideremos un poco más de cerca los ingredientes matemáticos de esta «complexificación». Una buena manera de entender lo que está implícito aquí es pensar primero en el plano euclídeo ordinario 2-dimensional $\mathbb{E}^2$, y su complexificación estándar $\mathbb{CE}^2 (= \mathbb{C}^2)$. El espacio real $\mathbb{E}^2$ se denomina a veces una *sección real* de $\mathbb{CE}^2$ (Figs. 30.5a,b; véase también la Fig. 18.2). Es una sección real euclídea, porque posee una métrica euclídea ordinaria. Pero $\mathbb{CE}^2$ tiene también secciones reales *lorentzianas* (véanse la Fig. 18.2 y §18.2), y podemos construir una de estas, $\mathbb{M}^2$, tomando la coordenada y de un par de coordenadas cartesianas estándar (x, y) para $\mathbb{E}^2$ (que son números reales) y exigir que y tome valores puramente imaginarios en lugar de reales. Entonces, $t = \mathrm{i}y$ sirve como una coordenada temporal para $\mathbb{M}^2$. (Este es simplemente el caso 2-dimensional de lo que ya hemos hecho en §18.1.) Consideremos ahora coordenadas *polares* (r, θ) para $\mathbb{E}^2$, en lugar de coordenadas cartesianas (x, y); véanse §5.1 y la Fig. 30.5a. El número real no negativo r mide la distancia al origen y el ángulo real θ da el ángulo entre el radio vector y el eje x, medido en sentido contrario a las agujas del reloj. ¿Cómo se extienden entonces estas coordenadas a nuestra sección lorentziana $\mathbb{M}^2$? Con tal de que restrinjamos la atención al cuadrante derecho $\mathbb{M}^R$, como se indica en la Fig. 30.5b, la cantidad r sigue siendo real y no negativa, pero θ es ahora puramente imaginaria, de modo que

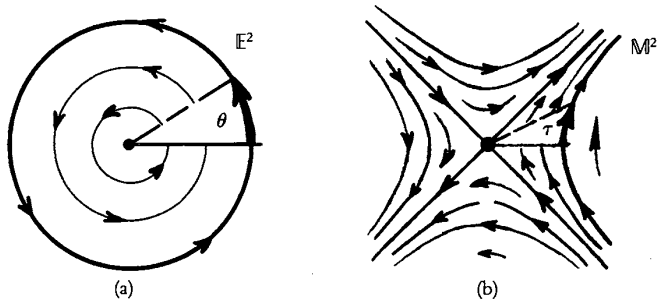


Fig. 30.5. Periodicidad en el tiempo imaginario, ilustrada por el 2-espacio de Minkowski $\mathbb{M}^2$, complexificado a $\mathbb{CM}^2 = \mathbb{CE}^2$. (a) El plano euclídeo $\mathbb{M}^2$ es una sección real del espacio complejo $\mathbb{CM}^2$. El vector de Killing $\partial/\partial\theta$ genera rotaciones en $\mathbb{E}^2$, donde tomamos coordenadas polares (r, θ). Cualquier función que es univaluada en $\mathbb{E}^2$ debe ser periódica en θ con período 2π. (b) En la sección real lorentziana $\mathbb{M}^2$ de $\mathbb{CM}^2$, la coordenada temporal de «Rindler» (aceleración uniforme) es $\tau = \mathrm{i}\theta$ (un análogo del tiempo de Schwarzschild que es natural para un agujero negro), y una función que es analítica en el origen O debe tener un período imaginario en τ, con período $2\pi\mathrm{i}$. (Vector de Killing $\boldsymbol{\kappa} = \partial/\partial\tau = -\mathrm{i}\partial/\partial\theta$.)

$$\tau = \mathrm{i}\theta$$

es real. La coordenada r mide ahora la distancia espacial lorentziana al origen, y τ es el «ángulo hiperbólico a partir de la horizontal».[30.4]

Vamos a interpretar que la coordenada τ (multiplicada por una constante r_0) mide el «tiempo» en el sentido espaciotemporal ordinario para un observador (en esta geometría espaciotemporal plana 2-dimensional) que «acelera uniformemente» alejándose del centro y cuya línea de universo está dada por $r = r_0$. («Coordenadas de Rindler».)[12] Se considera que el propio espaciotiempo es el $\mathbb{M}^2$ entero, pese al hecho de que el «tiempo» del observador se aplica solo al cuadrante $\mathbb{M}^R$. Supongamos que el observador está interesado en cantidades analíticas definidas sobre $\mathbb{M}^2$. Tales cantidades tendrán una extensión holomorfa a la complexificación del espaciotiempo (véanse §7.4 y §12.9), pero esto solo puede garantizarse para un entorno inmediato de la «sección

[30.4] Escriba estas coordenadas (r, τ) en términos de las coordenadas cartesiano-lorentzianas (x, t); vea por qué la parte real de θ se anula sobre $\mathbb{M}^2$.

real», lo que en este caso significa la sección lorentziana. Sin embargo, si una cantidad semejante es en realidad analítica en el origen O de esta sección, entonces también debe ser analítica en el origen de la sección euclídea (puesto que es exactamente el mismo punto O en cada caso). Pero una cantidad que es suave en el origen euclídeo debe ser periódica en θ, con período 2π, puesto que si incrementamos θ en 2π (para algún valor radial pequeño $r = \varepsilon$), simplemente damos una vuelta alrededor del origen para llegar exactamente al mismo punto del que hemos partido. De ello se sigue que la cantidad τ referida a las coordenadas espaciotemporales lorentzianas originales tiene un *período imaginario* de $2\pi i$ en τ (complexificado).

Esta es la base del argumento de Gibbons-Perry que ahora aplicamos a la geometría 4-espacial entera del agujero negro en lugar de a nuestro espaciotiempo simplificado 2-dimensional $\mathbb{M}^2$. La geometría relevante es ahora la de Schwarzschild (para el caso esféricamente simétrico y sin rotación, pero también podemos utilizar la de Kerr para un agujero en rotación). Para que el argumento funcione necesitamos un análogo del origen O en $\mathbb{M}^2$.Vemos esto en la Fig. 30.6a, donde he presentado un diagrama conforme estricto para lo que se conoce como un espaciotiempo de Schwarzschild «máximamente extendido» $\mathcal{K}$.[13] El espaciotiempo $\mathcal{K}$ se denomina a veces un «agujero negro eterno» porque no fue creado a partir de un colapso gravitatorio, sino que estuvo «siempre allí». El punto central O del diagrama representa una 2-*esfera*, de acuerdo con las convenciones de los diagramas conformes estrictos. $\mathcal{K}$ es el análogo de $\mathbb{M}^2$, pero también necesitamos un análogo del espacio euclídeo $\mathbb{E}^2$. Hay un espacio semejante, conocido a veces como el espacio de Schwarzschild «euclidianizado» $\mathcal{G}$ (y en ocasiones llamado incluso un «espacio euclídeo», ¡lo que me sorprende por ser tan confuso!), donde el «tiempo»[14] de Schwarzschild τ en $\mathcal{K}$ toma valores puramente imaginarios $\tau = i\beta\theta$ en $\mathcal{G}$, siendo θ la coordenada angular en $\mathcal{G}$ que se incrementa en 2π cada vez que se da una vuelta alrededor de O en una dirección positiva (Fig. 30.6b), y donde β es un número real (constante, a r constante) denominado «gravedad superficial». Cualquier cantidad que es regular (i.e., analítica) en O en $\mathcal{K}$ debe también ser regular en O en $\mathcal{G}$ (puesto que este es exactamente el mismo lugar «O», en el espacio de Schwarzschild complexificado, para cada una

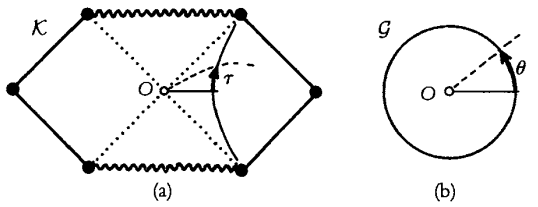


Fig. 30.6. Temperatura de Hawking a partir de una periodicidad en el tiempo imaginario (argumento de Gibbons-Perry). (a) Diagrama conforme estricto para un «agujero negro eterno», que es el espaciotiempo de Schwarzschild «máximamente extendido» $\mathcal{K}$, con tiempo de Schwarzschild τ y vector de Killing $\boldsymbol{\kappa} = \partial/\partial\tau$. El punto central O representa una 2-esfera. (b) El espacio de Schwarzschild «euclidianizado» $\mathcal{G}$, cerca de O, tiene una coordenada angular real θ, de modo que el «tiempo» τ toma valores puramente imaginarios $\tau = i\beta\theta$, siendo el número real constante β la «gravedad de superficie» del agujero. Aquí θ aumenta en 2π cuando rodeamos O una vez, de modo que cualquier función sobre $\mathcal{K}$, analítica en O (y que se extiende a $\mathcal{G}$), tiene período 2π en θ, i.e., período $2\pi i\beta$ en τ (cerca de O), donde β debe interpretarse como la temperatura de Hawking del agujero negro.

de las dos secciones reales $\mathcal{K}$ y $\mathcal{G}$). Al ser regular en O en $\mathcal{G}$, una cantidad semejante debe ser periódica, con período $2\pi i\beta$ en τ, porque θ es una coordenada angular ordinaria que cuando se incrementa en $2\pi(= 360°)$ simplemente nos devuelve al mismo lugar en el espacio(tiempo) del que hemos partido. Esta *periodicidad imaginaria* es característica de un «estado térmico de temperatura β» de acuerdo con los principios de la termodinámica estadística.

No es mi propósito aquí discutir estos principios termodinámicos. Esto nos llevaría demasiado lejos. El único punto de interés es si podemos confiar en el razonamiento para esta periodicidad compleja. Depende de tomar en serio la región O. ¿Está esto justificado? No está claro ni mucho menos. Para un agujero negro físico real, toda esa imagen «eterna» no es ciertamente adecuada. Un agujero físico tendría que haber sido creado a partir de un colapso gravitatorio (por ejemplo, de una estrella supermasiva o un cúmulo de material en un centro galáctico), a menos que fuera en cierto sentido una creación «primordial» del propio big bang. Incluso un agujero primordial —un agujero negro antes que su inverso temporal, a saber, un *agujero blanco*— seguiría representando en cierto sentido un «colapso» y, ya fuera negro o blan-

co, no está bien descrito por el modelo completo de la Fig. 30.6a. Sin embargo, cierta parte *exterior* de este modelo es apropiada para la descripción de un colapso a un agujero negro, a saber, esa parte de la Fig. 30.7 que está por encima y a la derecha de la línea fronteriza exterior indicada del material real que sufre el colapso. Por debajo y a la izquierda de esta frontera, la métrica espaciotemporal sería la de la materia, y sería diferente de la del agujero negro eterno. El colapso completo se esboza en la Fig. 30.7, lo que equivale a un ligero redibujado de la Fig. 27.16c. Notemos ahora que O está siempre fuera de la región donde se aplica la métrica de Schwarzschild (extendida). Parece que no hay ninguna justificación física para suponer que las cantidades físicas definidas en el espaciotiempo real tienen una regularidad en O, y es difícil ver por qué el argumento proporciona una justificación real para la temperatura de Hawking, pese a su elegancia matemática. (Cualquier modelo físicamente realista de un agujero negro presentaría desviaciones de la métrica de Schwarzschild —o de la de Kerr— exacta, y cabe esperar razonablemente que dichas desviaciones se hagan cada vez mayores, divergiendo finalmente a infinito cuanto más próximos a «O» nos extendamos.)[30.5]

Pero el modelo exacto de agujero negro estacionario representa el límite final de un colapso realista, donde se considera que todas las irregularidades se alisan a medida que avanza el tiempo. Es el espaciotiempo limitante el que tiene esta regularidad, y por consiguiente la periodicidad compleja requerida y, con ella, la temperatura requerida. Aunque no veo cómo se puede considerar este argumento como una *derivación* física real de la temperatura de Hawking (a pesar de que normalmente se toma así), lo cierto es que proporciona un «fuerte indicio» de una consistencia interna oculta de la idea global de esta «temperatura de agujero negro».

En esta tesitura, no puedo resistirme a hacer una comparación con otra observación, que se debe originalmente a Brandon Carter, que en un contexto diferente tiene una similaridad importante con el argumento que se acaba de dar, aunque nunca se ha presentado como una

[30.5] Vea si puede dar un argumento que justifique esta afirmación. *Sugerencia*: Considere pequeñas perturbaciones lineales. ¿Espera un comportamiento exponencial en el tiempo? (Considere modos propios de $\partial/\partial\tau$.)

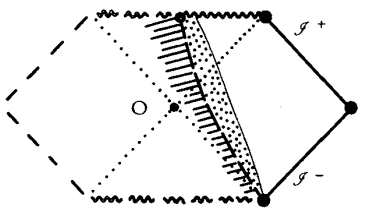


Fig. 30.7. Una historia de colapso gravitatorio esférico —un ligero redibujado de la Fig. 27.16c— representada en términos del espaciotiempo de Schwarzschild maximal K de la Fig. 30.6a. La región sombreada con líneas inclinadas debe ser borrada y en la sombreada con puntos la métrica difiere de la de K, debido a la presencia de materia. Nótese que O está siempre fuera de la región donde se aplica la métrica de K.

«derivación» de nada. Recordemos que un agujero negro estacionario libre de carga está descrito por los dos parámetros de Kerr m y a, donde m es la masa del agujero y am es su momento angular (y donde por conveniencia escojo unidades para las que $c = G = 1$, tales como las unidades de Planck, como en §27.10). Una generalización de la métrica de Kerr encontrada por Ezra T. Newman[15] (normalmente conocida como la métrica de Kerr-Newman) representa un agujero estacionario en rotación y eléctricamente cargado. Ahora tenemos tres parámetros: m, a y e. La masa y el momento angular son los de antes, pero ahora hay una carga eléctrica total e. También hay un momento magnético $M = ae$, cuya dirección coincide con la del momento angular. Carter notó que la *razón giromagnética* (dos veces la masa por el momento magnético dividido por el producto de la carga por el momento angular, que para un agujero negro cargado es $(2m \times ae)/(e \times am) = 2$), que está completamente fijada para un agujero negro (i.e., independiente de m, a y e), toma precisamente el valor que Dirac predijo originalmente para el *electrón*, a saber 2 (pues para el electrón de Dirac el momento angular es $\frac{1}{2}\hbar$ y el momento magnético es $\frac{1}{2}\hbar e/mc$, lo que de nuevo da una razón giromagnética 2, tomando $c = 1$). Véase §24.7. Newman (2001) ha ofrecido una interpretación de esta «coincidencia» en términos de un desplazamiento en una dirección compleja en el espacio.

¿Podemos considerar que este argumento proporciona una *deducción* de la razón giromagnética del electrón con independencia del argumento original de Dirac? Ciertamente *no* lo hace, en ningún sentido ordinario del término «deducción». Solo podría aplicarse si un

electrón pudiera considerarse en cierto sentido como un «agujero negro». De hecho, los valores reales para los parámetros a, m y e, en el caso de un electrón, violan flagrantemente una desigualdad

$$m^2 \geq a^2 + e^2$$

que es necesaria para que la correspondiente métrica de Kerr-Newman pueda representar un agujero negro. Por ello, este argumento está muy lejos de ser una deducción real de la razón giromagnética del electrón. Pese a todo, se parece algo al argumento de Gibbons-Perry para una temperatura de agujero negro en cuanto que revela cierta «naturalidad» de este valor por vía de extensiones a lo complejo.[16] El argumento de Gibbons-Perry tiene a su favor el punto adicional de que no está limitado a la mera consideración de la familia Schwarzschild/Kerr de métricas espaciotemporales. De todas formas, en mi opinión, esto difícilmente justifica su aceptación común como una deducción real.

30.6. Vectores de Killing, flujo de energía... ¡y viaje en el tiempo!

El «agujero negro eterno» ha llamado con frecuencia la atención de la gente por otras razones, pese al hecho de que tiene algunas curiosas propiedades globales que dificultan que se tome como modelo de un universo físicamente aceptable. Aunque varias de estas razones podrían pertenecer más al ámbito de la ciencia ficción que al de la realidad, el agujero negro eterno tiene un valor geométrico para nosotros, pues ilustra aspectos matemáticos interesantes que serán importantes en §§30.7,10. Vemos que tiene dos infinitos nulos pasados distintos ($\mathscr{I}^-$ y $\mathscr{I}^{-\prime}$) y dos infinitos nulos futuros distintos ($\mathscr{I}^+$ y $\mathscr{I}^{+\prime}$). A menudo se considera que este espaciotiempo representa una evolución temporal de dos universos diferentes que han llegado a conectarse por un «agujero de gusano» que posteriormente «se estrangula» dando una singularidad; véase la Fig. 30.8.

Con respecto a cada una de las dos regiones «externas», parecería que este universo contiene un agujero negro, pero el agujero negro es

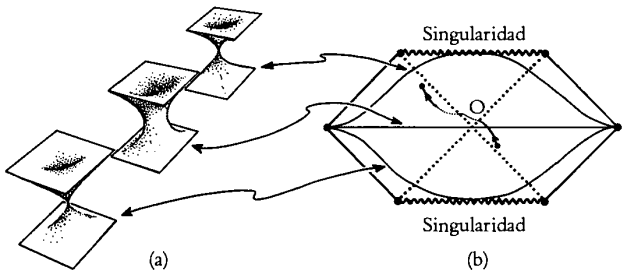


Fig. 30.8. El espaciotiempo $\mathcal{K}$ visto globalmente como un 3-espacio «en evolución temporal» que representa un «agujero de gusano» que conecta dos regiones asintóticamente planas. El agujero de gusano se estrangula de una manera singular, tanto en el futuro como en el pasado. Cualquier viajero espacial que se proponga atravesar el agujero de gusano desde una región a la otra no puede pasar antes de que «se estrangule» como se pone de manifiesto a partir del diagrama conforme, puesto que esto exigiría que la línea de universo del viajero tenga una porción de género espacio (superlumínica), la cual se muestra punteada.

singular en el sentido de que al mismo tiempo es también un «agujero blanco». Pueden escapar señales hacia cada universo externo $\mathcal{E}$ y $\mathcal{E}'$ desde la región interna *pasada* $\mathcal{B}^-$ (comportamiento de «agujero blanco») igual que es posible que se propaguen señales a la región interna *futura* $\mathcal{B}^+$ desde cada universo externo $\mathcal{E}$ y $\mathcal{E}'$ (comportamiento de «agujero negro»). El hecho de que el espaciotiempo es estacionario se expresa en la existencia de un vector de Killing $\boldsymbol{\kappa}$ (véanse §14.7, §19.5 y §30.4). He esbozado este vector de Killing en la Fig. 30.9. Notemos que el vector de Killing es de género tiempo en las dos regiones externas $\mathcal{E}$, $\mathcal{E}'$, pero es de género espacio en las regiones internas $\mathcal{B}^-$, $\mathcal{B}^+$. La naturaleza de género tiempo de $\boldsymbol{\kappa}$ en las regiones externas significa que $\boldsymbol{\kappa}$ expresa la estacionariedad del agujero blanco/negro. Una familia de observadores en $\mathcal{E}$, cuyas líneas de universo son tangentes al vector de Killing $\boldsymbol{\kappa}$ del campo, percibirá un universo que no cambia. Lo mismo se aplica a $\mathcal{E}'$. Sin embargo, la familia de observadores en $\mathcal{E}'$ con esta propiedad debe aplicar estas consideraciones al vector de Killing $-\boldsymbol{\kappa}$, en lugar de a $\boldsymbol{\kappa}$, porque necesitamos mantener las distinciones futuro/pasado consistentes para observadores locales a lo largo del espaciotiempo. En cierto sentido, la «dirección temporal» se ha invertido cuando

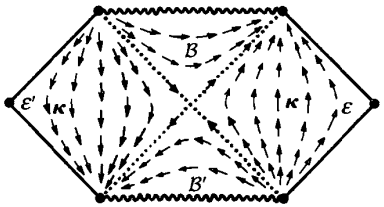


Fig. 30.9. El vector de Killing $\boldsymbol{\kappa}$ es de género tiempo en las dos regiones externas $\mathcal{E}$ y $\mathcal{E}'$, pero de género espacio en las regiones internas $\mathcal{B}$ y $\mathcal{B}'$. Comparando $\boldsymbol{\kappa}$ en $\mathcal{E}$ y $\mathcal{E}'$, observamos que invierte la orientación temporal, de modo que el concepto de densidad de energía conservada $T_{ab}\kappa^a$ invierte su signo.

pasamos de $\mathcal{E}$ a $\mathcal{E}'$. La magnitud *densidad de energía* conservada (§19.5) obtenida a partir de la contracción $T_{ab}\kappa^b$ del tensor energía-momento con el vector de Killing $\boldsymbol{\kappa}$ proporciona una densidad de energía positiva (para materia normal) en $\mathcal{E}$, pero una densidad negativa para materia normal en $\mathcal{E}'$ (puesto que $\boldsymbol{\kappa}$ apunta hacia el pasado en $\mathcal{E}'$, y $-\boldsymbol{\kappa}$ es ahora el vector de Killing ordinario que expresa la estacionariedad). Esto no es exactamente una contradicción, pero ilustra la extraña naturaleza del espaciotiempo en consideración.

De hecho, no es posible para un observador físico «pasar» realmente de $\mathcal{E}$ a $\mathcal{E}'$, pues ello implicaría una «línea de universo» que no es de género tiempo en todo lugar (Fig. 30.8). De todas formas, a menudo los teóricos intentan evitar este hecho, tratando de modificar el espaciotiempo de alguna forma aparentemente «menor». Sus razones proceden de una intención (en mi opinión, equivocada) de demostrar que se podría conseguir un tipo de viaje entre universos a través de un «agujero de gusano» —o (con una ligera modificación de la imagen, como en la Fig. 30.10) de una región del espaciotiempo a otra región distante— mediante alguna tecnología futura. Si una propuesta semejante pudiera tener éxito, entonces abriría la posibilidad de una forma de viaje espacial en la que se trascienden las limitaciones normales de la relatividad. En una auténtica línea *Star Trek*, se concibe un procedimiento «impulsado por curvatura» que permite a la nave espacial viajar a través del agujero de gusano hasta una región lejana que podría incluso ser «anterior» a cuando la nave espacial entró en el agujero de gusano.

Por extraño que pueda parecer, incluso profesionales de primera fila en relatividad general han tomado en consideración la posibilidad de tal «viaje en el tiempo».[17] Una razón seria para esto (al menos a veces) no

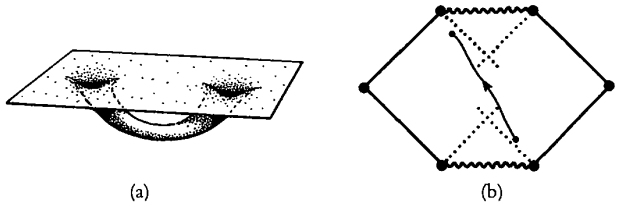


Fig. 30.10. Una sugerencia de ciencia ficción para el viaje espacial superlumínico basada en un espaciotiempo de agujero de gusano modificado. (a) «Identificando» partes distantes de las dos regiones espaciales externas de la Fig. 30.8, se obtiene un agujero de gusano que conecta regiones distantes del mismo espacio, pero atravesar de nuevo el agujero de gusano desde una a la otra no puede conseguirse por una curva de género tiempo. (b) Para que esto sea posible, se necesitaría algo similar a la versión «estirada» de $\mathcal{K}$, representada aquí (pero semejante modelo requiere densidades de energía negativas.

es tanto la posibilidad de que un viaje en el tiempo pudiera ser realmente factible dentro de los límites de la física actual (o imaginable) como el que podríamos aprender algo del hecho de que físicamente no debería serlo.[30.6] En la «descripción espacial» dada en la Fig. 30.8, el agujero de gusano «se estrangula hasta un tamaño cero» antes de que el viajero espacial pueda atravesarlo. No obstante, se contempla la posibilidad de que pudiera ser factible, dentro de los límites de la teoría, «mantener abierto el agujero de gusano» el tiempo suficiente para que el viajero pase al otro lado, si se admiten densidades de energía negativas. Tales densidades de energía negativa se consideran normalmente prohibidas en la teoría clásica, pero podrían ser admitidas bajo circunstancias especiales en la teoría cuántica de campos apropiada.

¿Realmente esperan los físicos relativistas que estas fantasiosas consideraciones podrían llevarnos a una idea de «impulso por curvatura», con la que podría conseguirse viajar a una parte distante del universo a través de dicho agujero de gusano apoyado en QFT? Imagino que muy pocos.[18] Parece que una cuestión más seria es que estas consideracio-

[30.6] Explique por qué, de acuerdo con los principios de la relatividad general, la posibilidad de viajar entre sucesos p y q, con separación de género tiempo, implica la posibilidad de viajar de p a un suceso en el pasado directo de p, en una línea de universo de género tiempo que pasa por p.

nes podrían proporcionar un «test» para ideas en gravedad cuántica. Si dichas ideas en QFT *sí* permiten realmente «mantenerlo abierto», entonces esto podría tomarse como una mala señal para esas ideas concretas acerca de la gravedad cuántica, y hay que replantearlas. De esta forma, se podría obtener una guía útil respecto a la plausibilidad de la teoría en consideración. (Al menos esta es mi propia lectura de tales propuestas. Quizá esté adoptando una posición demasiado «generosa» sobre esto, y más teóricos de los que yo imagino están pensando en realidad que ¡hay que tomar en serio dicho «impulso por curvatura»!)

30.7. Flujo de energía saliente de órbitas de energía negativa

Me he desviado demasiado de la tarea que tenía entre manos, que era considerar las implicaciones de la temperatura de Hawking de un agujero negro. ¿Podemos ver a partir de razones más físicas por qué, en el contexto de la mecánica cuántica, un agujero negro debería emitir radiación de acuerdo con que tiene una temperatura no nula? De hecho, Hawking también proporcionó una derivación «intuitiva» de la presencia de esta radiación de Hawking. Esta se ilustra en la Fig. 30.11. En la vecindad del horizonte del agujero se están creando continuamente pares virtuales partícula-antipartícula a partir del vacío, solo para aniquilarse mutuamente en un período de tiempo muy corto. (Este es el proceso que se ha considerado en §26.9, e ilustrado en las Figs. 26.9 y 26.10.) Sin embargo, la presencia de un agujero negro modifica esta actividad, porque de vez en cuando una de las partículas del par cae en el agujero y la otra escapa. Esto solo puede suceder cuando la partícula que escapa se convierte en una partícula real (i.e., «sobre la capa de masas», en oposición a la partícula virtual de partida «fuera de la capa»; véanse §26.8 y la Fig. 26.6), y por consiguiente la partícula que escapa debe tener energía positiva, de modo que (por la conservación de la energía) la partícula que cae en el agujero tiene que convertirse en una partícula real con energía negativa (siendo evaluadas dichas energías a partir del infinito). De hecho, energías negativas pueden darse para partículas reales dentro del agujero negro. Esta posibilidad surge porque el

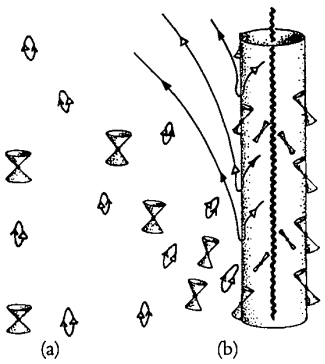


Fig. 30.11. Deducción «intuitiva» de Hawking de la radiación de Hawking. (a) Lejos del agujero se están produciendo continuamente pares virtuales partícula-antipartícula a partir del vacío, pero luego se aniquilan en un tiempo muy corto (véase la Fig. 26.9a). (b) Muy cerca del horizonte del agujero podemos imaginar que un miembro del par cae en el agujero, y el otro escapa al infinito externo. Por ello, ambas partículas virtuales se hacen reales, y la conservación de la energía exige que las partículas entrantes tengan energía negativa. Esto es posible porque el vector de Killing $\boldsymbol{\kappa}$ se hace de género espacio dentro del horizonte. (Si κ^a es de género espacio, la energía conservada $p_a\kappa^a$ puede ser negativa, donde p_a es el 4-momento de la partícula.)

vector de Killing κ^a se convierte en un vector de género espacio en la región interior $\mathcal{B}^+$ y un 4-momento p_a de género tiempo que apunta al futuro puede tener un producto escalar negativo $p_a\kappa^a$, siendo esto la energía (conservada) de la partícula; véase la Fig. 30.11b.[30.7] El proceso de Hawking ocurre porque una partícula real (en oposición a virtual) puede tener energía negativa si está dentro del horizonte de sucesos del agujero. La compañera real de dicha partícula tiene que tener energía positiva, de modo que puede salir energía positiva del agujero.

En este punto vale la pena comentar que algo muy similar sucede en la teoría de los agujeros negros *clásicos* cuando el agujero está en *rotación*. Y en contraste con el caso de la radiación de Hawking, para el que la emisión de energía por agujeros negros de tamaño plausible es ridículamente pequeña —y cuyo interés es puramente teórico—, lo que sucede con un agujero clásico en rotación parece tener enormes implicaciones astrofísicas. De hecho, las más potentes fuentes de energía conocidas en el universo (cuásares y radiogalaxias) parecen estar alimentadas por la energía rotacional de enormes agujeros negros.

El proceso tiene similitud con el que conduce a la radiación de

[30.7] Explique cómo un κ^a puede dar un valor de «energía» negativo $p_a\kappa^a$.

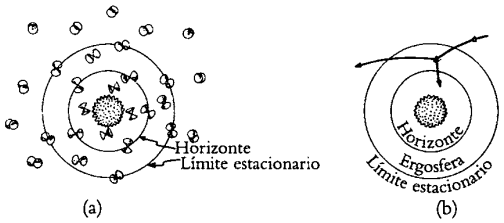


Fig. 30.12. Vistas desde «abajo» a lo largo del eje temporal de un agujero negro en rotación (de Kerr). (a) Para un agujero negro de Kerr, existe una región —llamada la «ergosfera»— dentro de la cual el vector de Killing $\boldsymbol{\kappa}$ de estacionariedad se hace de género espacio *fuera* del horizonte del agujero negro. Dentro de la ergosfera, las partículas pueden tener energías negativas conservadas (medidas desde el infinito), y otras partículas con las que chocan directamente pueden escapar al infinito, llevándose el exceso. (b) Según el denominado «proceso de Penrose», este hecho puede aprovecharse para extraer energía rotacional del agujero negro. En el más simple de dichos procesos, una partícula entra en la ergosfera, se divide en dos partículas, una de las cuales entra en el agujero llevando energía negativa, y la otra escapa al infinito llevándose más energía que la que ha traído la partícula original.

Hawking, en cuanto que la energía procede de partículas o campos de energía negativa que son engullidos por el agujero, lo que da como resultado una energía positiva que escapa del agujero al infinito. No obstante, la diferencia importante es que con un agujero negro en rotación la región dentro de la cual el vector de Killing $\boldsymbol{\kappa}$ se hace de género espacio se extiende a una región *fuera* del horizonte del agujero negro. Esta región se conoce como la *ergosfera* (Fig. 30.12a). Así pues, en la ergosfera, las partículas pueden tener energía negativa (medida a partir del infinito) al tiempo que siguen siendo capaces de comunicarse con el infinito. Se hace posible, por ejemplo, que una partícula entre en la ergosfera desde fuera y luego se divida en dos, una de las cuales tiene energía negativa, ¡de modo que la otra escapa llevando más energía que la que ha introducido la partícula original![19] El resultado neto es extraer energía del agujero, reduciendo ligeramente la energía almacenada en su movimiento rotacional (Fig. 30.12b). Una conclusión similar puede obtenerse si están implicados campos (electromagnéticos) en lugar de partículas.[20]

Habría que resaltar que la «partícula de energía negativa» que cae

en el agujero es, vista localmente, una partícula ordinaria (con un 4-momento de género tiempo ordinario del tipo descrito en §18.7). Lo que ocurre simplemente es que la cantidad $p_a\kappa^a$, que mide la energía conservada, vista desde el infinito, se hace negativa, lo que puede suceder muy bien cuando la partícula está dentro de la ergosfera. Este es un hecho notable y muy potente en relación con los agujeros negros, pero no hay nada matemáticamente inconsistente o físicamente irrazonable en ello. Sin embargo, es este hecho el que permite que la con frecuencia enorme energía rotacional de un agujero negro sea expulsada al mundo exterior.

De hecho, la explicación más plausible de la enorme emisión de energía de un cuásar (véase §27.9) es que dicha energía procede de la rotación de un enorme agujero negro. La inmensa energía rotacional del agujero es extraída poco a poco —y lanzada al espacio—, por lo que es en esencia el proceso antes mencionado; véase la Fig. 30.13. Se ha propuesto que la energía negativa engullida por el agujero puede estar básicamente en forma de un campo electromagnético (por ejemplo, Blanford y Znajek 1977 y Begelman *et al.*, 1984) antes que en forma de partículas reales (por ejemplo, Williams 1995, 2002, 2004). Pero la razón subyacente es la misma en cada caso.

30.8. Explosiones de Hawking

Volvamos ahora al proceso de Hawking mecanocuántico. La temperatura para un agujero negro de una masa solar 1 $M_\odot$ es extremadamente baja, como se ha señalado en §30.4 (aproximadamente 10^{-7} K). Para agujeros negros mayores, esta temperatura sería incluso más baja, en proporción inversa a la masa del agujero (para una razón $a : m$ dada; véase §27.10). No hay evidencia astrofísica de que haya ningún agujero negro de masa menor que 1 $M_\odot$, de modo que no se cree que las temperaturas de agujeros negros sean de interés astrofísico directo.

De todas formas, hay un considerable interés *teórico* en esta temperatura, como apuntó espectacularmente Hawking en 1974.[21] Por ejemplo, si el universo es del tipo en perpetua expansión (véanse §27.11 y §28.10), entonces llegará un momento en el que la temperatura am-

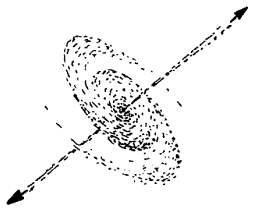


Fig. 30.13. La extraordinaria emisión de energía de un cuásar parece proceder de la energía rotacional de un enorme agujero negro en el centro de una galaxia. Este parece ser un proceso de la naturaleza general descrito en la Fig. 30.12, pero posiblemente mediante el engullido por el agujero negro de campos electromagnéticos de energía negativa antes que de partículas.

biente será menor que el valor para cualquier agujero negro dado. (Para un agujero negro de 1 $M_\odot$ en un universo $K = 0 = \Lambda$, esto llevaría unos 10^{16} años, que es alrededor de 10^6 veces la edad actual del universo.) Después de eso, el agujero negro empezaría a perder energía al irradiar más energía que la que absorbe del fondo. A medida que pierde energía, pierde masa, de modo que su radio se hace cada vez más pequeño y, por consiguiente, se calienta más. Imaginemos que empezamos con un agujero negro de 1 $M_\odot$. Seguiría radiando a un ritmo muy lento, perdiendo masa poco a poco durante unos 10^{64} años; su temperatura aumentaría lentamente al principio y luego lo haría a un ritmo cada vez más acelerado, hasta que alcanzara unos 10^9 K o 10^{10} K (las incertidumbres se deben a nuestra falta de conocimiento de la física de partículas a energías enormemente altas). En ese momento se produciría una inestabilidad desbocada, y habría una explosión en la que ¡la masa-energía restante del agujero se convertiría por completo, de forma casi instantánea, en radiación! (véase la Fig. 30.14).

Esta parece ser al menos la hipótesis más sencilla y de apariencia más natural, tal como originalmente la propuso Hawking. (Hawking había sugerido inicialmente que explosiones de esta naturaleza podrían ser detectables incluso ahora, ¡si el big bang hubiera sido tan amable de proporcionarnos un número significativo de «miniagujeros», de la masa de una montaña y el diámetro de un protón! Sin embargo, desde nuestra perspectiva actual, esto parece poco probable, y todavía no se ha identificado ninguna explosión semejante.) Otros físicos[22] han argumentado que aunque tuviera lugar una explosión final semejante, el agujero no desaparecería por completo, sino que dejaría algún «resto», o «pepita». La razón de que prefieran esto es que se sienten incómodos porque la «información» engullida por el agujero se perdería para el sistema, y ellos prefieren que quede «almacenada» en esta pepita rema-

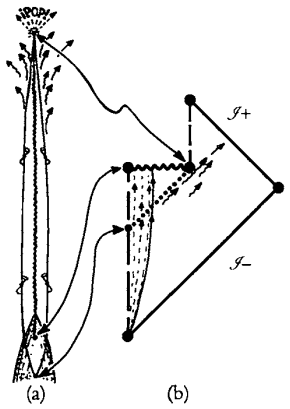


Fig. 30.14. Evaporación del agujero negro de Hawking. (a) Un agujero negro se forma por un colapso gravitatorio clásico. Luego, durante un período extraordinariamente largo, pierde masa-energía a un ritmo muy lento, a través de la radiación de Hawking, y se calienta poco a poco a medida que pierde masa. Finalmente, parece tener una desaparición explosiva (en una explosión que es pequeña para los niveles astrofísicos e independiente de la masa original del agujero). (b) Diagrama conforme estricto de este proceso (caso esféricamente simétrico). Esto parecería transmitir una clara imagen de acuerdo con la PÉRDIDA, donde el material que colapsa simplemente cae a través del horizonte llevando con él toda su «información», para ser finalmente destruido en la singularidad.

nente.[23] El problema es que resulta difícil ver cómo toda la información concerniente a los detalles de la materia que colapsó en el agujero —que podría incluso haber sido en su origen un agujero negro de tamaño estelar o incluso galáctico antes de que la radiación térmica (y por consiguiente, prácticamente «libre de información») se llevase casi toda la masa del agujero— podría estar almacenada en dicho resto. Como alternativa, algunos investigadores adoptan el punto de vista de que en la explosión final toda la información vuelve de nuevo «en el último minuto».

Estas tres alternativas son:

PÉRDIDA: la información se *pierde* cuando el agujero se evapora.
ALMACENAMIENTO: la información está *almacenada* en la pepita final.
RETORNO: toda la información se *recupera* en la explosión final.

El lector podría preguntarse por qué la gente siente la necesidad de llegar hasta el ALMACENAMIENTO o el RETORNO, cuando parece que la alternativa más obvia es la PÉRDIDA. La razón es que la PÉRDIDA parece implicar una violación de unitariedad, i.e., de la actuación de **U**. Si la filosofía de la mecánica cuántica de uno exige que la unitariedad sea inmutable, entonces uno está en dificultades con la PÉRDIDA. De ahí la

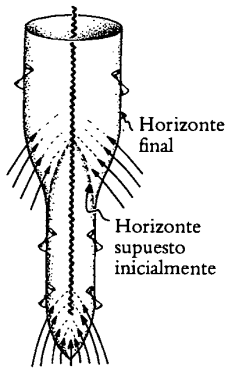


Fig. 30.15. La localización precisa del horizonte de un agujero negro está determinada «teleológicamente», pues depende de cuánto material cae en definitiva en el agujero.

popularidad entre muchos (y aparentemente la mayoría) de los físicos de partículas de las alternativas ALMACENAMIENTO o RETORNO, pese a la apariencia retorcida de las mismas.

Mi propia opinión es que la PÉRDIDA de la información es ciertamente la más probable. Un examen de la Fig. 30.14 transmite la imagen clara de que el material físico que colapsa cae simplemente a través del horizonte, llevando con él toda su «información», para ser finalmente destruido en la singularidad. Nada particular, de importancia física local, debería suceder en el horizonte. La materia ni siquiera «sabe» cuándo cruza el horizonte. Deberíamos tener en cuenta que podríamos estar considerando un agujero negro inicialmente muy grande, quizá como los agujeros que se cree que habitan en los centros galácticos, que podrían ser de un millón o más de masas solares. Cuando se cruza el horizonte no sucede nada en particular. La curvatura espaciotemporal y la densidad de materia no es grande: tan solo del tipo que encontramos en nuestro propio sistema solar. Ni siquiera la localización del horizonte está determinada por consideraciones locales, puesto que la localización depende de cuánto material caiga posteriormente en el agujero. Si cae más material posteriormente, entonces ¡el horizonte se habría cruzado en realidad antes! (véase la Fig. 30.15). Creo que es inconcebible que «en el instante inmediatamente anterior a que se cruce el horizonte» se emita alguna señal al mundo externo que transmita fuera los detalles completos de toda la información contenida en el material que colapsa. De hecho, una simple señal no sería

suficiente, puesto que el propio material es realmente, en cierto sentido, la «información» que a uno le interesa. Una vez que ha caído a través del horizonte, el material está atrapado y es inevitablemente destruido en la propia singularidad.

Esa es al menos la conclusión evidente si aceptamos la censura cósmica (§28.8). No veo que haya mucho margen en esto, incluso si no lo hacemos. La imagen es en esencia la de la Fig. 30.14. De acuerdo con esta imagen, el material que colapsa es destruido (y su «información» es destruida) solo cuando entra en la singularidad, no cuando cruza el horizonte. Si se va a mantener el punto de vista del RETORNO[24] —según el cual la información del material colapsante regresa de algún modo al suceso de la explosión final— indicado por la palabra «POP» en la Fig. 30.14, entonces hay que explicar de algún modo cómo se las arregla esta información para encontrar sigilosamente su vía de salida, para llegar a este punto desde la derecha a través de la singularidad (que, según una forma razonable de censura cósmica, debería ser esencialmente de género espacio; véase §28.8). No lo encuentro nada plausible.

La situación con el ALMACENAMIENTO no es mucho mejor, si es que la hay. Incluso si la pepita se forma, no sirve realmente para nada, puesto que la información está «encerrada» para siempre, y me parece que eso es lo mismo que perderse. Si el único propósito de la pepita es «salvar la unitariedad», entonces habría que formular una QFT consistente de pepitas, y hay varias dificultades para hacerlo.[25] Tal como lo veo, el razonamiento de Hawking nos ofrece un potente argumento por el que, de acuerdo con la PÉRDIDA, debe esperarse que la unitariedad sea simplemente violada en ciertas situaciones cuando la relatividad general entra en la imagen junto con los procesos mecanocuánticos.

¿Cuál es el punto de vista de Stephen Hawking con respecto a estas cuestiones? Desde el principio, él ha defendido con firmeza la PÉRDIDA, y me parece que el argumento para esto es tan poderoso ahora como cuando Hawking propuso estas ideas por primera vez. Por supuesto, la evaporación de un agujero negro es una noción completamente teórica, y podría darse el caso de que la propia naturaleza tenga otras ideas para el futuro remoto de los agujeros negros. No obstante, es difícil ver que pudiera ocurrir cualquiera de tales alternativas sin al-

gunos cambios radicales en la estructura de la QFT o de la relatividad general macroscópica (o de ambas). La posición de Hawking —al menos en 2003— era que la unitariedad debería violarse, pero solo en lo que yo consideraría un sentido más bien tibio. Hawking propuso que, en presencia de agujeros negros, el estado cuántico de un sistema evolucionaría realmente hacia una matriz densidad (estado no puro). De hecho, ya se ha aludido a esta idea en §29.6, al comentar el hecho de que si una parte de un estado cuántico entrelazado pudiera perderse genuinamente —aquí al caer parte de él en el agujero negro— frente a solo perderse FAPP, entonces podría ser razonable adoptar la postura ontológica de que la realidad cuántica tiene que describirse realmente por una matriz densidad antes que por un estado (puro). Hawking imaginó cierto tipo de evolución «superunitaria» que se aplica directamente a matrices densidad y permite que «estados puros» evolucionen hacia «estados mezcla».[26,[30.8]]

30.9. Una perspectiva más radical

Mi propio punto de vista es que mientras que estoy de acuerdo con Hawking en que probablemente sea correcta alguna forma de PÉRDIDA, creo que se necesita algo incluso más radical. Por ejemplo, la propuesta de Hawking, tal como se ha esbozado en el párrafo anterior, no incorpora ninguna característica con asimetría temporal.[27] Pero con simetría temporal sería admisible la imagen del «agujero blanco» de la Fig. 30.16a, que es el inverso temporal de la Fig. 30.4, como lo sería la inversa temporal del agujero negro en evaporación de la Fig. 30.14, tal como se muestra en la Fig. 30.16b. La «situación general con simetría temporal», en la que hay mucha destrucción de información con la misma creación de «nueva información» se ilustra en la Fig. 30.17. Todo esto viola la hipótesis de curvatura de Weyl (§28.8). El caso «simétrico» de la Fig. 30.17 implica la creación de un nuevo agujero blanco en el momento de la evaporación final del agujero negro original, y el agujero blanco crece hasta que alcanza el tamaño que tenía el agujero negro (véase la Fig. 30.17). ¡Nunca he visto sugerido seriamente un modelo de tan absurda apariencia! Una vez que se admiten situaciones como la de la Fig. 30.17, entonces no puedo ver por qué no proliferan en el big bang, llevando a una flagrante inconsistencia con el mensaje del capítulo 27.

[30.8] Utilice la notación de índices (por ejemplo ψ^a para $|\psi\rangle$) para indicar el tipo de transformación que podría conseguir esto. (*Sugerencia*: Eche una ojeada a la Fig. 29.5.)

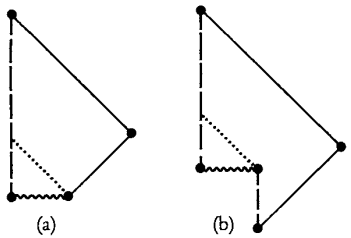


Fig. 30.16. Agujeros blancos: agujeros negros con el tiempo invertido. Estos violan la hipótesis de curvatura de Weyl. (a) Diagrama conforme de la inversión temporal de la formación de un agujero negro, como en las Figs. 27.11, 27.16. (b) Diagrama conforme de la inversión temporal de la formación de un agujero negro y posterior desaparición por radiación de Hawking, como en la Fig. 30.14.

No me propongo repetir aquí todos mis argumentos,[28] pero en líneas generales estos se basan en el hecho de que la naturaleza parece estar diciéndonos que algo muy parecido a la hipótesis de curvatura de Weyl sigue siendo cierto[29] para la estructura de aquellas singularidades espaciotemporales que ella permite realmente en su universo.* Si aceptamos esto, entonces hay una «pérdida de información» neta en las singularidades de los agujeros negros que no se recupera. La razón es que, según esta hipótesis, las singularidades finales del colapso pueden contener —y, por consiguiente, absorber— números enormes de grados de libertad (que residen en la propia curvatura de Weyl), mientras que estos grados de libertad están prohibidos para cualquier singularidad inicial.

Tratemos de poner este argumento en términos del *espacio de fases* de un sistema, incluyendo la formación y evaporación de agujeros negros. Estrictamente, para que funcione nuestro argumento del espacio de fases, deberíamos considerar un sistema cerrado que contiene una cantidad de energía finita dada. Para ayudar a nuestra imaginación, tratemos de pensar una caja enorme, de dimensiones más que galácticas,

* «Permite» se refiere simplemente a lo que sucede cuando las nociones de la presunta «geometría cuántica» apropiada cristalizan en un espaciotiempo clásico.

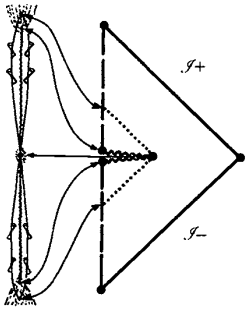


Fig. 30.17. (a) Situación con simetría temporal, donde hay creación de un agujero blanco en el momento de la evaporación final de un agujero negro, que se había formado por colapso gravitatorio. El nuevo agujero blanco crece hasta alcanzar el tamaño del agujero negro previo antes de desaparecer con la expulsión de una gran cantidad de material. (b) Diagrama conforme de esto.

con paredes que deben considerarse como espejos perfectos, de modo que ni información ni partículas materiales pueden cruzarlas en uno u otro sentido (véase la Fig. 30.18). Por supuesto, esto es absurdo en la práctica, ¡pero me apresuro a asegurar al lector que nuestro sistema constituye simplemente un «experimento mental», no uno real! Solo se contempla[30] para permitir que el razonamiento del espacio de fases se aplique a un sistema que implica la (aparente) pérdida de grados de libertad en el proceso de radiación de Hawking. El espacio de fases $\mathcal{P}$ en consideración describe todos los estados físicos posibles dentro de nuestra caja hipotética, con la energía total dada. La evolución dinámica se describe, en la Fig. 30.19, mediante una familia de flechas en $\mathcal{P}$, a la manera de la Fig. 20.5.

En esta situación («mental»), a medida que avanza el tiempo desaparecen grados de libertad que son absorbidos en las singularidades del agujero negro. La hipótesis de curvatura de Weyl obliga a que estos grados de libertad no reaparezcan en singularidades iniciales (tipo agujero blanco), pero mi tesis es que *sí* reaparecen vía el proceso **R**. La idea es que existe un equilibrio global entre la «pérdida de información» temporalmente asimétrica en los agujeros negros y el comportamiento temporalmente asimétrico de las probabilidades en los procesos **R** mecanocuánticos que se ha mostrado en §30.3. La naturaleza no determinista del proceso **R** nos dice que puede haber varias salidas alternativas para un mismo input, y esto va a compensar el hecho de que con los agujeros negros puede haber muchos inputs diferentes para la misma salida, al ser absorbida en las singularidades la «información»

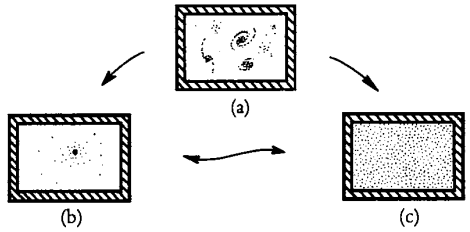


Fig. 30.18. Experimento mental de la «caja» de Hawking. (a) Imaginemos una «caja» enorme (de escala galáctica) de materia, cuyas paredes son espejos perfectos que no permiten que ninguna información o material entre o salga. (b) Un máximo de entropía local es un agujero negro que proporciona la mayor parte de la masa, pero con una pequeña cantidad de radiación a su alrededor en equilibrio térmico con el agujero. (c) Otro máximo de entropía local es simplemente la radiación térmica (y unas pocas partículas), pero sin agujero negro.

que distingue los diversos inputs. Recordemos que en el experimento hipotético de §30.3 (Fig. 30.1) teníamos diferentes salidas (fotones que llegan a D y fotones que llegan a C) para un input dado (fotón emitido desde S), mientras que para una salida dada (fotón que llega a D) había básicamente solo un input (fotón emitido en S). Así pues, obtenemos una ampliación efectiva del volumen del espacio de fases de acuerdo con el proceso **R**, mientras que la asimetría en la estructura de la singularidad espaciotemporal provoca una reducción efectiva del volumen del espacio de fases (véase de nuevo la Fig. 30.19). La tesis es que estos dos efectos deberían cancelarse en promedio.

Debería quedar claro que esta cancelación va a ser solo una característica *global* de los procesos físicos. No se está afirmando, por supuesto, que tenga que haber una presencia simultánea de un agujero negro acompañando cada caso de reducción del estado cuántico. La idea es solo que a lo largo del espacio de fases entero hay un equilibrio entre estos dos efectos. Por consiguiente, es la *potencialidad* para la formación de agujeros negros, y su capacidad para absorber información, lo que va a cancelar la aleatoriedad futura en **R**.

Podemos notar que ambos efectos violan el teorema (teorema de Liouville; véanse §20.4 y la Fig. 20.7) según el cual el volumen del espacio de fases tiene que conservarse en la evolución dinámica. Pero en cada caso tenemos algo que va más allá de la dinámica clásica ordina-

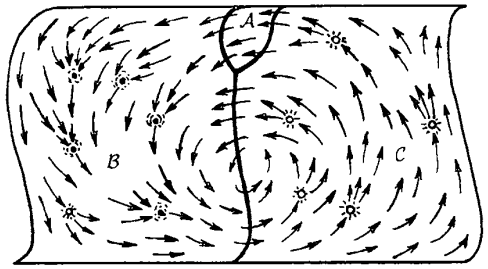


Fig. 30.19. Descripción en el espacio de fases de la caja de Hawking, donde las flechas describen la evolución hamiltoniana de los procesos involucrados en la Fig. 30.18. Las regiones $\mathcal{A}, \mathcal{B}, \mathcal{C}$ corresponden, respectivamente, a (a), (b), (c) en la Fig. 30.18. Por consiguiente, un agujero negro está presente en la región $\mathcal{B}$, pero no en la región $\mathcal{C}$. La presencia de un agujero negro da como resultado una confluencia de líneas de flujo (reducción del volumen del espacio de fases), de acuerdo con *c* PÉRDIDA, debido a la destrucción de información en la singularidad (futura) del agujero negro. Hay una creación compensatoria de líneas de flujo (aumento de volumen en el espacio de fases) implicada en la asimetría temporal del proceso **R** (que se supone objetivamente real) (véase la Fig. 30.1). La propuesta es que debería haber un equilibrio global entre estos dos procesos que violan el teorema de Liouville, dando finalmente una conservación del volumen del espacio de fases en el flujo.

ria que es el escenario de dicho teorema. De hecho, la misma noción de espacio de fases clásico no es aquí del todo adecuada, cuando estamos considerando efectos cuánticos y clásicos juntos. Para un sistema puramente cuántico, deberíamos pensar enteramente en términos de un *espacio de Hilbert*. Para quienes creen que la evolución cuántica **U** es toda la historia, la descripción mediante el espacio de Hilbert es la correcta. Pero entonces la destrucción de información (y por consiguiente de unitariedad) en la evaporación de agujeros negros presenta un serio problema. Mi punto de vista es que ninguna de las dos imágenes es enteramente adecuada, y deberían considerarse como una aproximación a algo más que todavía no sabemos cómo describir.[31]

Desde hace tiempo, mi intención ha sido obtener una estimación cuantitativa directa del ritmo de la reducción del estado cuántico investigando los detalles del balance entre estos dos procesos, tal como se ha esbozado en el argumento anterior (e ilustrado en la Fig. 30.19),

pero hasta el momento he sido incapaz de llevar a buen fin este argumento. Por lo tanto, es una suerte que exista una línea de razonamiento completamente diferente que puede utilizarse realmente para obtener esta estimación apropiada. Consideremos esto a continuación.

30.10. El bulto de Schrödinger

Volvamos al tipo de situación considerada en §29.7, conocida como «gato de Schrödinger». En la Fig. 29.7 he ilustrado cómo se podría establecer una superposición cuántica de un gato vivo y un gato muerto utilizando un divisor de haz para colocar un fotón en una superposición, donde la parte transmitida del estado del fotón dispara un dispositivo para matar al gato, mientras que la parte reflejada deja al gato vivo. Por supuesto, el uso de un gato real no solo sería inhumano, sino que requeriría un sistema físico innecesariamente complicado. De modo que en su lugar consideramos que el estado del fotón transmitido activa un dispositivo que desplaza un bulto de material en una pequeña distancia horizontal, mientras que la parte reflejada deja el bulto invariable; véase la Fig. 30.20. El bulto superpuesto desempeña ahora el papel del gato de Schrödinger, ¡aunque no tan espectacularmente como antes!

La cuestión que quiero plantear ahora es la siguiente: ¿es la superposición cuántica de las dos posiciones del bulto un estado estacionario? Esto es lo que sucedería en la mecánica cuántica convencional, si consideramos que cada posición del bulto por separado representa un estado estacionario y que la energía es la misma en ambos casos (de modo que la posición de reposo del bulto desplazado no está más alta ni más baja que su posición original). Esto es simplemente una aplicación elemental de las reglas que hemos aprendido en el capítulo 21 (véase también §24.3). Representando la posición del bulto original por $|\chi\rangle$ y la desplazada por $|\varphi\rangle$, tenemos las dos ecuaciones de Schrödinger que describen el carácter estacionario de cada una de las dos posiciones del bulto

$$i\hbar\frac{\partial|\chi\rangle}{\partial t} = E|\chi\rangle, \quad i\hbar\frac{\partial|\varphi\rangle}{\partial t} = E|\varphi\rangle,$$

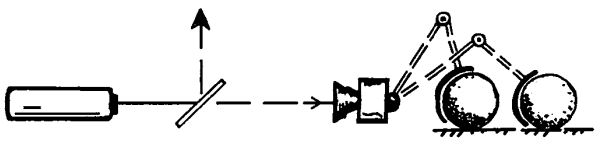


Fig. 30.20. El «gato» de Schrödinger de la Fig. 29.7, pero ahora la superposición cuántica resultante es entre dos localizaciones ligeramente diferentes de un bulto de materia.

que nos dan sendos autoestados de la energía, con el mismo autovalor E para la energía. Si la superposición se representa como el estado

$$|\Psi\rangle = w|\chi\rangle + z|\varphi\rangle,$$

entonces obtenemos directamente

$$i\hbar\frac{\partial|\Psi\rangle}{\partial t} = E|\Psi\rangle,$$

cualesquiera que pudieran ser los valores de las amplitudes (constantes) w y z.[30.9] Así pues, cada superposición cuántica $|\Psi\rangle$ es también un estado estacionario. Si cada uno de los estados $|\chi\rangle$ y $|\varphi\rangle$ estuviera allí individualmente para siempre, entonces también lo estaría toda superposición cuántica $|\Psi\rangle$ de ellos. Esta es simplemente la predicción de la mecánica cuántica estándar.

Empecemos ahora a introducir las lecciones que Einstein nos ha enseñado con su soberbia y ahora excelentemente comprobada teoría de la relatividad. En primer lugar, podríamos considerar importante introducir el campo gravitatorio que se expresa en la geometría espaciotemporal de fondo. Podemos imaginar que el experimento se está realizando en la Tierra, con las dos copias del bulto situadas en una plataforma horizontal. La geometría espaciotemporal de la Tierra no es exactamente plana, y debemos considerar qué efecto podría tener esta curvatura espaciotemporal en las consideraciones anteriores. De hecho, debemos preocuparnos un poco por el significado mismo del operador

[30.9] ¿Por qué? Explique qué propiedades de un campo vectorial $\boldsymbol{\kappa}$ se están utilizando cuando repetimos esta conclusión en el caso de un espaciotiempo de fondo estacionario.

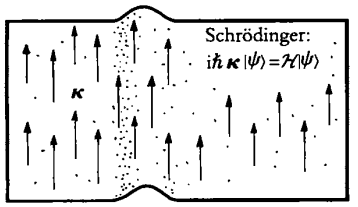


Fig. 30.21. El operador de diferenciación «$\partial/\partial t$» en la ecuación de Schrödinger debe considerarse (de forma invariante) como un campo vectorial $\boldsymbol{\kappa}$ en la variedad espaciotemporal (véase la Fig. 30.3), donde la estacionariedad del espaciotiempo se expresa en que $\boldsymbol{\kappa}(= \partial/\partial t)$ es un campo de Killing de género tiempo ($\pounds_{\boldsymbol{\kappa}}\mathbf{g} = 0$; véase §14.7).

«$\partial/\partial t$» que aparece en la ecuación de Schrödinger. En la relatividad general no solemos tener un sistema de coordenadas definido de forma natural y con respecto al cual estaría definido el concepto de «$\partial/\partial t$». Recordemos de §10.3 y §12.3 (véase la Fig. 10.5) que la forma «invariante» de pensar en un operador de derivadas parciales (como $\partial/\partial t$) es considerarlo como un campo vectorial en la variedad (espaciotemporal; véase la Fig. 30.21). Así pues, necesitaremos un campo vectorial en nuestro espaciotiempo para expresar nuestra noción requerida de «$\partial/\partial t$».

En la situación actual no estamos tan mal, porque estamos considerando la cuestión de «estados estacionarios», de modo que al menos debemos tener un espaciotiempo de fondo que es en sí mismo estacionario. De hecho, consideramos que el campo de la Tierra es estacionario. Como hemos visto (§§30.4,6, Fig. 30.3), un espaciotiempo estacionario se caracteriza por la existencia de un *vector de Killing* $\boldsymbol{\kappa}$ *de género tiempo*. ¿Cómo interviene este vector de campo concreto en la discusión? Nuestro espaciotiempo es estacionario en el sentido de ser «independiente de t», lo que nos dice que simplemente podemos hacer el reemplazamiento (Fig. 30.21) en las fórmulas anteriores

$$\frac{\partial}{\partial t} \mapsto \boldsymbol{\kappa}.$$

Puede plantearse una cuestión de un factor de escala global constante, pero esto no es muy importante aquí para nosotros. La forma habitual de fijar este factor global es exigir que $\boldsymbol{\kappa}$ se convierta en un desplazamiento temporal «ordinario» a grandes distancias, donde se considera que el campo gravitatorio decae a cero. Sin embargo, localmente la magnitud de $\boldsymbol{\kappa}$ puede cambiar de un lugar a otro, de una forma que tiene en cuenta los efectos de «frenado de reloj» del campo gravitato-

rio de la Tierra (§19.8).[30.10] Puesto que $\boldsymbol{\kappa}$ está tomando ahora el papel de $\partial/\partial t$, nuestras ecuaciones de Schrödinger estacionarias individuales son

$$i\hbar\boldsymbol{\kappa}|\chi\rangle = E|\chi\rangle, \quad i\hbar\boldsymbol{\kappa}|\varphi\rangle = E|\varphi\rangle,$$

e, igual que antes, deducimos que para cualquier superposición $|\Psi\rangle$ seguimos teniendo

$$i\hbar\boldsymbol{\kappa}|\Psi\rangle = E|\Psi\rangle.$$

Así pues, la presencia de un campo gravitatorio estacionario como fondo no altera el hecho de que cualquier superposición cuántica de los dos estados estacionarios $|\chi\rangle$ y $|\varphi\rangle$ es en sí misma estacionaria.

Pero veamos ahora lo que sucede cuando tenemos en cuenta el *propio* campo gravitatorio del bulto. Si consideramos cada uno de los estados $|\chi\rangle$ y $|\varphi\rangle$ por separado, entonces parece que no tenemos ningún problema real. Por supuesto, cada uno de los $|\chi\rangle$ y $|\varphi\rangle$ es un estado cuántico y, en ausencia de una teoría cuántica de la gravedad aceptada, no podemos saber cómo tratar su campo gravitatorio. Pero no importa realmente el hecho de que no sepamos cómo hacer esto en detalle. El punto de vista convencional afirmaría que la teoría cuántica de la gravedad correcta puede acomodar cosas que se parecen a bultos clásicos de material con campos gravitatorios que están descritos de forma muy precisa de acuerdo con los principios de la relatividad general clásica de Einstein, aunque no con toda exactitud. (En mi opinión, la validez de este «punto de vista convencional» podría cuestionarse, pero si creemos las hipótesis gemelas estándar —que no es necesario cambiar el formalismo cuántico, y también que la relatividad general clásica es válida para cuerpos macroscópicos—, entonces debemos aceptarlo. Después de todo, la naturaleza del argumento consiste en explorar los límites de compatibilidad de estas dos suposiciones.) Por consiguiente, debería haber un estado cuántico $|\chi\rangle$ y un estado cuántico $|\varphi\rangle$ que

[30.10] Vea si puede dar una explicación de esto utilizando la ley de conservación que proporciona un vector de Killing $\boldsymbol{\kappa}$, como se describe en §30.6 y teniendo en cuenta el hecho de que la norma $\kappa_a\kappa^a$ puede diferir de la unidad en la vecindad de un cuerpo gravitante, incluso si está normalizada a la unidad a grandes distancias del cuerpo. ¿Cómo afecta esto a la medida del tiempo?

describan muy aproximadamente el bulto de material, situado en la plataforma horizontal en la Tierra, en cada una de sus dos posiciones separadas, y donde cada copia del bulto esté acompañada por su campo gravitatorio einsteiniano clásico.[32] Puesto que cada uno de estos dos estados de posición del bulto se considera estacionario en su espaciotiempo acompañante, cada uno tendrá su respectivo vector de Killing asociado,[33] $\boldsymbol{\kappa}_\chi$ y $\boldsymbol{\kappa}_\varphi$, y satisfará su ecuación de Schrödinger apropiada con valor propio E:

$$i\hbar\boldsymbol{\kappa}_\chi|\chi\rangle = E|\chi\rangle \quad \text{e} \quad i\hbar\boldsymbol{\kappa}_\varphi|\varphi\rangle = E|\varphi\rangle.$$

En la situación anterior, cuando ignorábamos los campos gravitatorios de los bultos, éramos capaces de escribir la ecuación de Schrödinger para cualquier superposición $w|\chi\rangle + z|\varphi\rangle$ y averiguar que todas estas son estacionarias. No obstante, hay un problema, porque estos dos vectores de Killing $\boldsymbol{\kappa}_\chi$ y $\boldsymbol{\kappa}_\varphi$ son diferentes. ¿Qué vamos a hacer? Parece que necesitamos una noción invariante de «$\partial/\partial t$» que se aplique a espaciotiempos superpuestos, y ni $\boldsymbol{\kappa}_\chi$ ni $\boldsymbol{\kappa}_\varphi$ satisfacen esta necesidad. En la próxima sección veremos que este no es un problema menor, sino que nos presenta una dificultad fundamental, y nos lleva a un choque entre los principios fundacionales de la mecánica cuántica y los de la relatividad general.

30.11. Conflicto fundamental con los principios de Einstein

Es importante desarrollar este punto con más profundidad. Cuando digo que los vectores de Killing $\boldsymbol{\kappa}_\chi$ y $\boldsymbol{\kappa}_\varphi$ son diferentes, lo entiendo en un sentido profundo. ¡Son realmente campos vectoriales en espaciotiempos diferentes! Se podría tratar de adoptar el punto de vista de que estos dos espaciotiempos solo difieren en que tienen estructuras métricas muy ligeramente diferentes. En consecuencia, podríamos tratar de considerarlos como si fueran realmente uno y el mismo espacio, pero con campos tensoriales métricos especificados de forma ligeramente diferente, digamos $\boldsymbol{g}_\chi$ y $\boldsymbol{g}_\varphi$. Pero adoptar esta posición es abandonar uno de los principios básicos de la teoría de Einstein, a saber, el *principio de*

covariancia general (véase §19.6). Considerar que los conjuntos de puntos de estos dos espaciotiempos son, en cierto sentido, los «mismos» conjuntos de puntos sería, de hecho, especificar una identificación puntual entre los dos espaciotiempos. Eso sería como identificar un punto en un espaciotiempo con el punto en el otro espaciotiempo que tiene las mismas etiquetas coordenadas. Pero el principio de covariancia general niega cualquier significado a sistemas de coordenadas particulares. De hecho, afirma que no debería haber una identificación punto a punto preferida entre dos espaciotiempos diferentes.

¿Por qué causa dificultades esta falta de identificación entre los espaciotiempos de las dos posiciones del bulto? Necesitamos poder escribir la ecuación de Schrödinger. Pero sin un «$\boldsymbol{\kappa}$» único, ¿cómo vamos a hacerlo? La sugerencia más inmediata podría ser la de tratar de identificar $\boldsymbol{\kappa}_\chi$ con $\boldsymbol{\kappa}_\varphi$, pero eso sería cometer ciertamente una violación de un principio fundamental de la teoría de Einstein, puesto que implicaría que estamos considerando que estos dos vectores de Killing habitan el mismo espacio, ¡y eso es hacer trampa! Me parece que en tal situación, estamos empezando a ser testigos de un choque entre los principios fundamentales de la mecánica cuántica y los de la relatividad general.

De todas formas, no deberíamos «abandonar» en este punto. Aunque estrictamente hablando necesitaríamos la nueva teoría apropiada para saber qué hacer a continuación, creo que podemos hacer un progreso genuino si estamos dispuestos a aceptar este choque por el momento y buscar alguna medida del *error* implicado en nuestra «trampa». Adoptemos el punto de vista de que, en cierto sentido, lo que la naturaleza estaría dispuesta a hacer sería permitir la identificación de dos espaciotiempos localmente con tal de que la noción de «caída libre» sea la misma en cada uno de ellos. Esto refleja de algún modo el principio de equivalencia (véase §17.4). Nuestra identificación tentativa trataría de hacer que la noción de geodésica en un espacio coincida con la noción de geodésica en el otro. Ello no se puede conseguir normalmente excepto en la vecindad inmediata de un punto, de modo que, en su lugar, trataremos de calcular el error que está implicado en esto si en efecto identificamos los dos espaciotiempos. Una cosa así es difícil de hacer en la relatividad general completa, pero podemos aplicar también

la mayoría de estas ideas en la situación límite en la que se considera que la velocidad de la luz c es infinita, reteniendo, de todas formas, buena parte de la filosofía básica de la teoría de Einstein. Esta situación nos lleva a la formulación de Cartan de la gravedad newtoniana, como se ha discutido en §17.5.[34]

Recordemos del capítulo 17 que en el esquema gravitatorio de Newton-Cartan, el espaciotiempo es realmente un fibrado sobre el espacio euclídeo 1-dimensional $\mathbb{E}^1$ de los diferentes «tiempos» admisibles t. Las fibras son los diferentes 3-espacios euclídeos $\mathbb{E}^3$, cada uno de los cuales se refiere al «espacio» en un instante dado. Así pues, tenemos realmente un «tiempo absoluto», descrito por la coordenada temporal t. El lector podría ser excusado si quizá piensa que, puesto que ahora tenemos la misma noción de tiempo (medida por t) para los espaciotiempos de ambas posiciones del bulto, nuestros problemas deberían haber desaparecido. Pero la triste realidad es que conocer t no nos permite conocer $\partial/\partial t$, pues el operador $\partial/\partial t$ requiere saber que las variables coordenadas restantes (digamos x, y, z) se están manteniendo fijas. Esta es la cuestión de la «segunda confusión fundamental del cálculo», considerada en §10.3 (véase la Fig. 10.7). Podemos ver claramente la cuestión por referencia a la geometría que está implicada. Saber t nos dice dónde están las secciones $\mathbb{E}^3$, pero saber $\partial/\partial t$ nos diría un campo vectorial de Killing, que define una familia de curvas que cortan a esta familia de 3-superficies (véase la Fig. 30.22). De hecho, esta cuestión general de no ser capaces de especificar la $\partial/\partial t$ de Schrödinger se considera una cuestión profunda incluso en los enfoques más «convencionales» de la gravedad cuántica. Está relacionada con el denominado «problema del tiempo» en cosmología cuántica.[35]

En el contexto actual no estoy tratando de ser tan ambicioso como para resolver estas cuestiones. Todo lo que necesitamos es una estimación del error implicado si tratamos de hacer una identificación «ilegal» de los vectores diferentes $\boldsymbol{\kappa}_\chi$ y $\boldsymbol{\kappa}_\varphi$. Hacemos esto identificando realmente los $\mathbb{E}^3$ pero tomando luego el error total en la diferencia entre las aceleraciones gravitatorias (diferencias entre caídas libres, i.e., geodésicas) en los dos espacios. Supongamos que las aceleraciones gravitatorias están dadas, respectivamente, por los 3-vectores $\boldsymbol{\Gamma}_\chi$ y $\boldsymbol{\Gamma}_\varphi$. Entonces estimamos nuestro error tomando el cuadrado de la longitud de su

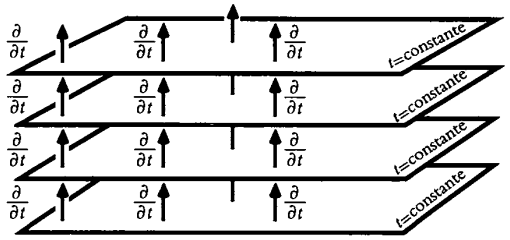


Fig. 30.22. El conocimiento de t no nos dice $\partial/\partial t$ («segunda confusión fundamental del cálculo»; véanse la Fig. 10.7 y §10.3); t nos dice dónde están las secciones $\mathbb{E}^3$, pero $\partial/\partial t$ define una familia de curvas que atraviesan esta familia de 3-superficies.

diferencia $(\mathbf{\Gamma}_\chi - \mathbf{\Gamma}_\varphi)^2$ e integrándolo sobre la totalidad de $\mathbb{E}^3$. Este error integrado se interpreta como una medida de la incertidumbre absoluta en la definición del operador «$\partial/\partial t$» necesario para la ecuación de Schrödinger, en el instante t que especifica dicha elección particular de $\mathbb{E}^3$. Esta incertidumbre conduce directamente, vía la ecuación de Schrödinger, a una incertidumbre absoluta E_G en la energía de los estados superpuestos en consideración. El paso siguiente es convertir esta expresión para E_G en otra forma matemática (equivalente) que podemos interpretar[30.11] como:

E_G = autoenergía gravitatoria de la diferencia entre las dos distribuciones de masa en los estados $|\chi\rangle$ y $|\varphi\rangle$.

La *autoenergía gravitatoria* en una distribución de masa es la energía que se gana al reunir dicha distribución de masa a partir de masas puntuales completamente dispersas en el infinito. La diferencia anterior podría considerarse como la distribución de masa en $|\chi\rangle$ tomada positivamente, junto con la distribución de masa en $|\varphi\rangle$ tomada negativamente (véase la Fig. 30.23). (La razón que hace que esto no nos dé simplemente cero es que la energía tiene que ver con el efecto del campo gravitatorio de cada distribución de masa sobre la otra.)

[30.11] Vea si puede confirmar esto. La demostración sigue líneas similares a las del ejercicio [24.3] en §24.3. Hacemos uso de la ecuación de Poisson $\nabla^2\Phi = -4\pi\rho$, donde Φ es el potencial gravitatorio (escalar) newtoniano. Aquí nuestra estimación del «error» es la integral espacial de $|\nabla\phi_1 - \nabla\phi_2|^2$.

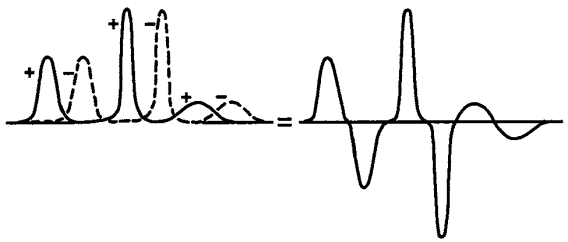


Fig. 30.23. Cada uno de los dos estados estacionarios en superposición, $|\chi\rangle$ y $|\varphi\rangle$, define un «valor esperado» para su distribución de densidad de masa. La diferencia entre estos dos (i.e., uno tomado positivamente y el otro negativamente) forma una distribución de densidad de masa positiva y negativa cuya autoenergía gravitatoria es la cantidad E_G.

Esta es una pequeña dificultad para que sea apreciada en términos ordinarios, debido especialmente a las distribuciones de masa negativa implicadas. Es una suerte que en la situación más habitualmente considerada, a saber, cuando el estado $|\varphi\rangle$ es solo un desplazamiento rígido del estado $|\chi\rangle$, la cantidad E_G puede interpretarse de forma más directa de otro modo. Consideramos la energía que costaría llevar una copia de nuestro bulto, originalmente en la posición $|\chi\rangle$, pero desplazada a la posición $|\varphi\rangle$, lejos del campo gravitatorio de la otra que se considera fija en la posición $|\chi\rangle$. Esta energía resulta ser la misma energía E_G que antes en el caso de un desplazamiento rígido,[30.12] pero no siempre en otras circunstancias.

De hecho, se podría considerar que se adopta esta segunda medida de energía (a saber, la energía de *interacción* gravitatoria) como una definición alternativa de E_G. La primera propuesta, en términos de autoenergía gravitatoria, parece estar mejor fundada, por lo que puedo ver, pero no habría que descartar otras posibilidades en nuestro estado de conocimiento actual. Diósi (1989) había considerado ambas propuestas anteriores, con un objetivo similar al que voy a plantear, aunque proponiendo también una dinámica (estocástica), algo que no

[30.12] ¿Puede ver por qué esto da la misma respuesta para E_G que antes en esta situación concreta? ¿Qué sucedería si la localización final del bulto desplazado está ligeramente elevada con respecto a su localización inicial? ¿Qué sucedería si se comprime?

hago aquí. Estas diferentes sugerencias (y algunas otras) deberían ser distinguibles experimentalmente, en experimentos del tipo de los que expondré enseguida. Sin embargo, habría que señalar que incluso las propuestas mejor fundadas tienen una motivación incompleta y no están del todo libres de controversias.[36]

De modo que ¿qué vamos a hacer con nuestra fundamental «incertidumbre en la energía» E_G? El próximo paso es invocar una forma del principio de incertidumbre de Heisenberg (la relación de incertidumbre tiempo-energía; véase §21.11). Un hecho familiar en el estudio de partículas inestables o núcleos inestables (tal como el uranio U_{238}) es que la *vida media* T, que tiene una incertidumbre temporal intrínseca, guarda una relación inversa con una incertidumbre en la energía, dada por $\hbar/2T$. Por ejemplo, como se ha advertido en §21.11, la vida media de un núcleo de U_{238} es de aproximadamente 10^9 años, de modo que hay una incertidumbre fundamental en la energía de cada núcleo de aproximadamente 10^{-51} julios, que se traduce, mediante la expresión $E = mc^2$ de Einstein, en una incertidumbre de masa de aproximadamente 10^{-44} de su masa total. Ahora vamos a considerar que nuestro estado superpuesto $|\Psi\rangle = w|\chi\rangle + z|\varphi\rangle$ es análogo a esto, siendo él mismo inestable, con una vida media T_G que está relacionada, mediante la fórmula de Heisenberg, con la incertidumbre de la energía fundamental E_G que se ha expuesto antes. De acuerdo con esta imagen,[37] cualquier superposición como $|\Psi\rangle$ decaería a uno u otro de sus estados constituyentes, $|\chi\rangle$ o $|\varphi\rangle$, en una escala de tiempo media de

$$T_G \approx \hbar/E_G.$$

30.12. ¿Estados de Schrödinger-Newton preferidos?

Parece que el resultado del argumento anterior es que una superposición cuántica de dos estados debería decaer a uno u otro de sus constituyentes en una escala de tiempo del orden $\hbar/E_G$. Pero en este punto el lector perspicaz muy bien podría quejarse de que *cualquier* estado cuántico $|\psi\rangle$ puede expresarse como una superposición lineal de un par de otros estados (por ejemplo, $|\psi\rangle = |\alpha\rangle + (|\psi\rangle - |\alpha\rangle)$, para cualquier $|\alpha\rangle$). No tendría ningún sentido considerar que todos estos estados se desintegran en tales «constituyentes», especialmente si para un $|\psi\rangle$ dado escogemos $|\alpha\rangle$, de modo que las distribuciones de masa en dichas alternativas difieren lo suficiente para que la desintegración tuviera que ser ¡casi instantánea!

Podría juzgarse que un absurdo de esta naturaleza es la conclusión de la discusión anterior, incluso si vamos a considerar que nuestra superposición $|\Psi\rangle = w|\chi\rangle + z|\varphi\rangle$ involucra meramente a un único electrón. Pues, en efecto, podríamos tomar $|\chi\rangle = |\alpha\rangle$ para representar el electrón en una posición (casi) precisa. La distribución de masa sería prácticamente una función delta (§21.10), que conduce a un valor esencialmente infinito para E_G, lo que parecería implicar una reducción casi instantánea del estado $|\Psi\rangle$ a uno u otro de $|\chi\rangle$ o $|\varphi\rangle$. Lo mismo sería válido para un sistema compuesto de entidades puntuales (por ejemplo, quarks). Es evidente que esto no tiene sentido; si semejante comportamiento fuera cierto, entonces no habría mecánica cuántica.

Lo que debemos hacer es ser mucho más cuidadosos sobre qué tipo de estados vamos a admitir como nuestros $|\chi\rangle$ y $|\varphi\rangle$. Recordemos que en el argumento anterior se ha considerado que $|\chi\rangle$ y $|\varphi\rangle$ son estados estacionarios. Un electrón en un estado de (casi) posición dada no es ciertamente estacionario. Por el principio de incertidumbre posición/momento de Heisenberg (§21.11), incluiría momentos muy grandes y se dispersaría instantáneamente. Por otra parte, parece que tenemos una dificultad en aplicar siquiera el argumento a partículas individuales si exigimos una estacionariedad *exacta* para ambos $|\chi\rangle$ y $|\varphi\rangle$, pues no existen soluciones estacionarias de la ecuación de Schrödinger ordinaria, para una única partícula libre (de masa positiva), que decaigan hacia el infinito espacial.[30.13] La respuesta a esta paradoja reside en el hecho de que debemos tener en cuenta el campo gravitatorio de la partícula cuando escribamos su ecuación de Schrödinger. No estoy pidiendo que el propio campo gravitatorio esté cuantizado en esta descripción, sino meramente que sus efectos estén resumidos en una función para el potencial gravitatorio newtoniano Φ, cuya fuente va a ser lo que se denomina el «valor esperado» de la distribución de masa en la función de onda. Quizá no sea oportuno que ofrezca en este libro una

[30.13] ¿Por qué? (*Sugerencia*: Eche una ojeada de nuevo al ejercicio [24.3]).

descripción completa de lo que está aquí involucrado,[38] pero esta receta parece dar respuestas razonables. Los detalles de esto son materia de activa investigación en la actualidad. Se concluye que para una única partícula, esta ecuación de Schrödinger modificada —que llamaré ecuación de *Schrödinger-Newton* (debido a que incorpora un campo gravitatorio newtoniano)— tiene realmente soluciones estacionarias de buen comportamiento para una única partícula que decaen adecuadamente hacia el infinito. (Sin embargo, para un único electrón la dispersión en la función de onda excedería la medida del universo observable; para un átomo de hidrógeno sería un poco menor que el universo observable, pues la dispersión decrece como el inverso del cubo de la masa de la partícula.)

Tenemos ahora lo que parece ser una propuesta plausible para una reducción de estado objetiva que se aplica, al menos, en situaciones en las que un estado cuántico es una superposición de otros dos estados, cada uno de los cuales es estacionario (en el sentido de Schrödinger-Newton antes mencionado). Según esta propuesta, tal estado superpuesto se reducirá espontáneamente a uno u otro de sus constituyentes estacionarios en una escala de tiempo media de aproximadamente $\hbar/E_G$, donde E_G es la autoenergía gravitatoria de la diferencia entre las dos distribuciones de masa. Llamaré a esta propuesta la ***RO** gravitatoria* (donde **RO** significa «reducción objetiva» del estado cuántico). Para cualquier par de tales estados estacionarios constituyentes, la magnitud autoenergía gravitatoria E_G está bien definida. Se refiere a la diferencia entre las dos distribuciones de masa, cada una de las cuales es esa misma expresión del «valor esperado» utilizada al definir la ecuación de Schrödinger-Newton.

Una característica de todas las demás propuestas para un esquema **RO** es que entran en dificultades con la *conservación de la energía*. En particular, la ingeniosa y rompedora propuesta de Giancarlo Ghirardi, Alberto Rimini y Tullio Weber (1986) tropieza precisamente con esta clase de dificultades, como lo han hecho otras propuestas.[39] Ha sido una actitud común la de «vivir con» este problema, con tal de que la no conservación de la energía pudiera reducirse a un nivel aceptablemente minúsculo. Mi propia perspectiva sobre esta cuestión consiste en tomarla más en serio. En el esquema gravitatorio **RO** que se ha propuesto antes existe la ventaja de que la incertidumbre de la energía en E_G

parecería cubrir dicha no conservación potencial, sin llevar a una violación real de la conservación de la energía. Sin embargo, esta es una cuestión que necesita más estudio. Parecería que hay algún tipo de «compromiso» entre las dificultades de energía evidentes en el proceso **RO** y la naturaleza decididamente no local (y curiosamente «resbaladiza») de la energía gravitatoria que se ha mencionado en §19.8.

Mi punto de vista con respecto a la reducción del estado cuántico es que se trata realmente de un proceso objetivo, y que es siempre un fenómeno *gravitatorio*. Este sería el caso incluso en situaciones en que ha habido una sustancial decoherencia por el entorno que conduce a una reducción de estado FAPP, por ejemplo en un sistema (tal como una molécula de ADN) que es demasiado pequeño para que le sea directamente aplicable una **RO** gravitatoria. En tales situaciones sería el desplazamiento de masa total en el entorno el que daría como resultado una **RO** gravitatoria. Creo que en las situaciones concretas que he considerado, donde el estado en cuestión es una superposición de dos estados estacionarios, este proceso de reducción está bien aproximado por el esquema **RO** que acabo de describir.

Ciertamente falta una teoría completa, y no he ofrecido ninguna dinámica real para la reducción del estado, según este proceso **RO**, ni siquiera en el caso de las superposiciones concretas que he considerado. En este aspecto, mi propuesta es una propuesta «minimalista» y no aspira a una dinámica más completa, tal como hacen las propuestas inspiradas de Károlyházy, Károlyházy y Frenkel, Pearle, Kibble, Ghirardi, Rimini y Weber, Ghirardi, Grassi y Rimini, Diósi, Percival, Gisin y otros.[40] De todas formas, mi propuesta minimalista parece tener claras consecuencias experimentales, y cerraré este capítulo presentando la idea subyacente en una clase de experimentos reales que tienen una capacidad definida para decidir si tal esquema **RO** gravitatorio es o no realmente respetado por la naturaleza.

30.13. La propuesta FELIX y otras relacionadas

El esquema básico es construir un «gato de Schrödinger» que consiste en un espejo minúsculo M, colocado en una superposición cuántica de

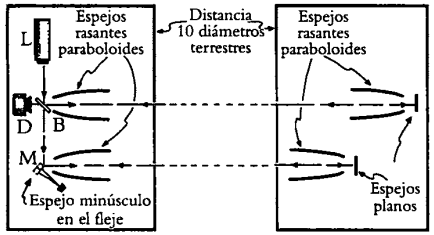


Fig. 30.24. FELIX (Free-orbit Experiment with Laser-Interferometry X-rays). Se indica un montaje esquemático. Un fotón, producido por el láser de rayos X L, se dirige al divisor de haz B. La parte transmitida del estado resultante del fotón se dirige a un espejo minúsculo M, aproximadamente de diez micras, y el impacto imparte un momento cuando el fotón es reflejado. Esto coloca a M en una superposición cuántica (gato de Schrödinger), que debe mantenerse durante, digamos, un segundo. Mientras, las dos partes de la función de onda del fotón deben mantenerse coherentes (aquí, por reflexión entre dos plataformas espaciales) hasta que ha transcurrido dicho período y el proceso entero se invierte. Un montaje perfecto (con trayectorias de igual longitud) y mecánica cuántica convencional haría que el detector responda el 0 % del tiempo. La **RO** gravitatoria lleva a una expectativa del 50 %.

dos posiciones ligeramente diferentes, desplazadas una respecto a otra en tan solo un diámetro nuclear.[41] Un tamaño razonable para este minúsculo espejo sería algo comparable a una mota de polvo, quizá de una décima parte del grosor de un cabello humano y que contenga algo del orden de 10^{14} a 10^{16} núcleos (de modo que su masa sería de aproximadamente 5×10^{-12} kg y su diámetro de aproximadamente 10^{-3} cm). Consideremos que el espejo M está colocado en su superposición por el impacto de un único fotón de rayos X que ha sido situado en una superposición de dos haces, uno de los cuales se dirige hacia M.

Un posible montaje experimental se indica en la Fig. 30.24. El fotón es producido por un láser de rayos X L y dirigido a un divisor de haz B. La parte transmitida (digamos) del estado resultante del fotón se dirige hacia M, y su impacto es tal que imparte un momento a este minúsculo espejo cuando es reflejado por el mismo. El espejo tiene que ser de alta calidad, de modo que es de una naturaleza «rígida» para que responda *como un todo* al impacto del fotón, sin que se exciten oscilaciones internas ni se descoloquen átomos. El espejo está suspendi-

do de tal modo que recuperará su posición original en aproximadamente una décima de segundo. Mientras tanto, las dos partes de la función de onda del fotón deben mantenerse coherentemente de algún modo, haciendo tiempo hasta que haya transcurrido este período, después del cual todo el proceso va a invertirse de modo que pueda establecerse si se ha perdido la coherencia de fase, como sería el caso si el minúsculo espejo en superposición cuántica se redujera espontáneamente a una posición u otra.

No obstante, mantener coherente un fotón de rayos X durante una décima de segundo no es una tarea desdeñable. (Por desgracia, se necesitan energías de rayos X para que pueda impartirse el momento suficiente para producir un movimiento adecuado en el espejo.) Una sugerencia para conseguir coherencia durante ese período de tiempo es realizar todo el experimento en el *espacio*, donde se mantiene la coherencia del fotón mediante reflexión entre grandes espejos situados en dos plataformas espaciales separadas quizá un diámetro terrestre. Un fotón necesita aproximadamente una décima de segundo para recorrer esta distancia de ida y vuelta. La parte de la función de onda del fotón que ha sido reflejada en M regresa entonces a M, mientras que la parte que se ha reflejado en el divisor de haz B es devuelta a B. La precisión cronométrica es tal que todo el proceso se invierte exactamente. Así, la parte de la función de onda del fotón responsable del movimiento de M choca de nuevo con M precisamente cuando M vuelve a su posición original, de modo que el fotón recupera el momento que había cedido a M y deja a M en reposo; además, las dos partes de la función de onda del fotón están sincronizadas para recombinarse en el divisor de haz B. Con tal de que no haya habido ninguna pérdida de coherencia de fase a lo largo de esta actividad, la función de onda del fotón se combinará en un único haz que vuelve al láser L. Así pues, un detector situado en la posición «alternativa» D, a la que podría haber llegado el fotón que sale del divisor de haz B (véase la Fig. 30.24), no detectará *nada*. Esta propuesta ha sido bautizada como FELIX, siglas de Free-Orbit Experiment with Laser-Interferometry X-rays (Experimento en órbita Libre con Interferometría de Rayos X).

Notemos que durante aproximadamente una décima de segundo el estado de M será una *superposición* de estar desplazado y no desplazado,

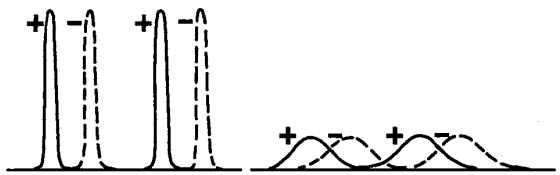


Fig. 30.25. Un factor importante sería la cantidad de «dispersión» en la distribución de masa de los núcleos en el espejo. Para una masa total dada, distribuciones de masa más estrechamente localizadas darían tiempos de reducción más cortos.

que es esencialmente la misma situación que teníamos con la masa de material descrita antes, ilustrada en la Fig. 30.20. Según el esquema **RO** gravitatorio, el estado de M debería reducirse espontáneamente a haber sido desplazado *o* no haber sido desplazado en una escala de tiempo de aproximadamente una décima de segundo. El estado del fotón está entrelazado con el de M, de modo que tan pronto como se reduce el estado de M, el estado del fotón se reduce con él. Entonces, el fotón está en un haz o en el otro, de modo que cuando finalmente vuelve al divisor de haz B tendrá la misma probabilidad de activar el detector D que de volver al láser L. Este procedimiento se repetiría entonces muchas veces. El efecto de **RO** sería que el detector responde aproximadamente en el 50 % de los ensayos; mientras que si, de acuerdo con la mecánica cuántica estándar, *no* se pierde coherencia de fase, entonces el detector no responde en absoluto.

Por supuesto, en cualquier situación práctica podría haber muchas otras maneras de perder coherencia de fase. Para que este experimento tenga éxito sería necesario que aquellas se mantuvieran en un nivel muy bajo, de modo que pudiera distinguirse la firma particular del **RO** gravitatorio. El experimento tendría que repetirse muchas veces, utilizando minúsculos espejos de tamaños y materiales diferentes, variando la escala de tiempo (utilizando quizá reflexiones repetidas entre plataformas espaciales). Un factor importante en el esquema **RO** concreto en consideración sería la cantidad de «dispersión» en la distribución de masa de los núcleos en el espejo. Para una masa total concreta, una distribución de masa más firmemente localizada daría un tiempo de reducción más corto (véase la Fig. 30.25).

La propuesta FELIX anterior es extraordinariamente difícil desde

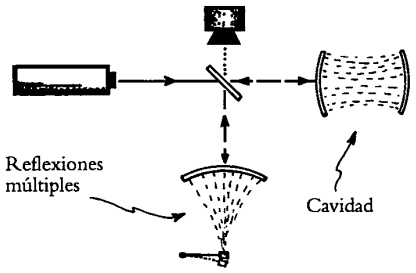


Fig. 30.26. Una versión más práctica de FELIX no utiliza rayos X, pero requiere unos 10^6 impactos de un fotón de luz visible en el espejo minúsculo, en lugar de un único impacto de un fotón de rayos X.

el punto de vista técnico, por varias razones. Un problema fundamental sería la precisión requerida para apuntar fotones de rayos X entre plataformas espaciales separadas aproximadamente 10.000 km. En cualquier caso, los experimentos espaciales son intrínsecamente difíciles y muy caros, y si hay una alternativa factible basada en tierra esta puede tener muchas ventajas. Afortunadamente, parece que tal alternativa es una posibilidad práctica. Debido a una ingeniosa sugerencia de William Marshall y algunas ideas asimismo ingeniosas para su implementación aportadas por Dik Bouwmeester y Christoph Simon, parece haber una posible alternativa basada en tierra que es ahora objeto de investigación activa. La propuesta[42] consiste en que en lugar de utilizar el impacto de un único fotón de rayos X para producir el movimiento deseado del minúsculo espejo, podría utilizarse un fotón de energía considerablemente menor (como un fotón de luz visible e incluso infrarroja), reflejado a un lado y a otro (digamos) $\sim 10^6$ veces, de modo que ahora hay 10^6 impactos del mismo fotón en el minúsculo espejo, en lugar del único impacto del fotón de rayos X que se proponía antes (véase la Fig. 30.26). En el momento de escribir esto, parece que no hay ningún obstáculo fundamental para que se pueda realizar un experimento preliminar de este tipo en un par de años más o menos. Si se realiza con éxito, este experimento preliminar podría quedarse todavía a cinco o seis órdenes de magnitud de lo que se necesita para un test definitivo de la **RO** gravitatoria. De todas formas, si puede mantenerse la coherencia gravitatoria para la superposición de las dos posiciones del mi-

núsculo espejo, esto representaría un avance (en términos de masa) sobre el actual «récord del gato de Schrödinger» (moléculas de fullereno C70)[43] en un factor de quizá 10^{12}. Parece probable que si se alcanzara con éxito esta etapa, donde el esquema **RO** gravitatorio «minimalista» de §§30.9-12 predice *acuerdo* con la mecánica cuántica estándar, entonces las mejoras adicionales necesarias para poner a prueba las nuevas predicciones de la **RO** gravitatoria pueden muy bien estar listas dentro de algunos años.

Es quizá notable que la extraordinariamente minúscula incertidumbre en la energía gravitatoria E_G que ocurre en esta clase de experimento —algo del orden de 10^{-33} de julios— es suficiente para dar esta «razonable» vida media de colapso de una décima de segundo o menos. La pequeñez de los efectos gravitatorios, en general, ha hecho que muchos físicos los desprecien por completo. Pese a todo, vemos que los efectos de introducir consideraciones gravitatorias en nuestra imagen podrían tener profundas consecuencias observacionales. Habría que señalar que la escala de tiempo $\hbar/E_G$ implica el *cociente* de las dos pequeñas cantidades $\hbar$ y G, y por ello no tiene por qué ser una cantidad pequeña en términos humanos ordinarios. Esto contrasta con las magnitudes características en gravedad cuántica, la longitud de Planck y el tiempo de Planck (§27.10 y §31.1), de tamaños 10^{-33} cm y 10^{-43} s, que son absurdamente pequeñas, y surgen del *producto* de $\hbar$ y G.

Imaginemos que se ha realizado con éxito un experimento para poner a prueba la **RO** gravitatoria. Si *no* se pierde coherencia de fase en las escalas de tiempo predichas por el esquema **RO** gravitatorio que se ha esbozado antes, dicho esquema tendrá que ser abandonado, o al menos seriamente modificado. Pero ¿qué pasa si los resultados de experimentos semejantes *apoyan* las predicciones de la **RO** gravitatoria? ¿Podemos entonces concluir que la reducción del estado cuántico es un efecto gravitatorio objetivo? Me temo que muchos preferirían seguir manteniendo uno de los puntos de vista más «convencionales» con respecto a esta cuestión. Podrían seguir argumentando, por ejemplo, que se mantiene la unitariedad estricta (**U**), mientras que parte del estado se hace inaccesible, perdido quizá en «fluctuaciones en el campo métrico» (véanse §29.6 y §30.14).

Personalmente, no tengo ningún deseo de resistirme a un cambio fundamental en una teoría física previamente aceptada, puesto que creo que, en el caso de la teoría cuántica, es realmente necesario un cambio fundamental, como he argumentado en extenso. Pero quizá no sea demasiado fantasioso hacer una comparación con las opiniones de muchos científicos altamente respetados, tales como Lorentz, que preferían considerar los efectos de la *relatividad especial* como meras «correcciones» a aplicar dentro de una visión del mundo del siglo XIX que acepta un estado de reposo absoluto. Sin duda habría un número similar de físicos reconocidos que podrían ser reacios a abandonar su visión del mundo de la mecánica cuántica duramente ganada en el siglo XX, si realmente resulta que las predicciones de la **RO** gravitatoria *son* apoyadas por una realización satisfactoria de un experimento del tipo FELIX. En mi opinión, semejante postura sería retrógrada, y significaría renunciar a la posibilidad de hacer nuevos progresos sobre la base de una nueva imagen cuántica del mundo ¡que realmente podría tener sentido!

Por supuesto, aquellos de nosotros que estamos esperando que la **RO** gravitatoria fortalezca nuestros puntos de vista menos convencionales debemos estar preparados para que eventualmente nuestras ideas puedan ser *contradichas* por un experimento semejante. Mi reacción ante esto sería de considerable desconcierto, pese al hecho de que muchos físicos cuánticos con los que he discutido esta cuestión han expresado su firme esperanza en que la mecánica cuántica convencional debe salir intacta. Mi desconcierto vendría básicamente de mi convicción en que la mecánica cuántica actual no tiene una ontología creíble, de modo que *debe* ser modificada en profundidad para que la física del mundo tenga sentido. Esto no implica en sí mismo que tenga que ser la **RO** *gravitatoria* la que venga en nuesto auxilio, ni es imperativo que la propuesta gravitatoria concreta esbozada aquí deba ser la correcta.[44] De todas formas, tengo la sensación de que la solidez y robustez de la teoría cuántica moderna no dejará que sea desplazada fácilmente. En mi opinión, cualquier desplazamiento exigiría la intervención de algo igualmente formidable, y ninguna otra cosa en la física conocida tiene esta envergadura salvo la teoría de la relatividad general de Einstein y sus profundos principios motivadores. Es esto lo que me lleva a prever un esquema de **RO** gravitatoria como el que he

sugerido. ¡Cualquiera que sea el resultado final de estas elucubraciones, preveo que muchas nuevas cuestiones mecanocuánticas, poderosas e intrigantes, serán planteadas y respondidas durante el curso del siglo XXI!

30.14. Origen de las fluctuaciones en el universo primitivo

Antes de poner fin a este capítulo, quiero plantear una de las numerosas cuestiones importantes que podrían quedar profundamente afectadas por un cambio en las reglas de la teoría cuántica, en la línea de las reflexiones de este capítulo. En §27.13 he llamado la atención sobre el estado extraordinariamente especial en que parece haber empezado el universo. Lo más especial de este estado, y lo que le da su entropía absurdamente baja, eran una isotropía y homogeneidad espaciales muy precisas, de modo que la geometría espaciotemporal está (aún) en un acuerdo notablemente próximo con uno de los modelos cosmológicos FLRW estándar (§27.11). Por supuesto, como se argumenta a menudo, el universo no puede haber sido absoluta y exactamente un modelo simétrico semejante. Si *alguna vez* existió esta alta simetría, debería haber permanecido para siempre, porque la dinámica de la relatividad general de Einstein —y del resto de la física clásica— conservará exactamente dicha simetría.

Pero ¿qué pasa con la física *cuántica*? La aleatoriedad inherente a los procesos de evolución cuántica, ¿no permite que aparezcan desviaciones de esta simetría exacta? En esta etapa se suele invocar la noción de «fluctuaciones cuánticas», como una forma de ofrecer las necesarias pequeñas desviaciones de la simetría exacta. La idea es que tales «fluctuaciones» podrían empezar siendo minúsculas, pero actuarían como semillas de irregularidad en la distribución de masa, que poco a poco aumentarían mediante la aglomeración gravitatoria, de modo que, finalmente, podrían desarrollarse estrellas, galaxias y cúmulos de galaxias, de acuerdo con la observación.

Pero ¿qué *son* las fluctuaciones cuánticas? Una característica de las relaciones de incertidumbre de Heisenberg (§21.11), aplicadas a magnitudes de campo (véanse §§26.2,3,9), es que si se intenta medir el valor de un campo cuántico en una región muy pequeña con gran precisión, esto llevará a una incertidumbre muy grande en otras magnitudes de campo relacionadas (canónicamente), y con ello a un valor esperado rápidamente cambiante de la cantidad que se está midiendo. Así, el propio acto de establecer el valor preciso de una magnitud de campo dará como resultado que dicha magnitud fluctúe de manera incontrolada. Dicha magnitud podría ser una componente de la métrica espaciotemporal, de modo que vemos que cualquier intento de medir la métrica de forma precisa dará como resultado cambios enormes en ella. En la década de 1950, consideraciones como estas llevaron a John Wheeler a argumentar que la naturaleza del espaciotiempo en la escala de Planck de 10^{-33} cm sería una «espuma» salvajemente fluctuante (véanse el final de §29.6 y la Fig. 29.6).

Para clarificar esta imagen, debemos recordar lo que en realidad establecen las relaciones de Heisenberg. Estas no nos dicen que haya algo intrínsecamente «borroso» o «incoherente» en el comportamiento de la naturaleza en las escalas más minúsculas. En su lugar, la incertidumbre de Heisenberg restringe la precisión con que pueden realizarse dos medidas no conmutativas. Recordemos que, para una única partícula, su posición y su momento en una dirección, que no son conmutativos, no pueden determinarse exactamente al mismo tiempo, pues el producto de sus respectivos errores no es menor que $\frac{1}{2}\hbar$ (§21.11). Sin embargo, hay un estado cuántico perfectamente bien definido, y si no se realiza ninguna medida real, el estado de la partícula evolucionará de forma precisa de acuerdo con la ecuación de Schrödinger (suponiendo válida la mecánica cuántica **U** estándar).

Análogamente, en mecánica cuántica estándar no todas las variables que definen un estado espaciotemporal pueden determinarse conjuntamente. La descripción cuántica del espaciotiempo debería estar, de todas formas, muy bien definida. Pero según el principio de Heisenberg, esa descripción no puede parecerse a una variedad (pseudo)riemanniana clásica, pues diferentes magnitudes geométricas espaciotemporales no conmutan entre sí. En su lugar, de acuerdo con la imagen de Wheeler, el estado consistiría en una vasta superposición de geometrías diferentes, muchas de las cuales se desviarán enorme-

mente de la planitud y por ello tendrán el carácter «espumoso» que él imagina.

Veamos cómo se aplica esto al estado del universo primitivo. ¿Pueden atribuirse realmente las desviaciones de la simetría exacta a las «fluctuaciones cuánticas» si todo el estado inicial entero posee simetría cosmológica FLRW exacta? La evolución **U** de este estado debe continuar para mantener la exacta simetría FLRW, con independencia de las «fluctuaciones cuánticas» o de cualquier otra manifestación de la incertidumbre de Heisenberg.[30.14] ¿Cómo es esto compatible con la geometría «espumosa» muy irregular que imagina Wheeler? Aquí no hay necesariamente una contradicción, pues el estado entero es una superposición de tales geometrías irregulares, no una geometría individual. La superposición propiamente dicha puede poseer una simetría que no tienen las geometrías individuales de las que está compuesta. Si contribuye una geometría irregular, así lo hacen todas las demás que se obtienen de ella por la aplicación de cada simetría FLRW.[45]

¿Cómo se supone entonces que esta vasta superposición cuántica FLRW simétrica de geometrías irregulares da lugar a algo que se parece a un universo específico «casi FLRW simétrico» que está perturbado solo de una forma mínima que es compatible con las observaciones? Al lector le debería quedar claro que no hay forma de que esto pueda suceder dentro de la evolución **U** de la mecánica cuántica estándar, puesto que esta debe conservar exactamente la simetría. Debe estar teniendo lugar algo de la naturaleza de un proceso **R**, que resuelve esta vasta superposición de geometrías en una única geometría o, mejor, en una superposición menor de geometrías que se parece más a una geometría única. La clave está en que las irregularidades que surgen de «fluctuaciones cuánticas» no pueden darse sin alguna acción tipo **R**, por la que el único estado cuántico inicial se resuelve de algún modo en una mezcla probabilista de estados diferentes. Esto nos lleva de nuevo a las cuestiones abordadas en el capítulo 29, donde se han discutido diferentes actitudes hacia la «realidad» de **R**.

Deberíamos tener en cuenta que lo que nos interesa aquí es el uni-

[30.14] ¿Puede ver por qué la conservación de esta simetría se sigue simplemente de la *unicidad* determinista de la evolución **U**, junto con una hipótesis general muy débil sobre la evolución **U**?

verso muy primitivo, en el que la temperatura podría haber sido del orden de 10^{32} K. No había experimentadores en esa época realizando «medidas», de modo que es difícil ver cómo puede aplicarse la perspectiva de «Copenhague» estándar ((a) de §29.1). ¿Qué pasa con la idea de los muchos mundos ((b) de §29.1)? En esta imagen no hay **R** real, y el estado FLRW simétrico del universo se mantendría hasta el presente, siendo representable como una gran superposición de muchas geometrías espaciotemporales constituyentes. Según esta idea, solo cuando observadores conscientes trataran de dar sentido al mundo se juzgaría apropiada la resolución en geometrías espaciotemporales alternativas, y habría entonces una superposición de observadores conscientes, cada uno de los cuales percibiría un único «mundo».[46] En la idea «FAPP» ((c) de §29.1), la presencia de decoherencia ambiental (suficiente) se considera la señal por la que se permite que nuestra superposición cuántica de geometrías diferentes sea considerada como *una mezcla probabilista* de geometrías diferentes.

Resulta ilustrativo hacer una comparación con un ejemplo en mecánica cuántica ordinaria.[47] Imaginemos un núcleo radiactivo en reposo, en un estado esféricamente simétrico (i.e., espín 0; véase §22.11) en un punto O situado en el centro de una cámara de burbujas[48] (Fig. 30.27). Supongamos que por fisión nuclear se divide en dos partes A y B, que son expulsadas desde O en direcciones opuestas. Podemos suponer que A y B están cargadas eléctricamente, de modo que dejan trazas en la cámara de burbujas. En este ejemplo empezamos con un estado con simetría esférica, centrado en O. Pero después de la desintegración, la simetría esférica se rompe por el eje a lo largo del que han salido las partes A y B. ¿Cómo vamos a entender esto en términos de la evolución **U** del estado original? Evidentemente, como se ha afirmado antes, la simetría esférica debe conservarse, pero el estado lo consigue al estar compuesto como una superposición lineal de todas las situaciones posibles dadas por direcciones diferentes del eje. La función de onda tiene la forma de una onda esférica centrada en O, aunque debemos tener en cuenta que el estado es un estado entrelazado que incluye a A y B, en el que cada localización de A está correlacionada con una localización de B en la dirección contraria. A medida que la influencia de las cargas en A y B empieza a ionizar el material en la cá-

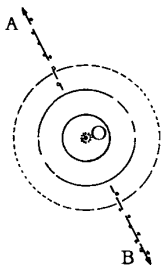


Fig. 30.27. Ruptura de simetría por **RO**. Un núcleo de espín 0 (esféricamente simétrico) se divide en dos partes, que se observa que siguen un par específico de caminos en direcciones opuestas. La evolución **U** del estado inicial conserva la simetría esférica, pero esta consiste en una superposición cuántica (entrelazada) de pares de caminos opuestos (Mott). R da como resultado que solo uno de estos es percibido. Este ejemplo debe tomarse como un modelo ilustrativo de lo que podría estar sucediendo en la creación de fluctuaciones de densidad en un estado cuántico del universo primitivo inicialmente muy simétrico.

mara de burbujas y se forman burbujas, el estado se entrelaza con dicho material, de modo que encontramos que el estado entero consiste en una superposición en la que cada componente incluye un par de trazas de burbujas en direcciones opuestas, una correspondiente al paso de A, y la otra, al paso de B.

La situación recién descrita no es esencialmente diferente de la del universo primitivo. Se necesita una versión de **R** para que la superposición cuántica simétrica pueda ser reemplazada por una mezcla probabilista de alternativas menos simétricas. Al parecer, en la práctica los teóricos tienden a adoptar una forma de interpretación FAPP ((c) de §29.1), en la que el tamaño del horizonte cosmológico es tomado de forma arbitraria (e ilógica) para proporcionar algún tipo de «corte» para los entrelazamientos cuánticos. La superposición cuántica es considerada entonces como una mezcla probabilista, aunque esta postura real apenas se deja clara alguna vez. Por ejemplo, en su libro de texto para licenciatura *The Early Universe*, los destacados cosmólogos inflacionarios Kolb y Turner (1994) afirman en la p. 286:

> Conforme cada modo cruza el horizonte hacia fuera, se desacopla de la microfísica y se «congela» como una fluctuación clásica.

Aquí el «modo» se refiere a una componente de la superposición cuántica, así que vemos que los autores intentan utilizar el horizonte como algo que permite el paso de una amplitud cuántica a una probabilidad de una alternativa clásica real. Esto parece estar en la línea de una propuesta FAPP (véase §29.6) y, como se ha argumentado en §§29,6,8, es, estrictamente hablando, ilógico.[49]

En mi opinión, está claro que la introducción de desviaciones de una simetría FLRW exacta vía fluctuaciones cuánticas requiere necesariamente una teoría de reducción de estado objetiva. Sin embargo, la propuesta «minimalista» para la **RO** gravitatoria que se presenta en §§30.9-12 no es suficientemente fuerte tal como está. Se necesita alguna propuesta **RO** más general, en la que puedan manejarse superposiciones cuánticas de grandes números de geometrías espaciotemporales, donde las geometrías individuales no tienen por qué ser estacionarias, como lo eran en §30.10. Cuando se disponga de un esquema semejante, será confrontado de inmediato con un conjunto impresionantemente creciente de datos observacionales, frente al que debe mantenerse o caer. Ya BOOMERANG, WMAP y otras observaciones han suministrado cantidades ingentes de datos concernientes a fluctuaciones densidad/temperatura en el universo primitivo, y habrá muchos más procedentes de otros experimentos ahora en proyecto.

Aquí es apropiado hacer un comentario final sobre esta situación. Advertimos cómo el estado simétrico en nuestro anterior ejemplo de la fisión nuclear era un estado muy entrelazado. Esto también sería cierto, y en un grado aún mayor, para las reducciones de estado que nos alejen de un estado inicialmente FLRW simétrico para un universo sometido a «fluctuaciones cuánticas». Así, de acuerdo con nuestra discusión de los estados EPR dada en §§23.3-6, tenemos «violaciones de la desigualdad de Bell» que proporcionan correlaciones entre sucesos distantes que parecen violar la causalidad clásica. Esta aparente violación de la causalidad no tiene por qué ser indicativa de un mecanismo tal como la inflación que sirviera para poner tales sucesos separados en contacto causal, sino que podría aparecer como resultado de cualquier esquema apropiado de reducción de estado objetiva (**RO**). Sin embargo, de la exposición que acabamos de hacer[50] vemos que, incluso dentro de las cosmologías FLRW estándar, esta aparente «violación de causalidad» puede ocurrir sin necesidad de inflación si las fluctuaciones iniciales resultan vía un esquema de reducción de estado objetiva.

Queda claro que estamos lejos de una teoría que pueda abordar con fiabilidad estas cuestiones. Pero al menos espero que haya sido capaz de persuadir al lector de la importancia fundamental de tener una mecánica cuántica con una ontología viable. Las cuestiones que se han

abordado en los capítulos 29 y 30 de este libro no son solo materias de interés filosófico. La importancia de tener una mecánica cuántica (mejorada) ontológicamente coherente no puede sobrestimarse. En esta sección he tocado solo una de las cuestiones fundacionales que podrían quedar profundamente afectadas por el conocimiento de una teoría semejante. Hay muchas más, entre ellas situaciones en biología (véanse §§34.7,10) en las que, como sucede en el universo primitivo, el punto de vista «de Copenhague» actual no puede aplicarse realmente, ya que no hay una división clara entre un sistema cuántico y un aparato de medida clásico.

Notas

Sección 30.1

30.1. Véase Roseveare (1982).

Sección 30.2

30.2. Véase Penrose (1980).

Sección 30.3

30.3. No hay barrera teórica o técnica para esto, al menos si no exigimos un 100 % de exactitud. Por ejemplo, se podría arreglar que el fotón inicial sea siempre uno de un par (producido, por ejemplo, por subconversión paramétrica; véase §23.10), de modo que el otro miembro del par dispara el dispositivo registrador.

30.4. Encuentro notable la dificultad que suele tener la gente con este argumento. Quizá la materia se clarifica si contemplamos numerosos ejemplos de este experimento que tienen lugar en varias localizaciones en el espaciotiempo. Hay cuatro rutas alternativas a considerar para el fotón, SBD, SBC, FBD y FBC. Para ver cuáles son las diversas probabilidades, preguntemos la proporción de SBD, dado S (situación hacia delante en el tiempo), o la proporción de SBD, dado D (situación hacia atrás en el tiempo). La regla del módulo al cuadrado da correctamente la respuesta real (50 %) en el primer caso, pero *no* da la respuesta real (casi 100 %) en el segundo caso.

30.5. Sin embargo, dependen de que el estado inicial sea el que se supone que es y no sea parte de un estado entrelazado (§23.3) que podría incluir también algo en el detector. También podríamos plantear la cuestión de si la inversa temporal de tales entrelazamientos podría ser responsable de que la regla del módulo al cuadrado invertida temporalmente dé respuestas del todo erróneas. Pero soy incapaz de ver cómo se construye una explicación plausible según estas líneas. Quizá algún lector con iniciativa pueda hacerlo mejor.

30.6. Véase Aharonov y Vaidman (1990).

30.7. Para una discusión de esta cuestión, véase Aharonov *et al.* (1964).

30.8. Véanse Aharonov y Vaidman (2001); Cramer (1988); Costa de Beauregard (1995); y Werbos y Dolmatova (2000).

Sección 30.4

30.9. Véase Unruh (1976); véase también Wald (1994).

30.10. Véanse Penrose (1968b, 1987b) y Bailey *et al.* (1982).

30.11. Véanse Kay (2000), Kay y Wald (1991), Kay, Radzikowski y Wald (1996), Hollands y Wald (2001) y Haag (1992).

Sección 30.5

30.12. Véase Wald (1984).

30.13. Véanse Wald (1984), Kruskal (1960) y Szekeres (1960).

30.14. Existe una discrepancia ligeramente confusa entre la interpretación de «τ» como tiempo *real* en el caso de Schwarzschild, que se ha considerado aquí, mientras que es $r_0\tau$ lo que mide el tiempo del observador acelerado en el caso plano (Rindler) de las Figs. 30.5a y b.

30.15. Véase Newman *et al.* (1965).

30.16. Esta razón giromagnética se refiere a una «partícula de Dirac pura», a la que un electrón es una excelente aproximación, pero un electrón real está sujeto a *correcciones radiativas* que vienen de la teoría cuántica de campos; véase el final de §24.7. Un protón o un neutrón están mucho más lejos de ser una partícula de Dirac, pero dicha noción se aplica de forma mucho más estrecha a sus quarks constituyentes.

Sección 30.6

30.17. Véanse Novikov (2001), Thorne (1995a) y Davies (2003).

30.18. Davies (2003) ofrece una exposición legible y divertida de tales posibilidades.

Sección 30.7

30.19. Véanse Penrose (1969a) y Floyd y Penrose (1971).
30.20. Véanse Blanford y Znajek (1977) y Begelman *et al.* (1984). Véase también Williams (1995, 2002, 2004).

Sección 30.8

30.21. Véanse Hawking (1974, 1975, 1976a, 1976b) y Kapuusta (2001).
30.22. Véase Preskill (1992).
30.23. Véanse Preskill (1992), o, para una perspectiva diferente, Kay (1998a, 1988b).
30.24. Véanse Preskill (1992) y Susskind *et al.* (1993).
30.25. Véase Gottesman y Preskill (2003) para una crítica de Horowitz y Maldacena (2003). Véase también Susskind (2003).
30.26. Hawking introdujo una generalización de la evolución unitaria en la que la descripción de matriz S de la QFT ordinaria (§26.8) se generaliza a lo que se conoce como un operador de «superdispersión» (no relacionado con la supersimetría; véase §21.2), denotado por un signo «$». Este opera entre estados de matriz densidad, más que entre los estados puros tratados por la matriz S. Véase Hawking (1976b). Habría que decir que, en 2004, Hawking se *retractó* (lamentablemente, en mi opinión) de su posición PÉRDIDA, en favor de RETORNO.

Sección 30.9

30.27. El principal desacuerdo entre Stephen Hawking y yo, durante veinte años, se ha centrado en esta cuestión de la asimetría temporal. A lo largo de estas discusiones, él ha mantenido categóricamente una física con simetría temporal, o bien la inmutabilidad de la mecánica cuántica **U**, o la tibia generalización antes mencionada (nota 30.26). Como explicaré, mi punto de vista sobre estas materias es muy diferente.
30.28. Véase Penrose (1979).
30.29. La hipótesis de curvatura de Weyl se refiere a la geometría *clásica*, de modo que dice algo sobre lo que sucede en el punto donde la «geometría cuántica» cristaliza en un espaciotiempo clásico.
30.30. Véanse Hawking (1976a, 1976b) y Gibbons y Perry (1978).
30.31. Quizá aquí se requiera una noción generalizada del espacio de Hilbert, que también podría asumir algunas de las propiedades de un espacio de fases (curvo); por ejemplo, Mielnik (1974), Kibble (1979), Chernoff y Marsden (1974), Page (1987) y Brody y Hughston (2001).

Sección 30.10

30.32. Estos podrían ser *estados coherentes* como los mencionados en §26.6.
30.33. Téngase en cuenta que los índices en $\boldsymbol{\kappa}_\chi$ y $\boldsymbol{\kappa}_\varphi$ son solo etiquetas y no «índices tensoriales» en el sentido de §12.8. Lo mismo se aplica a $\boldsymbol{g}_\chi$ y $\boldsymbol{g}_\varphi$.

Sección 30.11

30.34. Véase Christian (1995).
30.35. Véanse Isham (1992), Kuchar (1992), Rovelli (1991), Smolin (1991) y Barbour (1992).
30.36. Véase la nota 29.12 para muchas de las teorías de reducción de estado objetiva. Las de Diósi, Percival, Kibble, Pearle, Squires y yo mismo implican la gravitación de forma crucial.
30.37. Más recientemente, ha surgido una idea para una justificación más rigurosa de este tipo de propuesta para una **RO** gravitatoria. Recordemos del ejercicio [21.6] que para hacer la teoría cuántica compatible con el principio de equivalencia se necesita un factor de fase que incluye un término cúbico en el tiempo t, cuando se pasa de un sistema en caída libre a uno fijo en un campo gravitatorio. Por consiguiente, los dos sistemas describen estrictamente vacíos diferentes (véase §26.5), siendo esto el remanente del efecto Unruh, mencionado en 30.4, que sobrevive en el límite galileano. Así pues, si hay que respetar plenamente el principio de equivalencia, la superposición de dos campos gravitatorios implicará la superposición de dos vacíos diferentes, y por ello debería ser inestable, incluso en el límite galileano. Detalles de este argumento serán publicados más adelante.

Sección 30.12

30.38. Véanse §22.5 y Moroz *et al.* (1988). Para el «valor esperado», véase la nota 22.11.
30.39. Véase la nota 29.12 para muchas de las referencias seminales en este campo.
30.40. He aprendido mucho del estudio de tales propuestas. Quizá algunas de estas puedan proporcionar indicadores hacia una teoría **RO** gravitatoria más completa. Véase la nota 29.12.

Sección 30.13

30.41. Los detalles específicos de esta propuesta han tenido contribuciones importantes de varios colegas. Un ingrediente importante de la idea

original (que incluía el impacto sobre un cristal «tipo Mössbauer» por un fotón en un haz dividido) se debía a Johannes Dapprich; algunas ideas más específicas, entre ellas, los parámetros sugeridos para el tamaño del minúsculo espejo, la energía del fotón y muchas otras cosas, surgieron en conversaciones con Anton Zeilinger y otros en su grupo experimental (en aquella época) en Innsbruck. La idea de un experimento en el espacio (FELIX) surgió en discusiones con Anders Hansson. Las ingeniosas ideas que parecen proporcionar una alternativa más práctica basada en tierra se deben a Willian Marshall, Dik Bouwmeester y Christoph Simon.Véase Penrose (2000a) y Marshall *et al.* (2003).

30.42. Véase Marshall *et al.* (2003).

30.43. Véase Arada *et al.* (1999).

30.44. Por ejemplo, tanto el esquema **RO** gravitatorio original de Károlyházy (1974) como la propuesta más reciente de Percival (1994) hacen predicciones muy diferentes de las que se presentan aquí.

Sección 30.14

30.45. No obstante, aquí hay una sutileza porque se podría considerar que la acción de una simetría abstracta sobre una geometría espaciotemporal dé simplemente de nuevo la misma geometría (debido al principio de covariancia general; véase §19.6). Existen diferentes actitudes que pueden adoptarse sobre esta cuestión, pero en cualquier caso el punto general planteado en el texto no queda afectado.

30.46. En la propia variante de Wheeler, el «universo participatorio» (véase Wheeler, 1983), sería la presencia final de observadores conscientes la que de algún modo determina (teleológicamente) la selección concreta de geometría espaciotemporal que ocurrió en el universo primitivo.

30.47. Esto guarda cierta similitud con una discusión de la traza en una cámara de niebla de la emisión de una partícula α, que se debe a Neville Mott (1929).

30.48. Aparato estándar en el que el paso de una partícula cargada queda indicado por una cadena de burbujas minúsculas; véanse la nota 30.47 y Fernow (1989).

30.49. La razón real para un «corte» en el radio de Hubble (donde la recesión alcanza la velocidad de la luz) en esa época no tiene que ver directamente con el «tamaño del horizonte» real (que es, en cualquier caso, mucho más grande que el radio de Hubble en el esquema inflacionario; véase la Fig. 28.5) y no tiene nada que ver con el paso de la física

cuántica a la clásica. Es un efecto puramente clásico de la expansión del universo en un campo sometido a las ligaduras de la relatividad.

30.50. Parece que varias personas han sugerido que tales correlaciones tipo EPR «acausales» podrían haber estado presentes en las fluctuaciones del universo primitivo por tales razones. Por ejemplo, Bikash Sinha me propuso hace años una sugerencia de esa naturaleza.

31

Supersimetría, supradimensionalidad y cuerdas

31.1. Parámetros inexplicados

Probablemente la mayoría de los físicos tendrán ideas muy diferentes de las que se han esbozado en el capítulo anterior respecto a lo que puede depararnos la física del siglo XXI. Muy pocos de ellos parecen prever que haya cambios fundamentales en el armazón de la mecánica cuántica. En su lugar, discuten ideas que suenan extrañas, como la necesidad de dimensiones extra para el espaciotiempo, o de que las partículas puntuales sean reemplazadas por entidades extensas conocidas como «cuerdas», o por estructuras de dimensiones mayores llamadas «membranas», o *p*-branas o simplemente «branas», y donde curiosos objetos adicionales llamados «D-branas» parecen desempeñar un papel importante. Hay intrigantes extensiones de la idea de simetría, conocidas como «supersimetría», o de los «grupos cuánticos». Hay generalizaciones de la propia noción de geometría descrita como «no conmutativa», y hay imágenes del mundo en las que en los niveles más minúsculos impera la discretización en lugar de la continuidad, o donde el tejido del propio espacio consiste en nudos o lazos; incluso se ha sugerido que la propia noción de espaciotiempo tendrá que ser abandonada o reformulada en otros términos.

¿Cuáles son estas diversas ideas y qué vamos a hacer con ellas? Y lo que es más importante, ¿qué motiva a tantos físicos a describir una «realidad» tan poco parecida a lo que directamente percibimos en las escalas humanas ordinarias? Sin duda, parte de la razón para contemplar tales propuestas reside en el éxito de la mecánica cuántica y, en menor grado, de la relatividad general. Estas teorías del siglo XX nos han mostrado que nuestras intuiciones directas pueden confundirnos,

y que la «realidad» puede diferir profundamente de las imágenes que proporcionaba la física de los siglos precedentes. Pese a todo, el hecho de que se nos presente un esquema exótico o inusual para el mundo no nos proporciona una base para creerlo. Tendremos que tratar de entender algo de las motivaciones subyacentes que guían las investigaciones de los teóricos modernos cuando intentan sondear más profundamente el funcionamiento interno del universo.

Debemos retomar los hilos del razonamiento que hemos visto por primera vez en el capítulo 24 y hemos continuado en los capítulos 25 y 26, donde los requisitos combinados de la relatividad especial y la teoría cuántica nos han conducido al atolladero de la teoría cuántica de campos. Esta, a su vez, nos ha llevado a un campo minado de infinitos, y se ha necesitado gran ingenio para evitar la mayoría de estos, lo que ha conducido en definitiva al modelo estándar de la física de partículas, que encuentra buen acuerdo con el funcionamiento acompasado de la naturaleza. Pese a todo, el propio modelo estándar no está libre de infinitos, pues es meramente una teoría «renormalizable» más que una teoría finita. La renormalizabilidad solo permite hacer ciertos cálculos que dan respuestas finitas a la mayoría de las preguntas de interés dentro de la teoría, pero no nos ofrece ningún asidero para algunos de los parámetros más importantes, tales como los valores concretos de las masas o las cargas eléctricas de las partículas descritas por la teoría. Estas habrían resultado «infinitas» (o quizá «cero»), si no fuera por el propio procedimiento de renormalización que evita tales escalamientos infinitos mediante una redefinición de términos, y permite que se obtengan respuestas finitas para *otras* magnitudes. Básicamente, se «renuncia» a calcular masa y carga, cuyos valores se introducen en la teoría como parámetros no explicados; de hecho, hay unos diecisiete de estos parámetros, entre ellos constantes de acoplamiento de varios tipos, además de los valores de las masas de los quarks y leptones básicos, la partícula de Higgs, etc., que tienen que ser especificados.

Existe un misterio considerable en torno a los extraños valores de las masas y cargas de las partículas reales de la naturaleza. Tenemos, por ejemplo, la inexplicada «constante de estructura fina» α que gobierna la intensidad de las interacciones electromagnéticas, definida por la fórmula

$$\alpha = \frac{e^2}{\hbar c},$$

donde $-e$ es la carga del electrón. El recíproco de la constante de estructura fina toma un valor muy próximo a $\alpha^{-1} = 137$, pero más exactamente

$$\alpha^{-1} = 137{,}0359\ldots$$

Durante varios años, algunos físicos creyeron que α^{-1} podría tomar realmente el valor *exacto* 137. En particular, sir Arthur Eddington (1946) pasó la última parte de su vida tratando de elaborar una «teoría fundamental», una de cuyas consecuencias sería que «$\alpha^{-1} = 137$». Muchos de los físicos actuales podrían ser menos optimistas que sus predecesores respecto a encontrar una «fórmula» matemática directa para α, o para otras «constantes de la naturaleza». En nuestros días, los físicos tienden a considerar estas cantidades como funciones de la energía de las partículas implicadas en una interacción, y no solo como números, y se refieren a ellas como «constantes de acoplamiento móviles» (véase la nota 26.21). Los valores escalares observados que conocemos como «constantes de la naturaleza» serían entonces «límites a baja energía» de estos valores «móviles». Aunque todavía habría esperanzas de encontrar una razón puramente matemática para estos valores límite específicos, tales valores pueden parecer de algún modo menos «fundamentales» de lo que lo serían si no hubiera dependencia de la energía.

Suele ser revelador expresar cantidades como la carga y la masa en términos de las unidades *absolutas* (de Planck) introducidas en §27.10, para las que la constante gravitatoria de Newton G, la velocidad de la luz c, la forma de Dirac de la constante de Planck $\hbar$ y la constante de Boltzmann k se hacen iguales a la unidad:

$$G = c = \hbar = k = 1.$$

En estas unidades la carga del protón (o menos la del electrón) resulta ser aproximadamente $e = 1/\sqrt{137}$, y con más precisión[1]

$$e = 0{,}085\ 4246$$

y la carga del quark básico (menos la carga del quark down; véase §25.6) tiene un tercio de este valor. A veces, las unidades absolutas se

suelen conocer como unidades *de Planck* (o bien unidades *de Planck-Wheeler*), porque Max Planck (que debe su fama a la mecánica cuántica; véase §21.4) propuso una idea al respecto en 1906. Irónicamente, en su artículo utilizaba la carga eléctrica como una unidad básica, en lugar de su propia «constante de Planck», y en ese esquema tenemos simplemente $e = -1$. (El misterio de la carga no desaparece, por supuesto, porque en su propio esquema $\hbar = 137{,}036$). Fue John Wheeler (por ejemplo, 1973) quien más tarde resaltó en muchos de sus escritos la importancia de estas ideas (utilizando $\hbar$ en lugar de la carga eléctrica escogida por Planck).

Si eso fuera todo, las unidades de Planck podrían denominarse más apropiadamente unidades de *Stoney*, pues el físico irlandés George Johnstone Stoney (que midió por primera vez la carga del electrón) propuso, ya en 1881, la misma idea que Planck en 1906. Sin embargo, hay otro artículo de Planck —publicado en 1899, antes, de hecho, que su famoso artículo de 1900 que inició la teoría cuántica— en el que se utilizaba la «constante de Planck» para definir unidades absolutas. En consecuencia, me atendré a la terminología convencional que llama «unidades de Planck» a las unidades absolutas.

¿Qué pasa con los valores de las masas de las partículas? El problema de la masa es mucho más espinoso que el de la carga eléctrica. Al parecer, todas las partículas de la naturaleza tienen valores para la carga que son múltiplos enteros de una carga básica. Podemos considerar que esta es la carga del protón si estamos interesados solamente en partículas que pueden existir libres, o menos la carga del quark-down si queremos incluir los constituyentes internos de los hadrones. Aunque por ahora no se comprende bien este hecho, y ciertamente no hay una comprensión adecuada del valor 137,036, el problema parece ser mucho más tratable que el problema correspondiente para los valores de las masas. Uno de los aspectos misteriosos del problema de las masas es el tamaño absurdamente minúsculo que tienen los valores de las masas de las partículas ordinarias medidas en unidades absolutas. Por ejemplo, la masa m_e del electrón, en unidades absolutas, es aproximadamente

$$m_e = 0.000\ 000\ 000\ 000\ 000\ 000\ 000\ 043$$

y la del protón es solo unas 1.836 veces mayor que este valor. La masa del neutrino electrónico ν_e es menor que 10^{-5} del valor anterior. Otra forma de expresar el enigma de estos valores minúsculos de las masas es preguntar por qué la «masa de Planck» natural (de 10^{-5} g, aproximadamente la masa de una pulga) es mucho mayor que las masas de todas las partículas básicas que se encuentran en la naturaleza. Pero otra forma de expresar este enigma es preguntar por qué la distancia de Planck de $1{,}6163 \times 10^{-35}$ m es unos 20 órdenes de magnitud menor que las escalas más minúsculas que se encuentran normalmente en la física de partículas. Se considera que esta distancia tiene una relevancia profunda en la teoría de la gravitación cuántica, pues por debajo de esa escala de distancias las ideas normales de un espaciotiempo continuo parecen no tener un sentido real.[2]

Una forma de ver estos misterios sería considerar que los pequeños valores de la carga eléctrica o de la masa son el resultado de un proceso de renormalización, donde el valor desnudo (§26.9) podría ser algún número matemáticamente respetable como 1 o 4π. En consecuencia, los pequeños valores observados podrían ser resultado de algún factor de renormalización simplemente *grande* en lugar de infinito. Esto podría ocurrir si las sumas y las integrales divergentes de la QFT pudieran reemplazarse por algo convergente. Las divergencias (esto es, las divergencias «ultravioleta»; véase §26.9) se producen normalmente porque suponen sumar momentos cada vez mayores, sin límite, que se corresponden con distancias cada vez menores, sin límite. En consecuencia, los infinitos podrían eliminarse si hubiera un corte para las integrales divergentes a, digamos, la escala (gravitatoria) de Planck[3] de 10^{-35} m. De hecho, una idea de este tipo fue propuesta por Oskar Klein en torno a 1935. Todo esto sugiere que cuando la gravitación se introduzca de manera adecuada en los cálculos de QFT, el resultado podría ser una teoría finita en lugar de una meramente renormalizable, y que dentro de tal teoría finita podría haber lugar para entender estos números inexplicados.

Aunque se ha tenido esperanza en ello durante más de medio siglo, los problemas que surgen al introducir la gravedad en la imagen han empeorado las cosas en lugar de mejorarlas. Cuando se han aplicado técnicas estándar de cuantización a la teoría de Einstein, el resultado ha

sido una teoría no renormalizable y no una teoría finita. Esto ha llevado a muchos investigadores a intentar algo *no* estándar en sus búsquedas de una teoría cuántica de la gravedad. Por supuesto, uno de los mensajes de los capítulos anteriores de este libro (y muy en particular los capítulos 27-30) era que deberíamos buscar una unión no estándar entre una teoría cuántica (de campos) y la relatividad general. Pero mi opinión de que debería haber algún cambio en el lado cuántico de las cosas no se ha tomado en serio. Cuando se somete directamente la teoría de Einstein a los procedimientos estándar de la QFT, el resultado es una insatisfactoria gravedad cuántica no renormalizable,[4] y aunque muchos han propuesto un cambio en la teoría de Einstein, no han propuesto un cambio en la QFT.

31.2. Supersimetría

¿Qué tipo de cambios se han sugerido? En uno de estos se han adoptado las ideas de la *supersimetría*, y se han amalgamado con la teoría de Einstein (también con una torsión incluida; véanse §14.4 y la nota 19.10) para producir un esquema conocido como *supergravedad*. ¿Qué es la supersimetría? ¿Por qué muchos físicos la consideran «algo bueno», hasta el extremo de que las ideas supersimétricas subyacen en muchos desarrollos en las teorías fundamentales modernas, y de manera especialmente importante en la teoría de cuerdas? De hecho, se ha atribuido un notable estatus a los principios de la supersimetría,[5] pese al hecho de que las predicciones de este esquema de ideas parecen guardar poca o ninguna relación con lo que se ha observado hasta ahora en el propio esquema de cosas de la naturaleza.

En este punto debo de nuevo declarar mis prejuicios, y ofrecer la requerida advertencia reglamentaria al lector. Yo mismo soy escéptico respecto a la relevancia física del esquema de supersimetría, al menos de la forma en que se utiliza hoy día en la física de partículas y las teorías subyacentes. Por ahora, las observaciones no proporcionan mucho apoyo —probablemente ninguno en absoluto— a las afirmaciones de la supersimetría. El atractivo de las ideas procede de una evidente elegancia matemática y del indudable valor de la supersimetría para can-

celar grandes lotes de infinitos en aquellos modelos de QFT que caen bajo su paraguas. Supongamos que usted es un físico interesado en construir una QFT que esté libre de infinitos incontrolables. ¡Entonces su tarea será muchísimo más fácil si considera que su teoría es supersimétrica!

La idea básica que hay tras la supersimetría es que ofrece un medio mediante el cual los fermiones y los bosones pueden ser «emparejados» de acuerdo con un tipo de relación de simetría. Como hemos visto en §§25.5-8, los grupos de simetría normales de la física de partículas se limitan a «rotar» conjuntos de bosones entre sí y conjuntos de fermiones entre sí. No «rotan» bosones para dar fermiones, o viceversa. Por el contrario, la supersimetría hace precisamente esto. Recordemos de §26.2 que los bosones satisfacen leyes de conmutación mientras que los fermiones satisfacen leyes de anticonmutación. Un operador que envía unos a otros debe tener propiedades de anticonmutación. Pero los operadores que proceden de un grupo continuo ordinario son los generadores infinitesimales del grupo, que forman un álgebra de Lie; véase §13.6. Los elementos de un álgebra de Lie ordinaria satisfacen leyes de conmutación y no leyes de anticonmutación. Esto significa que los operadores necesarios no son generadores infinitesimales de un grupo ordinario, sino de una noción más amplia conocida como un *supergrupo*, donde se extienden las leyes de un álgebra de Lie, de modo que algunos de los generadores satisfacen leyes de anticonmutación tanto como leyes de conmutación.

En §§26.2,3 ya hemos visto objetos semejantes, a saber, ecuaciones como

$$\boldsymbol{ab} \pm \boldsymbol{ba} = \boldsymbol{c},$$

que son satisfechas por los operadores creación, aniquilación y de campo de la QFT (en ecuaciones tales como $\boldsymbol{\Psi^*\Phi} \pm \boldsymbol{\Phi\Psi^*} = i^k\langle\psi|\phi\rangle\boldsymbol{I}$ de §26.2). De acuerdo con esto, se construye una *superálgebra* de Lie de la misma manera que un álgebra de Lie ordinaria, salvo que ahora existe la posibilidad de un signo (+) en las relaciones que la definen. En §13.6 se ha señalado que las relaciones que definen un álgebra de Lie tienen la forma $[\boldsymbol{E}_\alpha, \boldsymbol{E}_\beta] = \gamma^\chi_{\alpha\beta}\boldsymbol{E}_\chi$, donde $\gamma^\chi_{\alpha\beta}$ son las constantes de estructura y $[\boldsymbol{E}_\alpha, \boldsymbol{E}_\beta] = \boldsymbol{E}_\alpha\boldsymbol{E}_\beta - \boldsymbol{E}_\beta\boldsymbol{E}_\alpha$. Estas relaciones tienen la forma de la ecuación que se ha mostrado antes cuando aparece el signo menos normal entre $\boldsymbol{ab}$ y $\boldsymbol{ba}$. Pero en el caso de una superálgebra de Lie admitimos también el signo más, que ocurre cuando ambas $\boldsymbol{a}$ y $\boldsymbol{b}$ son cantidades fermiónicas (en lugar de que ambas sean bosónicas o una fermiónica y otra bosónica). Suele utilizarse la notación $[\boldsymbol{a}, \boldsymbol{b}]_+$ para tales anticonmutadores, i.e., $[\boldsymbol{a}, \boldsymbol{b}]_+ = \boldsymbol{ab} + \boldsymbol{ba}$, para complementar la notación habitual de los paréntesis de Lie $[\boldsymbol{a}, \boldsymbol{b}] = \boldsymbol{ab} - \boldsymbol{ba}$. Esto nos hace ir más allá de la notación habitual de un álgebra de Lie.

Los generadores de supergrupos se suelen describir como construidos de una manera concreta. En lugar de partir de cantidades ordinarias en números reales, tomamos estos generadores como elementos de un álgebra de Grassmann, que, como hemos visto en §11.6, implica tanto propiedades de anticonmutación como de conmutación. Veremos con más detalle cómo funciona esto en §31.3.

Los supergrupos constituyen ahora un área respetable de las matemáticas puras. Además, las ideas de la supersimetría pueden aplicarse directamente en argumentos matemáticos para obtener resultados que no son tan fáciles de obtener por otros medios.[6] Sin embargo, esto no nos dice si la supersimetría, tal como ha sido utilizada, tiene una relevancia directa para la física. Por otra parte, hay varios casos en los que la supersimetría se ha mostrado útil para motivar o establecer resultados matemáticos que tienen relevancia física directa.[7] No obstante, una vez más, no creo que esto sea un argumento de peso para que los supergrupos tengan una relevancia directa para la física de partículas o la QFT.

¿Qué evidencia hay de que la supersimetría desempeñe un genuino papel en física de partículas? Recordemos el modelo estándar de la física de partículas tal como se ha descrito en el capítulo 25. Su renormalizabilidad debe mucho a un preciso «ajuste fino» de sus parámetros. Estas relaciones pueden entenderse básicamente en términos de su requisitos de simetría $SU(3) \times SU(2) \times U(1)/Z_6$ (§25.7). Pese a todo, según algunas afirmaciones,[8] el modelo estándar requiere algún otro ajuste fino muy preciso además de dichas relaciones. Podrían invocarse simetrías adicionales para arreglar esto, y se ha sugerido la supersimetría como un medio de conseguir tal ajuste fino. Por ello, con frecuen-

cia se han utilizado ideas semejantes en teorías de gran unificación (§25.8). Pero ¿tenemos razones para creer en tales GUT? Por el momento no hay ninguna evidencia observacional a favor de ello.

Al parecer, las características tentadoras de la supersimetría consisten en que ofrece una vía de interrelación entre los bosones y los fermiones, y que es mucho más fácil hacer que las QFT supersimétricas proporcionen respuestas finitas que hacerlo con las QFT no supersimétricas. Con un emparejamiento supersimétrico de bosones con fermiones, los infinitos de un conjunto pueden cancelarse completamente con los infinitos del otro conjunto. Ello hace que el trabajo del constructor de QFT sea mucho más fácil de lo que lo sería sin supersimetría. Pero no nos dice que la propia naturaleza siga este camino. ¡Esta muy bien podría guardar un as muy diferente en su manga!

Ahora bien, la dificultad principal con la supersimetría (tal como se utiliza hoy) es que exige que toda partícula en la naturaleza tenga lo que se denomina una «supercompañera» con un espín que difiere en $\frac{1}{2}\hbar$ del espín de la partícula original. Eso requiere que haya un «selectrón» de espín 0 como compañero del electrón, un «squark» de espín 0 para acompañar a cada variedad de quark, un «fotino» de espín 1/2 para acompañar el fotón, un «wino» y un «zino» de espín 1/2 como compañeros respectivos de los bosones W y Z. El problema es que nunca se ha encontrado ninguna de estas «compañeras supersimétricas». La explicación oficial para ello es que, debido a algún mecanismo de «ruptura de supersimetría» cuya naturaleza nunca ha sido propuesta de manera adecuada, cada una de estas presuntas compañeras supersimétricas debe ser enormemente más masiva que la partícula a la que acompaña. El tipo de masa que ahora se postula que deben tener estas partículas inobservadas es del orden de mil veces la del protón, o más. Tengo que decir que no soy el único que cree que esto parece un poco retorcido.

Parece postularse que, de las dos «compañeras», la que tiene el espín menor (en $\frac{1}{2}\hbar$) debe ser la extraordinariamente más masiva del par (excepto cuando ambos miembros del par carecen de masa). Presumiblemente, solo las partículas consideradas «elementales» (que aparentemente son el fotón, el gravitón, los bosones W y Z, los gluones, los leptones y los quarks) tienen supercompañeras. De lo contrario, tenemos problemas con las partículas de espín 0 como los piones. Si *existen* partículas elementales de espín 0, tales como el aún no descubierto bosón de Higgs, tendrían que ser más masivas que sus supercompañeras en este cálculo concreto (puesto que el espín negativo está excluido). Si esto es correcto, ¿por qué no se ha encontrado la supercompañera del bosón de Higgs? Pese a todo, una vez más los que creen en la supersimetría y la cosmología inflacionaria deben explicar cómo encaja la partícula φ escalar del último fenómeno (§28.4) en la imagen de «supercompañera».

La única evidencia «positiva» que se cita ahora con frecuencia en apoyo de la supersimetría tiene que ver con ciertas ideas respecto a la pretendida unión de las tres fuerzas de la física de partículas (fuerte, débil y electromagnética) en un esquema unificado muy simétrico cuando la temperatura del universo tenía un valor altísimo (aproximadamente 10^{28} K), unos 10^{-39} segundos (solo unos 10.000 tiempos de Planck) después del big bang.[9] La idea es que tal unificación requeriría que todas las intensidades de las interacciones fueran iguales a dicha temperatura. Habría que decir que en condiciones ordinarias hay un factor del orden de 10^{13} entre las intensidades de las interacciones fuerte y débil (aunque realmente no pueden ser comparadas de forma directa). El argumento es que cuando se toman en consideración los efectos de renormalización (y recordemos de §26.9 cuán diferente podría ser la carga observada de una partícula de su carga desnuda), entonces estas intensidades se igualarían, y se presentarían los valores «desnudos» a tales enormes temperaturas. Los argumentos afirman que sin supersimetría los valores no coinciden por completo, sino que «fallan» (véase la Fig. 31.1); pero cuando se introduce la supersimetría en la imagen, las curvas llegan a una gloriosa coincidencia, ¡y puede tener lugar la gran unificación de la física de partículas!

El lector notará mi escepticismo. (Ya en §28.3 he expresado algunas de mis dificultades con las teorías en las que se da «restauración de simetría» cuando la temperatura del universo es suficientemente alta.) Hay enormes extrapolaciones implicadas en este conjunto par-

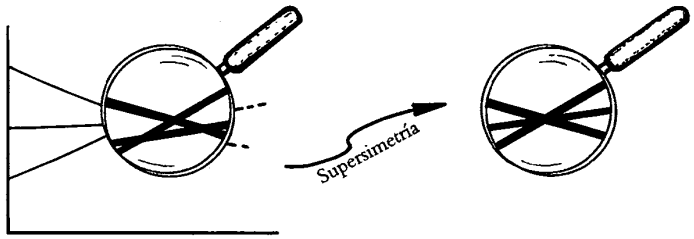


Fig. 31.1. Según cierta perspectiva de «gran unificación», las constantes de acoplamiento de las interacciones fuerte, débil y electromagnética, tratadas como «constantes de acoplamiento móviles» (véanse la nota 26.21 y §31.1), deberían alcanzar exactamente el mismo valor a temperaturas suficientemente grandes, aproximadamente 10^{28} K, que se habrían dado alrededor de 10.000 instantes de Planck después del big bang ($\sim 10^{-39}$ s). Se ha visto que la supersimetría es necesaria para que los tres valores coincidan exactamente.

ticular de ideas que se aducen como apoyo observacional en favor de la supersimetría. Una de ellas es la presuposición de que nada esencialmente nuevo va a manifestarse en el enorme intervalo de energías (temperaturas) entre 10^{28} K y los aproximadamente 10^{14} K que son accesibles en los aceleradores actuales. Esto parece en sí mismo una extrapolación poco razonable, y no veo cómo puede considerarse que tales argumentos ofrezcan cualquier apoyo observacional significativo en favor de la supersimetría, excepto quizá para los ya convencidos.

31.3. El álgebra y la geometría de la supersimetría

Volvamos a la teoría de supergravedad con la que he empezado este discurso. De acuerdo con lo anterior, debería haber un supercompañero de espín 3/2 del gravitón, conocido como un *gravitino*. Esta partícula postulada carecería de masa, como el propio gravitón, a menos que hubiera una severa ruptura de supersimetría. ¿Cómo se relaciona el gravitino con la geometría? Einstein nos ha enseñado que la gravitación está descrita por la curvatura espaciotemporal (§17.9 y §19.6). ¿Implica esto que el gravitino debería estar desempeñando un correspondiente papel (super)simétrico? Al desear un papel semejante, muchos teóricos de la supergravedad argumentarán que hay que generalizar la noción ordinaria de una *variedad* (como se ha descrito en los capítulos 10 y 12), y por ello se ha propuesto el concepto de una *supervariedad*. Podemos pensar que es algo definido de una manera muy formal, generalizando la noción ordinaria de *coordenadas* para incluir elementos anticonmutativos. En el caso de una variedad ordinaria, las coordenadas son en general números reales (o números complejos, si se considera una variedad compleja; véase §12.9). En el caso de una supervariedad, consideramos que son elementos de un álgebra de Grassmann (§11.6).

La mayoría de los teóricos de la supersimetría no adoptarían una actitud tan rigurosa respecto a la naturaleza de la «variedad» en la que viven sus magnitudes de campo supersimétricas (incluso si la naturaleza «geométrica» de la relatividad general estándar pareciera exigirlo en el caso de la supergravedad). En las descripciones que siguen no será necesario atenerse rigurosamente a un punto de vista de «supervariedad». Puede considerarse que las ideas de la «superálgebra» se refieren solo a magnitudes definidas en una variedad espaciotemporal ordinaria.

La más simple de tales álgebras se obtiene si añadimos un único elemento anticonmutativo ε al sistema de los números reales $\mathbb{R}$. La cantidad ε debe anticonmutar consigo misma, $\varepsilon\varepsilon = -\varepsilon\varepsilon$, de donde $\varepsilon^2 = 0$. Así pues, cada elemento del álgebra tiene la forma

$$a + \varepsilon b,$$

donde a y b son números reales que conmutan con ε. Nótese que la suma y el producto de dos números semejantes vienen dados por

$$(a + \varepsilon b) + (c + \varepsilon d) = (a + c) + \varepsilon(b + d),$$
$$(a + \varepsilon b)\,(c + \varepsilon d) = ac + \varepsilon(ad + bc).$$

Nótese también que si ignoramos los términos que multiplican a ε, entonces volvemos a obtener simplemente las reglas del álgebra ordinaria.

Esto se sigue aplicando si tenemos varios generadores de supersimetría diferentes, digamos $\varepsilon_1, \ldots, \varepsilon_N$, que anticonmutan

$$\varepsilon_i\varepsilon_j = -\varepsilon_j\varepsilon_i, \text{ de donde } (\varepsilon_i^2) = 0,$$

y el elemento general de la superálgebra tiene la forma[31.1]

$$a + b_1\varepsilon_1 + b_2\varepsilon_2 + \ldots + b_N\varepsilon_N + c_{12}\varepsilon_1\varepsilon_2 + c_{13}\varepsilon_1\varepsilon_3 + \ldots + f_{12\ldots N}\varepsilon_1\varepsilon_2 \ldots \varepsilon_N.$$

El álgebra se comporta de tal modo que si tomamos simplemente la parte «ordinaria» a de cualquier elemento (sin ninguna ε involucrada), entonces obtenemos el álgebra familiar de los números ordinarios (reales o complejos). La parte «súper» del álgebra es lo que queda. Es «nilpotente» en el sentido de que cuando se eleva a una potencia suficientemente alta, se anula por completo.[31.2] A veces se utiliza la extravagante terminología «cuerpo» y «alma» para estas partes «ordinaria» y «súper», respectivamente.

Al ser una persona a la que le gusta tener una «imagen» de lo que está pasando, siempre he encontrado insatisfactorias tales descripciones puramente formales de superálgebras y supervariedades. Por fortuna, hay una forma geométrica más convencional de considerar estas cosas. Consideremos, por el momento, el caso más fácil de un único generador supersimétrico ε. Puesto que tiene que ser una entidad anticonmutativa, podríamos tratar de considerarla como una 1-forma ε. Sin embargo, no puede ser solo una 1-forma ordinaria que remita al espacio ordinario —digamos la n-variedad $\mathcal{M}$— con el que estamos trabajando. Todas las formas diferenciales ordinarias dentro de $\mathcal{M}$ tienen ya significados que son aceptados (véase §12.4). Lo que debemos hacer es considerar $\mathcal{M}$ como una *hipersuperficie* inmersa en una variedad $(n + 1)$-dimensional $\mathcal{M}'$ (siendo una «hipersuperficie» una subvariedad de una dimensión menos que el espacio ambiente; véase la nota 27.11), donde ε va a ser una 1-forma que remite a la variedad mayor $\mathcal{M}'$, aunque restringida a puntos de $\mathcal{M}$. Se supone que no estamos interesados en $\mathcal{M}'$, excepto precisamente en puntos de $\mathcal{M}$, y que $\mathcal{M}'$ suministra una dimensión adicional que apunta fuera de $\mathcal{M}$. Véase la Fig. 31.2a. (De hecho, lo que nos interesa es solo lo que se conoce como la *primera* vecindad de $\mathcal{M}$ en $\mathcal{M}'$. Eso significa «derivadas primeras» fuera de $\mathcal{M}$, de modo que estamos interesados en las nociones de vectores tangente o cotangente o espacios que «apuntan» a $\mathcal{M}'$ fuera de $\mathcal{M}$, pero no las nociones de derivadas superiores tales como *curvatura* en direcciones que se salen de $\mathcal{M}$.) Lo que estamos haciendo es todavía algo n-dimensional, en el sentido de que todas nuestras cantidades pueden representarse como funciones de las n coordenadas independientes en la variedad $\mathcal{M}$. Las cantidades «alma» se refieren a direcciones que apuntan fuera de $\mathcal{M}$ y hacia $\mathcal{M}'$, mientras que las cantidades «cuerpo» se refieren simplemente a direcciones dentro de la propia $\mathcal{M}$.

[31.1] Escriba la suma y el producto de dos de estas cantidades cuando $N = 3$. ¿Cuál es el inverso multiplicativo de uno de estos elementos, donde $a \neq 0$?

[31.2] Demuéstrelo. ¿Qué potencia?

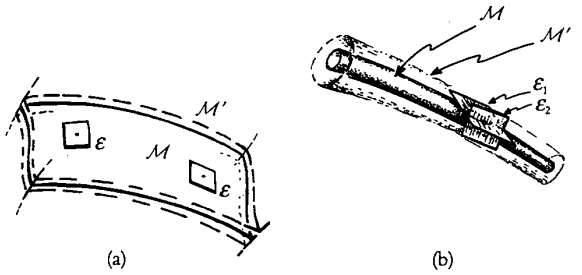


Fig. 31.2. Descripción geométrica de generadores de supersimetría. (a) Para un único generador ε, consideramos nuestra n-variedad $\mathcal{M}$ como una hipersuperficie en una $(n + 1)$-variedad $\mathcal{M}'$, donde ε es una 1-forma en $\mathcal{M}'$ definida en $\mathcal{M}$ (y ε define el n-plano tangente a $\mathcal{M}$ como en la Fig. 12.7 y §12.3). Estamos interesados en $\mathcal{M}'$, solo hasta «primer orden» en $\mathcal{M}$, pero $\mathcal{M}'$ suministra una dimensión adicional que apunta fuera de $\mathcal{M}$. (b) Con N generadores de supersimetría $\varepsilon_1, \ldots, \varepsilon_N$, la n-variedad $\mathcal{M}$ se considera ahora como una subvariedad de una $(n + N)$-variedad $\mathcal{M}'$, donde de nuevo estamos interesados en $\mathcal{M}'$ solo hasta primer orden fuera de $\mathcal{M}$. Las N 1-formas independientes $\varepsilon_1, \ldots, \varepsilon_N$, «sienten» las N direcciones extras que apuntan fuera de $\mathcal{M}$ y hacia $\mathcal{M}'$.

La situación no es esencialmente diferente si exigimos N generadores de supersimetría $\varepsilon_1, \ldots, \varepsilon_N$. Ahora consideramos que nuestra n-variedad $\mathcal{M}$ está inmersa en una $(n + N)$-variedad $\mathcal{M}'$, donde de nuevo lo que nos interesa de $\mathcal{M}'$ es solo su inmediata (primera) vecindad de $\mathcal{M}$. Ahora exigimos N 1-formas diferentes $\varepsilon_1, \ldots, \varepsilon_N$ para sondear las N direcciones extra[10] que apuntan fuera de $\mathcal{M}$ y hacia $\mathcal{M}'$. En mi opinión, esta imagen (que se debe a varias personas, entre ellas Abhay Ashtekar, que tiene un importante papel en el capítulo 32)[11]

hace las ideas subyacentes de supersimetría y supervariedades mucho más claras que los procedimientos formales (y de aspecto más bien misterioso) que se adoptan normalmente. Nótese que el «cuerpo» solo se refiere a cantidades que son enteramente intrínsecas a $\mathcal{M}$, mientras que el «alma» se refiere a cantidades con una componente que «apunta hacia fuera» en $\mathcal{M}'$, fuera de $\mathcal{M}$.Véase la Fig. 31.2b.

Incluso con esta interpretación geométrica clara, hay cosas extrañas en la forma en que se utiliza normalmente la «superálgebra», si se quiere mantener la consistencia con la imagen geométrica. Una p-forma ordinaria α en $\mathcal{M}$, para la cual p es un número impar, anticonmutaría con un generador de supersimetría ε, si consideramos que el producto de ε por α tiene que ser un producto cuña. Sin embargo, este no es el convenio normal en las aproximaciones estándar a la superálgebra, donde normalmente se consideraría que ε conmuta con α. Esto es básicamente una cuestión de notación, y si vamos a considerar productos de generadores de supersimetría con formas, podemos tomarlos formalmente como productos simétricos en lugar de productos cuña. Aunque esto tiene sentido matemático (formal), enturbia un poco la «limpia» imagen geométrica que estoy proponiendo aquí.

De hecho, en las aplicaciones de la supersimetría en teorías de física de partículas «ordinaria» es el caso más simple $N = 1$ el que se adopta normalmente. Al parecer, la razón es que la proliferación de supercompañeras es mucho mayor para un N grande, pues cada partícula básica pertenece a un 2^N-plete de «compañeras». ¡El problema observacional ya es suficientemente malo sin esto! En cada caso, puede considerarse que los supergrupos relevantes son transformaciones que implican «simetrías internas» (que remiten a las simetrías de las fibras de algún fibrado $\mathcal{B}$ sobre el espaciotiempo; §15.1) junto con las «rotaciones» que remiten a la extensión de $\mathcal{B}$ en la vecindad inmediata de $\mathcal{B}$ dentro de algún espacio $\mathcal{B}'$ de N dimensiones más altas.

31.4. Espaciotiempo de dimensiones más altas

Ahora que tenemos una idea mejor de lo que trata realmente la supersimetría y la «supergeometría», volvamos a la cuestión de la supergravedad. La excitación original que rodeó a esta idea a finales de la década de 1970 se debía a la esperanza en que, a diferencia de la relatividad general estándar de Einstein, la supergravedad podría ser renormalizable. En la teoría de vacío de Einstein habían aparecido divergencias no renormalizables «en el nivel de 2-lazos», donde los «lazos» se refieren a desarrollos en diagramas de Feynman,y el «número de lazos» se refiere al número de cortes que serían necesarios para reducir el diagrama de Feynman a un árbol (véanse §26.8, en especial el párrafo final, y la Fig. 26.8; véanse §§26.9, 10). Sin embargo, en presencia de materia tales divergencias aparecen ya en el nivel de 1-lazo, lo que puede considerarse un auténtico desastre. En supergravedad, estas divergencias de 1-lazo se cancelaban mágicamente para el tipo de materia que permitía la teoría, y muchos teóricos tenían grandes esperanzas en que esto continuaría para todos los órdenes de lazos. Desgraciadamente, no fue así, y se encontraron de nuevo divergencias no renormalizables en el nivel de 2-lazos en supergravedad.[12] Posteriormente se advirtió que si la dimensionalidad del espaciotiempo se incrementaba desde las cuatro dimensiones estándar a once, entonces las cosas parecían mucho más prometedoras. Pese a ello, seguía sin obtenerse una versión de la supergravedad completamente renormalizable y, como ha demostrado un trabajo mucho más reciente,[13] no se puede obtener así.

¿Cómo es posible que los físicos tomasen en serio la posibilidad de que la dimensionalidad del espaciotiempo fuera diferente de las cuatro que experimentamos directamente (una dimensión temporal y tres espaciales)? Como ejercicio matemático, estos objetos de dimensiones superiores están bien, pero se supone que esta va a ser una teoría física en la que el «espaciotiempo» significa realmente la combinación del espacio real con el tiempo. De hecho, como veremos en §31.7, la teoría de cuerdas (tal como se entiende hoy día) requiere que el espaciotiempo tenga más de cuatro dimensiones. En la teoría primitiva, el número de dimensiones era 26, pero innovaciones posteriores (que incluían las ideas de supersimetría; véase §31.2) llevaron a que la dimensión de dicho espaciotiempo se redujese a 10.

Antes de descartar esta idea como una fantasía total, debemos recordar de §15.1 el ingenioso esquema propuesto en 1919 por el entonces poco conocido matemático polaco Theodor Kaluza, y posterior-

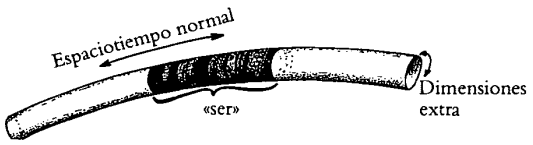


Fig. 31.3. Modelo de manguera de un espaciotiempo de dimensiones más altas de tipo Kaluza-Klein (véase la Fig.15.1), donde la dimensión a lo largo de la longitud de la manguera representa el 4-espaciotiempo normal y la dimensión alrededor de la manguera representa las dimensiones extra «pequeñas» (quizá a escala de Planck). Imaginamos un «ser» que habita en este mundo, que rebasa estas dimensiones extra «pequeñas», y por ello no es realmente consciente de ellas.

mente asumido por el físico matemático sueco Oskar Klein, al que ya se ha mencionado antes en este capítulo. Con tal de que las dimensiones extra (es decir, que exceden de 4) se consideren como dimensiones *pequeñas*, en un sentido apropiado, podríamos no ser directamente conscientes de ellas. ¿Qué significa «pequeñas» en este contexto? Recordemos la analogía de la «manguera» de la Fig. 15.1. Cuando se mira desde una gran distancia, la manguera parece ser 1-dimensional, pero si la examinamos más de cerca, encontramos una superficie 2-dimensional. La idea es que algún *ser* que habitara en el universo de la manguera no «sabría» que la dimensión extra que se enrolla alrededor de ella está realmente «allí», con tal de que las dimensiones físicas de dicho ser fueran mucho mayores que la circunferencia de la manguera. Comentarios similares se aplicarían a un «universo manguera» de $4 + d$ dimensiones, donde d de las dimensiones son «pequeñas» y no directamente percibidas por un ser mucho mayor que habita en el universo y que solo percibe las 4 dimensiones «grandes» (véase la Fig. 31.3).

¿Qué grado de «pequeñez» cabe esperar en el modelo de Kaluza-Klein, o en las modernas versiones en dimensiones aún más altas de esta idea? El propio Klein llegó a la conclusión de que la «escala» de la dimensión minúscula extra («circunferencia de la manguera») debería ser del orden de la distancia de Planck de 10^{-35} m. Este parece ser también el tipo de escala (o un poco mayor que ella) que se adopta en los esquemas más modernos, tales como la supergravedad en dimensiones más altas y la teoría de cuerdas. Es evidente que para «seres» como nosotros esto cuenta como «pequeño», ¡y cabría esperar que no tuviéra-

mos experiencia directa de dimensiones espaciotemporales extra tan minúsculas como eso!

De hecho, hay algunos estudios recientes (en teoría de cuerdas) en los que no se considera que las dimensiones extra sean (tan) pequeñas, sino que podrían referirse a algo incluso tan «grande» como de un milímetro de diámetro (o quizá ni siquiera cerrado). Una idea es que podría haber consecuencias observacionales de un esquema semejante que se manifiestan como una modificación de la ley de la inversa del cuadrado para la atracción gravitatoria a tales distancias. De hecho, recientemente se han realizado experimentos muy delicados para establecer si podrían ser detectables tales desviaciones de la teoría de Newton.[14] Por el momento se ha llegado a distancias de medio milímetro sin que se haya encontrado ninguna desviación.

Cualquiera que sea el estatus de estas ideas más recientes, esta sugerencia de una dimensionalidad más alta para el espaciotiempo no tiene, en el estado actual de nuestras deliberaciones, un estatus más convincente que el de ser una «idea curiosa», como lo era la sugerencia original de Kaluza-Klein. Cualquiera que pueda ser la elegancia matemática de esta idea, tenemos que abordar la cuestión de si existen buenas razones físicas para creer en un esquema semejante. En el caso del modelo de Kaluza-Klein original, la razón para adoptar esta perspectiva en dimensiones más altas era la de «geometrizar» el electromagnetismo. Como se ha dicho en §25.1, las únicas fuerzas de la naturaleza que se conocían (y entendían) en la primera parte del siglo XX eran la gravitatoria y la electromagnética. Einstein acababa de demostrar cómo se podía incorporar el campo gravitatorio en la curvatura de un espaciotiempo 4-dimensional. Tratar de incluir también el electromagnetismo en un marco geométrico semejante era una idea muy atractiva y que parecía natural. Más aún, había quizá algo más bien milagroso en la forma en que las mismísimas «ecuaciones de Einstein en vacío» —a saber, la anulación del tensor de Ricci ($R_{ab} = 0$; véase §19.6)— se aplicaban en la teoría 5-dimensional de Kaluza-Klein igual que en la relatividad general estándar 4-dimensional. En la teoría 4-dimensional, esta ecuación se refiere al *vacío*, es decir, a la ausencia de cualquier campo físico excepto la gravedad. En la teoría 5-dimensional, se refiere *casi* al estado en el que solo operan la gravedad

y el electromagnetismo, recogiendo así los campos físicos conocidos en aquel momento.

No obstante, el «casi» establece el caso de una manera demasiado fuerte. Pues —y esto es lo más importante— es esencial para el modelo de Kaluza-Klein que haya una *simetría* en la dimensión «pequeña», de modo que no hay infinitamente demasiados grados de libertad. Veamos por qué estos grados de libertad extra aparecerían si no fuera así. Recordemos la discusión de §16.7 respecto al «tamaño» del espacio de dimensión infinita de campos en cierto espacio dado. Para un campo que puede ser especificado por k componentes independientes libremente escogidas en una superficie q-dimensional de datos iniciales, la libertad es $\infty^{k\infty^q}$. En el caso de la relatividad general de Einstein, tenemos (por razones algo complicadas)[15] $k = 4$ y $q = 3$, de modo que esta cantidad resulta ser $\infty^{4\infty^3}$, y para la teoría de Maxwell obtenemos exactamente la misma libertad. En el caso de la teoría combinada de Einstein-Maxwell, el número efectivo de componentes por punto de la superficie de datos iniciales es la suma de los valores para cada campo por separado, de modo que tenemos $4 + 4 = 8$ componentes efectivamente independientes por punto de la 3-superficie inicial, y el valor correcto para la libertad completa es

$$\infty^{8\infty^3}.$$

Ahora, en una teoría 5-dimensional sujeta solo a la condición de Ricci-planitud (i.e., $R_{ab} = 0$), la superficie inicial es 4-dimensional (de modo que $q = 4$), y realmente resulta que $k = 10$. Esto nos daría una libertad para el campo $\infty^{10\infty^4}$ enormemente mayor que el valor requerido (como se ha mostrado antes), no porque 10 sea mayor que 8 (el valor k), sino porque 4 es mayor que 3 (el valor q). ¡Hay muchísimas más funciones de 4 variables que funciones de 3 variables!

En el modelo de Kaluza-Klein reducimos de nuevo el 4 a 3 imponiendo una simetría continua (de hecho, U(1); cf. §13.9) en la dimensión pequeña. Tiene que haber un vector de Killing (§14.7) que exprese dicha simetría y, en efecto, el 5-espacio de Kaluza-Klein es un S^1-fibrado $\mathcal{B}$ sobre el espaciotiempo 4-dimensional ordinario $\mathcal{M}$. Esto no parece tan alejado de la descripción fibrada convencional del electromagnetismo, como se ha descrito en §19.4 (y §15.8). Una diferencia básica es que al propio $\mathcal{B}$ se le asigna una (pseudo)métrica lorentziana Ricci-plana, en lugar de asignar meramente una métrica al espaciotiempo $\mathcal{M}$.[16] El hecho sorprendente en el modelo de Kaluza-Klein es que la imposición de Ricci-planitud en $\mathcal{B}$ (además de la simetría U(1)) está sorprendentemente próxima a ofrecernos las ecuaciones completas de la teoría de Einstein-Maxwell[17] en $\mathcal{M}$. Además de esto, todo lo que se necesita es que el vector de Killing tenga una norma constante no nula (de hecho negativa). Esto elimina un campo escalar indeseado, ¡y con ello se expresa la teoría de Einstein-Maxwell 4-dimensional exacta!

Por elegante que sea, la perspectiva de Kaluza-Klein sobre la teoría de Einstein-Maxwell no nos ofrece una imagen convincente de la realidad. Ciertamente, no hay una fuerte motivación para adoptarla que proceda de direcciones físicas. La supersimetría, por ejemplo, tiene ciertamente un argumento físico más contundente, debido a su valor indudable para reducir el problema de los infinitos en QFT. ¿Por qué son entonces tan populares las teorías tipo Kaluza-Klein en dimensiones más altas en los intentos modernos hacia una teoría más profunda de la naturaleza? Las principales razones proceden de la teoría de cuerdas que, en todas las versiones hoy activamente seguidas, emplean de una forma esencial tanto supersimetría como dimensiones más altas.[18]

31.5. La teoría de cuerdas hadrónica original

¿Qué *es* entonces la teoría de cuerdas? ¿Y por qué tiene un atractivo tan poderoso sobre tantos teóricos actuales? Una vez más, este esquema de cosas tiene su motivación más fuerte en el deseo de eliminar los infinitos de la QFT. En ese sentido, representa una continuación de las ideas impulsoras de los capítulos 24-26. Pero hay también otra motivación histórica importante, de carácter seminal, que tiene que ver específicamente con algunas curiosidades observacionales de la física hadrónica. En primer lugar, echemos una ojeada a estas.

Esta cuestión original tenía que ver con ciertas relaciones que se encontraron en la física de partículas de la dispersión de hadrones. En el capítulo 25 se ha mencionado que entre los hadrones existen mu-

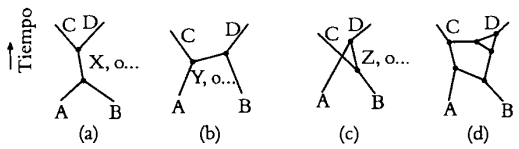


Fig. 31.4. Diagramas de Feynman de una dispersión hadrónica, donde dos partículas A, B se convierten en el par C, D. (a) En una familia de tales procesos, A y B se combinan para formar una partícula (resonancia), que casi inmediatamente se desintegra en C y D, habiendo muchos intermediarios posibles X, X′, X″, ... todos los cuales contribuyen al total. (b) En una familia de procesos alternativos se «intercambia» una partícula Y (o Y′, o Y″, o ...), convirtiendo A en C y B en D, y cada intermediario contribuye. (c) Un «intercambio» similar, que incluye Z (o Z′, o Z″, o ...), donde ahora A se convierte en D y B en C, y cada intermediario contribuye. Resulta que, a orden más bajo, las alternativas (a), (b), (c) son equivalentes, en lugar de tener que ser sumadas. (d) Otras maneras de conseguir esta transformación implican lazos cerrados.

chas «partículas» que tienen una vida tan corta (solo duran unos 10^{-23} segundos) que apenas merecen dicho nombre, y se les suele denominar *resonancias*. Recordemos ahora que las reglas de la QFT (§25.2, §§26.6,8) exigen que en cualquier proceso físico hay que sumar todas las diferentes actividades que pudieran tener lugar para obtener la amplitud cuántica total. Por consiguiente, todas las posibles partículas y resonancias deben ser tenidas en cuenta a tal efecto. Por ejemplo, podríamos tener un proceso de dispersión hadrónica en el que dos partículas A y B se acercan y, tras un instante fugaz, se convierten en el par de partículas C y D. Ahora bien, una manera en la que podrían hacer esto sería que A y B se combinen para formar una partícula (resonancia) X, que casi al instante se desintegra en las partículas C y D. Podría haber muchas de estas partículas intermedias posibles X, X′, X″, ..., y el efecto de cada una tendría que sumarse en el total. Los diagramas de Feynman para cada uno de estos procesos se muestran en la Fig. 31.4a. Ahora bien, una forma alternativa en que podría ocurrir la transformación es que se «intercambie» una partícula Y entre A y B, que convierte A en C y B en D. De nuevo podría haber una lista de posibles partículas intercambiadas Y, Y′, Y″, ..., cuyos diagramas de Feynman se muestran en la Fig. 31.4b. Existe una tercera familia de procesos mediante los cuales podría tener lugar la transformación, que difiere de

esto en que las partículas salientes C y D se toman en sentido opuesto, y los diagramas de Feynman para los mismos se muestran en la Fig. 31.4c. Podríamos imaginar otras maneras más complicadas mediante las cuales tiene lugar la transformación, que implican *lazos cerrados* (Fig. 31.4d), pero estos procesos de «orden superior» se considerarán por el momento poco importantes.

Para obtener la amplitud total para el proceso por el que el par (A, B) se convierte en el par (C, D), hay que sumar estas diferentes contribuciones, pero de forma bastante sorprendente lo que se encuentra es que cada una de estas tres posibilidades da la misma respuesta, y que esta única respuesta parece ser básicamente la respuesta correcta. Si sumáramos las tres respuestas, entonces obtendríamos algo demasiado alto. ¡De algún modo, cada uno de los tres conjuntos de diagramas de Feynman dados en las Figs. 31.4a,b,c, respectivamente, sumados por separado, representan físicamente la misma cosa! Esta «dualidad»[19] parece incomprensible desde la perspectiva de los diagramas de Feynman estándar, pero en 1970 el físico japonés-estadounidense Yoichiro Nambu,[20] basándose en una notable fórmula[21] hallada por el joven físico italiano Gabriele Veneziano en 1968, llegó a la conclusión de que se podía dar sentido a todo esto desde una perspectiva diferente, en la que las partículas hadrónicas individuales se modelaban por *cuerdas* en lugar de por partículas puntuales. Una historia de cuerda es una superficie 2-dimensional, de modo que los procesos descritos en los diagramas de Feynman de las Figs. 31.4a,b,c, respectivamente, pueden ahora representarse como las diversas «fontanerías» alternativas que se muestran en las Figs. 31.5a,b,c, respectivamente. Lo que choca en esta «perspectiva de cuerdas» es que estos tres procesos, que parecen tan diferentes desde la perspectiva de los diagramas de Feynman estándar, son ahora topológicamente equivalentes, y pueden considerarse como tres formas diferentes de mirar *el mismo* proceso. Así pues, la imagen de «cuerda» sugiere una forma de dar sentido a un hecho enigmático de la física hadrónica.

Este es solo un comentario cualitativo, pero la imagen de cuerdas también presentaba un modelo físico que proporcionaba una deducción matemática de la fórmula de Veneziano. Además, el modelo de cuerdas —en el que las cuerdas se comportan como minúsculas gomas

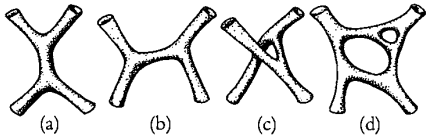


Fig. 31.5. La imagen de una historia de cuerda de los respectivos procesos de la Fig. 31.4 ofrece una explicación de la equivalencia entre (a), (b) y (c), puesto que pueden transformarse una y otra, ya que son topológicamente iguales. (d) Los procesos de orden más alto corresponden a una topología más complicada, en la que el género topológico corresponde al número de lazos (compárese con la Fig. 8.9).

elásticas, con una *tensión de cuerda* que aumenta en proporción al estiramiento de la cuerda— ofrecía una explicación de otra característica observada de la física hadrónica, a saber, la rectitud de las *trayectorias de Regge*. Las trayectorias de Regge son líneas que se ven cuando para una clase de hadrones concreta representamos el valor del espín frente al cuadrado de la masa. Estas líneas resultan ser extraordinariamente rectas. Un ejemplo se muestra en la Fig. 31.6. Por lo que sé, no hay todavía ninguna otra explicación de este sorprendente hecho observado concerniente a los hadrones.[22]

Además, el modelo de cuerdas ofrecía una esperanza razonable y significativa de que con esta imagen podría obtenerse una teoría finita de la física hadrónica. En términos generales, servía para «suavizar» las divergencias ultravioletas del enfoque convencional de Feynman (§26.8). Puede considerarse que dichas divergencias aparecen de efectos a pequeñas distancias cuando partículas puntuales se aproximan más y más sin límite. Las cuerdas no son partículas puntuales, de modo que esto ofrece una posibilidad de aliviar el problema. De hecho, los lazos cerrados en la imagen estándar de los diagramas de Feynman son los que provocan las divergencias. En la imagen de cuerdas, los lazos cerrados están simplemente dominados por superficies con topología más alta, como se indica en la Fig. 31.5d, que es una versión en cuerdas de la Fig. 31.4d. Esto debería dar algo finito en lugar de las integrales divergentes que salen de los diagramas de Feynman. Además, una imagen de una única historia de cuerda puede englobar muchos diagramas de Feynman diferentes, lo que nos da una oportunidad mucho mejor para representar la respuesta física total a un problema —que debería ser finita— en lugar de partes no físicas que podrían divergir individualmente pero que se supone que se cancelan mutuamente. Además, diferentes variedades de partículas podrían incorporarse simplemente como diferentes modos vibracionales de las cuerdas. Por último, las historias de cuerdas espaciotemporales 2-dimensionales tienen la notable propiedad adicional de que pueden tomarse como *superficies de Riemann*, que, como recordamos del capítulo 8, tienen propiedades analíticas y geométricas extraordinariamente ricas (el hecho que en realidad subyace en la notable fórmula de Veneziano). Aquí hay lugar, de hecho, para la magia compleja que parece ser parte del diseño de la naturaleza en el nivel cuántico de la realidad.

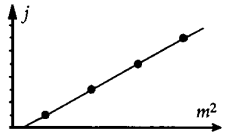


Fig. 31.6. «Trayectorias de Regge» rectas de resonancias de partículas de espín creciente, representadas frente a la masa al cuadrado. La imagen de cuerda elástica ofrece una explicación.

Indudablemente, esta es una imagen matemática extraordinariamente elegante de lo que podría estar pasando en cierto nivel de descripción física más profundo que el de las partículas ordinarias. Cuando oí hablar por primera vez de esta imagen (alrededor de 1970, en boca de Leonard Susskind, que fue uno de los primeros investigadores en este campo), quedé extraordinariamente turbado por la belleza y potencia de este conjunto de ideas. Me pareció que aquí había algo nuevo, que era a la vez matemáticamente excitante y aparentemente de relevancia directa para un área importante de la física de partículas. En aquella época, mi interés se centraba sobre todo en la teoría de twistores (a la que llegaremos en el capítulo 33), y me pareció que debía intentar forjar algún eslabón entre lo que yo estaba haciendo y estas nuevas ideas que sonaban muy prometedoras. La teoría de twistores hace un uso crucial de estructuras complejas (holomorfas), y con la teoría de cuerdas básica parece que estamos viendo que estas estructuras controlan el comportamiento físico mediante el uso esencial de superficies de Riemann, que en realidad son curvas complejas.[23]

Llama la atención que un reciente trabajo de Witten (2003) quizá esté haciendo realidad ahora algunas de estas primeras aspiraciones. Volveré a estos nuevos desarrollos muy positivos, que no utilizan espaciotiempos de dimensiones más altas, en §31.18. Pero de momento estos no representan un nueva teoría de cuerdas completa, y mis comentarios en las secciones intermedias se refieren, en su lugar, a lo que puede llamarse la «corriente principal» de la teoría de cuerdas.

31.6. Hacia una teoría de cuerdas del universo

¿Cómo han soportado el test del tiempo estas ideas iniciales tras los más de treinta años transcurridos desde entonces? Los estudios que se han llevado a cabo en el tema durante estos años, ¿han mantenido o superado la promesa inicial? Son preguntas a las que varias personas podrían dar respuestas muy diferentes. A veces, la teoría de cuerdas ha llegado a ser incluso un tema muy emocional. Para sus defensores acérrimos, la teoría de cuerdas (con su última transmutación) es la física del siglo XXI, y representa una revolución en el pensamiento físico comparable con, si no mayor que, las de la relatividad general o la mecánica cuántica. Para sus detractores más extremos, no ha conseguido absolutamente nada de eso, y tiene pocas posibilidades de desempeñar cualquier papel significativo en la física del futuro.

Sería imposible intentar siquiera ser desapasionado en la exposición de estos desarrollos, pero al menos trataré de ser razonablemente preciso y dar razones para las ideas que me he formado. Como antes, debo ofrecer la advertencia reglamentaria al lector de que mis propias opiniones están lejos de ser compartidas por numerosos físicos teóricos muy activos y excepcionalmente capaces. Pero no puedo sino presentar las cosas tal como yo las veo.

Puesto que mi punto de vista será menos que positivo acerca de muchos aspectos del programa actual de la teoría de cuerdas, debería al menos dar al lector la oportunidad de corregir este posible desequilibrio. En primer lugar, presento puntos de vista de dos de las figuras más importantes en el desarrollo del tema. En palabras de Michael Green,[24] de la Universidad de Cambridge:

> En el momento en que uno encuentra la teoría de cuerdas y se da cuenta de que casi todos los desarrollos importantes en física durante los últimos cien años surgen —y lo hacen con gran elegancia— de un punto de partida tan simple, comprende que esta teoría increíblemente convincente no tiene igual.

Y es famoso el comentario de Edward Witten del Instituto de Estudios Avanzados de Princeton:[25]

> Se ha dicho [Danielle Amati] que la teoría de cuerdas es una parte de la física del siglo XXI que cayó por azar en el siglo XX.

En cuanto a una exposición divulgativa, que es muy accesible, elocuente y entusiasta, y en absoluto crítica —pero que no profundiza en las ideas matemáticas—, véase Greene (1999).[26]

Para presentar una perspectiva consistente, aunque no necesariamente imparcial, del tema de la teoría de cuerdas desde mi punto de vista, propongo dar una exposición histórica aproximada de cómo ha incidido el tema en mi forma de pensar. De este modo, no solo trataré de indicar algo de los sucesivos desarrollos que se han producido en la teoría, sino también cuál ha sido mi reacción ante ellos. Lo que hace de la teoría de cuerdas algo tan difícil de evaluar desapasionadamente es que se apoya y elige sus direcciones de desarrollo casi por entero a partir de juicios estéticos guiados por desiderata matemáticos. Creo que es importante tomar nota de cada uno de los giros que ha sufrido la teoría, y señalar que casi todos nos han alejado de hechos establecidos observacionalmente. Aunque la teoría de cuerdas tuvo sus inicios en aspectos experimentalmente observados de la física hadrónica, luego se apartó drásticamente de ellos, y más tarde no ha tenido prácticamente ninguna guía de datos observacionales concernientes al mundo físico.

Imaginemos a un turista que trata de localizar un edificio concreto en una ciudad enorme y desconocida. No hay nombres de calles (o al menos ninguno que tenga sentido para el turista), no hay mapas y el cielo, totalmente cubierto, no ofrece indicio alguno de cuáles sean las direcciones norte, sur, o cualquier otra. A menudo hay una bifurcación

en la carretera. ¿Debe el turista girar a la izquierda o a la derecha, o quizá tomar ese atractivo pasadizo oculto a un lado? Frecuentemente, los giros no son en ángulo recto y las carreteras apenas son rectas. En ocasiones, la ruta es un callejón sin salida, de modo que hay que desandar el camino y tomar otro derrotero. A veces puede abrirse una ruta que nadie ha advertido antes. No hay nadie para preguntarle por el camino; en cualquier caso, la lengua local es desconocida. Al menos el turista sabe que el edificio que está buscando tiene una elegancia sublime, con un jardín muy bello. Esa, después de todo, es una de las razones principales para buscarlo. Y algunas de las calles que elige el turista tienen un atractivo estético más obvio que otras, con una arquitectura más atractiva y bellos patios adornados con soberbios arbustos y flores, que a veces, tras un examen riguroso, pueden resultar de plástico. Hay muchas elecciones implicadas en la ruta a seguir, y para cada elección la única guía del turista es el atractivo estético de la zona, junto con cierta sensación de una consistencia global, de estilo, o de algún tipo de pauta subyacente imaginada para la ciudad.

Ahora supongamos que *usted* es el turista, pero forma parte de un grupo conducido por un guía turístico de una inteligencia, conocimiento y sensibilidad impresionantes; el único problema es que, en este caso, el guía no tiene ningún conocimiento previo de la ciudad y nunca antes ha oído hablar la lengua del lugar. Quizá crea que el guía tiene mejores intuiciones estéticas que usted, y de hecho llega antes que usted a valorar estas cosas. En ocasiones, la sensibilidad del guía hacia las pautas ocultas localiza un edificio de una elegancia particularmente sofisticada. Pero, en esencia, los criterios no son de un tipo muy diferente de los que usted mismo podría utilizar. Si usted sigue al grupo, al menos tendrá la compañía de los otros, y puede hablarles de la arquitectura que les rodea y compartir la excitación de la búsqueda de su objetivo común. Incluso si no espera encontrar dicho objetivo, usted disfruta con la búsqueda. Pero tal vez, por el contrario, usted prefiera ir a su aire, cuando empieza a sospechar cada vez con más fuerza que el guía no sabe más que usted acerca de cómo encontrar el objetivo. Cada elección sucesiva de rumbo es una apuesta, y en muchas ocasiones quizá sienta que una elección diferente ofrecía más promesas que aquella que realmente ha elegido el guía...

Por supuesto, en los capítulos anteriores hemos sido testigos de varios ejemplos en los que grandes físicos han mostrado la potencia de sus intuiciones especiales, que a menudo son de una naturaleza típicamente matemática. Una de las más impresionantes debe ser seguramente el hallazgo de Dirac de la ecuación para el electrón, tal como se ha descrito en §24.7. Pese a todo, el salto estético fue en esencia solo un paso majestuoso en lo desconocido, a partir del cuerpo sólido de conocimiento matemático que había surgido de los hallazgos experimentales de la mecánica cuántica. La predicción que hizo Dirac de la antipartícula del electrón supuso otro salto semejante. Pero lo dio con gran cautela, y fue rápidamente confirmado por la observación. La relatividad general de Einstein también estuvo impulsada en parte por consideraciones estéticas matemáticas, y la fuerza de la relatividad general procede, en muy gran medida, de su estructura matemática profundamente bella. Cuando Einstein formuló la teoría por primera vez, no había ninguna exigencia clara sobre bases observacionales. Pese a todo, difícilmente puede decirse que Einstein estuviera impulsado solo por consideraciones estéticas matemáticas. Su guía procedía sobre todo de la física, y residía en su convicción de que el principio de equivalencia (§17.4) debe ser fundamental para la comprensión de la gravedad.

Por el contrario, la teoría de cuerdas ha tenido un impulso casi en exclusiva matemático. Antes de nada, debería dejar claro que esto no es necesariamente malo. Todas las teorías físicas satisfactorias tienen sólidas bases matemáticas. La consistencia matemática es, en efecto, una característica importante para una teoría física, si esta va a tener un sentido global. Y una vez que se ha establecido un marco matemático concreto, los desarrollos matemáticos rigurosos dentro de dicho marco pueden tener importantes consecuencias para el mundo físico. (Los desarrollos lagrangiano y hamiltoniano en física clásica descritos en el capítulo 20 ofrecen ejemplos impresionantes de ello.) No obstante, las dificultades aparecen cuando para superar una inconsistencia hay que cambiar una teoría previamente aceptada, y la forma concreta en que podría cambiarse una teoría depende del conocimiento matemático y las preferencias estéticas concretas del teórico. A menudo, el cambio será solo una idea —quizá incluso una «idea brillante»— que proba-

blemente tendrá aún fallos de consistencia matemática, aunque tal vez diferentes de los de la teoría que sustituye. Entonces podrían ser necesarios cambios adicionales, y así sucesivamente. Si hay demasiados de estos, la probabilidad de hacer la conjetura correcta en cada ocasión puede llegar a ser extraordinariamente pequeña.

31.7. Motivación de cuerdas para dimensiones espaciotemporales extra

Una primera inconsistencia de la imagen de la teoría de cuerdas fue la aparición de una anomalía grave. Recordemos de §30.2 que aparecen anomalías cuando las reglas de conmutación clásicas, que expresan una simetría o propiedad de invariancia clásica, no pueden ser completamente realizadas por conmutadores cuánticos, de modo que la teoría cuántica pierde una cualidad de la teoría clásica que podría haberse considerado esencial. En el caso de la teoría de cuerdas, esta anomalía se refería a una invariancia de parametrización esencial en la descripción de la cuerda. La presencia de la anomalía produjo efectos que se consideraban desastrosos. No obstante,[27] se vio que al incrementar de 4 a 26 el número de dimensiones espaciotemporales la anomalía desaparecía.[28] Por consiguiente, la teoría de cuerdas solo parecía ser mecanocuánticamente consistente en un espaciotiempo de 26 dimensiones.

Mi reacción a esto fue básicamente: «Debería haber una vía diferente para evitarlo», aunque nunca consideré el problema lo suficiente para apreciar la fuerza del razonamiento que había tras esta conclusión «26-dimensional». Sospecho que muchos otros reaccionaron de la misma forma, porque en ese momento la teoría perdió mucha de su popularidad anterior. Pero mis razones para descartar un modelo de universo 26-dimensional tenían un ingrediente adicional, que venía de la teoría de twistores. Como veremos en §§33.2,4,10, una consecuencia esencial de mi perspectiva «twistorial» es que el espaciotiempo tiene en realidad los valores directamente observados de una dimensión temporal y tres dimensiones espaciales (i.e., «1 + 3 dimensiones»).

Además del problema de qué hacer con estas dimensiones extra —que presumiblemente iban a ser tratadas de acuerdo con alguna receta de tipo Kaluza-Klein—, este modelo de cuerdas de apariencia simple para los hadrones tropezaba con otras dificultades, tales como la aparición del comportamiento *taquiónico* (propagación más rápida que la luz). Asimismo, el éxito creciente del modelo estándar, tal como se ha descrito en el capítulo 25, llevó a los físicos a interesarse menos que antes en tales sugerencias «desencaminadas» como modelos de cuerdas. Las características enigmáticas de la física hadrónica mencionadas antes y que pusieron a Veneziano, Nambu y otros en el camino de las cuerdas, encontraron una explicación (parcial) alternativa en QCD en términos de la imagen quark-gluón.

Más en concreto, la naturaleza «puntual» de los constituyentes de los hadrones se estaba haciendo experimentalmente manifiesta, siendo compatible con la imagen de los quarks del modelo estándar pero no con la imagen de cuerdas, tal como era entonces. El tamaño típico de un lazo de cuerda estaría relacionado con la intensidad del acoplamiento de la cuerda, y para las cuerdas hadrónicas originales (con una tensión de cuerda compatible con la intensidad de la constante de acoplamiento para la interacción fuerte), esto daría una escala de lazo promedio de unos 10^{-15} m. Esto difícilmente es «puntual» en la escala del protón, pues es comparable al «tamaño» del propio protón.

Tras casi una década durante la que hubo poco interés en la teoría de cuerdas, tuvo lugar un desarrollo que dio como resultado lo que a veces se conoce como «la primera revolución de las supercuerdas». En 1984, Michael Green y John Schwarz propusieron un esquema (que recogía algunas sugerencias anteriores hechas por Schwarz y Joël Scherk) en el que la supersimetría era incorporada en la teoría de cuerdas (para proporcionarnos «supercuerdas» en lugar de solo «cuerdas»), y la dimensionalidad espaciotemporal[29] se reducía con ello de 26 a 10. Esto eliminaba el «problema taquiónico» que se ha mencionado antes. Además, con un cambio radical en la escala y naturaleza de la tensión de la cuerda, la teoría de cuerdas se iba a considerar ahora fundamentalmente una teoría de gravitación cuántica en lugar de una teoría de interacciones fuertes. Ya se había reconocido que debería haber una partícula/campo sin masa de espín 2 que aparece de

un modo de vibración de las cuerdas. Esto había sido algo embarazoso para la versión «hadrónica» original de la teoría de cuerdas, puesto que no hay ninguna partícula hadrónica de esta naturaleza. Pero con las nuevas cuerdas, con su tensión de cuerda mucho mayor, sería apropiado identificar este campo sin masa con la gravedad. Ahora el tamaño de un lazo de cuerda típico es algo del orden de la minúscula longitud de Planck (gravitatoria) —aproximadamente, 20 órdenes de magnitud menor que antes, y ciertamente puntual en la escala hadrónica.

Debería mencionar otra diferencia técnica en la naturaleza de la tensión de la cuerda que se introduce con las nuevas cuerdas de «escala gravitatoria» (que normalmente no se resalta en las exposiciones divulgativas). Las cuerdas hadrónicas originales eran como gomas elásticas, en las que la tensión aumenta cuando se estira la cuerda, proporcionalmente a la longitud de estiramiento.[30] Sin embargo, las nuevas supercuerdas de escala gravitatoria ejercen una tensión constante $\hbar c/\alpha'$, que es así independiente del estiramiento, donde α' es un número muy pequeño (una medida de área) conocida como la *constante de cuerda*. A este respecto, la cuerda hadrónica original era mucho más parecida al tipo de entidad que había sido familiar en la física ordinaria en la que tiene sentido físico una versión clásica de ella. (Una versión clásica de la nueva supercuerda, con su tensión constante, ¡se contraería casi instantáneamente hasta una singularidad de tamaño nulo!)

31.8. ¿La teoría de cuerdas como gravedad cuántica?

Estos desarrollos transformaron por completo la percepción general que se tenía de la teoría de cuerdas, que rápidamente adquirió gran popularidad. Eran frecuentes las afirmaciones de que la teoría de cuerdas proporcionaba una «teoría completa y consistente de la gravedad cuántica», en donde la no renormalizabilidad de la relatividad general estándar (véase §31.1) es reemplazada por una teoría de cuerdas de la gravedad cuántica completamente finita.[31] Aunque, si se les insta, algunos defensores de la teoría de cuerdas podrían admitir que

no todas las reivindicaciones de finitud estaban perfectamente demostradas, esto se consideraría una cuestión de poca importancia. Como ha comentado un prominente físico teórico y teórico de cuerdas[32]

> La teoría de cuerdas es tan obviamente finita que si alguien publicara una demostración, yo no estaría interesado en leerla.

Además, los teóricos de cuerdas tendían a considerar la teoría de cuerdas de la gravedad cuántica como «el único juego en la ciudad», como ilustra el siguiente comentario sobre las aproximaciones a la gravedad cuántica distintas de la teoría de cuerdas realizado por Joseph Polchinski (1999):

> ... no hay alternativas ... todas las buenas ideas son parte de la teoría de cuerdas.

Sospecho que fue el carácter categórico de las primeras afirmaciones de finitud (aunque véase §31.13) lo que dio a la teoría buena parte del impulso que adquirió entonces. De hecho, si el pretendido descubrimiento de la «gravedad cuántica» buscada —la unión que faltaba entre las dos grandes revoluciones del siglo XX— fuera reconocido realmente, esto establecería la teoría de cuerdas no solo como uno de los principales logros intelectuales del siglo, sino también como un marco básico revolucionario para el futuro progreso en la física fundamental.

Creo que incluso muchos de los teóricos de *cuerdas* actuales considerarán que las afirmaciones que se hicieron en los años ochenta en el sentido de que la teoría «resolvía» por completo el problema de la gravedad cuántica eran exageradas. Esas personas preferirán ahora asumir una posición más sobria que la que habían defendido antes porque la «teoría de cuerdas» de hoy ha avanzado y difiere de manera sustancial del esquema de 1984. De todas formas, probablemente estarían dispuestos a aceptar la idea de que la teoría de cuerdas de 1984 había proporcionado al menos el paso más impresionante hacia este objetivo de gravedad cuántica.

¿Cuál fue mi respuesta a estas afirmaciones? Me temo que muy negativa, como lo fue la reacción de la mayoría de mis colegas más cercanos. Sin duda, buena parte de la razón para esta reacción negativa podría atribuirse a diferencias en la base cultural entre quienes, como yo mismo y mis colegas, tenían una perspectiva enraizada en un profundo interés por la relatividad general de Einstein, y aquellos cuyo empuje procedía más del lado de la QFT. El efecto principal de esta diferencia de perspectiva era que partíamos de ideas completamente diferentes respecto a las cuestiones centrales que tenían que resolverse en una unión cuántico/gravedad. Aquellos que venían del lado de la QFT tendían a tomar la renormalizabilidad —o, más correctamente, la *finitud*— como el objetivo fundamental de esta unión. Por el contrario, quienes veníamos del lado de la relatividad considerábamos que las cuestiones fundamentalmente importantes que había que resolver eran los profundos conflictos conceptuales entre los principios de la mecánica cuántica y los de la relatividad general, y a partir de esa resolución esperábamos avanzar hacia una nueva física del futuro. Nuestra reacción negativa ante las firmes afirmaciones que los teóricos de cuerdas hacían en esa época no se debían solo a cuestiones de detalle o incredulidad general (aunque estas eran también muy importantes), sino a una frustración por el hecho de que los mismos problemas que pensábamos que eran fundamentales para la cuestión general cuántico/gravedad ¡no parecían ser reconocidos en absoluto por los teóricos de cuerdas!

Algunas de estas cuestiones han sido abordadas en §30.11 (y se mencionarán otras en §33.2). No obstante, habría que mencionar que las cuestiones planteadas en dichas secciones apenas escarban en la superficie de los profundos conflictos que plantea[33] el principio de covariancia general en relación con la QFT (§19.6). Está también la cuestión básica de lo que debe ser realmente una «geometría espaciotemporal cuántica». La teoría de cuerdas opera básicamente con un espaciotiempo de fondo «clásico» suave, que no está siquiera directamente influenciado por la presencia de una cuerda, puesto que la propia cuerda básica no lleva energía, y así no «curva» directamente el espaciotiempo de fondo. Muchas personas en la comunidad relativista piensan que la verdadera «geometría cuántica» debería asumir algunos elementos de dis-

cretización, o al menos diferir profundamente de la imagen clásica de variedad suave.

Estas cuestiones profundas serán abordadas más directamente en los dos capítulos siguientes, como lo serán ciertas aproximaciones a su resolución. En particular, los «lazos» que encontraremos en el capítulo siguiente (§32.4), aunque con un parecido superficial a las cuerdas, son completamente diferentes de estas en muchos aspectos. En particular, la geometría espaciotemporal *está* profundamente influida —en realidad, esencialmente *creada*— por la presencia de dichos lazos, pues la métrica espacial está totalmente concentrada a lo largo de ellos y se anula por completo en cualquier otro lugar. Por otra parte, en la teoría de cuerdas, se considera que ya hay presente un espaciotiempo suave como fondo para las cuerdas, y las restricciones a su geometría métrica proceden solo de influencias indirectas de las cuerdas, de una manera a la que pronto llegaremos (en §31.9). Pero de momento dejemos a un lado la cuestión de si los puntos realmente importantes de la gravedad cuántica han sido o no adecuadamente abordados por la teoría de cuerdas. Consideremos las afirmaciones de los teóricos de cuerdas en el sentido de que tienen una teoría cuántica finita de la gravedad. ¿La tienen? Trataré de abordar esta cuestión en lo que queda de esta sección y en las cinco que la siguen.

Un punto importante está quizá contenido en las palabras utilizadas. Lo que afirman los teóricos de cuerdas es que tienen una «teoría cuántica de la gravedad», no de la relatividad general o de la teoría de Einstein. ¿Qué entienden ellos por «gravedad» si no es la relatividad general soberbiamente confirmada de Einstein? Recordemos, en primer lugar, que el espaciotiempo de los teóricos de cuerdas es ahora 10-dimensional (o, como pronto aprenderemos —§31.14—, *aproximadamente* 10-dimensional; ¡pero tratemos de que esto no nos preocupe por el momento!). ¿Qué es la «gravedad» en 10 dimensiones? El cálculo tensorial funciona igual de bien en 10 dimensiones que en 4 (véanse §§14.4,8), de modo que aún podemos construir el tensor de Ricci R_{ab}, igual que antes. Como hemos visto en §19.6, la condición para un *vacío* en la gravedad einsteiniana ordinaria es la Ricci-planitud, de modo que podríamos conjeturar que la «ecuación del vacío» gravitatoria de los teóricos de cuerdas es la misma, a saber

$$R_{ab} = 0,$$

excepto que ahora estamos en 10 dimensiones. También cabría esperar, por analogía con lo que sucede en la teoría Kaluza-Klein 5-dimensional donde el «5-vacío» incluye gravitación y electromagnetismo, que esta ecuación en 10 dimensiones, i.e., el «10-vacío», va a acomodar también todos los campos no gravitatorios además de la gravedad.

Bien, esto *es* lo que básicamente quieren decir los teóricos de cuerdas, al menos aproximadamente. Más exactamente, ellos consideran que la Ricci-planitud es consecuencia solo del primer término en una serie de potencias infinita en la constante de cuerda α', donde los términos de órdenes más altos nos proporcionan «correcciones cuánticas» a la Ricci-planitud. (El coeficiente de $(\alpha')^r$ podría incluir derivadas superiores de tensores de curvatura y expresiones polinómicas en tales tensores.) Recordemos que la constante de cuerdas es muy pequeña. Más aún; además de la métrica en el espaciotiempo 10-dimensional, hay también otros campos que aparecen en esta discusión. Uno de estos es un campo tensorial antisimétrico, y también hay un campo escalar conocido como *dilatón*[34] (que tiene que ver con escalamientos globales) bastante similar al (indeseado) campo escalar de la teoría de Kaluza-Klein original. (Recordemos que este escalar se elimina normalizando el vector de Killing; véase el penúltimo párrafo de §31.4.) El dilatón tendrá importancia para discusiones posteriores (véase §31.15). Recordemos que la constante de cuerda es muy pequeña. Ahora se considera que es solo algo mayor que el cuadrado de la longitud de Planck (siendo α' un área minúscula), con

$$\alpha' \approx 10^{-68}\ \mathrm{m}^2.$$

Así pues, se considera que la Ricci-planitud es una excelente aproximación para la métrica 10-espaciotemporal.

31.9. Dinámica de cuerdas

Usted podría preguntarse de dónde proceden realmente estos enunciados acerca de la curvatura espaciotemporal, puesto que la teoría de

cuerdas es realmente tan solo una teoría acerca de estas pequeñas cuerdas que se mueven en algún espaciotiempo de fondo (aunque de 9 dimensiones espaciales). De hecho, todavía no he sido muy específico acerca de las ecuaciones que controlan la dinámica de cuerdas. Veámoslo a continuación.

Como es normal en la teoría de campos, existe un lagrangiano (§§20.5,6 y §26.6), y el lagrangiano de cuerdas se define como $1/2\alpha'$ multiplicado por el área superficial de la historia 2-superficial —*la hoja de universo*— que describe la cuerda en el espaciotiempo. La métrica sobre la hoja de universo tiene que coincidir con la inducida a partir del espaciotiempo; y clásicamente la dinámica consistiría simplemente en que la hoja de universo es una especie de «película de jabón», o «superficie mínima» (de signatura métrica apropiada) en el fondo espaciotemporal dado. Clásicamente, el fondo no está sujeto a ninguna ligadura. La cuerda solo se agita de acuerdo con esta dinámica especificada. Sin embargo, en la mecánica cuántica la cuestión de la anomalía se agrava, y observamos que incluso la condición de que el fondo tenga 10 dimensiones con supersimetría no es ahora suficiente, sino que también se necesitan las condiciones anteriores sobre la curvatura 10-espacial para proporcionar condiciones de consistencia en la métrica de fondo para las cuerdas cuánticas.

Además de este requisito de consistencia tipo ecuación de Einstein, recordamos el «modo más bajo de excitación» de una cuerda cerrada, mencionado en §31.7, que parecía describir una partícula sin masa de espín 2. Su carácter de «espín 2» surge porque el modo tiene una estructura de cuadripolo (o $\ell = 2$) para su oscilación (véanse §22.11 y §32.2) y carece de masa esencialmente porque es el modo más bajo de una cuerda muy «rígida». Aunque el modo había presentado un serio problema para las cuerdas hadrónicas originales, ahora, en su nuevo contexto gravitatorio, se veía favorablemente porque en la física (4-dimensional) ordinaria un *gravitón* (cuanto del campo gravitatorio) es una partícula sin masa de espín 2. En el análisis convencional, esto sale de un examen de las perturbaciones en el campo métrico (descrito por un tensor simétrico «h_{ab}», que da el desplazamiento infinitesimal en una métrica desde g_{ab} a $g_{ab} + \varepsilon h_{ab}$, donde ε es infinitesimal; véase también §32.2). Parece que en la nueva teoría de cuerdas —en vista del anterior

requisito de consistencia tipo ecuación de Einstein (aunque en 10 dimensiones en lugar de 4), y de este modo de excitación de cuerda «tipo gravitón»— el punto de vista es que «la teoría de cuerdas incluye la gravedad». En palabras de Edward Witten (1996):

> La teoría de cuerdas tiene la extraordinaria propiedad de *predecir la gravedad*,

y Witten también ha comentado:[35]

> el hecho de que la gravedad es una consecuencia de la teoría de cuerdas es una de las más grandes intuiciones teóricas de cualquier época.

Sin embargo, se debería resaltar que, además de la cuestión de la dimensionalidad, la aproximación de la teoría de cuerdas está restringida (hasta ahora en casi todos los aspectos) a ser meramente una teoría perturbativa, expresada en términos de una serie de potencias (digamos en la «ε» mencionada antes, aunque la mayoría de los cálculos en la teoría de cuerdas se refieren a series de potencias en la constante de cuerda α'). Esta restricción es considerada como una seria limitación por la mayoría de los que trabajan en la relatividad, para quienes las consideraciones anteriores no bastan para ofrecernos una teoría con los mismos profundos principios subyacentes que la relatividad general de Einstein.

Una versión de una «filosofía de cuerdas» de las que tengo conocimiento es que deberíamos tratar de considerar la física como si fuera «realmente» una QFT *2-dimensional*, y que la noción geométrica de espaciotiempo 10-dimensional es una cuestión secundaria respecto a la «realidad» más primitiva de la propia hoja de universo de la cuerda 2-dimensional. Todo tiene que describirse en términos de «excitaciones de cuerdas», y estas deben considerarse como cantidades que son meramente funciones de las 2 coordenadas en la hoja de universo. Las 10 dimensiones espaciotemporales son sentidas por dichas excitaciones, pero todo es un tipo de «campo en la hoja 2-dimensional».

Tengo grandes dificultades con este punto de vista para una teoría que pretende describir la gravitación, donde habría grados de libertad dinámicos en la geometría espaciotemporal. Recordemos de §16.7 (y véase la discusión en §31.4 y §§31.10-12,15-17) que hay muchísimas más funciones o campos en un espacio de muchas dimensiones que en un espacio de menos dimensiones, cualquiera que sea el número de componentes independientes que pueda tener la función (campo) en cada punto, con tal de que este número de componentes sea finito. Además, para cualquier noción ordinaria de «excitación de cuerda», este número de componentes por punto sería finito (porque cada punto de la hoja de universo solo puede desplazarse en un número finito de direcciones independientes en el espacio ambiente). Parece que esta «filosofía de cuerdas» concreta sería una forma errónea de considerar las cosas. Aunque dudo de que sea en realidad mantenida de forma rigurosa por los teóricos de cuerdas, el hecho de que algunos de ellos hayan estado dispuestos a defender un punto de vista semejante es quizá indicativo de la actitud aparentemente displicente de muchos teóricos de cuerdas con respecto a la dimensionalidad espaciotemporal. La 4-dimensionalidad de nuestro espaciotiempo *observado*, quizá pensada como un «efecto de baja energía», parece considerarse con frecuencia como una cuestión de importancia menor.

En cualquier caso —incluso en 10 dimensiones—, la «ecuación de vacío de Einstein» de la teoría de cuerdas se considera una mera consecuencia de la *consistencia* de la propia hoja de universo 2-dimensional de la cuerda. Y se considera que la Ricci-planitud debe seguir manteniéndose ¡incluso en aquellas localizaciones espaciotemporales donde la propia hoja de universo de la cuerda *no* está localizada! Si la teoría cuántica estuviera describiendo realmente la dinámica cuantizada del sistema clásico acoplado especificado por

> «9-espacio de fondo que contiene una cuerda en movimiento»,

la consistencia sobre la curvatura de fondo necesitaría mantenerse solo donde está localizada la cuerda. Así que tenemos que aceptar la idea de que no es este sistema clásico el que se está cuantizando. De hecho, pese a que pretende ser una teoría de la gravedad, la teoría de cuerdas no trata apropiadamente el problema de describir los grados de libertad dinámicos en la métrica espaciotemporal. El espaciotiempo pro-

porciona simplemente un fondo fijo, limitado de ciertas maneras para permitir que las propias cuerdas tengan plena libertad.

31.10. ¿Por qué no vemos las dimensiones espaciales extra?

Si ahora vamos a tomar en serio la dinámica plena del espaciotiempo 10-dimensional, tenemos que afrontar el problema opuesto de cómo reducir la enorme libertad funcional extra en el espacio 10-dimensional a la que sería adecuada para una teoría física ordinaria en cuatro dimensiones espaciotemporales. La Ricci-planitud en 10 dimensiones permite una libertad funcional $\infty^{70\infty^9}$ (véase §16.7) que es enormemente mayor que la mera $\infty^{N\infty^3}$ que tenemos para una teoría de campos en el 4-espacio ordinario, tomando N componentes independientes por punto (§16.7 y §31.4). (Una teoría 10-dimensional «Ricci-plana» tendría, de hecho, 70 funciones independientes como datos libres en una superficie inicial 9-dimensional.)[36] La enormidad extra procede de que 9 es mayor que 3. En comparación, el tamaño relativo de 70 y N contribuye de forma despreciable a esta explosión de libertad funcional.[37] Una teoría *clásica* de campos ordinaria en un espaciotiempo 10-dimensional (sin una restricción como la simetría que decreta el vector de Killing de la teoría de Kaluza-Klein original; véase §31.4) entraría en conflicto con nuestro universo observado, debido a su tremendo exceso de libertad funcional. Volveremos a esta cuestión en §§31.12,16.

¿Por qué los teóricos de cuerdas no parecen estar especialmente molestos por esta excesiva libertad funcional? Al parecer, una razón es que tienen la esperanza de que en una teoría de cuerdas adecuadamente cuantizada puede haber ligaduras espaciotemporales adicionales, derivadas de requisitos de consistencia para las cuerdas cuantizadas, que reducen efectivamente la libertad funcional de las cuerdas. Echaremos una ojeada a estas esperanzas en §31.6. Pero el argumento principal que se suele esgrimir procede de la expectativa de que, si se supone que las seis dimensiones «extra» son extraordinariamente «pequeñas» (digamos del orden de la escala de Planck de 10^{-35} m), entonces —para energías que hoy son accesibles en el mundo físico—

consideraciones mecanocuánticas vendrían en auxilio y «eliminarían» efectivamente los grados de libertad que conciernen a las dimensiones espaciales extra.

¿Cómo funciona esto? Como se ha comentado antes, casi todas las consideraciones en la teoría de cuerdas se llevan a cabo en un marco perturbativo, en el que meramente se examinan perturbaciones pequeñas respecto a algún modelo básico concreto. Aquí vamos a considerar un «espaciotiempo» básico que es el producto $\mathbb{M} \times \mathcal{Y}$ del 4-espacio de Minkowski normal por un 6-espacio riemanniano dado $\mathcal{Y}$ de género espacio y compacto, donde el «tamaño» total de este $\mathcal{Y}$ particular es muy pequeño, digamos de la escala de Planck de 10^{-35} m. Vamos a examinar pequeñas perturbaciones de $\mathbb{M} \times \mathcal{Y}$.

En primer lugar, necesitamos una imagen más clara de lo que es una «variedad producto» $\mathcal{A} \times \mathcal{B}$, donde tanto el m-espacio $\mathcal{A}$ como el n-espacio $\mathcal{B}$ se toman como variedades (pseudo)riemannianas. Recordemos de §15.2 (y la Fig. 15.3a) que los puntos de $\mathcal{A} \times \mathcal{B}$ se describen como pares (a, b), donde a pertenece a $\mathcal{A}$ y b pertenece a $\mathcal{B}$, de modo que la dimensión de $\mathcal{A} \times \mathcal{B}$ es $m + n$ (véase el ejercicio [15.1]). ¿Cómo vamos a definir la métrica (pseudo)riemanniana en $\mathcal{A} \times \mathcal{B}$? Esta es la suma directa de las métricas en $\mathcal{A}$ y en $\mathcal{B}$. Podemos utilizar coordenadas locales $(x^1, \ldots, x^m, y^1, \ldots, y^n)$ para $\mathcal{A} \times \mathcal{B}$, donde $(x^1, \ldots, x^m)$ y $(y^1, \ldots, y^n)$ son coordenadas locales para $\mathcal{A}$ y $\mathcal{B}$, respectivamente. Entonces, las componentes métricas g_{ij} para $\mathcal{A} \times \mathcal{B}$ tienen «forma diagonal por bloques» (similar a la mostrada en §13.7 para las matrices de una representación completamente irreducible) que describe la suma directa de las componentes métricas para $\mathcal{A}$ y para $\mathcal{B}$, respectivamente: la distancia métrica al cuadrado en $\mathcal{A} \times \mathcal{B}$ es la suma de las distancias en $\mathcal{A}$ y $\mathcal{B}$ por separado (Fig. 31.7).

Un hecho clave de posterior relevancia (véase §31.14) es que si la métrica de $\mathcal{A}$ y la métrica de $\mathcal{B}$ son ambas Ricci-planas (se anula el tensor de Ricci; véase §19.6), entonces la métrica suma directa de $\mathcal{A} \times \mathcal{B}$ es también Ricci-plana.[31.3] El espacio $\mathcal{Y}$ en nuestro producto $\mathbb{M} \times \mathcal{Y}$ se considera Ricci-plano, y el propio $\mathbb{M}$, siendo plano, es ciertamente

[31.3] ¿Por qué? *Sugerencia*: Examine la forma de las expresiones explícitas en los ejercicios [14.26] y [14.27], en §14.7.

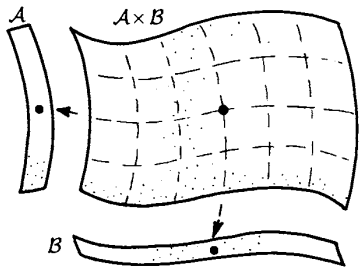


Fig. 31.7. La variedad producto $\mathcal{A} \times \mathcal{B}$ (véanse la Fig. 15.3a y §15.2) de dos espacios (pseudo)riemannianos $\mathcal{A}$ y $\mathcal{B}$ es ella misma (pseudo)riemanniana. Si $\mathcal{A}$ y $\mathcal{B}$ son Ricci-planas, entonces también lo es $\mathcal{A} \times \mathcal{B}$.

Ricci-plano. El producto $\mathbb{M} \times \mathcal{Y}$ es así también Ricci-plano, como se requería.

Vamos a considerar que el espacio *compacto* $\mathcal{Y}$ tiene un tamaño espacial total del orden general de la escala de Planck, o quizá un poco mayor. (Recordemos el significado de la compacidad, descrito en §12.6 y la Fig. 12.13.) ¿Cómo se van a describir las perturbaciones de $\mathbb{M} \times \mathcal{Y}$? Estas se darán mediante campos (tensoriales) en $\mathbb{M} \times \mathcal{Y}$, como las h_{ab} de §31.9, que nos proporcionan cambios infinitesimales en la métrica de $\mathbb{M} \times \mathcal{Y}$.

Para estudiar campos en $\mathbb{M} \times \mathcal{Y}$, es útil pensar en términos de un problema de valor inicial; así, representamos $\mathbb{M}$ como $\mathbb{M} = \mathbb{E}^1 \times \mathbb{E}^3$, donde el 1-espacio euclídeo $\mathbb{E}^1$ se refiere a una coordenada temporal t, y el 3-espacio euclídeo $\mathbb{E}^3$ se refiere al espacio. Entonces analizamos estos campos en términos de los modos normales en $\mathbb{E}^3 \times \mathcal{Y}$.Véase la Fig. 31.8. (Recordemos el concepto de «modo normal», tal como se ha descrito clásicamente en §20.3 y en el contexto cuántico en §§22.11,13.) ¿Qué aspecto tienen estos modos normales? Debido a la estructura «producto» de $\mathbb{E}^3 \times \mathcal{Y}$, podemos representar cada uno de estos modos simplemente como el producto ordinario de un modo de $\mathbb{E}^3$ por un modo de $\mathcal{Y}$. Los modos de $\mathbb{E}^3$ son precisamente estados de momento (§21.11) y forman una familia continua. En cuanto a los modos normales de $\mathcal{Y}$, la compacidad asegura que forman una familia discreta, cada uno de ellos caracterizado por un conjunto finito de autovalores. (Recordemos la discusión al final de §22.13.) ¿Cómo «excitaríamos»

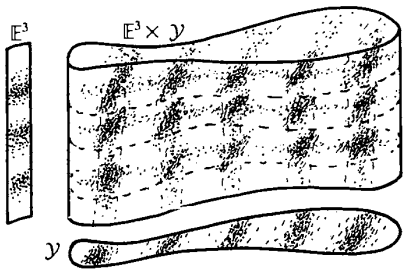


Fig. 31.8. Los modos de perturbación en $\mathbb{E}^3 \times \mathcal{Y}$ (para la ecuación de Laplace) son productos de modos en $\mathbb{E}^3$ por modos en $\mathcal{Y}$.

uno de estos modos, de manera que la simple geometría $\mathbb{E}^3 \times \mathcal{Y}$ se convierta en alguna otra cosa?

El argumento habitual de los teóricos de cuerdas según el cual podemos despreciar las perturbaciones de $\mathcal{Y}$, al menos en la época cosmológica actual, depende de una expectativa de que la energía necesaria para excitar cualquiera de los modos de $\mathcal{Y}$ sería enorme, excepto para cierto conjunto particular de modos de energía cero (que tendrán importancia en §31.14) que ignoraré por el momento. ¿Por qué se espera que esta energía sea tan grande? El razonamiento se basa en la muy minúscula escala del propio $\mathcal{Y}$. Una «onda estacionaria» en $\mathcal{Y}$ tendría una longitud de onda minúscula, comparable con la distancia de Planck de $\sim 10^{-35}$ m, y por lo tanto tendría algo parecido a una frecuencia de Planck de $\sim 10^{-43}$ s. La energía requerida para excitar un modo semejante sería del orden de la energía de Planck, a saber, unos 10^{12} julios, ¡que es unos 20 órdenes de magnitud mayor que las mayores energías implicadas en las interacciones ordinarias entre partículas! Por consiguiente, se aduce que los modos que afectan a la geometría de $\mathcal{Y}$ no serán excitados en ninguno de los procesos de física de partículas que son de relevancia para las acciones físicas que hoy son accesibles. Se presenta la imagen de que en las etapas muy primitivas del universo seis de sus dimensiones se asentaron en la configuración $\mathcal{Y}$ descrita aproximadamente por una escala de Planck, mientras que las tres dimensiones espaciales restantes se expandieron enormemente para dar la imagen casi espacialmente plana de un universo 3-dimensional de acuerdo con la cosmología actual. Los $\mathcal{Y}$-espacios habrían

permanecido básicamente imperturbados desde un tiempo no muy posterior a los primeros instantes de Planck de la existencia del universo.

Examinemos este argumento con un poco más de detalle. Para simplificar las cosas, consideremos una situación en la que, como sucede con la teoría de Kaluza-Klein original de §31.4 y la analogía de la manguera mostrada en la Fig. 15.1, $\mathcal{Y}$ es simplemente un *círculo* S^1, que tomamos con un radio muy pequeño ρ. Podemos escoger una coordenada real θ para S^1 (con θ identificada con $\theta + 2\pi$), donde ρ^θ mide la distancia real alrededor del círculo. Los modos de $\mathcal{Y}$ son ahora simplemente las cantidades $e^{in\theta}$, donde n es un entero, a saber, los modos de Fourier que hemos visto en §9.2. En $\mathbb{E}^3$ podemos escoger coordenadas cartesianas ordinarias (x, y, z). Recordemos de §22.11 que una manera de abordar la cuestión de encontrar «modos» es buscar autoestados del (apropiado) operador laplaciano. En el contexto presente, esto puede considerarse como una aproximación (o simplemente un «modelo»). De manera más correcta, deberíamos interesarnos por los autoestados del hamiltoniano $\mathcal{H}$ para la evolución de la geometría. Para espacios Ricci-planos (nuestras requeridas perturbaciones de $\mathcal{M} \times S^1$) necesitaríamos la formulación hamiltoniana apropiada de la relatividad general pentadimensional, que es complicada. No obstante, el término dominante de esta es en esencia un laplaciano, y esto bastará para nuestra presente discusión.

Encontramos por primera vez el laplaciano 2-dimensional $\nabla^2 = \partial^2/\partial x^2 + \partial^2/\partial y^2$ en §10.5. Aquí necesitamos la generalización a cuatro dimensiones, pero la métrica de nuestro espacio $\mathbb{E}^3 \times S^1$ sigue siendo plana, de modo que no exigimos una expresión más elaborada como la de §22.11. Todo lo que necesitamos aquí es aumentar el número de variables a cuatro, de modo que podemos escribir el laplaciano como[31.4]

$$\nabla^2 = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2} + \frac{1}{\rho^2}\frac{\partial^2}{\partial \theta^2},$$

donde nuestra coordenada cuarta es la $\rho\theta$ necesaria para S^1. Para encontrar nuestros «modos», buscaríamos autoestados de este ∇^2. Más

[31.4] ¿Por qué?

concretamente, el procedimiento consiste meramente en ocuparse del análisis de modos para la parte S^1 de $\mathbb{E}^3 \times S^1$ y dejar la parte $\mathbb{E}^3$ como un campo ordinario. Por consiguiente, separamos nuestros campos en contribuciones diferentes, cada una de ellas con un entero n diferente, que da una dependencia en θ de la forma específica $e^{in\theta}$, como se ha descrito antes. Así pues, para un modo de S^1 de n-ésimo orden, podemos escribir

$$\Psi = e^{in\theta}\psi,$$

sobre nuestra 4-superficie inicial $\mathbb{E}^3 \times S^1$, donde ψ es una función de las coordenadas espaciales ordinarias x, y, z. Para cualquiera de estos modos de orden n-ésimo Ψ, el término $\rho^{-2}\partial^2/\partial\theta^2$ en nuestro laplaciano anterior puede reemplazarse[31.5] simplemente por $-n^2/\rho^2$:

$$\frac{1}{\rho^2}\frac{\partial^2}{\partial\theta^2} \mapsto -\frac{n}{\rho^2}.$$

Con respecto a las variables restantes x, y, z, nuestro laplaciano vuelve ahora al 3-espacio ordinario, pero tenemos el término constante $-n^2/\rho^2$ sumado a este laplaciano espacial.

Recordemos la ecuación de campo de una partícula ordinaria (sin espín) de masa μ, en el espaciotiempo de Minkowski ordinario $\mathbb{M}$, que es la ecuación de onda de «Klein-Gordon» (véase §24.5)

$$\left(\Box + \frac{\mu^2}{\hbar^2}\right)\psi = 0,$$

donde $\Box = \partial^2/\partial t^2 - \partial^2/\partial x^2 - \partial^2/\partial y^2 - \partial^2/\partial z^2$. Podemos considerarla como una «partícula incidente libre» (como sería apropiado en una aproximación de matriz S a la QFT; véase §26.8). Sin embargo, en el espacio $\mathbb{M} \times S^1$ tendríamos un término adicional $-\rho^{-2}\partial^2/\partial\theta^2$ en el operador de ondas $\Box$. Si tomamos esta partícula 5-espacial como un autoestado n-modo para S^1, este término queda reemplazado por n^2/ρ^2, como antes. Por consiguiente, desde el punto de vista del 4-espacio de Minkowski ordinario, nuestra partícula n-modo 5-espacial de Klein-Gordon satisface la ecuación 4-espacial

$$\left(\Box + \frac{\mu^2}{\hbar^2} + \frac{n^2}{\rho^2}\right)\psi = 0.$$

[31.5] ¿Por qué podemos hacerlo?

Esta es simplemente la ecuación de Klein-Gordon de nuevo, pero con $\mu^2/\hbar^2 + n^2/\rho^2$ en lugar de $\mu^2/\hbar^2$. Así pues, tenemos la ecuación de Klein-Gordon 4-espacial para una nueva partícula, pero donde la masa ha aumentado de μ a $\sqrt{(\mu^2 + \hbar^2 n^2/\rho^2)}$.

Ahora cualquiera de las partículas observadas en la naturaleza tendría una masa μ que es enormemente menor (véase §31.1) que el valor original de Planck de aproximadamente $\hbar/\rho$ (para nuestro valor escogido de ρ). Suponiendo que $n \neq 0$, esta nueva partícula tendría una masa que es al menos del orden de Planck (pues $\hbar n/\rho$ es mucho mayor que μ), de modo que estaría mucho más allá del alcance de los aceleradores de partículas viables hoy día. Por consiguiente, los teóricos de cuerdas argumentan que ningún modo $n \neq 0$ puede ser accesible en ningún proceso de física de partículas que sea viable en la época cosmológica actual.

En esencia, el mismo argumento se aplicaría al espacio compacto completo de tamaño de Planck $\mathcal{Y}$. En la situación de relativa baja energía en la que nos encontramos hoy, los modos de excitación de $\mathcal{Y}$ para los cuales $n \neq 0$ son experimentalmente inaccesibles para nosotros, o eso mantienen los teóricos de cuerdas. Por consiguiente, se argumenta que no hay conflicto entre las hipótesis de dimensiones espaciales extra y la física observacional actual.

31.11. ¿Deberíamos aceptar el argumento de la estabilidad cuántica?

Pero ¿es este razonamiento realmente apropiado? Creo que hay profundas razones para cuestionarlo.[38] Incluso si dejamos aparte el interrogante no respondido de por qué tres de las dimensiones espaciales deberían comportarse de forma tan diferente de las seis restantes, debemos ser muy cautos sobre este razonamiento de la «física de partículas» que se basa en que la geometría de $\mathcal{Y}$ está inmunizada contra el cambio durante la evolución posterior del universo.

No obstante, antes de entrar en la cuestión de la energía de Planck (o superior) de $\mathcal{V}$, debería volver a los modos de $\mathcal{V}$ de *energía cero* que he decidido ignorar en §31.10. En §31.14 veremos que estos modos

suelen considerarse favorablemente en el programa de la teoría de cuerdas, pues ofrecen esperanzas para establecer un contacto genuino con los grupos de simetría de la física de partículas estándar (§§25.5,7). Pese a todo, matemáticamente conducen a una seria dificultad que se ha hecho conocida como el *problema de los moduli*. Como sucede con la superficie de Riemann (véase §8.4), hay ciertos parámetros conocidos como *moduli* que definen la forma específica del tipo de espacio bajo consideración. (En §31.14 veremos que los $\mathcal{Y}$ preferidos son ciertas variedades complejas conocidas como espacios de «Calabi-Yau», cuyos moduli constituyen generalmente una familia de números complejos.) Los modos de energía cero remiten a la variación de dichos moduli. Podemos dejar que esta variación tenga una $\mathbb{E}^3$-dependencia espacial, pero esto nos da solo un aceptable $\infty^{N\infty^3}$, donde N se refiere al número real de moduli (reales) independientes. Sin embargo, sucede que hay modos en que los moduli se contraen rápidamente a cero dejando un $\mathcal{Y}$-espacio singular. Esta inestabilidad aparentemente catastrófica es en esencia el «problema de los moduli» de los teóricos de cuerdas (véase también §31.14).[39] Parece estar sin responder; pese a todo, normalmente es ignorado.

Supongamos que también decidimos ignorarlo. Entonces, ¿son los modos de vibración de energía positiva (a escala de Planck) de las seis dimensiones extra inmunes a la excitación? Aunque la energía de Planck es realmente muy grande comparada con las energías normales de la física de partículas, aún no lo es tanto, pues es comparable a la energía liberada en la explosión de aproximadamente una tonelada de TNT. Hay, por supuesto, una energía enormemente mayor que esta disponible en el universo conocido. Por ejemplo, la energía que recibe la Tierra desde el Sol en un segundo es 10 veces mayor. Solo en términos de energía, ¡eso sería mucho más que suficiente para excitar el espacio $\mathcal{Y}$ para el *universo entero*!

En el razonamiento de los teóricos de cuerdas, esta energía se libera en una interacción de partículas local, y tendemos a imaginar que está administrada en una región minúscula del espacio ordinario. Pero los modos reales de excitación de $\mathcal{Y}$ que se suponen inaccesibles están uniformemente dispersos sobre la totalidad de $\mathbb{E}^3$, en nuestras perturbaciones de $\mathbb{E}^3 \times \mathcal{Y}$. Recordemos que los modos de excitación de

$\mathbb{E}^3 \times \mathcal{Y}$ son simplemente productos de los modos de $\mathbb{E}^3$ por los modos de $\mathcal{Y}$. Los que estamos considerando aquí son solo constantes sobre $\mathbb{E}^3$. No hay nada que diga que necesiten (o incluso deberían) ser inyectados en una región localizada en el espacio físico ordinario.

Sin embargo, esto en sí mismo no es un argumento contra el hecho de que las interacciones de partículas locales sean la forma apropiada para excitar tales modos. El que estén dispersos sobre la totalidad de $\mathbb{E}^3$, no es un argumento contra una perspectiva de la física de partículas. Recordemos de la discusión de los operadores de creación y aniquilación en §26.2 y de las gráficas de Feynman en §§26.7,8 que las partículas y sus interacciones en QFT se describen normalmente en términos de estados de momento. Tales estados están «dispersos» sobre la totalidad de $\mathbb{E}^3$, como se ha señalado sobre todo en §21.11. Las «partículas cuánticas» no necesitan estar localizadas espacialmente. Quizá la mejor manera de considerar estas cosas es referirse a «cuantos» en lugar de partículas. La cuestión es si es o no razonable esperar que un solo cuanto de energía de Planck pudiera ser inyectado por cualquier medio en un modo $\mathcal{Y}$. Pero no creo que tengamos que pensar que tales «medios» sean necesariamente interacciones de partículas locales, y no otra cosa como una perturbación no lineal de la geometría espaciotemporal entera.

¿Existen razones para creer que debería haber algún otro medio? En mi opinión, hay razones para preocuparse por esto. Volvamos a nuestra analogía de la manguera (Fig. 31.3). Consideremos que la manguera es esencialmente recta en su dimensión «grande» (análoga a $\mathbb{E}^3$), y tiene una sección transversal S^1 constante (análoga a $\mathcal{Y}$) que es un círculo de radio minúsculo ρ. Los modos de excitación de la manguera pueden estar compuestos de varias ondas que viajan en una u otra dirección a lo largo de su longitud («modos $\mathbb{E}^3$») y de varias distorsiones de su sección transversal de forma circular («modos $\mathcal{Y}$»). Como hemos visto, cualquiera de estos últimos modos ocurre simultáneamente a lo largo de toda la manguera. En mecánica cuántica, la energía en un único cuanto de excitación de un modo semejante —un *excitón*— de frecuencia vibracional ν es $2\pi\hbar\nu$ (véase §21.4) e independiente de la longitud de la manguera.

Para una manguera de longitud infinita, esto da una densidad de

energía nula, para cada excitón individual, de modo que quizá sea menos confuso si imaginamos que la manguera está curvada en un círculo muy grande, de radio R, digamos, con $R \gg \rho$. Ahora consideremos un modo de vibración particular de $\mathcal{Y}$, con una frecuencia particular ν. La energía total $2\pi\hbar\nu$ en este excitón es independiente de R. Esto puede parecer enigmático, puesto que implica que cuanto mayor tomemos R, menor es la energía que existe localmente en la vibración, en proporción a $1/R$. Pero no es una inconsistencia, sino que nos dice que la amplitud de la vibración en un excitón, para un modo de vibración determinado de $\mathcal{Y}$, es menor cuanto mayor es la longitud de la manguera. Si tomamos el límite $R \to \infty$, la energía almacenada localmente en el modo va a cero. De esto aprendemos que, cualquiera que sea la manera concreta en que pueda vibrar localmente la manguera, en el límite en que la manguera se hace infinita debe incluir números cada vez más grandes de cuantos, y el efecto de cada cuanto individual se hace cada vez menor, de modo que nos preguntamos si no sería apropiada una descripción clásica en lugar de una descripción cuántica del comportamiento de la manguera.[40]

Esto plantea la cuestión del *límite clásico* de un sistema cuántico para grandes números cuánticos, y el tema relacionado de la reducción de estado **R** para semejante configuración clásica. Hemos visto, en particular de la discusión en el capítulo 29, que la cuestión **R** no puede resolverse por completo dentro del marco de la teoría cuántica actual.[41] De todas formas, un buen físico debería saber cuándo una descripción cuántica es apropiada y cuándo tiene más sentido físico utilizar una clásica. Recordemos el caso del momento angular ordinario, como se ha discutido en §22.10. Un cuerpo con un momento angular muy grande suele ser considerado como un sistema clásico, de modo que obtemos un eje de rotación muy bien definido. Tratándolo como un sistema cuántico con un valor muy grande de j, obtenemos una descripción de Majorana, con muchas direcciones de espín, que normalmente apuntan en todas direcciones. En la práctica, la descripción clásica sería la utilizada para un momento angular muy grande, y esto proporciona una buena imagen de la realidad física. De manera más general, las descripciones clásicas suelen considerarse físicamente apropiadas cuando los números cuánticos se hacen extraordinariamente

grandes. En el caso del momento angular, el número cuántico relevante, a saber, j, se mide en términos de unidades de $\hbar$, de modo que se podría imaginar un criterio razonable que nos diga cuándo estamos lejos del régimen cuántico: el valor de j es muy grande en unidades de $\hbar$. En el caso de la manguera, vemos que la pequeñez de la distancia ρ no es por sí misma una medida apropiada para decirnos que una descripción «cuántica» es más adecuada que una clásica. Para ρ fijo, la descripción de las vibraciones locales de la manguera parece hacerse cada vez más «clásica» cuanto mayor tomamos R, puesto que necesitamos incluir números cada vez mayores de excitones que incluyen números cuánticos vibracionales (modos de $\mathcal{Y}$) cada vez mayores.[42]

En ausencia de una teoría que nos diga cómo los sistemas «grandes» llegan a describirse bien clásicamente, mientras que los «pequeños» se comportan de acuerdo con las reglas cuánticas (lo que en mi opinión requiere un cambio en la estructura misma de la mecánica cuántica según las líneas apuntadas en §§30.9-12), podría parecer que no podemos llegar a ninguna conclusión definitiva concerniente a la supuesta inaccesibilidad de las excitaciones de $\mathcal{Y}$. (Las consideraciones del capítulo 30 no parecen proporcionarnos una respuesta inequívoca, y desde luego no una indiscutible.) De todas formas, en vista del hecho de que perturbaciones reales de $\mathcal{Y}$ nos llevan a una imagen cuántica de números muy grandes de cuantos, donde cada cuanto individual apenas afecta a la geometría de $\mathcal{Y}$, y a grandes números cuánticos, parecería que podemos sacar más ideas sobre cómo se comportan las perturbaciones de un universo $\mathbb{M} \times \mathcal{Y}$ con $\mathcal{Y}$ «pequeño» si las estudiamos *clásicamente* en lugar de mecanocuánticamente. Consideramos esto a continuación.

31.12. Inestabilidad clásica de las dimensiones extra

¿Qué podemos decir si tomamos el modelo 10-espacial como enteramente clásico? Esto debería darnos al menos alguna guía hacia el comportamiento real del modelo cuántico completo. Hemos visto al principio de §31.10 que en un (1 + 9)-espaciotiempo clásico (i.e., de una dimensión temporal y nueve dimensiones espaciales) habría un diluvio

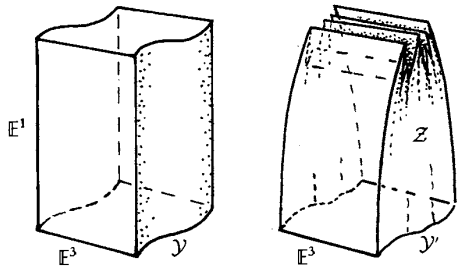


Fig. 31.9. Un teorema de singularidad se aplica a perturbaciones de $\mathbb{M} \times \mathcal{Y}$, donde $\mathcal{Y}$ es un espacio de Calabi-Yau «pequeño». (a) El caso canónico no singular $\mathbb{M} \times \mathcal{Y}$, donde expresamos $\mathbb{M} = \mathbb{E}^1 \times \mathbb{E}^3$, con $\mathbb{E}^1$ que se refiere al tiempo. (b) Una perturbación general $\mathcal{Y}'$ de $\mathcal{Y}$ evoluciona hasta un espacio $\mathcal{Z}$ que es singular, de modo que perturbaciones generales de $\mathbb{E}^3 \times \mathcal{Y}$ que no afectan a $\mathbb{E}^3$ evolucionan hasta espacios $\mathbb{E}^3 \times \mathcal{Z}$ que son singulares.

inaceptable de excesivos grados de libertad ($\infty^{M\infty^{\gamma}} \gg \infty^{M\infty^{3}}$). Esto es bastante grave pero, en mi opinión, las cosas son realmente mucho peores. Encontraremos que un universo clásico $\mathbb{M} \times \mathcal{Y}$ —sujeto a Ricci-planitud— es muy inestable frente a pequeñas perturbaciones. Si $\mathcal{Y}$ es compacto y de un tamaño de Planck, entonces hay que esperar que se den singularidades espaciotemporales (§27.9) en una minúscula fracción de segundo.

Consideremos, en primer lugar, perturbaciones de $\mathbb{M} \times \mathcal{Y}$ que perturban solo la geometría $\mathcal{Y}$ y que, por consiguiente, no se «filtran» al $\mathbb{E}^3$ espacial. Es decir, examinemos un (1 + 6)-espaciotiempo «genérico» Ricci-plano $\mathcal{Z}$ (la evolución perturbada de $\mathcal{Y}$), siendo $\mathcal{Z} \times \mathbb{E}^3$ el (1 + 9)-espaciotiempo entero. Consideremos que $\mathcal{Z}$ es la evolución temporal de un 6-espacio que (en un instante concreto) es «próximo» a $\mathcal{Y}$, de modo que $\mathcal{Z}$ empieza próximo a la (inalterada) «evolución temporal» de $\mathbb{E}^1 \times \mathcal{Y}$ de $\mathcal{Y}$, aunque $\mathcal{Z}$ puede desviarse fuertemente de $\mathbb{E}^1 \times \mathcal{Y}$ en instantes posteriores (Fig. 31.9). Aquí estoy expresando $\mathbb{M}$ como $\mathbb{M} = \mathbb{E}^1 \times \mathbb{E}^3$, como en §31.10 más arriba (donde $\mathbb{E}^1$ describe la dimensión temporal y $\mathbb{E}^3$ las dimensiones espaciales), de modo que consideramos el espaciotiempo $\mathbb{M} \times \mathcal{Y}$ como $(\mathbb{E}^1 \times \mathcal{Y}) \times \mathbb{E}^3$ (colocando la evolución temporal de $\mathcal{Y}$ primero en el producto); véase la Fig. 31.9a.

Ahora bien, a finales de de la década de 1960, Stephen Hawking y yo demostramos un teorema de singularidad que muestra que debemos esperar que $\mathcal{Z}$ sea singular.[43] Tal como establecimos explícitamente, este teorema se aplica tanto a (1 + 6)-espaciotiempos (y a (1 + 9)-espaciotiempos) como a los (1 + 3)-espaciotiempos convencionales que consideramos originalmente. Una de las consecuencias de este teorema es que cualquier espaciotiempo Ricci-plano que (como $\mathbb{E}^1 \times \mathcal{Y}$ o $\mathcal{Z}$) contiene una hipersuperficie compacta de género espacio, que es «genérica» en cierto sentido específico[44] (y libre de curvas cerradas de género tiempo; véanse §17.9 y la Fig. 17.18), debe ser singular. El $\mathbb{E}^1 \times \mathcal{Y}$ original se libra de ser singular porque la condición genérica falla en este caso. Pero el $\mathcal{Z}$ genéricamente perturbado tiene que ser singular.

Debería advertirse que en tales «teoremas de singularidad» no se establece directamente que la curvatura diverge a un valor infinito, sino simplemente que hay un impedimento de algún tipo para que geodésicas de género tiempo o nulas sean extensibles dentro del espaciotiempo a longitud infinita (o a longitud afín infinita, en el caso de geodésicas nulas; véase §14.5). La previsión normal sería que este impedimento se debe a la presencia de curvatura divergente, pero el teorema no muestra esto directamente. No obstante, este teorema nos dice que $\mathcal{Z}$ se hará singular de esta o de otra manera. Si la perturbación de $\mathcal{Y}$ es de la misma escala general que la propia $\mathcal{Y}$ (i.e., la escala de Planck), entonces debemos esperar que las singularidades en $\mathcal{Z}$ ocurran en una escala de tiempo comparable ($\sim 10^{-43}$ s), pero esta escala de tiempo podría hacerse algo mayor si las perturbaciones son de un tamaño proporcionalmente más pequeño que la propia $\mathcal{Y}$.

Concluimos que si queremos tener una oportunidad de perturbar $\mathcal{Y}$ de una manera genérica de modo que obtengamos una perturbación no singular del (1 + 9)-espacio completo $\mathbb{M} \times \mathcal{Y}$, entonces debemos considerar perturbaciones que se desbordan significativamente hacia la parte $\mathbb{M}$ del espaciotiempo. Pero en ciertos aspectos tales perturbaciones son incluso más peligrosas para nuestra imagen «ordinaria» del espaciotiempo que las que afectan solo a $\mathcal{Y}$, puesto que las grandes curvaturas a escala de Planck[45] que es probable que estén presentes en $\mathcal{Y}$ se desbordarán hacia el espacio ordinario, en grave conflicto con la ob-

servación, y darán como resultado singularidades espaciotemporales en un tiempo muy corto.[46]

Por supuesto, las singularidades inaceptables en una teoría clásica no nos dicen necesariamente que tales manchas persistirán en la versión cuántica apropiada de esa teoría. Como ya hemos visto en §22.13, la mecánica cuántica remedia la inestabilidad catastrófica de los átomos clásicos ordinarios, por la que los electrones caerían en espiral al núcleo con la emisión de radiación electromagnética. Sin embargo, la mera introducción de «procedimientos de cuantización» no asegurarán necesariamente que se eliminen las singularidades clásicas. Hay muchos ejemplos (tales como en la mayoría de los modelos de juguete de la gravedad cuántica)[47] donde las singularidades persisten después de la cuantización.

Deberíamos también tomar nota del hecho —véase §31.8— de que la Ricci-planitud (1 + 9)-dimensional no es precisamente el requisito que exige la teoría de cuerdas. Recordemos que la Ricci-planitud se considera una excelente aproximación a dicho requisito, que se da cuando se ignoran términos del orden superior al orden más bajo en la constante de cuerda α'. Quizá el requisito «exacto», que incluye todos los órdenes en la constante de cuerda α', podría evitar el teorema de singularidad anterior. Sin embargo, si este requisito nos proporciona una condición sobre el tensor de Ricci para la que se satisfacen las demandas usuales de positividad de energía local (véanse especialmente la nota 27.9 y §28.5), entonces el teorema de singularidad seguiría aplicándose. Por otra parte, violaciones de tales condiciones de energía local pueden darse ciertamente en QFT (§24.3), de modo que estas cuestiones están lejos de ser concluyentes.

Más grave, en mi opinión, es el hecho de que el requisito pleno, que incluye todos los órdenes en la constante de cuerda α', es en realidad un sistema infinito de ecuaciones diferenciales de orden ilimitado. Por consiguiente, los datos que serían necesarios en una 9-superficie inicial incluirían derivadas de todos los órdenes en las cantidades del campo (y no solo la primera o segunda derivadas que son necesarias en las teorías de campos ordinarias). El número de parámetros por punto en la 9-superficie es entonces infinito, de modo que obtenemos una libertad funcional *mayor que* $\infty^{M\infty^9}$, para cualquier entero positivo

M. Esto parecería empeorar el problema de la libertad funcional excesiva aún más que antes. No conozco ninguna discusión seria de la forma matemática de este requisito completo, ni de qué tipo de datos iniciales podría ser apropiado para ello.

31.13. ¿Es finita la QFT de cuerdas?

El tipo de argumento que he presentado antes ilustra por qué tengo graves dificultades para convencerme de que los modelos de teoría de cuerdas puedan reproducir la relatividad general (1 + 3)-dimensional de Einstein en cualquier tipo de «límite clásico» razonable. ¿Qué pasa con la otra parte de la afirmación de la teoría de cuerdas, según la cual es una QFT *consistente y finita* (cualquier cosa que signifique, físicamente, la teoría resultante)? Creo que el hecho de que se obtiene una amplitud finita para una determinada topología de hoja de universo de cuerda podría ser la parte más contundente del argumento pro cuerdas, pues esta conclusión parece proporcionar un verdadero reflejo de las virtudes originales de la idea de cuerdas. Pese a todo, incluso aquí hay preguntas fundamentales a las que hay que dar respuesta.

Para empezar, siempre he sentido cierta preocupación con respecto a la teoría de cuerdas. Se presenta como una teoría física de estructuras tipo cuerda cuyas hojas de universo son de género tiempo, y cuya métrica inducida es, por consiguiente, una métrica lorentziana (1 + 1)-dimensional. Pese a todo, las matemáticas se hacen con hojas de universo de cuerdas que poseen una métrica (definida positiva), de modo que puede apelarse a las elegantes ideas de la teoría de superficies de Riemann (como en el capítulo 8); véase la Fig. 31.10. De acuerdo con lo primero, se habla de modos de perturbación que viajan a lo largo de la hoja de universo de género tiempo, ya sean levógiros o dextrógiros, con la velocidad de la luz. (Estos se propagan a lo largo de curvas nulas en la hoja de universo lorentziana.) En la versión riemanniana de la teoría, estos modos «levógiro» o «dextrógiro» se convierten en funciones «holomorfa» y «antiholomorfa» sobre la superficie de Riemann. Parece que existe la tendencia de que los cálculos se hagan realmente con la imagen de superficie de Riemann definida positiva, y luego se reali-

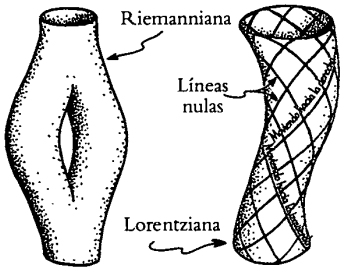


Fig. 31.10. Las matemáticas de la teoría de cuerdas utilizan «historias de cuerdas» que son superficies de Riemann, que tienen métricas riemannianas (definidas positivas). Pero, físicamente, las historias de cuerdas son lorentzianas. Pasar de una a la otra implica un tipo de «rotación de Wick».

ce una «rotación de Wick» (véase §28.9) para obtener finalmente la deseada teoría de cuerdas lorentziana. Es muy posible que este proceso sea satisfactorio aquí, pero no se puede suponer sin una justificación específica. Depende críticamente, por ejemplo, de que no se hagan aproximaciones en el cálculo de las amplitudes. De otra forma, podrían haber planteado serios interrogantes del tipo que hemos encontrado antes en relación con la aproximación de Hawking para la gravedad cuántica y otras aproximaciones a la QFT que implican prolongación analítica (véase §28.9). Entiendo que los cálculos para una superficie de Riemann pretenden ser exactos, de modo que hay razones para fiarse de la rotación de Wick. De todas formas, la justificación explícita para una rotación de Wick depende de que el espaciotiempo de fondo sea plano, lo que no sería el caso si vamos a hacer relatividad general seria (no perturbativa), de modo que sigue estando poco claro hasta dónde nos lleva esto en la dirección de una teoría cuántica de la gravedad real.

Incluso si confiamos en la validez de tales consideraciones de espacio plano, ¿debemos aceptar las afirmaciones de que para cada topología de superficie de Riemann determinada (i.e., género g determinado; véase §8.4, donde g corresponde al «número de lazos» para una gráfica de Feynman ordinaria —véanse §26.8, Fig. 31.5d y §31.5— la amplitud total es finita? Esto no se ha establecido de hecho. A pesar de repetidas promesas, todavía no se ha demostrado matemáticamente esta finitud. Las afirmaciones de finitud se refieren solo a las divergencias ultravioletas (momento grande, distancia pequeña) que los propios teóricos de campos cuánticos encuentran más problemáticas, pero incluso estas solo han sido establecidas hasta ahora en el nivel de 2 lazos. Además, parece que no hay ningún argumento que afirme que se eliminan

las divergencias infrarrojas (momento pequeño, distancia grande). Aunque tales divergencias se suelen considerar menos graves que las ultravioletas, ciertamente no pueden ignorarse y hay que tratarlas de alguna manera si se quiere justificar la afirmación de «finitud». Esto nos deja cierta incertidumbre con respecto al programa entero, y la finitud es la clave de la idea de cuerda.[48]

Quizá estos sean solo tecnicismos irritantes que se superarán en futuros desarrollos matemáticos. Sin embargo, incluso si se acepta que tenemos amplitudes finitas para cada topología fija, estamos lejos de haber acabado. Las expresiones tienen que *sumarse*. Ahora existe el problema de que aparentemente esta suma realmente diverge.[49] ¡La pretendida teoría *finita* no es realmente finita después de todo! Sin embargo, esta divergencia concreta parece no preocupar a los teóricos de cuerdas, pues consideran que esta *serie* es una realización poco adecuada de la amplitud total. Dicha amplitud sería cierta cantidad analítica, pero la serie de potencias trata de encontrar una expresión para ella «desarrollando en torno al punto equivocado», i.e., en torno a un punto que es singular para la amplitud (algo parecido a tratar de encontrar una serie de potencias para $\log z$, desarrollada en torno a $z = 0$, en lugar de desarrollar en términos de potencias de $z - 1$ —véase §7.4—, aunque en ese caso concreto esta serie tendría realmente coeficientes infinitos). Esto podría ser correcto, aunque se ha demostrado que la divergencia encontrada aquí es de un tipo bastante incontrolable («no Borel sumable»). Para dar sentido al tipo de razonamiento euleriano requerido (tal como el que da $1 + 2^2 + 2^4 + 2^6 + 2^8 + \ldots = -\frac{1}{3}$; véanse §4.3 y §26.9), parecen necesarios procedimientos más sofisticados.[50] Además, si los cálculos (perturbativos) con teoría de cuerdas son realmente desarrollos «en torno al punto equivocado», entonces no está claro hasta qué punto podemos confiar en tales cálculos perturbativos. Así pues, no sabemos todavía si la QFT de cuerdas es finita o no, por no hablar de si la teoría de cuerdas, con todos sus indudables atractivos, nos proporciona *realmente* una teoría cuántica de la gravedad.

31.14. Los mágicos espacios de Calabi-Yau; la teoría M

Las reservas particulares que he mantenido acerca de la teoría de cuerdas, expresadas en §§31.8-13, no son, sin embargo, las que parecen haber preocupado a los propios teóricos de cuerdas. Se han preocupado por otras cuestiones a las que ni siquiera me he referido todavía, a saber, la cuestión de la unicidad de la teoría. Originalmente se había considerado que una de las mayores esperanzas/triunfos de la teoría de cuerdas era que podría proporcionar un esquema único para el universo, y había mucho encerrado en esta supuesta unicidad. Una cuestión evidente tenía que ver con las variedades 6-dimensionales compactas a escala de Planck $\mathcal{Y}$, en las que se suponía que estaba fundamentalmente enrollado el universo 10-dimensional. ¿Qué *son* estas 6-variedades? ¿Por qué el universo tiende a enrollarse en estas en lugar de otras? Al principio, había parecido que los requisitos restrictivos de supersimetría, dimensionalidad y Ricci-planitud podrían llevar a respuestas únicas, pero luego resultó que, en realidad, era posible un gran número de alternativas.

Unas primeras sugerencias eran que el espacio $\mathcal{Y}$ podría ser un *hipertoro* $S^1 \times S^1 \times S^1 \times S^1 \times S^1 \times S^1$ con curvatura nula (véase la nota 31.45; recordemos el término «toro» para $S^1 \times S^1$, véanse las Figs. 8.9, 8.11 y 15.3). Pero entonces se vio que una teoría de cuerdas basada en un hipertoro no podía incorporar los aspectos quirales[51] del modelo estándar, y se necesitaba algo más sofisticado. Los «requisitos estrictos» conducían entonces a que estas 6-variedades eran lo que se denominaba *espacios de Calabi-Yau*.[52] Estos son espacios de considerable interés en matemática pura, y por esa razón habían sido estudiados previamente por Eugenio Calabi y Shing-Tung Yau. Son ejemplos de lo que se denominan variedades de *Kähler*, lo que significa que tienen métricas riemannianas y estructuras complejas (y, por consiguiente, pueden interpretarse como 3-variedades *complejas*), donde estas dos estructuras son compatibles (en el sentido de que la conexión métrica conserva la estructura compleja, de lo que se sigue que también son variedades simplécticas;[31.6] véanse §12.9 y §§14.7,8 para los conceptos relevantes).

[31.6] ¿Puede ver por qué deben ser, en consecuencia, variedades simplécticas? *Sugerencia*: Construya S_{ab} a partir de la métrica g_{ab} y la estructura compleja J_b^a, y verifique entonces que $dS = 0$ (suponga que S_{ab} no es singular).

Los espacios de Calabi-Yau tienen propiedades adicionales que se estiman esenciales para el programa de las cuerdas: poseen métricas que son Ricci-planas y están dotados de *campos espinoriales* que son constantes con respecto a la conexión métrica. Estos campos espinoriales constantes desempeñan necesariamente papeles como generadores de supersimetría. Sin ellos, la supersimetría no sería posible. Los diversos campos espinoriales, para una elección dada de espacio de Calabi-Yau, pueden ser (formalmente) «rotados unos en otros» por la acción de un grupo de simetría. Este grupo va a desempeñar entonces el papel que desempeñan los grupos de simetría de la física de partículas.

Debería quedar claro que esta simetría no se aplica directamente a los propios espacios de Calabi-Yau, a la manera de una simetría aplicada a la fibra $\mathcal{F}$ de un fibrado, descrito en la discusión del capítulo 15. De hecho, los espacios de Calabi-Yau no poseen simetrías (continuas), y no podemos considerar nuestro 10-espacio como un fibrado (no trivial) de 6-espacios de Calabi-Yau sobre el 4-espaciotiempo ordinario. Las simetrías (internas) de la partícula se refieren, en su lugar, a la «rotación» de los campos espinoriales constantes entre sí mismos. Los espacios de Calabi-Yau no se ven afectados por la acción de la simetría.[53]

Así pues, la teoría de cuerdas conduce a un tipo muy concreto aunque inusual de teoría GUT (véanse §25.8 y §§28.1-3). Se pretende que toda la física de partículas se encuentre dentro del esquema de la teoría de cuerdas apropiado. Los grupos de simetría que surgen de este modo son mucho mayores que los del modelo estándar (§§25.5-7), pero, como sucede con otras teorías GUT (§25.8), se considera que una forma de ruptura de simetría es responsable de reducir los grupos a aquellos de relevancia más directa para el modelo estándar, aunque todavía no se ha alcanzado este programa de forma satisfactoria.

¿Qué pasa con la cuestión de la unicidad? Por desgracia, hay dece-

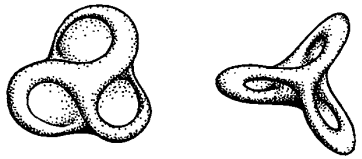


Fig. 31.11. La «forma» de un espacio de Calabi-Yau $\mathcal{Y}$ está descrita por un número de moduli (compárese con las Figs. 8.10 y 8.11). Las variaciones de estos moduli proporcionan modos de oscilación de $\mathcal{Y}$ de energía cero.

nas de miles de clases de alternativas posibles cualitativamente diferentes para los espacios de Calabi-Yau, de modo que el esquema, tal como se ha descrito, no es ni mucho menos único. De hecho, dentro de una clase particular de espacio de Calabi-Yau hay infinitos diferentes, que se distinguen por el valor de ciertos parámetros, llamados *moduli* (véase §31.11) que describen su forma (Fig. 31.11), igual que en el caso de las superficies de Riemann (§8.4, Fig. 8.11). La presencia de estos moduli se considera algo bueno, porque su variación proporciona los modos de oscilación de energía cero del espacio $\mathcal{Y}$ (mencionados en §31.11), que se toman como físicamente realizables y proporcionarían la ruta necesaria hacia la física de partículas y las consecuencias observacionales de la teoría de cuerdas.

Sin embargo, existen otros tipos de no unicidad que dan inicialmente la impresión de ser más graves incluso que la no unicidad de Calabi-Yau. Resulta que hay 5 posibles esquemas globales completamente diferentes para la forma detallada en la que la supersimetría interrelaciona los modos «bosónico» y «fermiónico» de vibración de la cuerda. Así pues, hay 5 teorías de cuerdas diferentes que se conocen como tipo I, tipo IIA, tipo IIB, heterótica O(32) y heterótica $E_8 \times E_8$. Los grupos O(32) y $E_8 \times E_8$ son aquellos que aparecerían de la forma esbozada en el párrafo anterior. (El lector puede reconocer la notación para estos grupos de §13.2, siendo E_8 el mayor de los grupos de Lie *excepcionales*.) Las teorías de tipo I emplean cuerdas abiertas tanto como lazos cerrados, mientras que todas las demás operan solo con lazos cerrados. En todos los modelos, las perturbaciones pueden viajar hacia la derecha o hacia la izquierda.[54] El tipo IIA y el tipo IIB difieren en cómo se relacionan mutuamente estas perturbaciones hacia la derecha o hacia la izquierda. Las cuerdas heteróticas son particularmente extrañas en cuanto que las perturbaciones que se mueven a izquierda y derecha parecen pertenecer a dos espaciotiempos de diferentes dimensiones (26 y 10, respectivamente). Es difícil que esto tenga un buen sentido geométrico —¡desde luego no para mí!—, aunque parece tener el sentido formal apropiado. Al parecer, la idea consiste en que la imagen 10-dimensional es la apropiada, pero las perturbaciones que se mueven hacia la izquierda se comportan de la misma manera que las de las más antiguas «cuerdas bosónicas» (no supersimétricas) de §§31.5,7

que iban a habitar en un espaciotiempo ambiente 26-dimensional. Como hemos visto antes, los teóricos de cuerdas no parecen estar muy preocupados por aparentes inconsistencias en dimensión espaciotemporal, al considerar esta dimensionalidad como un efecto «dependiente de la energía» (§31.10), y por consiguiente, no de importancia fundamental. Veremos más sobre esto enseguida en §§31.15,16.

Durante un tiempo, esta proliferación de diferentes modelos de cuerdas hizo que muchos teóricos desesperaran de ser capaces de avanzar mucho más. Pero empezaron a producirse algunos desarrollos notables, que indicaban ciertas profundas interrelaciones posibles entre estos modelos aparentemente muy diferentes. En 1995, Edward Witten pronunció una famosa conferencia[55] que inició lo que se conoce como «la segunda revolución de las supercuerdas». En esa conferencia, Witten esbozó un programa para el desarrollo de la teoría de cuerdas que ha transformado por completo la forma de ver el tema. La nueva característica esencial es que, invocando ciertos tipos de misteriosas «operaciones de simetría» (que se conocen como «dualidad fuerte-débil» o «simetría espejo»;[56] también llamadas a veces dualidades S o S-, T-, y U-dualidades), se ha puesto de manifiesto que estas teorías de cuerdas diferentes tienen tan profundas relaciones mutuas que aparentemente pueden considerarse teorías de cuerdas *equivalentes*. El límite a pequeña escala de algunas de estas teorías parece ser idéntico (en cierto sentido adecuado) al límite a gran escala de otras, y hay otros tipos de relaciones de simetría de este tipo general que surgen de una «dualidad» de la teoría de Yang-Mills (§25.7) que es análoga a la dualidad entre electricidad y magnetismo en la teoría electromagnética ordinaria. (Compárese también la «dualidad» que surge en la teoría de Chan-Tsou, como se ha descrito brevemente en §25.8.) Además, diferentes espacios de Calabi-Yau resultan ser «duales» uno de otro en varios aspectos. De momento, por lo que conozco del tema, no todas estas relaciones están demostradas como resultados matemáticos.[57] Pero las conjeturas originales, que proceden de la teoría de cuerdas con alguna impresionante evidencia circunstancial en su favor, han impulsado una investigación muy considerable en matemática pura que ha llevado a conocimientos más profundos de las variedades de Calabi-Yau y las relaciones entre ellas.[58]

Vale la pena señalar un ejemplo particularmente sorprendente de esta «evidencia circunstancial». Está relacionado con un problema matemático muy concreto en el que ciertos matemáticos puros (los geómetras algebraicos) se habían interesado con anterioridad durante varios años. Este problema no tenía evidentemente nada que ver con la física, sino con contar el número de curvas *racionales* en ciertas 3-variedades complejas.[59] Una curva racional es una curva compleja (i.e., una superficie de Riemann; véase el capítulo 8) de *género cero*, es decir, que tiene la topología de una esfera S^2. Estas 3-variedades complejas resultan ser espacios de Calabi-Yau que por las demandas de la teoría de cuerdas —según las propuestas de la «segunda revolución de supercuerdas»— deberían estar relacionados, vía simetría espejo, con otros espacios de Calabi-Yau. La simetría espejo, en cierto sentido, intercambia estructura compleja con estructura simpléctica; por consiguiente, el problema de contar curvas racionales holomorfas (que es técnicamente un problema muy difícil) se transforma en un problema de recuento mucho más fácil, y de apariencia completamente diferente, en el espacio de Calabi-Yau «espejo». Dos matemáticos noruegos, Geir Ellingstrud y Stein Arilde Strømme, habían desarrollado métodos para contar directamente el número de curvas racionales (de órdenes sucesivos[60] 1, 2, 3, ...) en sus espacios, llegando a los números sucesivos

$$2875,\quad 609\ 250,\quad 2\ 682\ 549\ 425,$$

para los tres primeros casos. Utilizando la hipótesis de que es válida la relación de simetría espejo, de modo que puede aplicarse un método de recuento mucho más simple, Philip Candelas y sus colaboradores pudieron llegar a los números

$$2875,\quad 609\ 250,\quad 317\ 206\ 375.$$

Puesto que la simetría espejo era en esa época solamente una «conjetura de los físicos» no demostrada, se supuso que el acuerdo en los dos primeros casos era accidental y que no había ninguna razón para aceptar el número 317 206 375 al que habían llegado Candelas y sus colaboradores. Pero entonces sucedió que, debido a un error en el programa del ordenador, el número de los matemáticos noruegos era incorrecto y el valor correcto ¡era exactamente el que se obtenía me-

diante el argumento de la simetría espejo! Luego se calcularon muchos números siguientes utilizando esta simetría espejo, tales como la extensión de la secuencia anterior al recuento de curvas racionales de órdenes más altos (4, 5, 6, ..., 10):

242 467 530 000
229 305 888 887 625
248 249 742 118 022 000
295 091 050 570 845 659 250
375 632 160 937 476 603 550 000
503 840 510 416 985 243 645 106 250
704 288 164 978 454 686 113 488 249 750

Este es un notable ejemplo que indica sin duda que hay «algo entre bastidores». Tal como están las cosas, esto es muy enigmático. Ciertamente hay algo que está lejos de ser obvio y que hay que descubrir en las matemáticas, y algunos desarrollos matemáticos recientes han avanzado algo para conseguirlo.[61] Pero la cuestión más importante tiene que ver con el significado físico de estos resultados. ¿Estamos autorizados a inferir del hecho indudable de que la teoría de cuerdas ha proporcionado intuiciones profundas y previamente inesperadas en *matemáticas* que también debe ser correcta en *física*? La respuesta a este enigma está lejos de ser obvia. Witten ha argumentado que la teoría de

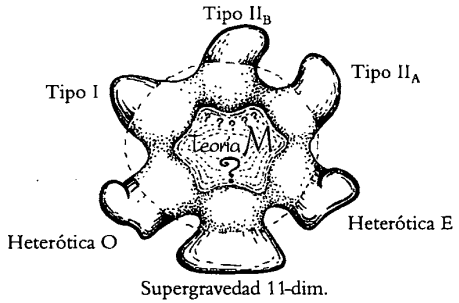


Fig. 31.12. La enigmática «teoría M» afirma que los cinco diferentes tipos de teoría de cuerdas, todos ellos relacionados mediante S-, T- y U-dualidades, y la supergravedad 11-dimensional, son seis aspectos diferentes de una y la misma estructura todavía no descubierta.

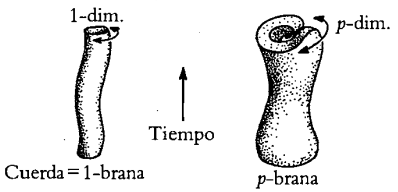


Fig. 31.13. Las membranas (o p-branas, o simplemente branas), tienen p dimensiones espaciales y 1 dimensión temporal, siendo $(1 + p)$-dimensional la hoja de universo. Estas estructuras están incluidas, junto con las cuerdas ordinarias (1-branas), como parte de la teoría M no definida.

cuerdas, tal como había sido entendida hasta ese momento, era solo la punta de un iceberg, o, mejor, que representaba cinco puntas de alguna misteriosa y todavía desconocida teoría que él bautizó como «teoría M»; véase la Fig. 31.12. Esta nueva teoría, cuando se encuentre, sustituirá y superará las diferentes teorías de cuerdas que se hayan propuesto previamente.

No solo se pretende que la misteriosa teoría M englobe todas estas teorías de cuerdas, sino que también va a incorporar otras muchas ideas relacionadas con las cuerdas y con la supersimetría. Las cuerdas se consideran ahora tan solo un caso especial de una noción más general que incluye estructuras de dimensiones mayores (e incluso menores). Estas se conocen como *membranas* (o p-branas, o solo branas) que tienen p dimensiones espaciales y una dimensión temporal, siendo la hoja de universo $(p + 1)$-dimensional. Véase la Fig. 31.13. También pueden incluirse unas estructuras de género tiempo relacionadas llamadas D-branas; diré algo sobre estas en §31.17.

En otro desarrollo, se supone que la teoría M engloba también la teoría de la supergravedad 10-dimensional que hemos dejado atrás en este capítulo (§31.4). De hecho, parece que la propia teoría M se considera aproximadamente como una teoría 11-dimensional, de modo que quizá el misterio dimensional reside más en su relación con las diversas teorías de cuerdas 10-dimensionales que con la supergravedad 11-dimensional. El hecho de que ahora parezcan estar «permitidas» 11 dimensiones para una teoría consistente de tipo cuerda parece ser una conclusión que se debe a Witten; según esta, el argumento original de que se necesitan «una más nueve» dimensiones para eliminar la anomalía de cuerdas mencionada en §31.7 debe considerarse realmente como una aproximación (debido, en parte, a la inclusión de estas «bra-

nas» de dimensiones más altas), y la respuesta más correcta es en realidad 11 (= 1 + 10, i.e., 1 dimensión temporal y 10 dimensiones espaciales).[62] Pero ni siquiera 11 dimensiones satisfacen quizá a los teóricos de cuerdas. Se ha hecho alguna sugerencia de que habría que pasar a dimensiones aún más altas, a la incluso más misteriosa (e incluso más desconocida) teoría F que tiene 12(= 10 + 2) dimensiones (¡de modo que hay 2 dimensiones temporales!)[63]

¿Cómo es posible que una teoría con un «espaciotiempo» 11-dimensional (o quizá 12-dimensional) pueda ser algo que se traduce, en ciertos límites de alta energía o baja energía, en varias teorías cada una de las cuales (salvo una) tiene un espaciotiempo 10-dimensional? De nuevo, esta discrepancia en la dimensionalidad espaciotemporal parece considerarse un «efecto de energía» (§31.10) y no particularmente fundamental. Cabe imaginar que se percibirán más y más dimensiones cuando se sondee con energías mayores. De esta manera, los teóricos de cuerdas parecen justificar su actitud aparentemente displicente respecto a las dimensiones espaciales. Ya he expresado mi desacuerdo con argumentos de este tipo en §§31.11,12. En mi opinión, las dificultades con las enormes diferencias en libertad funcional en diferentes números de dimensiones[64] no han sido convenientemente abordadas. Esta cuestión también surge como una amenaza en otros temas que en la actualidad ocupan el interés de muchos teóricos de cuerdas; a continuación haré una breve mención de ellos.

31.15. Cuerdas y entropía de agujero negro

Recordemos de §27.10 y §30.4 que la fórmula de Bekenstein-Hawking asigna una *entropía* a un agujero negro que es proporcional al área de su horizonte. Aunque se han dado argumentos diferentes en apoyo de esta conclusión, ninguno de ellos equipara inequívoca[65] y explícitamente la entropía del agujero negro con el logaritmo de un volumen del espacio de fases, como exige la fórmula de Boltzmann (§27.3). Esto equivaldría a un recuento directo de los grados de libertad «perdidos en el agujero», recuento realizado de acuerdo con la teoría de la gravitación cuántica apropiada. En 1966, Andrew Strominger y Cumrun

Vafa presentaron un cálculo,[66] utilizando cuerdas y membranas, que apoyaba una interpretación de la fórmula de entropía de Bekenstein-Hawking como «recuento de grados de libertad». Esto fue aclamado por los teóricos de cuerdas en términos tales como: «Un enigma que duraba un cuarto de siglo ha sido resuelto».[67]

Como parece ser habitual en las proclamas de los teóricos de cuerdas, esta conclusión es exagerada. Por ejemplo, el cálculo original de Strominger-Vafa se refería solo a agujeros negros en un espaciotiempo 5-dimensional. Resultados posteriores se aplican al 4-espaciotiempo ordinario, pero la expectación inicial que condujo a afirmaciones como las anteriores se debieron, al parecer, al cálculo 5-dimensional original. Además, estos resultados de teoría de cuerdas se refieren solo al caso límite de un «agujero extremal» (o a perturbaciones a partir del mismo), para los que la temperatura de Hawking (véase §30.4) es cero, y donde el agujero incluye campos supersimétricos adicionales de tipo Yang-Mills que no tienen una clara justificación en la física conocida. Además, los cálculos reales se realizaron en un espacio plano, donde no hay horizonte de sucesos real, y es una extrapolación argumentar que deberían aplicarse también a una métrica de agujero negro significativamente curva.

Permítanme que haga una aclaración. Hemos visto (§27.10) que en la teoría del vacío de la relatividad general ordinaria, donde el espaciotiempo es tetradimensional, un agujero negro aislado estacionario está descrito por la métrica de Kerr, caracterizada por los valores de solo dos parámetros reales (no negativos) m y a, donde m es la masa y $a \times m$ el momento angular (en unidades naturales). Para que la geometría de Kerr describa realmente un agujero negro, antes que una singularidad desnuda (§28.8), exigimos $m \geq a$. (Nótese, por ejemplo, que esta desigualdad es necesaria para la fórmula $A = 8\pi m[m + (m^2 - a^2)^{1/2}]G^2/c^4$, de §27.10, para el área A del horizonte de un agujero de Kerr.) El caso *extremal* de un agujero negro ordinario ocurre cuando $m = a$, que lo cualifica como un «agujero negro». Es realmente un caso límite (inalcanzable astrofísicamente) de un agujero negro ordinario, con un valor cero para su temperatura de Hawking.

Esto puede compararse con lo que sucede con el «agujero negro» explícito considerado por los teóricos de cuerdas, que es también «ex-

tremal», en el sentido de que su temperatura de Hawking se anula. Pero casi todos estos cálculos de cuerdas se refieren a un tipo de animal completamente diferente. En lugar de rotación —como se permite por la presencia del parámetro a de Kerr—, se introducen campos físicos adicionales. El «agujero» de los teóricos de cuerdas está configurado en su lugar sobre la solución de Reissner-Nordstrøm de la ecuación de Einstein, que, a diferencia de la solución de Kerr, es esféricamente simétrica. En lugar del parámetro a de Kerr, hay un parámetro e que mide la carga eléctrica total del agujero, pues la métrica de Reissner-Nordstrøm es una solución de las ecuaciones de Einstein-Maxwell (§31.4), a saber, las ecuaciones de Einstein con el tensor energía-momento igual al de un campo de Maxwell libre de fuentes.[68] El área del horizonte está dada ahora por una fórmula de aspecto muy similar

$$A = 8\pi[m + (m^2 - e^2)^{1/2}]G^2/c^4.$$

La condición para que esta métrica represente un agujero negro en lugar de una singularidad desnuda es $m \geq |e|$, y la condición para que el agujero sea extremal (temperatura cero) es $m = |e|$. (Las barras de módulo —véase §6.1— solo permiten una e negativa.)

El tipo de «agujero negro» en el que están interesados principalmente los teóricos de cuerdas es, en esencia, el mismo que el del caso Reissner-Nordstrøm, pero donde una familia supersimétrica de campos de Yang-Mills (§25.7) reemplaza al campo de Maxwell. La solución completa es, en efecto, un ejemplo particular de lo que se denomina un *estado BPS* (donde «BPS» significa «Bogomoln'yi-Prasad-Sommerfeld»), en el que los requisitos de supersimetría, estacionariedad y energía mínima determinan la solución. No me molestaré con los detalles de lo que significa esto. Aunque estas cuestiones tienen claro interés para los teóricos de cuerdas y otras personas interesadas en la supersimetría, no hay todavía evidencia de su relevancia para el mundo físico real (véase §31.2).

¿Qué hay sobre el hecho de que los cálculos específicos de grados de libertad en cuerdas se realizan en el espacio plano, donde no hay horizonte de sucesos? Al contar con una formación en la teoría de la relatividad general de Einstein, creo que este es uno de los as-

pectos más enigmáticos de las afirmaciones de los teóricos de cuerdas. Es difícil ver cómo se puede llegar a conclusiones rigurosas concernientes a los agujeros negros sin un profundo respeto por la geometría espaciotemporal muy curvada del agujero negro, donde la «información» en su formación está encerrada detrás de un horizonte de sucesos.

Tratemos de hacernos una idea aproximada de cómo procede el argumento de la cuerda.[69] Para empezar, imaginemos que se estima el número de grados de libertad física contando el número de posibles lazos de cuerda diferentes, de longitud l, en un retículo a escala de Planck, digamos dentro de un volumen esférico de un radio fijo, al que estas cuerdas aportan cierta masa-energía total M. Vamos a suponer que el valor real de la constante gravitatoria de Newton está a nuestra disposición. Para G suficientemente pequeña, no habrá ningún agujero negro, y en el límite en que G tiende a cero, el espacio tiempo se hace realmente plano. Pero si imaginamos una G gradualmente creciente («aumentando la constante de Newton»), finalmente llegamos a una situación en la que, según la relatividad general, se formaría un agujero negro (recordemos la expresión de Michell $2MG/c^2$ para el radio de un «agujero negro newtoniano»; véase §27.8). En la teoría de cuerdas, G depende de un parámetro g_s, llamado *constante de acoplamiento de cuerda*, y resulta que G se haría más grande cuando g_s se hace más grande ($G \sim g_s^2$). En el límite de G pequeña (g_s pequeña), el logaritmo de la cuenta de grados de libertad de cuerda (§27.3) da una entropía que es la misma que el valor de Bekenstein-Hawking para un agujero negro, incluso si no hay agujero negro. Se ofrecen argumentos de escala para demostrar que esta relación persiste cuando G aumenta, de modo que la fórmula de Bekenstein-Hawking real se obtiene cuando se llega a la etapa de agujero negro.

Esto da solo una correspondencia cualitativa entre una cuenta de grados de libertad de cuerdas y la fórmula de Bekenstein-Hawking $S_{BH} = \frac{1}{4}A \times kc^3/G\hbar$, que confirma una proporcionalidad aproximada entre la entropía de cuerda y el área A de un agujero negro. Para obtener el valor preciso de $\frac{1}{4}A$ en la fórmula, Strominger y Vafa apelaron a

una consideración de estados BPS, donde los requisitos de supersimetría permiten fijar la masa en términos de los diversos valores de la «carga» (para los campos de Yang-Mills supersimétricos), y donde, en lugar de enumerar configuraciones de cuerda, el «recuento» enumera ahora todos los diferentes estados BPS[70] que contribuyen al total (todos los estados BPS con el conjunto de cargas dado). Esto se puede hacer de forma bastante explícita, y el logaritmo de este número da de forma notable precisamente el valor de $\frac{1}{4}A$ (en el caso extremal). No solo esto, sino que (como demostró un trabajo posterior) perturbaciones fuera de la extremalidad (i.e., a temperatura de Hawking infinitesimalmente por encima de cero), la entropía aún sale correctamente, como lo hacen ciertas ligeras correcciones a la naturaleza puramente del «agujero negro» de la radiación de Hawking. Además, lo mismo resulta válido cuando hay rotación, y en el espaciotiempo 4-dimensional.

Al lector atento tal vez le intrigue por qué parece haber «otra constante» g_s en la teoría de cuerdas cuando todo debería haber estado determinado por la constante de cuerda α' (§§31.8,9), el único parámetro real que aparece en el lagrangiano. La respuesta a esto reside en el hecho de que las cosas no han sido realmente determinadas, porque hay que establecer un valor para el campo dilatón (§31.8). El valor de g_s está dado por el valor esperado[71] de este campo dilatón, que se supone constante. En general, el campo dilatón no tiene por qué ser constante, pero se suele tratar como tal (como en la discusión anterior) por conveniencia. Su valor dependería de varias cosas, tales como la elección específica del espacio $\mathcal{Y}$ de tamaño de Planck y la elección del tipo de teoría de cuerdas (i.e., tipo I, IIA, IIB, heterótica O(32), o heterótica $E_8 \times E_8$; véase §31.14). De hecho, esta dependencia subyace en las dualidades fuerte/débil de §31.14, siendo la g_s de una teoría el recíproco de la g_s de la teoría «dual».

El lector percibirá que encuentro que los argumentos anteriores están muy lejos de una deducción real de S_{BH} mediante la teoría de cuerdas, pese a algunos acuerdos notables. Sé que otros teóricos de la relatividad general también se sienten muy incómodos —muy particularmente con el hecho de que la esencial propiedad de horizonte de un agujero negro, para determinar su vasta entropía, no parece haber

desempeñado ningún papel en absoluto. (Esto contrasta bastante con la discusión de variables de lazo de la entropía de un agujero negro que veremos en §32.6.) De hecho, en la imagen de la teoría de cuerdas, la entropía apenas aumenta en el momento de formación de un agujero negro, lo que da un punto de vista completamente diferente del habitual (descrito en §27.10).

Además, debería mencionar una dificultad técnica específica con el argumento de cuerdas tal como se ha dado.[72] Esta se relaciona con la propiedad termodinámica peculiar de los agujeros negros normales por la que los agujeros con pequeño momento angular tienen un *calor específico negativo* (véase la nota 27.2). El calor específico de un cuerpo se mide por el aumento su temperatura cuando se le suministra una pequeña cantidad de energía térmica. Para un cuerpo ordinario, este es un número *positivo*, pues nuestra experiencia normal es que cuando se aplica calor a un cuerpo su temperatura aumenta. Pero en el caso de un agujero negro solemos encontrar que su temperatura *disminuye*. La energía térmica proporciona masa al agujero (por $E = mc^2$), de modo que el agujero se hace más masivo, y por la $T_{BH} = 1/8\pi m$ de Hawking para un agujero de Schwarzschild (§30.4) su temperatura disminuye, de modo que el calor específico es negativo. Esta curiosa negatividad del calor específico de un agujero negro podría presentar una dificultad para el argumento anterior de la teoría de cuerdas para la entropía de un agujero negro si se aplicara a agujeros negros que no están próximos al caso extremal, puesto que para el argumento parece ser necesario un calor específico positivo, y el calor específico de un agujero negro solo es positivo cuando está próximo a ser extremal. De hecho, encontramos que el calor específico es positivo solo cuando $m > a > > (2\sqrt{3} - 3)^{1/2}m$ en el caso de Kerr, y solo cuando $m > e > m\sqrt{3}/2$ en el caso de Reissner-Nordstrøm, y lo mismo es válido para todos los demás campos de Yang-Mills cargados.

Parece haber algunas relaciones sorprendentes que emergen de estos cálculos de la teoría de cuerdas. Pero, en mi opinión, quedan muy lejos de ofrecer una justificación independiente de la fórmula de la entropía de Bekenstein-Hawking. La aproximación de variables de lazo (brevemente considerada en §32.6) parece ofrecer un ataque mucho más impresionante a este problema basado en la gravedad cuántica.

31.16. El «principio holográfico»

Los argumentos de cuerdas mencionados antes, como prácticamente todos los cálculos de cuerdas, son de naturaleza perturbativa. Sin embargo, más recientemente se han propuesto ciertas ideas con la intención de ofrecer resultados exactos. Estas dependen de varios ejemplos de algo denominado la «conjetura holográfica», que de algún modo se ha promocionado a la categoría de *principio holográfico*. Parece que la idea de este «principio» consiste en que, en ciertas circunstancias apropiadas, los estados de una teoría (cuántica) de campos de cuerdas definida en un espaciotiempo $\mathcal{M}$ pueden ponerse en correspondencia directa 1-1 con los estados de otra teoría cuántica de campos, donde la segunda QFT está definida en otro espaciotiempo $\mathcal{E}$ de dimensión menor. A menudo, se presenta $\mathcal{E}$ como si fuera una frontera (de género tiempo) de $\mathcal{M}$, o al menos una subvariedad conformemente suave de género tiempo de $\mathcal{M}$ (véase la Fig. 31.14). Sin embargo, no es este el caso en el ejemplo habitual que examinaremos enseguida. El principio holográfico se considera, en cierto sentido, análogo a un *holograma*, donde se percibe una imagen 3-dimensional cuando se mira una superficie (básicamente) 2-dimensional.[73]

La forma más familiar de este «principio holográfico» deriva del trabajo de Juan Maldacena en 1988, y que a veces se conoce como la *conjetura de Maldacena*, o también la conjetura ADS/CFT. Aquí $\mathcal{M}$ va a ser un producto (1 + 9)-dimensional $AdS_5 \times S^5$, donde AdS_5 es el *espacio anti-De Sitter* (véanse §28.4 y las Figs. 28.8 y 28.9b, si bien aquí hay cuatro dimensiones espaciales). El S^5 es una esfera de género espacio cuyo radio es de dimensión cosmológica, igual a $(-\Lambda')^{\frac{1}{2}}$, donde Λ' es la constante cosmológica (negativa) de AdS_5 (§19.7). El espacio más pequeño $\mathcal{E}$ va a ser el «scri» 4-dimensional (i.e., infinito conforme; véase §27.12) de AdS_5; véase la Fig. 31.15. Notemos que $\mathcal{E}$, siendo 4-dimensional, no es ciertamente la frontera de $\mathcal{M}$ en este caso, puesto que $\mathcal{M} = AdS_5 \times S^5$ es 10-dimensional. En su lugar, la «frontera» —i.e., el «scri»— de $\mathcal{M}$ puede considerarse (pero no conformemente) como $\mathcal{E} \times S^5$. La conjetura de Maldacena propone que la teoría de cuerdas en $AdS_5 \times S^5$ va a ser equivalente a cierta teoría de Yang-Mills supersimétrica en $\mathcal{E}$.

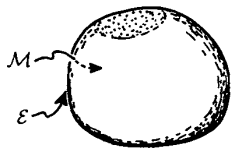


Fig. 31.14. ¿«Principio holográfico»? Un espaciotiempo $\mathcal{E}$ es una frontera (de género tiempo) de otro espaciotiempo $\mathcal{M}$. Se conjetura que una QFT adecuada definida sobre $\mathcal{E}$ puede ser equivalente a una QFT de cuerdas en $\mathcal{M}$.

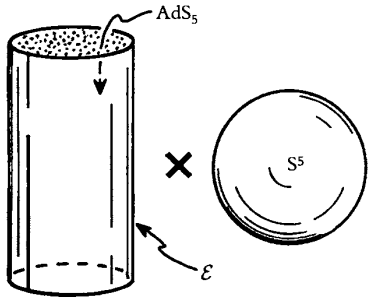


Fig. 31.15. Conjetura ADS/CFT (de Maldacena). Aquí $\mathcal{E}$ va a ser el «scri» 4-dimensional (infinito conforme; §27.12) del 5-espacio anti-De Sitter AdS_5 (véase la Fig. 28.9), antes que el 10-dimensional $\mathcal{M} = AdS_5 \times S^5$, pero se conjetura que la teoría de cuerdas sobre $\mathcal{M}$ es equivalente a una teoría Yang-Mills supersimétrica en $\mathcal{E}$.

Aquí no hay oportunidad de apelar al tipo de argumento de «energía cuántica» propuesto en §31.10 para explicar la gran discrepancia entre la libertad funcional de un campo ordinario $\mathcal{M}$, a saber $\infty^{M\infty^9}$ y un campo ordinario en $\mathcal{E}$, a saber $\infty^{E\infty^3}$. Puesto que las dimensiones extra de $\mathcal{M}$ no son «pequeñas» de ninguna manera —pues son de escala cosmológica—, el diluvio de grados de libertad adicionales, de la dependencia de los campos en la parte S^5 de $\mathcal{M}$, eliminaría cualquier posibilidad de un acuerdo entre las dos teorías de campos. Lo mismo se aplicaría a QFT ordinarias en $\mathcal{M}$ y $\mathcal{E}$, puesto que los propios estados de una partícula están descritos simplemente por «campos ordinarios» (véase §26.2). La única oportunidad de que el principio holográfico sea realmente verdadero para dichos espacios es que las QFT en consideración estén lejos de la «ordinaria».

En el caso de la teoría de cuerdas en $\mathcal{M}$, es ciertamente concebible que haya condiciones de consistencia muy fuertes que reduzcan drásticamente la libertad funcional $\infty^{M\infty^9}$. Pero, a la hora de la verdad, parece muy poco probable. Recordemos del capítulo 21, §22.8 y §16.7 que el estado cuántico de una única partícula en un espaciotiempo (1 + 9)-dimensional tiene la libertad funcional $\infty^{P\infty^8}$, donde P es un entero positivo que describe el número de grados de libertad internos o

rotacionales (por ejemplo, el espín) de la partícula. El estado cuántico de una única cuerda parecería tener una libertad funcional mucho *mayor*, puesto que una cuerda *clásica* tiene infinitos grados de libertad. Si el número $\infty^{P\infty}$ se reduce de alguna manera, entonces debe haber ligaduras enormes, quizá del tipo que llevan a las restricciones en la dimensión y curvatura espaciotemporal mencionadas en §31.7, pero no soy consciente de que se hayan sugerido tales ligaduras, que, en cualquier caso, afectarían drásticamente al recuento de estados de cuerdas, tal como en §31.15.

La posibilidad restante es encontrar una manera de aumentar enormemente la libertad funcional en los campos de Yang-Mills supersimétricos en $\mathcal{E}$. La única manera que puedo ver de conseguir esto sería tener un número infinito de tales campos, que podría alcanzarse tomando el límite $N \to \infty$ (siendo N el número de generadores de supersimetría). Sin embargo, en la forma habitual de esta conjetura se toma $N = 4$ para que haya un «grupo interno» SO(6) que actúe sobre las compañeras supersimétricas pero deje inalterados los potenciales de Yang-Mills.[74] Esta simetría interna se toma de modo que encaje con la simetría SO(6) del S^5 que interviene en $AdS_5 \times S^5$. En mi opinión, es fundamentalmente erróneo tratar de encajar una «simetría espaciotemporal» en un grupo interno de este tipo, a menos que, como sucede con la teoría de Kaluza-Klein original (§31.4), la simetría espaciotemporal se especifique como exacta, por la existencia de campos de Killing, y sea también respetada por todos los campos físicos en el espaciotiempo. Los grados de libertad en exceso en $\infty^{M\infty}$ resultan precisamente por la razón de que no hay tal simetría especificada en la parte S^5 de $\mathcal{M}$, que debe ser respetada por campos en $\mathcal{M}$.

Mi opinión es que la importancia de este tipo de discrepancia en libertad funcional ha sido profundamente subestimada. Los «tamaños» de los espacios de Fock (véanse §26.6 y la nota 26.12) serán completamente diferentes cuando quiera que la libertad funcional en los campos clásicos sea del todo diferente. Debería advertirse que la condición de *frecuencia positiva*, exigida por los estados de una partícula en QFT, no cambia la libertad «$\infty^{M\infty N}$» para los campos clásicos. Simplemente compensa el hecho de que estos campos clásicos tienen que ser complexificados cuando pasamos a una descripción de QFT; véase §26.3.

¿Por qué se toma tan en serio la conjetura ADS/CFT? El apoyo para ella parece deberse a una correspondencia entre estados BPS en los dos lados, que había sido advertida por Maldacena, y a otras correspondencias. Muchas de las últimas pueden entenderse puramente a partir de la correspondencia entre los grupos de simetría (a saber, $SO(2,4) \times SO(6)$) de las teorías de campos en los dos lados, pero hay también algunas «coincidencias» adicionales que necesitan explicación. Parece ser que una razón para esperar que ADS/CFT sea cierta es que podría proporcionar un asidero para ver cómo sería una teoría de cuerdas sin recurrir a los métodos perturbativos habituales, con todas las graves limitaciones que tienen tales métodos.

Los cálculos en el lado $\mathcal{E}$ resultan más fáciles por el hecho de que el espacio $\mathcal{E}$ es conformemente plano (y a veces conocido como «plano», aunque no tiene una métrica real asignada al mismo, sino solo una métrica conforme de signatura $+---$). Es el espacio de recubrimiento universal (nota 15.9) del «espacio de Minkowski compactificado» al que llegaremos en §33.3,[75] que tiene la topología de $S^3 \times \mathbb{E}^1$. El espacio *métrico* $S^3 \times \mathbb{E}^1$ se denomina a veces «cilindro de Einstein» o «universo de Einstein», pues era el modelo cosmológico defendido por este durante el período 1917-1929, cuando había incorporado una constante cosmológica en su ecuación de campo (§19.7).[76]

La conjetura ADS/CFT surgió como otra manera de considerar la «deducción» de cuerdas de la fórmula de la entropía de agujero negro de Bekenstein-Hawking (§31.15). Para ello, un «agujero negro» se representa como un «estado térmico» en $\mathcal{E}$. Esto solo sería de relevancia para agujeros negros de tamaño cosmológico y, en el mejor de los casos, proporciona una «conjetura» basada en algunos acuerdos notables entre «cálculos de entropía» hechos de diferentes maneras, en lugar de una deducción real de la expresión de Bekenstein-Hawking.

31.17. La perspectiva de la D-brana

En varios apartados de la exposición anterior, en especial en §§31.11, 12,15,16, he expresado mis reparos al uso de espaciotiempos de di-

mensiones más altas en la teoría de cuerdas. Una de mis mayores dificultades con ello es el enorme incremento de la libertad funcional en las teorías de dimensiones más altas ($\infty^{P\infty^M}$ para un espaciotiempo $(1 + M)$-dimensional), donde hay que idear algún medio de congelar esta libertad extra. Recordemos la analogía de la «manguera», como se ilustra en las Figs. 15.1 y 31.3. Lo que percibimos como un «punto en el espaciotiempo» es una entidad plenamente dinámica, descrita aquí como un círculo, pero en general con alta dimensionalidad, y es esta la que nos proporciona la enorme libertad extra. Recordemos que Kaluza y Klein recurrieron a eliminar esta libertad por decreto, afirmando la presencia de un vector de Killing que nos lleva alrededor de la «manguera», reduciendo en efecto el espaciotiempo a uno 4-dimensional. Este es, matemáticamente, un procedimiento perfectamente respetable, pero nunca parece haber sido propuesto con seriedad en la teoría de cuerdas. En lugar de ello, como en §31.10, los teóricos de cuerdas parecen confiar normalmente en la enormidad de la «energía» que se requeriría para excitar oscilaciones en dichas dimensiones extra. Como hemos visto en §§31.11,12, hay muchas razones para dudar de esta línea argumental.

Pese a todo, la naturaleza de la actividad en la teoría de cuerdas es tal que resulta difícil de atacar por medio de argumentos específicos, tales como los indicados en §§31.11,12,16. En efecto, siempre existe la posibilidad de que el tema se metamorfosee en una forma diferente, momento en el cual tales ataques se estiman irrelevantes.[77] De hecho, y por lo que puedo ver, la filosofía entera puede haber dado un vuelco con algunas maneras recientes de mirar las dimensiones más altas, sin ninguna indicación pública de que haya sucedido un cambio importante. Aunque no sea necesariamente el punto de vista claro de los teóricos de cuerdas de la «corriente principal», la introducción de cierta «filosofía de la D-brana» parece tener este carácter.

En primer lugar, ¿qué son las *D-branas*? ¿Por qué las requiere la teoría de cuerdas? La respuesta básica a la segunda pregunta es que dan sentido a las cuerdas abiertas que intervienen en la teoría tipo I, como se ha mencionado en §31.14: cada uno de los dos extremos de una cuerda abierta debe residir en una D-brana (Fig. 31.16). En respuesta a la primera pregunta, una D-brana (o una D-q-brana) es una estructura

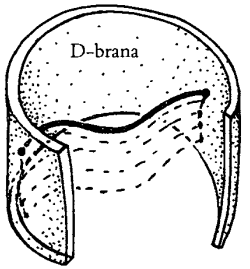


Fig. 31.16. Se supone que los dos extremos de una cuerda abierta residen en un subespacio $(q + 1)$-dimensional de género tiempo llamado una D-brana, o D-q-brana. Una D-brana es una entidad esencialmente clásica (aunque posee propiedades de supersimetría), que representa una solución de la teoría de la supergravedad 11-dimensional (un tipo de «estado BPS»).

de género tiempo de $1 + q$ dimensiones espaciotemporales (i.e., q dimensiones espaciales y 1 dimensión temporal), que es una solución estable de la supergravedad 11-dimencional. (Invocando una de las dualidades de la teoría M, alternativamente podemos considerar una D-brana como una solución de las ecuaciones de alguna otra versión de la teoría M de cuerdas.) Básicamente esta es una «brana» (como se ha descrito en §31.14) de una dimensión (0, 1, 2, ..., o 9) que es un estado BPS (véanse §§31.15,16), de modo que posee una colección supersimétrica de «cargas» de Yang-Mills y tiene una energía mínima sujeta a esto. Las D-branas aparecen en muchas discusiones modernas relacionadas con las cuerdas (por ejemplo, en la entropía de los agujeros negros; véase §31.15). Suelen tratarse como si fueran objetos clásicos que yacen dentro del espaciotiempo completo de $1 + 9$ (o $1 + 10$) dimensiones. La «D» viene de «Dirichlet», por analogía con el tipo de problema de valor de frontera conocido como un problema de *Dirichlet*, en el que hay una frontera de género tiempo sobre la que se especifican los datos (según Peter G. Lejeune Dirichlet, un eminente matemático francés que vivió entre 1805 y 1859; recordemos la «serie de Dirichlet» de §7.4).

No intentaré discutir aquí las D-branas en detalle. Lo único que quiero plantear es que con la introducción de tales «D-branas» varios teóricos han expresado una «filosofía de cuerdas» que parece representar un profundo cambio respecto a lo anterior. En efecto, se afirma con cierta frecuencia que podríamos «vivir en» esta o esa D-brana, lo que significa que nuestro espaciotiempo percibido podría yacer realmente dentro de una D-brana, de modo que la razón de que no se perciban

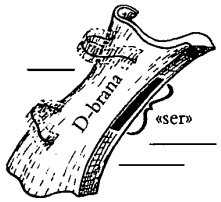


Fig. 31.17. Un punto de vista alternativo al de la Fig. 31.3, a menudo expresado en el contexto de las D-branas, es que un «ser» en un espacio de dimensiones más altas no necesita rebasar todas las dimensiones extra, sino que puede considerarse que «vive» en un subespacio, quizá en una frontera D-brana.

ciertas «dimensiones extra» se explicaría por el hecho de que «nuestra» D-brana no se extiende a esas dimensiones extra.

La última posibilidad sería la postura más económica, por supuesto, de modo que «nuestra» D-brana (una D-3-brana) sería de 1 + 3 dimensiones. Esto no *elimina* los grados de libertad en las dimensiones extra, pero los reduce drásticamente. ¿Por qué es así? Nuestra perspectiva ahora es que no somos «conscientes» de los grados de libertad que están implicados en el interior profundo del espacio de mayores dimensiones entre las D-branas, y es en esto donde se está dejando sentir la excesiva libertad funcional. Solo vamos a ser conscientes de dimensiones extra allí donde inciden directamente sobre la D-brana en la que «vivimos». Volvamos a nuestra analogía de la manguera. Más que una imagen de tipo «espacio cociente»[78] que evoca la analogía de Kaluza-Klein original (Fig. 31.3), nuestro espaciotiempo observado aparece ahora como un *subespacio* 4-dimensional del espacio de dimensiones más altas. Para visualizarlo, pensemos que una tira dibujada a lo largo de la longitud de nuestra manguera representa el subespacio D-brana que es ahora «nuestro universo 4-dimensional observado» (Fig. 31.17).

¿Cuánta libertad funcional esperamos ahora? La situación es ahora algo parecida a la imagen geométrica que se ha adoptado en §31.3 para obtener una perspectiva más convencional con respecto a la «supergeometría» (véase la Fig. 31.2). Puesto que ahora estamos interesados solo en el comportamiento en la D-brana (que suponemos que es geométricamente una (1 + 3)-superficie ordinaria), podemos imaginar que nuestra libertad funcional se ha convertido en una aceptable $\infty^{M\infty^3}$, aunque para un M bastante grande. Sin embargo, incluso esto supone que la restricción de la dinámica en el 10-espacio (u 11-espacio) completo nos proporciona ecuaciones dinámicas dentro de «nuestra» D-brana

4-dimensional que son del tipo convencional, de modo que bastarán los datos iniciales en una 3-superficie para determinar el comportamiento en todo el 4-espacio. Esto es difícilmente probable, en general, de modo que aún cabe esperar un excesivo $\infty^{M\infty^4}$. ¡El problema no ha desaparecido todavía!

Tal actitud hacia las D-branas se ha utilizado para intentar resolver el «problema de la jerarquía» que se ha mencionado en §31.1. En concreto, esta es la cuestión de por qué las interacciones gravitatorias son tan minúsculas comparadas con las demás fuerzas importantes de la naturaleza o, de manera equivalente, por qué es la masa de Planck tan enormemente mayor que las masas de las partículas elementales de la naturaleza (en un factor de aproximadamente 10^{20}). La aproximación de la D-brana a este problema parece requerir la existencia de más de una D-brana, una de las cuales es «grande» y la otra «pequeña». Hay un factor exponencial involucrado en cómo se estira la geometría desde una D-brana hasta la otra, y esto se considera una ayuda para abordar la discrepancia en 10^{40}, más o menos, entre las intensidades de la fuerza gravitatoria y las otras fuerzas.[79] Se puede decir que este tipo de imagen de un espaciotiempo de dimensiones más altas, que se estira desde la frontera de una D-brana hasta la otra, es uno de los tipos de geometría sugeridos por las teorías 11-dimensionales, tales como la teoría M, donde la undécima dimensión tiene la forma de un segmento abierto, y la geometría de cada frontera tiene la forma topológica (por ejemplo, $\mathcal{M} \times \mathcal{Y}$) de los 10-espacios considerados antes. En otros modelos, la undécima dimensión es topológicamente S^1.

31.18. ¿El estatus físico de la teoría de cuerdas?

¿Qué vamos a hacer de todo esto con respecto al estatus de la teoría de cuerdas como una teoría física para el futuro? Para mí, la situación tiene aspectos muy enigmáticos y notables, así como otros que suenan implausibles, y sería erróneo ser completamente dogmático en esta etapa. Pese a todo, muchas de las afirmaciones de los teóricos de cuerdas se hacen con gran seguridad y aparente confianza. Es indudable que estas deben ser suavizadas y tomadas con reservas antes de que sean de

consumo general. Creo que es justo decir que algunas de las afirmaciones con más peso pueden ser descartadas (tales como que la teoría de cuerdas ha proporcionado una teoría completa y consistente de la gravedad cuántica). No obstante, tengo que admitir que parece haber algo de auténtica trascendencia «entre bastidores» en algunos aspectos de la teoría M de cuerdas. Como me comentó en un correo electrónico el matemático Richard Thomas del Imperial College de Londres

> No puedo destacar suficientemente cuán profundas son algunas de estas dualidades; constantemente nos sorprenden con nuevas predicciones. Muestran una estructura que nunca se creyó posible. Los matemáticos predijeron varias veces con seguridad que estas cosas no eran posibles, pero personas como Candelas, De la Ossa, *et al.* han demostrado que esto era falso. Cada predicción hecha, apropiadamente interpretada matemáticamente, ha resultado ser correcta. Y *no* por ninguna razón *conceptual* matemática hasta ahora; no tenemos ninguna idea de por qué son verdaderas, solo calculamos ambos lados independientemente y encontramos las mismas estructuras, simetrías y respuestas en ambos lados. Para un matemático estas cosas no pueden ser coincidencias, deben proceder de una razón superior. Y esa razón es la *suposición de que esta gran teoría matemática describe la naturaleza*...

Pero bien podría ser que este «algo» sea de interés puramente matemático, sin que haya ninguna razón real para creer que nos acerca más a los secretos de la naturaleza. Creo que esta es una postura perfectamente sostenible, aunque en mi fuero interno estoy dispuesto a aceptar que la naturaleza podría tener quizá un interés en estas cuestiones (probablemente a lo largo de líneas algo diferentes de las que se han sugerido hasta ahora). La fuerza del argumento a favor de la teoría de cuerdas parece residir en varias relaciones matemáticas notables entre «situaciones físicas» en apariencia diferentes (normalmente, algo alejadas de la física del mundo real de la naturaleza). ¿Son una «coincidencia» estas relaciones, o hay alguna razón más profunda tras ellas? Me parece que para muchas de ellas hay realmente una razón, todavía no descubierta, pero eso aún no nos garantiza que los teóricos de cuerdas están haciendo física. O, si la hacen, ¿qué área de la física están explorando realmente?

No creo que pueda hacerse una valoración adecuada de estas cuestiones sin abordar el papel concreto de Edward Witten. Él es aceptado generalmente como la figura con más responsabilidad en la dirección de la investigación en la teoría de cuerdas (y la teoría M) desde finales de la década de 1980. He mencionado su papel en lanzar la «segunda revolución en supercuerdas» en 1995 (§31.14), pero ya entonces había establecido su preeminencia al iniciar varios desarrollos importantes en la teoría de cuerdas, y en muchas otras áreas que tienen cierta relación (no siempre obvia) con la teoría de cuerdas. La teoría de cuerdas ha tenido varios «guías turísticos» (§31.6) a lo largo de sus aproximadamente treinta años de historia, pero sin duda Witten ha sido en muchos aspectos el más impresionante de estos. Como ejemplo de ello, se puede mencionar que el artículo original de Maldacena, que inició buena parte de la actividad discutida en §31.16, quedó esencialmente inadvertido en los archivos hasta que fue continuado por Witten en 1998. Inmediatamente se convirtió en el artículo más citado por los teóricos de cuerdas.[80]

Es interesante que en un nuevo trabajo que parece bastante importante[81] Witten ha vuelto a consideraciones dentro de un espacio-tiempo 4-dimensional estándar (aunque sigue habiendo supersimetría). Combinando ideas de la teoría de twistores y la teoría de cuerdas, Witten es capaz de obtener algunos resultados fascinantes concernientes a las interacciones Yang-Mills de varios gluones (§25.7). Este trabajo es particularmente importante desde mi propia perspectiva orientada a los twistores (véase el capítulo 33), y muy bien podría llevar a nuevos desarrollos.

No hay duda sobre la extraordinaria calidad de los logros intelectuales de Witten. Puedo hablar por experiencia directa. En numerosas ocasiones he asistido a seminarios en el Instituto de Matemáticas de Oxford (en la serie de geometría y análisis), en los que se ha anunciado algún enfoque nuevo y muy original a algún problema, y ha resultado que la idea seminal procedía en realidad de Witten. A menudo, tales enfoques han abierto un nuevo campo, donde estas nuevas ideas imprevistas han arrojado una potente luz original sobre problemas matemáticos difíciles —a veces problemas que previamente parecían intratables—. Sin duda, Witten posee una extraordinaria intuición y unos

conocimientos matemáticos de primer orden. (De hecho, ganó una Medalla Fields en 1990, que entre los matemáticos tiene la misma categoría que un premio Nobel en el mundo de la ciencia. Ciertamente esta es una estraordinaria hazaña para un físico.) Pese a todo, creo que el propio Witten negaría que sus capacidades estén tanto en el lado matemático. Tal como entiendo sus propias ideas, sus éxitos se deben a su profunda observación de los modos de la naturaleza, sacando ideas de la estructura de la QFT con sus integrales de camino y espacios funcionales de dimensión infinita, de las ideas de la supersimetría y de la misma naturaleza de la teoría de cuerdas. Si él tiene razón, entonces quizá este sea uno de los argumentos más contundentes para aceptar sus opiniones de que la supersimetría y la teoría de cuerdas encuentran un profundo favor en la naturaleza. Por otra parte, ¡quizá sea un matemático más notable de lo que él admite!

¿Hasta qué punto me impresiona que las muy sorprendentes relaciones matemáticas que Witten y sus colegas han descubierto indiquen realmente alguna cercanía profunda a la naturaleza? No estoy del todo seguro de cómo ver esta cuestión, y ciertamente no estoy convencido de ello. Recordemos la notable hazaña del matemático Andrew Wiles al demostrar el famoso «último teorema de Fermat» tras tres siglos y medio de intentos fallidos (§1.3). Lo que Wiles estableció en realidad era que, en un caso importante, dos cálculos de aspecto muy diferente dan en realidad siempre la misma lista de respuestas, una notable afirmación cuya forma general se conoce como la conjetura de Taniyama-Shimura. (De hecho, la demostración de Wiles estableció solo parte de la conjetura T-S completa —una parte que era suficiente para establecer la afirmación de Fermat—, pero sus métodos proporcionaron un ingrediente esencial para la prueba completa, posteriormente completada por Breuil, Conrad, Diamond y Richard Taylor.)

Hay quizá cierta similitud entre esta conjetura y las relaciones de «simetría espejo» de espacios de Calabi-Yau mencionada antes (§31.14). En cada caso se tienen dos listas infinitas de números que misteriosamente resultan ser la misma. Este tipo de cosas está lejos de ser único en matemáticas, y podrían pasar muchos años, en cualquier caso concreto, antes de que salgan a la luz las razones subyacentes para la igual-

dad de las listas. Tal como lo entiendo, muchas de las relaciones deducibles utilizando «simetrías espejo» han sido ahora establecidas mediante argumentos puramente matemáticos.[82] Por lo que yo sé, esas misteriosas relaciones no se suelen presentar como apoyo a propuestas de teorías *científicas* (en oposición a las matemáticas). Esta cuestión será abordada de nuevo en §34.9.

Hemos visto el mismo tipo de «coincidencias» en los argumentos de la teoría de cuerdas para la entropía del agujero negro, como se presentan en §31.15 (e incluso en los anteriores argumentos «sin cuerdas» de §30.5). ¿Son estos meras coincidencias matemáticas, o consideramos que tales argumentos proporcionan deducciones reales? Permítanme terminar este capítulo considerando otro ejemplo de una coincidencia matemática sorprendente, tomado de la física de principios del siglo XX. En 1912, Woldemar Voigt formuló una teoría de las líneas espectrales basada en un modelo de oscilador incorrecto. Quince años después, Heisenberg y Jordan encontraron la que hoy consideraríamos la aproximación correcta de este problema, y vale la pena citar a Heisenberg en sus reminiscencias de la obra de Voigt:[83]

> Él fue capaz de disponer el acoplamiento de los osciladores entre sí, y con el campo externo, de tal modo que en campos magnéticos débiles el efecto Paschen-Back se representaba también de forma correcta. Para la región intermedia de campos moderados, obtuvo para las frecuencias e intensidades raíces cuadradas largas y complejas; es decir, fórmulas que eran básicamente incomprensibles, pero que obviamente reproducían los experimentos con gran exactitud. Quince años más tarde, Jordan y yo nos tomamos la molestia de calcular el mismo problema por los métodos de la teoría de perturbaciones en la mecánica cuántica. Para nuestro asombro, llegamos exactamente a las viejas fórmulas voigtianas en lo que se refería a las frecuencias y las intensidades, y esto también en el área compleja de los campos moderados. Más tarde pudimos percibir la razón para esto; era una razón puramente formal y matemática.

En el capítulo 34 volveré a esta desconcertante cuestión de las relaciones matemáticas como fuerza impulsora tras la teoría de cuerdas y otras propuestas para el desarrollo de una teoría física fundamental.

Notas

Sección 31.1

31.1. Por supuesto, habría que ser cuidadoso en no confundir esta *e* con la base de los logaritmos naturales e = 2,718281828459...; véase §5.3.

31.2. Esta enorme discrepancia entre las intensidades de las interacciones, fuerte, electromagnética, débil, y especialmente la gravitatoria, aproximadamente caracterizadas por las respectivas constantes de acoplamiento 10, $\frac{1}{137}$, $\sim 10^{-6}$, $\sim 10^{-39}$, se conoce a veces como «el problema de la jerarquía». La Universidad del Estado de Georgia tiene una preciosa página que explica los puntos más importantes para la comparación de dichas constantes: http://hyperphysics.phy-astr.gsu.edu/hbaseforces/couple.html.

31.3. El factor de renormalización relativamente «moderado» para la carga podría resultar debido a la naturaleza logarítmica de la divergencia electrodinámica. El lector perspicaz advertirá, por supuesto, que el rompecabezas de los valores minúsculos de las masas de las partículas no ha desaparecido, sino que se ha reexpresado en términos de una escala de distancias absurdamente minúscula.

31.4. Recordemos, sin embargo, el comentario de 't Hooft citado en §26.9.

Sección 31.2

31.5. Para una colección muy útil, con exposiciones de la historia, personalidades e ideas básicas subyacentes en la introducción de la supersimetría, véanse Kane (2001) en un nivel profano, y Kane (1999) para algo un poco más técnico.

31.6. Véanse Witten (1982) y Seiberg y Witten (1994) sobre la teoría de Yang-Mills supersimétrica, que lleva a una gran simplificación en la teoría de Donaldson de 4-variedades; véase Donaldson y Kronheimer (1990). Según John Baez, la teoría de Seiberg-Witten acorta algunas demostraciones en la teoría de Donaldson hasta una milésima parte de su longitud original.

31.7. Véanse Witten (1981) y Deser y Teitelboin (1977) para demostraciones de energía positiva utilizando supersimetría; véase Gibbons (1997) para desigualdades interesantes sobre agujeros negros.

31.8. Véase Greene (1999), nota 5, p. 399.

31.9. Véanse Lawrie (1998), o, para más detalles, Mohapatra (2002).

Sección 31.3

31.10. Hay tendencia a que N sea múltiplo de 2, por ser el número de componentes de un espinor (véanse §11.5 y §33.4). Esto no debería confundirse con el número 2^N de elementos en el álgebra de la supersimetría. Véase Wess y Bagger (1992) para una descripción de la simetría por uno de sus creadores.

31.11. Para más información sobre supervariedades, véase DeWitt (1984); Rogers (1980).

Sección 31.4

31.12. Véase el artículo de revisión por Bern (2002) para una amplia exposición de todo esto. Véase también Deser (1999, 2000).

31.13. Véanse Deser (1999, 2000) sobre la «última esperanza» para la renormalizabilidad de la supergravedad, y Deser y Zumino (1976).

31.14. Véase, por ejemplo, Hoyle *et al.* (2001), p. 1.418.

31.15. Para una vía rápida a estos recuentos, véase Penrose y Rindler (1984).

31.16. En una descripción convencional de fibrado, la métrica sobre un espacio base $\mathcal{M}$ podría estar «elevada» en un fibrado $\mathcal{B}$, sobre él, si se desea, para proporcionar en general una métrica «degenerada» canónica sobre $\mathcal{B}$, pero quizá una métrica no degenerada si puede hacerse uso de una estructura métrica sobre las fibras. Pero este no es un aspecto esencial de la estructura de un fibrado.

31.17. Estas son las ecuaciones de Einstein con el tensor energía-momento como fuente, junto con las ecuaciones de Maxwell libres en el espaciotiempo curvo de fondo.

31.18. No obstante, hay algunas aplicaciones recientes que combinan la teoría de cuerdas con la teoría de twistores que utiliza el espaciotiempo 4-dimensional normal. Véanse §31.31, §33.14 y la nota 31.81.

Sección 31.5

31.19. Compárese con la nota 12.14.

31.20. Véase Schwarz (2001) para una historia general de la teoría de cuerdas; en particular, véanse Veneziano (1968), Nambu (1970), Susskind (1970), Nielsen (1970) y Goddard *et al.* (1973).

31.21. Esto describe las cosas en términos de la función β, encontrada por el

gran Euler en 1777.Véase Goddard *et al.* (1973) para la primera exposición importante de la dualidad.

31.22. Veneziano ideó inicialmente el modelo para explicar los polos de Regge.Véanse Collins (1977) para la teoría de Regge en general, así como Penrose *et al.* (1978).

31.23. Posteriormente se obtuvieron algunos éxitos limitados al reunir las ideas de twistores y las de teoría de cuerdas, pero estos eran básicamente de naturaleza matemática y no ofrecían un punto de vista físico unificado; véanse Shaw y Hughston (1990) y la nota 31.76.

Sección 31.6

31.24. Citado en Greene (1999), p.139, de una entrevista a Michael Green, realizada por Brian Greene, 10 de diciembre de 1997.

31.25. Véase Witten (1996).

31.26. Trabajos autorizados más detallados y técnicos son los de Green *et al.* (1987), Polchinski (1998) y Green (2000).

Sección 31.7

31.27. Véanse Green *et al.* (1987), Polchinski (1998) o Green (2000) para un argumento que conduce a 26 dimensiones.

31.28. El número relevante que aparece de forma anómala en los conmutadores cuánticos (y debe establecerse en cero) es 24-σ, donde σ es el número de dimensiones espaciales *menos* el número de dimensiones temporales.

31.29. Con supersimetría, la anomalía se elimina cuando 8 - σ se fija a cero, con σ igual que en la nota anterior.

31.30. Una cuerda hadrónica muestra una diferencia menor respecto a una goma elástica ordinaria en cuanto que la segunda tiene una longitud *natural* finita para la que la tensión se hace cero. En el caso de una cuerda hadrónica, esta longitud natural sería cero.

Sección 31.8

31.31. Muchas de estas afirmaciones se encuentran en Greene (1999).

31.32. Citado por Abhay Ashtekar en una conferencia en el NSF-ITP Quantum Gravity Workshop en la Universidad de California en Santa Bárbara.

31.33. Aunque no todos los que forman parte de la comunidad relativista me acompañarían hasta el final, en la idea de que la unión cuántico/gravedad buscada debe implicar un cambio en las reglas de la QFT, continuamente encuentro apoyo de dicha comunidad para este punto de

vista. ¡La respuesta de la comunidad de la QFT suele ser mucho menos favorable!

31.34. El término «dilatón» no es una mala transcripción de «dilatación», sino que se refiere a una versión *cuántica* de dicha noción que surge a partir de los grados de libertad disponibles en un cambio de escala en la métrica. Recordemos del capítulo 26 que, según las reglas de la QFT, los propios grados de libertad cuantizados pueden manifestarse como un tipo de partícula.

Sección 31.9

31.35. Citado en Greene (1999), p. 210 de una entrevista a Edward Witten realizada por Brian Greene, 11 de mayo de 1998.

Sección 31.10

31.36. El número 70 procede de la fórmula $n(n-3)$ para el número de componentes independientes por punto de una $(n-1)$-superficie inicial en un n-espacio Ricci-plano; véanse Wald (1984), Lichnerowicz (1994), Choquet-Bruhat y DeWitt-Morette (2000).

31.37. Véanse Penrose (2003), Bryant *et al.* (1991), así como también Gibbons y Hartnoll (2002).

Sección 31.11

31.38. Véase Penrose (2003).

31.39. Véase Dine (2000) para reflexiones sobre los moduli.

31.40. Uno podría preferir quedarse dentro de un marco QFT y utilizar un estado coherente (§26.6) en lugar de una descripción clásica. Sin embargo, esto no evita las cuestiones que se plantean aquí.

31.41. Aunque dudo que a muchos teóricos de cuerdas les gustara hacer de $\mathbf{R}$ un proceso dinámico, hay algunas notables excepciones; véase Ellis *et al.* (1997a, 1997b).

31.42. Los excitones se comportan como bosones en una descripción por la teoría cuántica de campos de las vibraciones de la manguera (§22.13, §23.8 y §26.2), de modo que puede haber muchos cuantos en cualquier modo particular de $\mathcal{Y}$. Un sistema físico real para el que tal descripción cuántica puede ser apropiada sería una guía de ondas ópticas larga y estrecha (por ejemplo, una fibra óptica).

Sección 31.12

31.43. Véase Hawking y Penrose (1970).

31.44. La condición es que cada geodésica de género tiempo o nula encuentra curvatura «genérica», en el sentido de que en alguna parte a lo largo de cada una de dichas geodésicas, $k_{[a}R_{b]cd[e}k_{f]}k^c k^d \neq 0$, donde el vector nulo k^a es tangente a la geodésica. Una simple evaluación directa de los grados de libertad muestra que esta condición es ciertamente satisfecha en cualquier espaciotiempo «genérico». Habría que mencionar que el teorema se aplica en circunstancias más generales que la Ricci-planitud. Solo necesitamos que el tensor de Ricci satisfaga una apropiada «condición de energía no negativa» (véase §27.9, especialmente la nota 27.20 y §28.5). Para la «hipersuperficie compacta», véanse §12.6 y la nota 27.36.

31.45. Hay casos excepcionales de una $\mathcal{Y}$ de curvatura cero con la topología de un «hipertoro» $S^1 \times S^1 \times S^1 \times S^1 \times S^1 \times S^1$. No obstante, estos no son los modelos para $\mathcal{Y}$ que hoy prefieren los teóricos de cuerdas (§31.14). Además, la mayoría de las perturbaciones del hipertoro no serían planas.

31.46. Esta conclusión se sigue de otra aplicación del antes mencionado teorema de singularidad, que se aplica directamente al espaciotiempo entero $\mathcal{M}$. En esta aplicación, la condición de que existe una hipersuperficie de género espacio compacta es reemplazada por la existencia de un punto p cuyo cono de luz futuro $\mathcal{C}$ «se enrosca y se corta a sí mismo» en todas direcciones. El lugar geométrico $\mathcal{C}$ es el barrido por la familia de rayos de luz ℓ (i.e., geodésicas nulas; véase §28.8), con un punto extremo pasado p que se extiende indefinidamente hacia el futuro. Técnicamente, la condición requerida es satisfecha si cada ℓ semejante contiene un punto q para el que hay una curva estrictamente de género tiempo en el futuro desde p a q. En los modelos $\mathcal{M} \times \mathcal{Y}$ exactos recién descritos la condición falla (como debe ser, puesto que $\mathcal{M} \times \mathcal{Y}$ puede ser no singular), pero solo por poco. Lo que sucede esencialmente es que entre la familia 8-dimensional de rayos de luz ℓ hay solo una subfamilia 2-dimensional que falla al entrar en la «parte $\mathcal{Y}$» del espaciotiempo y volver, enroscándose así en el interior de $\mathcal{C}$. Puede demostrarse que, con una perturbación genérica pero pequeña encontrada por $\mathcal{C}$, esta propiedad se destruirá, y el teorema de singularidad que se ha mencionado antes se aplicará. Los detalles de este argumento se presentarán en otro lugar.

31.47. Véase, por ejemplo, Minassian (2002), que remite a un campo de investigación relevante.

Sección 31.13

31.48. Véanse Smolin (2003) y Nicolai (2003).

31.49. Véanse Smolin (2003), Gross y Periwal (1988) y Nicolai (2003).

31.50. La serie $1 + 2^2 + 2^4 + 2^6 + 2^8 + \ldots$ no es Borel sumable, incluso si el valor «euleriano» $-\frac{1}{3}$ para su suma es inequívoco, como se puede ver utilizando la continuación analítica (§7.4). No sé si tales procedimientos se han aplicado a las amplitudes de cuerdas totales.

Sección 31.14

31.51. Este comentario no se aplica a las cuerdas heteróticas, a las que llegaremos enseguida, para las que el marco de cuerdas básico es ya quiral.

31.52. La referencia más reciente es Gross *et al.* (2003). Smolin (2003) ofrece más referencias para estas variedades en la teoría de cuerdas; Polchinski (1998) también las discute.

31.53. Tengo cierta dificultad con esto, puesto que los campos espinoriales tienen realmente una interpretación geométrica. No pueden ser «rotados» (y, por consiguiente, no pueden ser «calibrados», estrictamente hablando; véanse §§15.2,7) sin que esto se aplique al propio espacio ambiente; véase Penrose y Rindler (1984).

31.54. Aplicando una «rotación de Wick» para obtener una superficie de Riemann, la distinción es entre holomorfa y antiholomorfa, como se ha mencionado en §31.13.

31.55. Greene (1999); Smolin (2003) da una lista de prácticamente todas las dualidades conocidas, su estatus y referencias.

31.56. La noción de «simetría espejo» es completamente diferente de la de simetría de reflexión espacial (paridad), denotada por P, que se ha expuesto en §25.4.

31.57. Véase Cox y Katz (1999), que ofrece una excelente cobertura de tales ideas.

31.58. Véanse, por ejemplo, Kontsevich (1994) y Strominger *et al.* (1996); para algunos de los estudios más recientes, véase Yui y Lewis (2003).

31.59. Estas variedades particulares son 3-superficies complejas llamadas «quínticas», lo que significa que tienen «orden 5». El orden de una n-superficie es el número de puntos en que corta a un plano complejo general.

31.60. Para el «orden» de una curva compleja, véase la nota 31.59 anterior. Aquí $n = 1$.

31.61. Véanse Cox y Katz (1999), Candelas *et al.* (1991) y Kontsevich (1995).
31.62. Véanse Smolin (2003), en particular la referencia 171, y Witten (1995); para una exposición divulgativa, Greene (1999), p. 203.
31.63. Véanse Vafa (1996), o Bars (2000).
31.64. Véase Bryant *et al.* (1991), nota 31.37.

Sección 31.15

31.65. Para un argumento sugerente, sin embargo, véase Thorne (1986).
31.66. Véase Strominger y Vafa (1996).
31.67. Véase Greene (1999), p. 340.
31.68. Al lector podría preocuparle cómo un campo de Maxwell libre de fuentes pueda conducir a una carga no nula. No hay aquí inconsistencia, puesto que el agujero negro podría haber aparecido por el colapso de un cuerpo de material cargado, desapareciendo todas las fuentes cargadas dentro del agujero.
31.69. Para una revisión fácilmente legible de estos temas, véase Horowitz (1998).
31.70. La cuenta incluye estructuras denominadas «D-branas» que consideraremos en §31.17.
31.71. Véase la nota 22.11.
31.72. Me lo señaló Abhay Ashtekar.

Sección 31.16

31.73. Véase Kasper y Feller (2001) para una referencia sobre hologramas «reales».
31.74. Este es un conjunto de ideas bastante difícil. Para un reto real, véanse Maldacena (1997) y Witten (1998).
31.75. Gary Gibbons ha señalado una intrigante geometría relacionada con esta imagen que incluso parece tener conexiones con la teoría de twistores. Varios temas de relevancia para esta construcción pueden encontrarse en Penrose (1968a).
31.76. Einstein (1917).

Sección 31.17

31.77. Véase Ashtekar y Das (2000) para un ejemplo de este fenómeno.
31.78. Un «espacio cociente» es como un espacio base de un fibrado; véanse §§15.1,2.
31.79. Véase Randall y Sundrum (1999a); véase también Randall y Sundrum

(1999b) para ideas más generales sobre estas cuestiones. Johnson (2003) es la referencia estándar sobre la «tecnología» de D-branas. Una de las aplicaciones más fantásticas de esta tecnología ha sido el modelo «ekpirótico» del origen del universo, propuesto por Steinhardt y Turok (2002), en el que se propone que el big bang surgió de la colisión de dos D-branas en una fase previa del universo. A pesar de invocar elementos tan exóticos, los autores de este modelo no intentan explicar el rompecabezas principal que presenta el big bang, a saber, su carácter extraordinariamente especial, como se ha descrito en 27.13.

Sección 31.18

31.80. Véase la nota 31.74.
31.81. Véanse Nair (1998), Witten (2003), Cachazo *et al.* (2004a, 2004b, 2004c) y Brandhuber *et al.* (2004).
31.82. Véanse las notas 31.37 y 31.58.
31.83. Esta cita es de la alocución de Heisenberg en la Sociedad Alemana de Física en 1975: «¿Qué es una partícula elemental?». (Estoy especialmente agradecido a Abhay Ashtekar por este ejemplo.) Véase Heisenberg (1989).

32

El sendero más estrecho de Einstein; variables de lazo

32.1. Gravedad cuántica canónica

A pesar de la popularidad de la teoría de cuerdas, sería absurdo adoptar el punto de vista, tal como han hecho algunos,[1] de que es «el único juego en la ciudad» (véase §31.8). Se están siguiendo muchas otras ideas interesantes, que tienen diferentes virtudes y diferentes dificultades. Por desgracia, no puedo entrar en una discusión de muchas de estas ideas alternativas para unir la teoría cuántica con la estructura del espaciotiempo. En lugar de ello, en este capítulo y en el siguiente me concentraré en algunas de las áreas más activas que están más cerca de mi pensamiento respecto a las líneas que van a ser probablemente fructíferas en la búsqueda de la verdadera unión entre la relatividad general y la mecánica cuántica. Como puede deducirse de mis comentarios en el capítulo anterior, soy de la opinión de que tendremos que adoptar una postura más controlada que aquellas que permiten un crecimiento en la dimensionalidad del espaciotiempo o se aventuran en la supersimetría (aunque estoy menos en contra de la última que de la primera, que, como hemos observado en §§31.11,12, encuentra serios problemas de estabilidad). Por ello, en los dos capítulos siguientes veremos algunas ideas que se relacionan concretamente con un espaciotiempo lorentziano 4-dimensional, y donde se trata de abordar la ecuación de campo de Einstein,[2] sin supersimetría, en un contexto genuinamente cuántico. Veremos que incluso aquí las «imágenes de la realidad física» que encontramos siguen estando muy alejadas de lo que nos es familiar; no hasta el grado que hemos visto en el capítulo anterior en algunos aspectos, pero sí más alejadas en otros. En este capítulo seremos testigos de algunas de las ideas que hay tras las variables de Ashtekar, las variables de lazo y las redes de espín. En el capítulo siguiente nos familiarizaremos con la teoría de twistores. También se mencionarán en estos dos capítulos otras ideas que han ganado aceptación, en particular el espaciotiempo discreto, las estructuras q-deformadas («grupos cuánticos») y la geometría no conmutativa.

Una de las maneras más directas de enfocar la cuantización de la teoría de Einstein es ponerla de una forma hamiltoniana y tratar de aplicar entonces los procedimientos de cuantización canónica que se describieron en §§21.2,3. Hay muchas dificultades en esto y no quiero entrar en detalles. Muchas de ellas derivan del hecho de que la teoría de Einstein tiene «covariancia general» (§19.6), de modo que las coordenadas concretas que se utilicen no tienen importancia. Recordemos de la discusión de §21.2 que la «receta de cuantización» estándar por la que el momento p_a es reemplazado por el operador $i\hbar\partial/\partial x^a$, donde x^a es la variable de posición conjugada (clásicamente), no es siempre correcta, ni siquiera en el espaciotiempo plano si utilizamos coordenadas curvilíneas. Así pues, hay que tener muchísimo cuidado cuando se lleva a cabo este tipo de procedimiento de cuantización.

Otra dificultad es la complicada estructura no polinómica que se encuentra para el hamiltoniano estándar para la relatividad general. También deberíamos tomar nota del hecho de que además de tener ecuaciones de evolución que nos apartan de una 3-superficie inicial $\mathcal{S}$ de género espacio, que están gobernadas por el hamiltoniano, existen otras ecuaciones que actúan *dentro de* $\mathcal{S}$, que se denominan *ligaduras*.[3] Estas nos proporcionan ecuaciones de compatibilidad para los datos *en* $\mathcal{S}$, y la satisfacción de las ecuaciones de ligadura es necesaria (y suficiente) para una evolución satisfactoria de los datos *fuera de* $\mathcal{S}$ (al menos localmente) y esta evolución conserva la satisfacción de las ligaduras.

El enfoque canónico para cuantificar la relatividad general tiene una historia larga y distinguida, que se remonta a Dirac en 1932, que tuvo que desarrollar un marco de cuantización completamente nuevo para manejar las ligaduras complicadas que se dan en la teoría de Einstein.[4] Durante muchos años este tipo de enfoque fue seguido por varios investigadores diferentes, cada vez con mayor sofisticación,[5] pero la naturaleza no polinómica y complicada del hamiltoniano hacía que los

progresos resultasen difíciles. Más tarde, en 1986, el físico indo-estadounidense Abhay Ashtekar hizo un avance importante. Mediante una sutil elección de las variables utilizadas en la teoría (relacionada en parte con ideas que habían sido propuestas antes por Amitabha Sen),[6] con las que las ligaduras podían reducirse a la forma polinómica, él fue capaz de simplificar de manera espectacular la estructura de las ecuaciones y se eliminaron los difíciles denominadores en el hamiltoniano, lo que llevó a una estructura polinómica relativamente simple.

32.2. El ingrediente quiral de las variables de Ashtekar

Una de las características sorprendentes de las «nuevas variables» originales de Ashtekar, tal como (aún) se las llama, es que son asimétricas con respecto al tratamiento de las partes dextrógira y levógira del gravitón (el cuanto de gravitación).[7] Recordemos de §§22.7,9 que una partícula sin masa (no escalar) tiene dos estados de espín, que puede ser dextrógiro o levógiro respecto a su dirección de movimiento. Se les conoce, respectivamente, como los estados de *helicidad* positiva y negativa de la partícula. El gravitón tiene que ser una partícula de espín 2, de modo que sus dos estados respectivos de helicidad deberían ser $s = 2$ y $s = -2$ (tomando $\hbar = 1$), donde «s» representa la helicidad (véase también §22.12). El enfoque original de Ashtekar trata estos dos estados de forma *diferente*. Así pues, ¡el formalismo es asimétrico respecto a la izquierda/derecha!

Aquí hay que hacer un comentario respecto a por qué el gravitón debe considerarse una entidad de espín 2, mientras que el fotón tiene espín 1 (véanse §22.7 y §33.7). ¿Qué significa esto? El valor del espín de una partícula cuántica tiene que ver con las simetrías (y ecuaciones de campo) de la cantidad de campo que la describe (y, como veremos en §33.8, es más manifiesta con las ecuaciones escritas en forma 2-espinorial). Pero es bueno tener una forma directamente *geométrica* de ver la diferencia entre el carácter de espín 2 de la gravedad frente al carácter de espín 1 del electromagnetismo. Examinemos las ondas apropiadas para cada campo, a saber, las ondas electromagnéticas que constituyen la luz, por una parte, y las ondas gravitatorias por la otra.

En el caso del electromagnetismo, hemos visto la naturaleza geométrica de las ondas en la Fig. 22.12 y en §22.9. El punto clave es que los vectores eléctrico y magnético son en realidad magnitudes vectoriales, de modo que una rotación de la onda un ángulo π (i.e., 180°) alrededor de su dirección de movimiento transforma la magnitud del campo en su *negativa*, y necesitamos una rotación de 2π para restituirla. En el caso de la gravedad, se trataría de una onda de distorsión espaciotemporal, como se ha ilustrado en las Figs. 17.8a, 17.9a, 28.15 y 28.16. Ahora, si rotamos la onda solo un ángulo π, la distorsión coincide consigo misma, y sería una rotación de $\frac{1}{2}\pi$ la que la transformaría en su negativa. Podemos señalar que esto es resultado de que la curvatura de Weyl es una entidad *cuadripolar*, como se ha ilustrado por las elipses de distorsión en la Fig. 28.15, y en §31.9 advertimos que esto corresponde a una entidad de espín 2. Para un valor de espín σ, una rotación de ángulo π/σ alrededor de la dirección de la onda transformaría la magnitud de campo en su negativa, mientras que sería necesaria una rotación de $2\pi/\sigma$ para restituirla. (Nótese que esto también funciona si σ es semiimpar, y la magnitud de campo es necesariamente espinorial en dicho caso; véase §11.3.)

Para un campo sin masa, como es la situación aquí, podemos ir más lejos y considerar que las ondas planas están compuestas de partes circularmente polarizadas levógira y dextrógira, estando ilustrada la polarización circular electromagnética en la Fig. 22.12b. Para un campo cuantizado, las partículas relevantes tendrían en correspondencia helicidad positiva o negativa (Fig. 22.13). Para el espín σ, estos valores de helicidad serían $\pm\sigma$ (con una descripción correspondiente a la Fig. 22.13, pero con $q^{2\sigma} = z/w$ en lugar de $q^2 = z/w$). Así pues, para la gravedad obtenemos los dos valores de helicidad posibles +2 y −2.

Para ver cómo se describen los estados de helicidad tendremos que examinar las matemáticas de una forma un poco más explícita. De hecho, esta asimetría izquierda/derecha a la que me he estado refiriendo es una característica importante de la teoría de twistores que examinaremos en el capítulo siguiente, y el enfoque de Ashtekar parece haber obtenido parte de su inspiración de las ideas twistoriales. Por el mo-

mento, todo lo que necesitaremos de ideas relacionadas con los twistores es la forma de expresar matemáticamente esta asimetría izquierda/derecha en términos espaciotemporales ordinarios. Recordemos las dos cantidades tensoriales que describen los dos campos sin masa conocidos de la naturaleza, el electromagnetismo y la gravitación. Estas cantidades son el tensor de campo de Maxwell $\boldsymbol{F} = F_{ab}$ (§19.2) y el tensor conforme de Weyl $\boldsymbol{C} = C_{abcd}$ (§19.7). Cada uno de ellos tiene lo que se denomina un tensor *dual*, definido en la notación de índices por

$$*F_{ab} = \frac{1}{2}\varepsilon_{abpq}F^{pq} \text{ y } *C_{abcd} = \frac{1}{2}\varepsilon_{abpq}C^{pq}{}_{cd}$$

(donde ε_{abpq} es el tensor de Levi-Civita antisimétrico, escogido aquí de modo que $\varepsilon_{0123} = 1$, en una base ortonormal dextrógira estándar; véanse §12.7 y §19.2). Ya hemos visto el tensor de Maxwell dual $*\boldsymbol{F}$ en §19.2. El tensor de Weyl dual $*\boldsymbol{C}$ se ve como una cantidad análoga. También podríamos pensar en «dualizar» en su lugar el par restante *cd* de los índices del tensor de Weyl. Pero resulta que esto es lo mismo que dualizar en *ab*.[32.1]

Recordemos que un elemento 2-plano en un punto en un espaciotiempo 4-dimensional puede describirse mediante una 2-forma $\boldsymbol{f}$ (o si no, un bivector) que es *simple* (§12.7). Como hacíamos con el tensor de Maxwell (2-forma) $\boldsymbol{F}$, podemos construir su dual $*\boldsymbol{f}$ y dar sentido a la noción de que, para un $\boldsymbol{f}$ complejo, podría ser *autodual*, i.e., $*\boldsymbol{f} = \mathrm{i}\boldsymbol{f}$, o *anti-autodual*, i.e., $*\boldsymbol{f} = -\mathrm{i}\boldsymbol{f}$. Por consiguiente, el elemento 2-plano (complejo) correspondiente a $\boldsymbol{f}$ se denomina «autodual» o «anti-autodual». Esta noción tiene gran importancia en la teoría de twistores (§33.6).[32.2]

Ahora bien, en la teoría cuántica se permite que las cantidades de campo tomen valores complejos, al menos cuando tienen la interpretación de funciones de onda. De hecho, hay varias formas diferentes, pero matemáticamente equivalentes, de considerar estas cosas (véase el capítulo 26). Será más apropiado para nuestros propósitos presentes si pensamos en un tensor de Maxwell complejo, o incluso en un

[32.1] Explique por qué. (Puede encontrar útiles la notación diagramática y las identidades de §12.8.)

[32.2] Demuestre que si los dos índices de f_{ab}, que describen una $\boldsymbol{f}$ autodual, se contraen con cualquier par de índices de un tensor de Weyl (o de Maxwell) anti-autodual, entonces el resultado es cero.

tensor de Weyl complejo, como algo que representa alguna forma de función de onda para el fotón o el gravitón, respectivamente. En lugar de la condición de realidad que es característica de una cantidad de campo clásica, nuestras funciones de onda complejas deben ser de frecuencia positiva (de acuerdo con los requisitos especificados en §24.3 y §26.2). No nos preocuparemos demasiado por lo que *significa* esto en realidad en el caso de la curvatura de Weyl. (Provisionalmente, podemos considerar que estamos examinando espacios curvos que solo difieren infinitesimalmente de los planos, en cuyo caso $\boldsymbol{C}$ podría considerarse como un campo en el espacio de Minkowski, y entonces la condición de frecuencia positiva no es problemática. Pero resulta que se puede hacer algo mejor que esto, como se indicará en §§33.10-12.[8] Ahora los fotones y gravitones *dextrógiros* se describen por las cantidades *autoduales* (de frecuencia positiva) ${}^{+}\boldsymbol{F}$ y ${}^{+}\boldsymbol{C}$, donde

${}^{+}\boldsymbol{F} = \frac{1}{2}(\boldsymbol{F} - \mathrm{i}*\boldsymbol{F})$ y ${}^{+}\boldsymbol{C} = \frac{1}{2}(\boldsymbol{C} - \mathrm{i}*\boldsymbol{C})$, de modo que tenemos

$$*({}^{+}\boldsymbol{F}) = \mathrm{i}{}^{+}\boldsymbol{F} \text{ y } *({}^{+}\boldsymbol{C}) = \mathrm{i}{}^{+}\boldsymbol{C},$$

y los levógiros por las cantidades *autoduales* (de frecuencia positiva) ${}^{-}\boldsymbol{F} = \frac{1}{2}(\boldsymbol{F} + \mathrm{i}*\boldsymbol{F})$ y ${}^{-}\boldsymbol{C} = \frac{1}{2}(\boldsymbol{C} + \mathrm{i}*\boldsymbol{C})$, para las que

$$*({}^{-}\boldsymbol{F}) = -\mathrm{i}{}^{-}\boldsymbol{F} \text{ y } *({}^{-}\boldsymbol{C}) = -\mathrm{i}{}^{-}\boldsymbol{C}.$$

En el marco original de Ashtekar, las partes autodual y anti-autodual de la curvatura de Weyl desempeñan papeles diferentes.

Esto puede parecer extraño desde el punto de vista físico, porque no hay evidencia alguna de asimetría izquierda/derecha en el campo gravitatorio, y por supuesto no hay ninguna en la teoría estándar de la relatividad general de Einstein. Por lo que puedo ver, podemos adoptar dos actitudes diferentes respecto a esta cuestión. Por una parte, podríamos considerar la asimetría como una característica sin importancia de las matemáticas concretas que solo resulta ser útil para simplificar el hamiltoniano. Por otra parte, podríamos adoptar la posición de que hay algo profundamente asimétrico izquierda/derecha en la naturaleza, y el formalismo asimétrico está realmente sondeando esto de alguna forma. De hecho, sabemos que la naturaleza *es* asimétrica izquier-

da/derecha, como se manifiesta claramente en las interacciones débiles (véase §25.3). En cierto sentido, el electromagnetismo contiene restos de esta asimetría, pero esto solo se ve indirectamente a través de su unificación con las interacciones débiles en la teoría *electrodébil*. En ausencia de una unificación similar (conocida) que incluya la gravedad, no hay ninguna razón para esperar que la propia gravedad debiera poseer directa o indirectamente tales rasgos asimétricos. Pese a todo, como sucede con la perspectiva que tienen los teóricos de cuerdas de su propia teoría, podríamos adoptar la idea de que una teoría de la gravedad cuántica apunta a algo que es mucho más que la mera gravitación; está poniendo el marco básico para toda la física, donde el marco actual del espaciotiempo clásico va a verse como una conveniencia o aproximación a algo más fundamental. Si ese «algo» tiene una quiralidad intrínseca (como lo tienen los aspectos heteróticos de la teoría de cuerdas —véase §31.14—, así como los marcos de Ashtekar o de twistores), entonces las asimetrías de las interacciones débiles se entienden mucho más fácilmente.

32.3. La forma de las variables de Ashtekar

¿Cuáles son entonces estas variables quirales de Ashtekar? La quiralidad procede de una elección asimétrica de uno de los dos tipos de 2-espinores que se aplican a un 4-espacio lorentziano. Podemos recordar estos objetos de §25.2, donde hemos visto que la función de onda del electrón ψ podía considerarse consistente en un par de entidades 2-complejo componentes α_A y $\beta_{A'}$, una de ellas con un índice no primado y la otra con un índice primado. Notemos que el índice no primado se refiere a la parte *zig* o *levógira* (helicidad negativa) del electrón, y el índice primado a la parte *zag* o *dextrógira* (helicidad positiva). Hemos visto también (§25.3) que la interacción débil presta atención a la parte zig α_A, pero no a la parte zag $\beta_{A'}$. Un formalismo espaciotemporal que seleccione o bien el espinor no primado, o bien el primado como «más fundamental» que el otro, incorpora así una quiralidad básica, y establece un marco que tiene la capacidad de distinguir estas dos helicidades en un nivel fundamental.

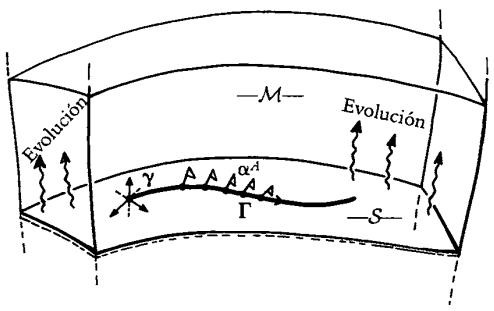


Fig. 32.1. Las variables canónicas originales de Ashtekar, definidas en una 3-superficie $\mathcal{S}$ de género espacio en el espaciotiempo $\mathcal{M}$, toman como parámetros de «posición» $\boldsymbol{\Gamma}$ las componentes de la conexión de espín 4-espacial, restringida a $\mathcal{S}$ (para espinores α_A, indicados por semiflechas). Los parámetros de «momento» son básicamente las componentes de la métrica intrínseca (inversa) $\boldsymbol{\gamma}$ de $\mathcal{S}$ (expresados como 2-formas y referidos a una base ortonormal en cada punto de $\mathcal{S}$).

De hecho, esto es precisamente lo que sucede en los formalismos de Ashtekar (y twistorial). En el enfoque de Ashtekar, las *variables canónicas*, escogidas con respecto a una 3-superficie de género espacio $\mathcal{S}$, son básicamente las componentes de la 3-métrica (inversa) $\boldsymbol{\gamma}$ *intrínseca* a $\mathcal{S}$, y las componentes de la *conexión de espín* (no primada) $\boldsymbol{\Gamma}$ tomadas en $\mathcal{S}$. Para ser algo más precisos, estas son componentes métricas inversas referidas a una base espinorial local y representadas como 2-formas. Además, la conexión de espín $\boldsymbol{\Gamma}$ se refiere al transporte paralelo de espinores α^A definidos en el 4-espacio completo y no solo a «espinores intrínsecos a $\mathcal{S}$».[9] Así pues, $\boldsymbol{\Gamma}$ nos dice cómo transportar un espinor 4-espacial no primado α^A (un 2-espinor) paralelo a sí mismo con respecto a la conexión métrica del 4-espacio (§§14.2,8) a lo largo de cierta curva que resulta estar dentro del 3-espacio $\mathcal{S}$.[10] El campo de cantidades (densidad) 3-métricas $\boldsymbol{\gamma}$ desempeña el papel del conjunto de variables de *momento*, y el campo de cantidades de conexión $\boldsymbol{\Gamma}$ desempeña el papel de las correspondientes *posiciones* conjugadas; véase la Fig. 32.1. Al pasar a la teoría cuántica, al reemplazamiento del momento p_a por $i\hbar\partial/\partial x^a$ en la x^a-representación (§21.2), le corresponde el reemplazamiento de los campos $\boldsymbol{\gamma}$ por $i\hbar\delta/\delta\boldsymbol{\Gamma}$ en la $\boldsymbol{\Gamma}$-representación (donde el «$\delta/\delta\boldsymbol{\Gamma}$» se refiere a la noción de *derivada funcional* que se ha men-

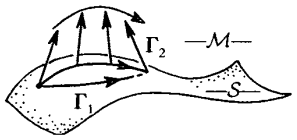


Fig. 32.2. Podemos expresar $\boldsymbol{\Gamma} = \boldsymbol{\Gamma}_1 + i\boldsymbol{\Gamma}_2$, donde $\boldsymbol{\Gamma}_1$ se refiere a la curvatura intrínseca y $\boldsymbol{\Gamma}^2$ a la curvatura extrínseca de $\mathcal{S}$ (de modo que $\boldsymbol{\Gamma}_2$ mide cómo se «curva» $\mathcal{S}$ dentro de $\mathcal{M}$.

cionado en §20.5). Correspondientemente, en la $\boldsymbol{\gamma}$-representación $\boldsymbol{\Gamma}$ viene representada por $-i\hbar\delta/\delta\boldsymbol{\gamma}$.

La conexión $\boldsymbol{\Gamma}$ tiene una parte $\boldsymbol{\Gamma}_1$ que se refiere solo a la curvatura *intrínseca* de $\mathcal{S}$ y otra parte $\boldsymbol{\Gamma}_2$ que se refiere a la curvatura *extrínseca* (i.e., a cómo se «dobla» $\mathcal{S}$ dentro del espaciotiempo $\mathcal{M}$ (véase la Fig. 32.2). La cantidad total $\boldsymbol{\Gamma}$ puede expresarse como

$$\boldsymbol{\Gamma} = \boldsymbol{\Gamma}_1 + i\boldsymbol{\Gamma}_2$$

(donde habríamos tenido $\boldsymbol{\Gamma} = \boldsymbol{\Gamma}_1 - i\boldsymbol{\Gamma}_2$ si hubiéramos escogido la quiralidad opuesta para el formalismo). La cantidad $\boldsymbol{\Gamma}$ define una conexión fibrada (en el sentido de §15.8), donde el espacio base es $\mathcal{S}$ y la fibra es el espacio de espín (no primado) $\mathbb{S}$ (un espacio vectorial complejo 2-dimensional). El grupo relevante de la fibra es SL(2,$\mathbb{C}$) (véase §13.3).[11]

En este punto debería mencionar una dificultad técnica en el enfoque original de Ashtekar. Esta surge del hecho de que el grupo de la fibra SL(2,$\mathbb{C}$) es no compacto y tiene representaciones irreducibles no deseadas de dimensión infinita, la mayoría de las cuales son no unitarias (véase §13.7). Todo esto crea graves problemas para la construcción de la teoría de la gravedad cuántica requerida. Por ello, para hacer progresos se ha utilizado la conexión modificada

$$\boldsymbol{\Gamma}_\eta = \boldsymbol{\Gamma}_1 + \eta\boldsymbol{\Gamma}_2$$

donde η es un número real no nulo, conocido como el parámetro de Barbero-Immirzi. Esto tiene la ventaja puramente técnica de que el grupo es ahora SU(2), que es compacto, y todas sus representaciones (irreducibles) son de dimensión finita y unitarias. La teoría *clásica* definida por cada $\boldsymbol{\Gamma}_\eta$ difiere de la definida por $\boldsymbol{\Gamma}$ solo en lo que se denomina una «transformación canónica», lo que significa que las teorías clásicas son equivalentes (al tener la misma estructura simpléctica; véanse §14.8 y §20.4), aunque están descritas por «coordenadas generalizadas» diferentes en el espacio de fases (§20.2). Sin embargo, las teorías cuánticas resultantes no tienen por qué ser equivalentes. Esta es la cuestión planteada en §21.2: el proceso de cuantización no es normalmente invariante bajo cambios de las coordenadas generalizadas. La cuestión de cuánto «daño» podría hacer el reemplazamiento de $\boldsymbol{\Gamma}$ por $\boldsymbol{\Gamma}_\eta$ parece ser una cuestión no resuelta. En cualquier caso, se puede aprender mucho estudiando primero el caso «más fácil» de $\boldsymbol{\Gamma}_\eta$. Aunque uno podría preocuparse porque la teoría resultante es solo un «modelo de juguete» en lugar de la pretendida aproximación a la gravedad cuántica (donde valores de η diferentes de $\pm i$ o 0 no parecen tener justificación física), se supone que la desviación de la pretendida versión cuantizada de la teoría de Einstein quizá no sea grande.

En el caso de $\boldsymbol{\Gamma}_\eta$, las cosas son relativamente simples, porque las diferentes representaciones irreducibles necesarias de SU(2) son matemáticamente idénticas a los diferentes estados de *espín* (de una partícula masiva) en la mecánica cuántica ordinaria (no relativista). Recordemos de §22.8 que estos espines diferentes se etiquetan mediante los números naturales $n = 0, 1, 2, 3, 4, 5, \ldots$, donde $\frac{1}{2}n\hbar$ es el valor del espín. (En §22.8 hemos visto que el espacio de representación para cada n consiste en tensores de espín simétricos $\psi_{AB\ldots D}$ con n índices.) Veremos enseguida cómo hay que utilizar estos diferentes «valores de espín».

32.4. Variables de lazo

¿Cómo expresar las cosas de un modo que haga más manifiesta la covariancia general (§19.6), al menos dentro de la 3-superficie inicial $\mathcal{S}$? Esto se hace mediante el ingenioso artificio de describir nuestro estado cuántico general en términos de una familia particularmente simple de estados de gravedad cuántica *básicos* que pueden describirse de una forma esencialmente discreta y para la que la covariancia general dentro de $\mathcal{S}$ se trata de forma muy simple. (Llegaremos a estos estados básicos enseguida.) El estado *general* se expresa entonces en términos de tales estados básicos por medio de la superposición lineal. Para entender lo que se necesita, consideremos un lazo cerrado dentro de $\mathcal{S}$, e

imaginemos el efecto de utilizar nuestra conexión Γ para permitirnos transportar un espinor no primado α^A «paralelo a sí mismo» alrededor de dicho lazo. Cuando volvemos al punto de partida, encontramos que se ha efectuado una transformación lineal del espacio de espín $\mathbb{S}$. Esta se define en forma de componentes por una matriz (2×2) compleja $T^A{}_B$ (véase §13.3), y los elementos de esta matriz dependen de cierta elección de base en $\mathcal{S}$. Sin embargo, la *traza* $T^A{}_A$ de esta matriz es un número complejo independiente de la base,[32.3] de modo que es simplemente una propiedad de la conexión de espín Γ relacionada con la elección del lazo. (Este es un ejemplo de una noción más general, conocida como *lazo de Wilson*, por Kenneth Wilson, que utilizó esta idea por primera vez en las teorías gauge.[12] En 1988, Carlo Rovelli, Lee Smolin y Ted Jacobson desarrollaron esta idea en la relatividad general, llamando a estas trazas dependientes del lazo *variables de lazo* para la relatividad general. Tomando estas variables de lazo como operadores cuánticos, los «estados básicos» mencionados al principio de este párrafo son esencialmente sus *autoestados*.

¿Cuál es el carácter geométrico de estos estados básicos de gravedad cuántica? Resulta que son muy peculiares desde el punto de vista de la geometría métrica ordinaria con la que estamos familiarizados, y están muy lejos de las «geometrías suaves» de la relatividad general clásica. De hecho, veremos que estos estados básicos son muy «singulares» como estados geométricos, análogos a las *funciones delta* (de Dirac) que hemos visto en §9.7 y §21.10. En primer lugar, consideremos $\mathcal{S}$ como una variedad sin rasgos distintivos.[13] A continuación, consideremos una familia de lazos cerrados en $\mathcal{S}$. Hay que tener en cuenta que cada estado de lazo tiene toda su geometría concentrada de algún modo a lo largo del lazo. No es realmente la *curvatura* la que está concentrada a lo largo del lazo.Esto sería análogo a la geometría del cono de base plana ilustrado en la Fig. 32.3, donde hay una función delta (§9.7) en la curvatura a lo largo del borde de la base del cono, y también en el vértice,[32.4] que es el tipo de situación que aparece en una aproximación diferente a la gravedad cuántica, conocido como cálculo de

[32.3] ¿Puede explicar por qué?
[32.4] ¿Puede explicarlo? *Sugerencia*: Utilice las ideas de §14.5.

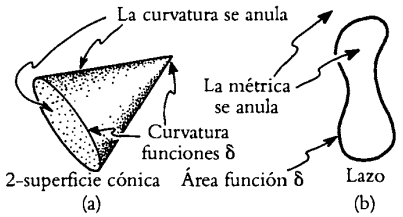


Fig. 32.3. (a) Este ejemplo no suave de una 2-superficie cónica tiene curvatura cero en todo lugar excepto en el vértice y en el borde de la base, donde hay funciones δ para la curvatura. (La aproximación del cálculo de Regge para la gravedad cuántica opera con 4-espacios análogos, con funciones δ para la curvatura en 2-espacios; véase la Fig. 33.3.) (b) Sin embargo, no es esto lo que sucede con la gravedad cuántica de lazo. Aquí hay una función δ para el área a lo largo de los lazos, y la propia métrica se anula en todos los demás lugares.

Regge; véase §33.1. En su lugar, es la *métrica* entera la que debe estar concentrada a lo largo del lazo en un tipo de función delta y ser nula fuera del lazo. Hay un «grado» de esta concentración que se mide por un valor de «espín» que se asigna al lazo, correspondiendo los diferentes valores de espín $j = \frac{1}{2}n$ a diferentes representaciones irreducibles de SU(2) (donde utilizamos Γ_η para nuestra conexión, en lugar de la aparentemente más «correcta» Γ).

Estas afirmaciones necesitan una clarificación adicional. La noción de «métrica» que aparece aquí es realmente la que asigna un *área* a cualquier elemento de 2-superficie de prueba que corta el lazo. En esencia, hay una función delta en la medida del área que está concentrada por entero a lo largo de cada lazo de la familia. Imaginemos una superficie 2-dimensional $\mathcal{T}$ (no necesariamente cerrada) en la 3-superficie $\mathcal{S}$. Esta puede intersectar varios lazos en varios lugares. Cada vez que $\mathcal{T}$ corta uno de estos lazos, hace cierta medida de área; no hay contribución a su área *excepto* donde corta el lazo. Así pues, el carácter de «función delta» de la métrica se manifiesta aquí en el hecho de que cada lazo asigna una medida de área solo donde $\mathcal{T}$ intersecta el lazo. Cada punto de intersección proporciona el valor

$$8\pi G\eta\hbar\sqrt{j(j+1)},$$

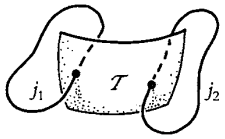


Fig. 32.4. Superficie test 2-dimensional $\mathcal{T}$ en 3-superficie $\mathcal{S}$. Cada intersección de $\mathcal{T}$ con un lazo contribuye con un valor $8\pi G\eta\hbar\sqrt{j(j+1)}$ al área (siendo j el valor de «espín» del lazo).

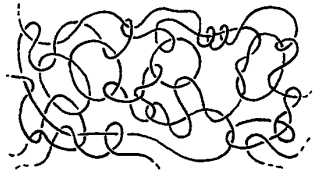


Fig. 32.5. La descripción de variable de lazo original de un espaciotiempo aproximadamente clásico podría presentarse como una superposición de difusiones casi uniforme de «entramados».

donde $j = \frac{1}{2}n$ es el valor de «espín» del lazo concreto; véase la Fig. 32.4. Sumamos estas contribuciones de área procedentes de todos los lazos de la familia.

Hay un contraste interesante entre la teoría de cuerdas y la teoría de variables de lazo. Mientras que la teoría de cuerdas es una aproximación casi enteramente *perturbativa* a la gravedad cuántica, la aproximación de variables de lazo es fundamentalmente no perturbativa. En la teoría de cuerdas los cálculos se realizan casi invariablemente en un fondo que es el espaciotiempo *plano*, i.e., un producto de un espacio de Minkowski $\mathbb{M}$ por, digamos, algún 6-espacio de Calabi-Yau (véase §31.14), y solo estamos interesados en campos débiles en dicho fondo. La idea es que se contempla «perturbar» a partir de este límite de campo débil (i.e., considerar series de potencias en algún parámetro pequeño ε); véanse §26.10 y §31.9. Por el contrario, en el caso de las variables de lazo, los estados de lazo básicos (o estados de red de espín; véase §32.6) están muy lejos de ser planos (o clásicos), pues tienen *funciones delta* en la medida de área a lo largo de los lazos (o líneas de red de espín). Para obtener la descripción de variables de lazo de un espaciotiempo aproximadamente clásico, necesitamos considerar algo parecido a una difusión casi uniforme de «entramados», como se indica en la Fig. 32.5.

Nótese que esta es una descripción muy topológica. No supone ninguna diferencia cuán «cerca» podría estar un lazo de otro (puesto que la noción de «métrica» no tiene sentido fuera del propio lazo). Las

únicas cosas que tienen importancia son las relaciones topológicas de «enlazamiento» y «anudamiento» (o intersección) entre lazos, y los valores de «espín» discretos que se asignan a ellos. Así pues, se cuida muy bien de la covariancia general (dentro de $\mathcal{S}$) con tal de que retengamos meramente esta imagen topológica discreta.

32.5. Las matemáticas de nudos y enlaces

La imagen de variables de lazo de la gravedad cuántica nos lleva al campo de las matemáticas que se interesa en la topología de nudos y enlaces. Esta es una disciplina sorprendentemente sofisticada, considerando la naturaleza tópica de sus ingredientes: ¡básicamente desenredar cabos de cuerda! Tenemos que hacer uso de criterios matemáticos que están a nuestra disposición para decidir si un «lazo de cuerda» cerrado está o no realmente *anudado* (donde «anudado» significa que es imposible deformar el lazo hasta convertirlo en un círculo ordinario mediante movimientos suaves dentro del 3-espacio euclídeo ordinario, y donde no se permite que unos trozos del lazo atraviesen a otros; véase la Fig. 32.6). Análogamente, podemos pedir criterios que decidan si dos o más lazos distintos pueden separarse por completo unos de otros, de modo que están *desenlazados*. Desde principios del siglo XX se conocen varias expresiones matemáticas ingeniosas (tales como el «polinomio de Alexander») que proporcionan respuestas completas a esta cuestión, pero en años más recientes se han encontrado varios procedimientos

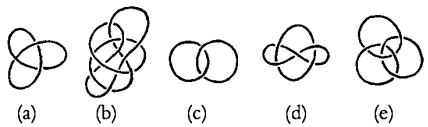


Fig. 32.6. Nudos y enlaces. (a) Un nudo trébol, un ejemplo de lazo anudado. (b) Un ejemplo de un lazo desanudado (que no es obvio para el ojo). (c) Un enlace simple entre dos lazos. (d) Un enlace de Whitehead, donde los dos lazos no pueden separarse aunque tienen «número de enlace» cero (el número neto de veces que cada uno de ellos corta una superficie que llena el otro). (e) Anillos borromeanos, que no pueden separarse, a pesar de que ningún par de ellos están enlazados.

Fig. 32.7. El tipo de álgebra tensorial diagramática de las Figs. 12.17, 12.18 y las Figs. 13.6-13.9, etc. puede generalizarse para producir un álgebra para nudos y enlaces. La característica adicional es que ahora supone diferencia el que una «línea de índice» pase por encima o por debajo de otra línea semejante cuando se cruzan.

fascinantes y más refinados que se inspiran básicamente en ideas procedentes de la física. Llevan nombres como «polinomio de Jones», «polinomios HOMFLY», «polinomio de Kauffmann», etc.[14]

Una forma de pensar en estas nuevas estructuras matemáticas es considerar que surgen de una especie de «álgebra diagramática», que es una generalización de la descripción diagramática del álgebra tensorial introducida en §12.8 (véanse las Figs. 12.17 y 12.18) y utilizada extensamente en el capítulo 13 (véanse las Figs. 13.6-13.19). En esta generalización, supone una diferencia si una «línea de índice» pasa por encima o por debajo de otra línea semejante, cuando quiera que se cruzan en un diagrama; véase la Fig. 32.7. Hay varias «identidades algebraicas» que pueden imponerse al álgebra, tal como la que se muestra en la Fig. 32.8, que sirve para especificar el álgebra de Kauffmann. Estas proporcionan elegantes generalizaciones del esquema combinatorio que subyace en la teoría de *redes de espín* a la que llegaremos enseguida.

La cantidad A en la Fig. 32.8 es un número complejo, escrito a veces en términos de una cantidad $q = A^2 = e^{i\pi/r}$. (El caso $A = -1$ da el «cálculo binorial» que subyace en la teoría de redes de espín; Penrose 1969b, 1971.) Hay análogos de los simetrizadores y antisimetrizadores

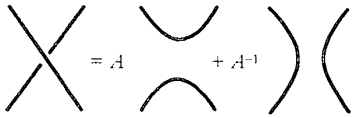


Fig. 32.8. Se ilustra la identidad algebraica básica para el álgebra de Kauffmann, donde $q = A^2 = e^{i\pi/r}$, lo que da una versión q-deformada del álgebra «binorial» subyacente en la teoría de redes de espín (para la que $A = -1$, y no se dan los cruces de la Fig. 32.7).

de la Fig. 12.17. Se ha desarrollado una notable teoría de tales objetos, conocidos a veces como estructuras *q-deformadas*. A menudo se utiliza el término «cuánticas», de forma más bien equívoca, en lugar de «q-deformadas», tal como sucede con la noción de un «grupo cuántico». No obstante, no hay una relación muy clara entre un «grupo cuántico» y la teoría cuántica, y la existencia de una aplicación importante de los grupos cuánticos a la física en el nivel fundamental, aunque posible, es de momento básicamente una conjetura.

A modo de aparte, vale la pena mencionar que existe otra conexión posible entre estas estructuras matemáticas recién encontradas (polinomios de Jones, etc.) y la física, que ha sido desarrollada en especial por Edward Witten.[15] Se trata de la noción de una *teoría cuántica de campos topológica*. En una teoría semejante, las ecuaciones de campos desaparecen por completo, pero sigue habiendo información en la estructura global y en «fallas» que puede considerarse que proporcionan «fuentes» para el campo (que se anula localmente). Un buen ejemplo sería la relatividad general en 1 + 2 dimensiones. En 1 + 2(= 3) dimensiones, el tensor de Weyl se anula idénticamente, de modo que toda la curvatura está en el tensor de Ricci. Así pues, en el «espacio vacío» (Ricci-planitud), la curvatura entera se anula. No obstante, el campo gravitatorio de una «fuente puntual» no es trivial, porque la fuente proporciona una «falla» que se manifiesta en la geometría global. Esto se ilustra en la Fig. 32.9. La geometría es muy parecida a la geometría de una cuerda cósmica, que se ilustra en la Fig. 28.4, excepto que aquí la imagen representa un espaciotiempo (1 + 2)-dimensional, en lugar del espacio 3-dimensional. Se ha eliminado un segmento, con eje a lo largo de la línea de universo (de género tiempo) de la fuente, y se han empalmado las dos fronteras planas resultantes. En la imagen clásica, las líneas de universo de tales fuentes tienen que ser rectas, pero para una QFT basada en este modelo clásico —una QFT *topológica*, puesto que el campo (aquí el campo de curvatura) se anula— permite líneas de fuentes curvadas y, de hecho, anudadas o enlazadas. Esto permite que se extraigan ideas[16] concernientes a las matemáticas de nudos y lazos utilizando la idea QFT topológica. Se advertirá que las variables de lazo proporcionan un sistema algo parecido al esquema general de la «QFT topológica», puesto que la contribución a la medida de área se anula excepto en las «fallas» que son

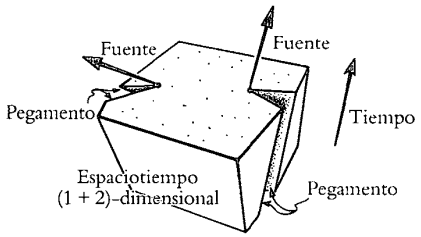


Fig. 32.9. La «relatividad general» en 2 + 1 dimensiones exige espaciotiempo plano donde quiera que no hay fuentes, puesto que el tensor de Ricci se anula allí, y el tensor de Weyl siempre se anula en 3 dimensiones. Pero una línea de universo fuente proporciona una «falla» en el espaciotiempo plano (una singularidad cónica) reminiscente de la «cuerda cósmica» cuya 3-geometría espacial se ilustra en la Fig. 28.4 y en §28.2. Las líneas de universo fuente son siempre rectas, clásicamente, pero ello se relaja en la versión cuántica de esta teoría, un ejemplo de una teoría cuántica de campos topológica.

los propios lazos. De todas formas, hay una diferencia, pues con las variables de lazo las ecuaciones de campo no desaparecen.

Aunque las QFT topológicas son interesantes como estructuras matemáticas, es difícil verlas desempeñando papeles directos como modelos de teorías físicas serias, debido a la completa desaparición de las ecuaciones de campo. La mayor parte de la física conocida depende de la no trivialidad de tales ecuaciones para que los campos se propaguen hacia el futuro de una manera controlada. No obstante, hay otra posibilidad de que las ideas de la QFT topológica puedan utilizarse en unión con la *teoría de twistores*. Como veremos en el capítulo 33 (al final de §33.11), en la descripción del espacio twistorial las ecuaciones de campos desaparecen localmente. Por el momento, la aplicación de ideas QFT topológicas a la teoría de twistores no se ha llevado muy lejos,[17] pero sería interesante ver lo que puede lograrse en esta área.

32.6. Redes de espín

Por notables que sean los estados de lazo, como configuraciones límite («funciones δ») de 3-geometría, todavía no proporcionan una *base*

(ortonormal) apropiada para esta geometría. Para ello es necesario pasar a una generalización en la que los lazos puedan cortarse. Esto nos lleva a considerar un tipo de red de «líneas de lazo intersecantes», pero tenemos que preguntar: ¿qué se encontrará en estos puntos de intersección? La respuesta se hallará en ciertos tipos de estructura que están formalmente muy próximas a las *redes de espín* que yo mismo había estudiado hace casi cincuenta años con un propósito diferente, aunque ligeramente relacionado.

¿Qué son las redes de espín y por qué me habían interesado en la década de 1950? Mi objetivo era tratar de describir la física en términos de cantidades combinatorias discretas, puesto que en aquella época yo tenía la firme convicción de que la física y la estructura espaciotemporal deberían estar basadas, de raíz, en la *discretización* antes que en la continuidad (véase §3.3). Una motivación guía era una forma de *principio de Mach* (§28.5),[18] por el que la noción del propio espacio sería una noción *derivada*, y no inicialmente presente en el esquema. Todo tenía que expresarse en términos de la *relación* entre objetos y no entre un objeto y cierto espacio de fondo.

Había llegado a la conclusión de que la mejor perspectiva para satisfacer estos requisitos era considerar la magnitud mecanocuántica del *espín total* de un sistema. El «espín total» se define por la cantidad escalar j ($= \frac{1}{2}n$), que mide la cantidad de espín como un todo, en lugar de la componente concreta del espín en una dirección, medida por una cantidad m. (Las letras «j» y «m» son las que se utilizan normalmente en la discusión del momento angular mecanocuántico, tomado en unidades de $\hbar$, donde m va, por pasos enteros, entre los enteros o semienteros $-j$ y j; véanse §§22.8,10,11.) La *magnitud* real del espín total (obtenida como la raíz cuadrada de la suma de los cuadrados de los valores m en tres direcciones perpendiculares) es $\hbar\sqrt{j(j+1)}$, que es la misma cantidad que aparece en la expresión del área anterior. Los valores permitidos de $n = 2j$ son simplemente números naturales (pares para los bosones e impares para los fermiones; véase §23.7). Además, aunque independiente de la dirección, n está en cualquier caso íntimamente *relacionado* con aspectos direccionales del espacio. Me parecía que el espín total, medido por el número natural n, era una cantidad ideal en la

que centrar la atención si, partiendo de cero, se estuviera interesado en construir alguna estructura combinatoria discreta que llevara a una noción del espacio físico real. Como ingrediente adicional, si se establecen las cosas de forma correcta, se podría mostrar que las probabilidades mecanocuánticas son *puras probabilidades*, no dependientes de la forma detallada en que puedan estar orientadas unas partes de un aparato físico con respecto a otras.

¿Cómo funciona esto? Llamemos a una cantidad de espín total $\frac{1}{2}n\hbar$ una *n-unidad*. Para que quede más claro, consideremos esta unidad como si fuera una partícula, aunque no tiene por qué ser una partícula elemental. Por ejemplo, un átomo de hidrógeno completo serviría perfectamente. Simplemente tiene que tener un valor bien definido para su espín total (que, en el caso de un átomo de hidrógeno vendría dado por $n = 0$ o 2, los casos del ortohidrógeno o el parahidrógeno, respectivamente).[19] ¿Cómo obtenemos una probabilidad pura? Podríamos, por ejemplo, tomar una pareja de pares EPR-Bohm de 1-unidad (A, B) y (C, D), partiendo cada uno de ellos del estado 0-unidad. (Esto es solo un par de configuraciones, cada una de ellas como la de la Fig. 23.3 o §23.4; véase la Fig. 32.10.) Ahora, si juntamos B y D y les dejamos combinarse en una única unidad, existen las dos posibilidades de que den como resultado una 0-unidad o una 2-unidad, con probabilidades respectivas 1/4 y 3/4.[32.5] Si alternativamente juntamos A y C, entonces los resultados posibles y sus probabilidades serían exactamente los mismos. Sin embargo, estas dos probabilidades no son ni mucho menos independientes una de otra, puesto que si obtenemos una 0-unidad en un caso no podemos obtener una 2-unidad en el otro, y viceversa.

Este era el tipo de idea que yo tenía para obtener probabilidades puras, y creía que una probabilidad semejante tenía que ser un número racional (puesto que equivaldría a que la naturaleza hiciera una elección aleatoria de algún tipo entre un número finito de posibilidades discretas). El ejemplo que se acaba de considerar es muy simple, pero empieza a ilustrar la idea general. Se considera que todas las unidades en una red de espín concreta están inicialmente producidas de la for-

[32.5] ¿Puede ver por qué?

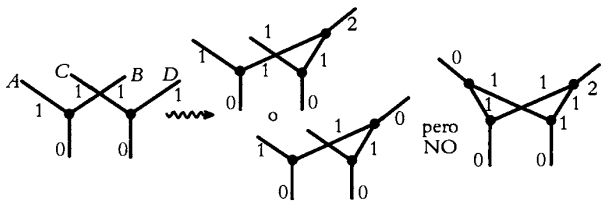


Fig. 32.10. Redes de espín. Cada segmento de línea, etiquetado por un número natural n, representa una partícula o subsistema de espín total $\frac{n}{2} \times \hbar$, llamada una n-unidad. En este ejemplo muy simple tenemos dos pares de 1-unidad EPR-Bohm, (A, B) y (C, D), cada uno los cuales parte de un estado 0-unidad (como en la Fig. 23.2). Si B y D se combinan para formar una única unidad, existen dos posibilidades: puede ser una 0-unidad o una 2-unidad, con probabilidades respectivas 1/4 y 3/4. Las mismas probabilidades se dan si alternativamente combinamos A y C. Pero estas probabilidades no son independientes, puesto que no podemos tener 0-unidad en un caso y 2-unidad en otro.

ma anterior a partir de 0-unidades iniciales (aunque normalmente esto se expresaría de forma explícita en el diagrama), de modo que no hay sesgo respecto a ninguna dirección espacial concreta. Posteriormente, varios pares de unidades podrían juntarse para formar unidades únicas, y anotarse los valores de espín para las unidades resultantes. A las unidades individuales también se les permite dividirse en pares de unidades. Un ejemplo de ello se ilustra en la Fig. 32.11. Podemos visualizar todo esto como algo que sucede dentro de cierto espaciotiempo. Sin embargo, en el caso de la teoría de redes de espín original, no iba a presuponerse ningún fondo espaciotemporal real. La idea era construir todas las nociones espaciales requeridas simplemente a partir de la red de espines y de las probabilidades que aparecen (y estas pueden calcularse utilizando reglas mecanocuánticas) cuando se juntan dos unidades para formar una tercera. Una característica particular de estas redes de espín es que en cada uno de estos vértices se juntan exactamente tres líneas. Esto conduce a una unicidad en los cálculos de probabilidad. Todo lo que se necesita es la estructura (grafo) topológica de la red de espín, junto con la especificación de todos los números de espín en las líneas.

Desarrollé un procedimiento («recuento») enteramente combinatorio para calcular las probabilidades requeridas (que, de hecho, *son*

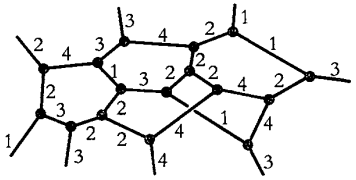


Fig. 32.11. Ejemplo de una red de espín, como fue concebida originalmente. No se presupone ninguna variedad espaciotemporal de fondo. Todas las nociones espaciales van a aparecer a partir de la red de espines y de los valores de la probabilidad (cuando dos unidades se combinan para dar una tercera). En cada vértice se encuentran tres líneas, lo que especifica unívocamente la conexión.

todas racionales). Las reglas proceden originalmente de la mecánica cuántica estándar del espín, pero podemos «olvidarnos» de dónde proceden y considerar simplemente que el sistema de la red de espín proporciona un tipo de «universo combinatorio». Entonces es posible extraer las nociones de geometría (en este caso, la 3-geometría euclídea ordinaria) considerando redes de espín que son «grandes» en un sentido apropiado. La imagen es que podría considerarse que una unidad de espín grande define una «dirección en el espacio» (semejante al eje de giro de una pelota de golf, por ejemplo). Podemos imaginar que se mide el «ángulo entre los ejes de rotación» de dos de estas unidades grandes, pongamos por caso, desgajando una 1-unidad de una y añadiéndola a la otra. La probabilidad conjunta de que un espín aumente mientras que el otro disminuya en esta operación da una medida del ángulo entre los ejes de giro.[32.6]

Esto casi funciona tal como está, pero no del todo, y se requiere un ingrediente adicional. Se necesita un medio de distinguir la «probabilidad cuántica» —que procede del ángulo entre los ejes de giro de unidades grandes— de la «probabilidad por ignorancia» que puede resultar simplemente de las conexiones insuficientes entre las dos unidades grandes. (Recordemos el juego sutil entre estas dos nociones de probabilidad que ocurre con las matrices densidad, tal como se han expuesto en §§29.3,4). Resulta que este «factor de ignorancia» puede eliminarse repitiendo la transferencia de una 1-unidad desde una unidad grande a la otra, y seleccionando solo las situaciones en las que las probabilidades

[32.6] ¿Cuál es esta medida de ángulo en términos de las probabilidades?

resultan iguales la segunda vez. Para familias de unidades grandes en las que se da esto, es posible demostrar un teorema geométrico, según el cual la «geometría de ángulos», definida así por las probabilidades cuánticas, es precisamente la geometría de los ángulos entre direcciones en el 3-espacio euclídeo ordinario.[20] De este modo, vemos que nociones de la geometría euclídea ordinaria surgen meramente de la combinatoria cuántica de redes de espín.

Se verá que mi motivación original subyacente para las redes de espín es muy diferente de la que subyace en el enfoque de las variables de lazo para la cuantización espaciotemporal, pues no hay ningún papel real para la gravedad en las redes de espín tal como se proponían originalmente. Por consiguiente, fue para mí una sorpresa considerable observar que las redes de espín desempeñaban un papel tan importante en esta aproximación a una teoría de la gravitación cuántica. Por supuesto, los dos programas tienen mucho en común, pues en ambos casos se está tratando de descomponer la noción de espacio en algo más discreto y mecanocuántico. Sin embargo, existe una diferencia importante: en el contexto de las variables de lazo, la cantidad n es realmente una medida de *área*, y no la medida de espín de las redes de espín originales. Estas son dimensionalmente diferentes, como se refleja en la aparición de la constante gravitatoria G en la expresión de variables de lazo. Pronto volveré a esta cuestión y su posible importancia.

Ahora bien, ¿cómo intervienen las redes de espín en la gravitación cuántica de variables de lazo? Como he dado a entender antes, los nodos de las redes de espín van a resultar, en efecto, de la intersección de un par de lazos. Esto también permite que el valor j en el lazo cambie en dicho lugar. Por consiguiente, tendremos nodos donde se juntan *cuatro* líneas (o quizá más), en lugar de las tres que ocurrían en mis redes de espín original. Esto da como resultado ambigüedades, puesto que la unicidad de la interpretación ocurre solo con los nodos «trivalentes» originales. En consecuencia, hay una especificación adicional requerida (un «operador de entrelazamiento») en cada nodo. Una manera de expresar esta especificación se ilustra en la Fig. 32.12, donde podemos representar una de estas «X», en la que se juntan cuatro líneas, como una combinación lineal de pares de vértices de 3 aristas de tipo «H». La especificación de los coeficientes elimina las ambigüedades.

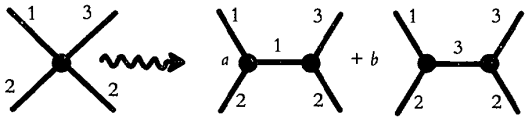


Fig. 32.12. Todos los nodos de redes de espín en la teoría de variables de lazo son 4-valentes (o más) y necesitan información extra. Esta puede estar codificada en vértices tipo «X» que se expresan como combinación lineal de pares de vértices tipo «H» de tres aristas. Los coeficientes específicos eliminan esta ambigüedad.

Hay otra diferencia bastante más importante entre estas redes de espín de variables de lazo y las que yo había propuesto antes: las primeras eran estructuras enteramente combinatorias, mientras que las redes de variables de lazo adquieren una estructura topológica adicional de su inmersión en la variedad $\mathcal{S}$. Por ejemplo, las líneas de red podían estar anudadas o podrían estar enlazadas con otras de varias maneras, y esto proporciona información adicional (véase la Fig. 32.13). No obstante, esta información es aún de naturaleza combinatoria discreta, pues es de un carácter enteramente topológico, pero resulta más difícil expresar que especificar lo que sucede en los nodos individuales.

Hasta ahora nuestras descripciones de lazos nos han dado, en efecto, tan solo una descripción *estática*, sin ninguna dinámica implicada. De hecho, los lazos y las redes de espín que hemos considerado estaban relacionadas con la resolución de las ecuaciones de ligadura de la relatividad general —i.e., las condiciones que tienen que satisfacerse dentro de la superficie $\mathcal{S}$— respetando plenamente el principio de covariancia general. Este no es un logro menor, pero parece que el formalismo no ha resuelto todavía el problema más difícil de la evolución dinámica lejos de $\mathcal{S}$ (a veces conocido como la «ligadura hamiltoniana»), para que pueda acomodarse plenamente la ecuación de Einstein (véase §32.1). Un trabajo importante de Thomas Thiemann ha proporcionado una respuesta posible a este problema de evolución hamiltoniana, pero quedan algunas dudas acerca de si esta es realmente la apropiada para la teoría de Einstein.[21]

A falta de una solución plenamente aceptada para estas difíciles cuestiones dinámicas, ha sido posible, de todas formas, utilizar el formalismo de las variables de lazo para llegar a resultados impresionantes

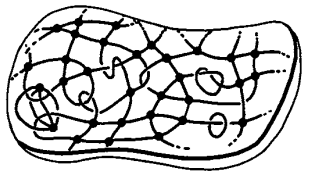


Fig. 32.13. Las redes de espín de la aproximación de variables de lazo estándar no son ya entidades combinatorias totalmente incorpóreas, sino que deben estar inmersas en una 3-superficie (como $\mathcal{S}$) sin estructura (pero quizá analítica), siendo importantes sus propiedades topológicas de enlazado y anudado.

en otras direcciones. En particular, estas ideas de redes de espín se han mostrado útiles para proporcionar una aproximación a la cuestión de la entropía de los agujeros negros mucho más directa y realista que la que proporciona la teoría de cuerdas, tal como se ha mencionado en §31.15. Aquí la geometría de los agujeros negros es directamente la de una solución de vacío de Schwarzschild o de Kerr de la teoría de Einstein 4-dimensional. El recuento de estados cuánticos gravitacionales puede llevarse a cabo explícitamente utilizando redes de espín y aproximaciones adecuadas. Cuando el agujero negro empieza a hacerse razonablemente grande, la respuesta para la entropía resulta estar en acuerdo con la fórmula de Bekenstein-Hawking $S_{\text{BH}} = \frac{1}{4} A$ (donde $k = c = G = \hbar = 1$), pero para obtener el factor 1/4 exacto de Hawking parece necesario tomar el curioso valor

$$\eta = \frac{\log 2}{\pi\sqrt{3}}$$

para el parámetro de Barbero-Immirzi. Aunque es cierto que es un valor extraño, esta elección da correctamente la entropía de Bekenstein-Hawking para todas las situaciones a las que se ha dado una respuesta inequívoca por otros medios, en las que puede haber carga, rotación o constante cosmológica.[22]

En relación con esto, hay varias «coincidencias» numéricas aparentes, de las que parece depender la teoría. Dos series infinitas independientes, calculadas de dos maneras completamente diferentes, tienen que coincidir término a término, lo que, de hecho, hacen. Esto parece reflejar una profunda consistencia interna de algunas de las ideas de la geometría cuántica.

De todas formas, los resultados de agujeros negros parecen haber provocado un cambio de punto de vista con respecto al parámetro de

Barbero-Immirzi. Previamente, la introducción de η parece haber sido solo un medio de avanzar, donde la teoría «geométricamente correcta» parecía exigir $\eta = \pm i$. Tomar un valor real para η era solo una conveniencia matemática para que apareciera el grupo compacto SU(2) en lugar del grupo no compacto SL(2, $\mathbb{C}$). Los impresionantes éxitos al obtener los valores correctos de la entropía para una clase muy amplia de horizontes —siendo dependiente de la única elección para η (dada antes como $\eta = \log 2/\pi\sqrt{3}$), para el parámetro de Barbero-Immirzi— ha llevado a varios defensores de la aproximación de variables de lazo a asumir la idea de que, después de todo, quizá dicho valor real sea realmente el «correcto» para la gravedad cuántica.

Esta es, por supuesto, una posibilidad, aunque personalmente la encuentro difícil de creer, pues no parece haber una razón geométrica clara para tal elección. Debería comentar que con cualquier valor real para η, como el que se ha mostrado antes, ha desaparecido el aspecto quiral de la teoría que he señalado en §32.3 para introducir el tema. El transporte espinorial con el que está relacionado $\mathbf{\Gamma}_\eta$, con un valor real de η, es una mezcla peculiar pero equilibrada de partes intrínseca y extrínseca, cuyo significado encuentro particularmente oscuro. Quizá futuros estudios arrojen alguna luz sobre esta cuestión.

32.7. ¿El estatus de la gravedad cuántica de lazo?

Debería tratar de dar mi valoración de los logros de la aproximación de variables de lazo de Ashtekar-Rovelli-Smolin a la gravedad cuántica y su potencial de desarrollo futuro en una teoría generalizada. Una vez más, debo alertar al lector sobre un posible sesgo que podría ser relevante para tal valoración. En este caso debo declarar un compromiso personal, pues no solo las personas responsables de ese programa son buenos amigos míos, sino que también he ejercido regularmente como profesor visitante en las dos universidades estadounidenses (Syracuse y del Estado de Penn) donde se han llevado a cabo importantes estudios en este campo. Debo añadir, además, mi propio interés en la teoría de redes de espín; es natural que encontrara gratificante que estas viejas ideas encontraran ahora un nuevo valor en esta aproximación. De todas formas, mi propia implicación en el programa de gravedad cuántica con variables de Ashtekar/variables de lazo ha sido algo tangencial al trabajo principal en esta área, así que espero mantenerme al margen y ser razonablemente objetivo.

Para empezar, debería comentar que tanto las variables de Ashtekar originales como las descripciones posteriores en términos de variables de lazo me impresionan como desarrollos potentes y bastante originales en la búsqueda de una teoría de la gravedad cuántica. Abordan directamente la teoría de la relatividad de Einstein en el contexto de la QFT, y proporcionan ideas muy innovadoras que son relevantes para el problema en cuestión. De hecho, tengo pocas dudas al decir que tales estudios son los más importantes en la aproximación canónica a la gravedad cuántica desde que Dirac y otros iniciaron esta disciplina hace aproximadamente medio siglo. Los estados de lazo parecen abordar al menos algunos de los profundos problemas planteados por la covariancia general. Además, estos desarrollos parecen haber llevado la discusión en una dirección fascinante, y quizá no completamente prevista, en donde empiezan a aparecer algunos gratificantes elementos de discretización en el espaciotiempo. Además, en un trabajo reciente la teoría original puramente gravitatoria se ha desplazado en la dirección de incorporar interacciones físicas distintas de la gravedad, de modo que la teoría puede ahora proclamarse como una aproximación a la física fundamental en general.[23]

En contra, está el hecho algo perturbador de que la teoría parece haberse visto obligada a desviarse al adoptar la conexión $\mathbf{\Gamma}_\eta$ (con un valor indeterminado de η), en lugar de la «geométricamente correcta» $\mathbf{\Gamma}$. En mi opinión, un enfoque completamente creíble de la gravedad cuántica no dará resultado hasta que se haya encontrado una forma de superar las dificultades que parecen surgir con la adopción de la $\mathbf{\Gamma}$ original. Además, sigue existiendo la dificultad fundamental de que el hamiltoniano de Einstein completo tiene aún que ser adecuadamente englobado dentro del marco de las variables de lazo, incluso si las ecuaciones de ligadura se tratan mediante el uso de las redes de espín.

Tengo la impresión de que es probable que estas dificultades estén relacionadas con otra característica menos que satisfactoria (en mi opinión) de la teoría de Ashtekar/variables de lazo. En común con todos

los demás enfoques *canónicos* convencionales para la gravedad cuántica, su formulación es directamente dependiente de una descripción 3-espacial (i.e., en términos de $\mathcal{S}$), en lugar de ser una descripción espacio-temporal más global. Como hemos visto, la parte 3-espacial del problema de la «covariancia general» es tenida en cuenta en los estados lazo/redes de espín, pero la extensión de esto a una covariancia general en un 4-espacio completo implica toda una «caja de Pandora» de problemas. Por lo que puedo ver, estos no son mucho mejor abordados por ahora en el enfoque de variables de lazo que en otros enfoques canónicos.[24]

La dificultad tiene que ver con la cuestión de cómo debe expresarse adecuadamente la evolución temporal, de acuerdo con la ecuación de Einstein, en un formalismo 4-espacial con covariancia general. Está relacionada con lo que se conoce como el «problema del tiempo» en gravedad cuántica (o, a veces, el problema del «tiempo congelado»). En la relatividad general no se puede distinguir la evolución temporal de un mero cambio de coordenadas (i.e., tan solo el reemplazamiento de una coordenada temporal por otra). Un formalismo con covariancia general debería ser ciego a un mero cambio de coordenadas, de modo que el concepto de evolución temporal se convierte en algo muy problemático. Como se ha indicado en §30.11, mi propia perspectiva sobre esta cuestión es que hay pocas probabilidades de resolverla sin abordar satisfactoriamente el problema de la reducción del vector de estado **R**, y que esto, a su vez, exigirá una drástica revisión de los principios generales.

En relación con estas cuestiones, hay otro tema que encuentro menos que satisfactorio, aunque es más un problema con las recetas de covariancia *per se* que con la aproximación de variables de lazo en particular. ¡En cierto sentido, esta aproximación es víctima de su propio éxito! En efecto, aunque los estados base de las redes de espín tienen individualmente una agradable descripción geométrica independiente de las coordenadas, está menos claro cómo interpretar las *superposiciones* cuánticas de tales estados base. Debido a la covariancia general, no hay correspondencia entre la «localización» de una red de espín y la de otra con la que va a superponerse. (Esta es una versión más seria de la cuestión abordada en §30.11, que he uitlizado para justificar la **RO**

gravitatoria.) ¿Cómo esperamos comprender de qué forma va a emerger de todo esto un mundo casi clásico?

Como el lector ya habrá deducido, en particular por la discusión del capítulo 30, considero que una característica necesaria de la unión cuántico/gravedad correcta es que debe apartarse de la mecánica cuántica estándar de alguna manera esencial, de modo que **R** llegue a ser un proceso físico realista (**RO**). ¿Hay lugar para esto en la aproximación de variables de lazo? Posiblemente. Con las variables de lazo, los números $n = 2j$ en las aristas de la red de espín se refieren a un *área* en unidades de longitud de Planck al cuadrado; pero mi utilización original de las redes de espín no abordaba tales cuestiones métricas ni, de hecho, ningún aspecto de la gravedad en absoluto, y los números de espín n se referían al *momento angular*. Sin embargo, mis ideas originales exigían que cada uno de estos números debe ser, en efecto, el resultado de una medida individual del valor de espín total (acción de **R** en cada arista), donde las probabilidades surgen al reunir las dos unidades para hacer una tercera. Si **R** es un proceso gravitatorio objetivo, entonces la implicación con procesos gravitatorios tendría que entrar en esta etapa, que es el mensaje proclamado del capítulo 30. En ese caso, no es posible separar la gravedad de las cuestiones de probabilidad de la teoría de redes de espín. Puede ser que la combinación completa de ideas de variables de lazo y redes de espín necesite incorporar reducción de estado en el formalismo. Si se demuestra que es así, entonces podría proporcionar un camino hacia un esquema **RO** gravitatorio apropiado, como se recomendaba en el capítulo 30. Pero en ausencia de tal formalismo, esas ideas deben seguir en el terreno de la especulación.

Para finalizar, debería comentar otro trabajo que se relaciona con la teoría de variables de lazo. Ahora no es por entero una teoría gravitatoria pura, pues el electromagnetismo ha sido abordado[25] en este formalismo.[26] Existen también formas aparentemente menos radicales que las del párrafo anterior de hacer más 4-dimensionales las redes de espín de la gravedad cuántica de lazo. Una de estas incluye una ingeniosa versión de redes de espín en más dimensiones conocida como *espumas de espín*. En estas, existen 2-superficies que llevan los «valores de espín» $n = 2j$, y podemos imaginar esta espuma de espín como una red de es-

pín que evoluciona en el tiempo. Tales ideas, que se deben originalmente a Louis Crane, John Barrett y otros,[27] han sido desarrolladas y modificadas por algunos,[28] pero las conexiones correctas con las ideas de la gravedad cuántica no han sido todavía establecidas del todo. Hay también posibles conexiones con la teoría de twistores, y será interesante ver si estas pueden desarrollarse completamente. En el capítulo siguiente, veremos algunas de las nociones básicas que están implicadas en la teoría de twistores.

He tratado de resaltar en otros capítulos de este libro, en especial en el 30, que la cuestión de la reducción del estado cuántico está íntimamente relacionada con la estructura de las singularidades espaciotemporales y su asimetría bajo inversión temporal. Es interesante que ya se haya empezado a examinar lo que tiene que decir la aproximación de variables de lazo sobre el efecto de la teoría cuántica en las singularidades espaciotemporales.[29] No soy capaz de comentar este trabajo en detalle, excepto para decir que no veo ningún indicio de aparición de la necesaria asimetría temporal.

Notas

Sección 32.1

32.1. Véase Greene (1999).

32.2. Aquí se da la ironía de que el propio Einstein, con su aventura en las teorías de campos unificados de sus últimos años, no siempre se atuvo al estrecho sendero que él había establecido previamente.

32.3. Véanse Dirac (1964), Ashtekar (1991) y Wald (1984) para las nociones de ligaduras y formulación hamiltoniana.

32.4. Véanse Dirac (1950, 1964), Pirani y Schild (1950), Bergmann (1956) y Arnowitt *et al.* (1962), por ejemplo. Véase DeWitt (1967) para la ecuación de Wheeler-DeWitt, que es básicamente la ecuación de Schrödinger de la gravedad cuántica en el espacio de 3-geometrías compactas.

32.5. Véanse Isham (1975) y Kuchar (1981).

32.6. Véanse Sen (1982) y Ashtekar (1986, 1987).

Sección 32.2

32.7. Como veremos de inmediato, dificultades técnicas han desviado el enfoque original de Ashtekar respecto a esta descripción manifiestamente quiral. Pero tal como entiendo las motivaciones que hay tras el programa de Ashtekar, esta «desviación» debe tomarse como una exploración temporal de modelos que son próximos, pero no idénticos, a la pretendida teoría de la gravitación cuántica.

32.8. Para la discusión original de la gravedad linealizada, véase Fierz y Pauli (1939); para más información sobre la gravedad linealizada, véase Sachs y Bergmann (1958). Para el gravitón no lineal, véase Penrose (1976a y 1976b).

Sección 32.3

32.9. Véase Ashtekar (1986 y 1987); véanse también Ashtekar y Lewandowski (2004) para una revisión, y Rovelli (2003) ¡para un libro de texto!

32.10. Quienes deseen entender cómo trabajan los «índices» deberían advertir primero que la 3-métrica inversa es una magnitud γ^{rs} con índices 3-dimensionales elevados. Formando el dual en s (§12.7), dentro de $\mathcal{S}$, obtenemos γ'_{tu} (una densidad) que es antisimétrica en t, u. Podemos leer estos subíndices como índices de 2-formas en el 4-espacio y tomar la parte anti-autodual que nos proporciona un par de subíndices espinoriales simétricos, lo que nos da una cantidad γ'_{PQ}, o equivalentemente $\gamma'^{\ Q}_{P}$, que es libre de traza en P, Q. En el caso de Γ, esta es una cantidad $\Gamma_{r\ Q}^{\ P}$, que es una matriz 1-forma en el espacio de espín, que es necesariamente libre de traza (porque es una conexión SL(2, $\mathbb{C}$)). Nótese que la estructura de índices de Γ es contraria a la de γ, como debería ser, para variables canónicas. Para más detalles, véanse Ashtekar (1986, 1987), Ashtekar y Lewandowski (2004) y Rovelli (2003).

32.11. Esta fue la conexión que introdujo Sen en 1981, reduciendo las ligaduras de la relatividad general a una forma polinómica. Es interesante señalar que Witten utilizó la misma conexión, independientemente (también en 1981), en su método de demostración de la positividad de la energía total en la relatividad general; véanse §19.8 y Witten (1981). Por lo que sé, el primer uso de esta conexión fue en la construcción del *espacio twistorial hipersuperficial* para $\mathcal{S}$ (Penrose y MacCallum 1972; Penrose 1975), que es la parte de la teoría de twistores más estrechamente relacionada con las variables de Ashtekar.

Sección 32.4

32.12. Véase Wilson (1976); para esta técnica en la teoría cuántica de campos, véase Zee (2003). Véase también Rovelli y Smolin (1990).

32.13. Algunos podrían considerar incluso que poder especificar su estructura de variedad es más de lo que tenemos derecho, puesto que esto le proporciona no solo una topología sino también una estructura «suavizada» o diferenciable; véanse §10.2 y §12.3. Compárese también con §33.1. De hecho, por conveniencia técnica, a S se le asigna una estructura *analítica*, en la aproximación de Ashtekar-Lewandowski. (Ashtekar y Lewandowski, 1994), lo que significa que es C^ω, en el sentido de §6.4.

Sección 32.5

32.14. Para una introducción muy accesible a la teoría de nudos, véase Adams (2000). Para una obra más técnica, véanse Rolfsen (2004) y Kauffmann (2001).

32.15. Véase Labastida y Lozano (1997) para un artículo de revisión; véase Witten (1988) para la «TQFT» original.

32.16. No obstante, las QFT topológicas no conducen directamente a la demostración de teoremas matemáticos rigurosos, debido a divergencias incontrolables.

32.17. Véase Penrose (1988a). Puede advertirse que las nuevas ideas twistor-cuerda propuestas por Witten (2003) proporcionan desarrollos de este tipo general; véase la nota 31.81.

Sección 32.6

32.18. Mi interés en el principio de Mach se debía a las discusiones con mi colega, amigo y mentor Dennis Sciama.

32.19. Véase Levitt (2001).

32.20. Véanse Penrose (1969b) y Moussouris (1983).

32.21. Véase Thiemann (1995, 1998a, 1998b, 1998c y 2001).

32.22. Véanse Ashtekar *et al.* (1998) y Ashtekar *et al.* (2000). Sin embargo, recientemente se ha demostrado que este valor para η tiene un error, y ahora se ha asignado a η un valor con una apariencia más complicada. Véanse Domagala y Lewandowski (2004) y Dreyer, Markopoulou y Smolin (2004). El resto de mi discusión en §§32.6,7 no parece afectada.

Sección 32.7

32.23. Véase Thiemann (1998c).

32.24. Véanse Hawking y Hartle (1983) y las notas 32.3-5. No puedo resistirme a remitir a Smolin (2003) y al excelente artículo de Ashtekar y Lewandowski (2004), del que se han tomado muchas de estas referencias. Agradezco profundamente a Abhay Ashtekar su ayuda con las referencias de este capítulo.

32.25. Véase Varadarajan (2000).

32.26. Véase Varadarajan (2001).

32.27. Véanse Barrett y Crane (1988), Baez (1998) y Reisenberger (1997, 1999).

32.28. Véanse Barrett y Crane (2000), Baez (2000), Reisenberger y Rovelli (2001, 2002) y Perez (2003).

32.29. Véanse Bojowald (2001) y Ashtekar *et al.* (2003).

33

Perspectivas más radicales: la teoría de twistores

33.1. Teorías donde la geometría tiene elementos discretos

¿Han sido suficientemente radicales las teorías descritas en los capítulos precedentes en sus intentos por descifrar el esquema *real* de la naturaleza en el que la física cuántica de lo pequeño está unida de algún modo con la geometría del espacio curvo de lo grande? Quizá deberíamos buscar algo de un carácter fundamentalmente diferente de la variedad real que constituye el escenario del espaciotiempo continuo del que dependen la teoría de Einstein y la mecánica cuántica estándar. La cuestión se ha planteado en §3.3, y ahora debemos preguntarnos si esa continuidad espaciotemporal, expresada mediante números reales, que es asumida casi universalmente en las teorías físicas, es *realmente* la matemática apropiada para describir los constituyentes últimos de la naturaleza.

Hemos visto que el enfoque de las variables de lazo para la gravedad cuántica empieza a alejarnos de la imagen estándar de un espaciotiempo continuo y de variación suave, y nos lleva hacia algo con un carácter topológico más discreto. Pese a todo, algunos teóricos argumentarán enérgicamente que se necesita una revisión mucho más radical de las ideas de espacio y tiempo si se quiere llegar a intuiciones más profundas acerca de la naturaleza de un «espaciotiempo cuántico». De hecho, la original (aunque limitada) propuesta de la red de espín de §32.6 tenía un carácter completamente discreto, pero la imagen de las variables de lazo sigue siendo dependiente de la naturaleza continua de la 3-superficie en donde se supone que están inmersas las «redes de es-

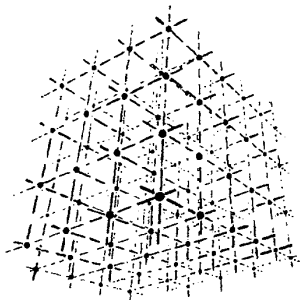
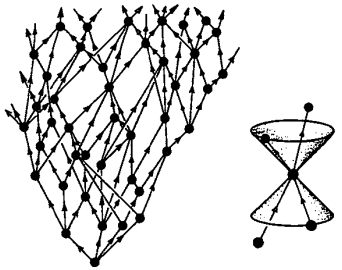


Fig. 33.1. Un espaciotiempo tipo Snyder-Schild es un retículo periódico, como los vértices de un conjunto de cubos apilados unos sobre otros de una forma regular. ¡Esto puede tener bastante más invariancia Lorentz de lo que cabría esperar!

pín». En el último esquema no se ha obtenido en realidad el armazón enteramente discreto y manifiestamente «combinatorio» que algunos piensan que es necesario para poder entender el funcionamiento de la naturaleza en las escalas más minúsculas.

Se han propuesto varias ideas, muy diferentes en enfoque de los esquemas iniciales de la red de espín o la espuma de espín, cuya intención es ofrecer una imagen completamente combinatoria/discreta del mundo. Entre las más extravagantes de estas ideas (ya mencionadas en §16.1) figuraba una propuesta de Ahamavaara (1965).[1] Este sugirió que el sistema de los números reales, fundamental para las matemáticas de la física convencional, debería ser reemplazado por un campo finito $\mathbb{F}_p$, donde p es algún número primo extraordinariamente grande. Recordemos (de §16.1) que $\mathbb{F}_p$ se obtiene tomando el sistema de los enteros módulo p. Otras sugerencias consideran que el espaciotiempo tiene una estructura *reticular*, periódica y discreta, como la que forman los vértices de un conjunto de cubos apilados uno sobre otro de un modo regular[2] (Fig. 33.1). Bastante más plausibles físicamente son los esquemas como la geometría de conjuntos causales de Raphael Sorkin[3] (o algunas ideas anteriores íntimamente relacionadas),[4] en la que se considera que el espaciotiempo consiste en un conjunto, posiblemente finito, de puntos discretos para el que la noción básica es la noción de *conexión causal* entre puntos. En términos clásicos ordinarios, esta «conexión causal» se refiere a la posibilidad de enviar una señal de un punto a otro, de modo que uno de estos puntos yace sobre o dentro del cono de luz del otro (véase la Fig. 33.2). La naturaleza fundamentalmente aleatoria de las conexiones causales en el esquema de Sorkin



(a) (b)

Fig. 33.2. (a) Modelo de universo discreto, descrito por una geometría de conjunto causal. (b) Las relaciones entre puntos están modeladas sobre la causalidad lorentziana, donde las flechas apuntan o bien dentro, o bien sobre los conos nulos.

permite que surja algo similar a la invariancia Lorentz de la relatividad especial, mientras que existen dificultades más serias para la invariancia Lorentz con las estructuras de tipo reticular (aunque hay más de esta simetría que lo que podría pensarse en un principio para retículos como el de la Fig. 33.1). Otras ideas que conducen a estructuras espaciotemporales exóticas surgen de la teoría cuántica de conjuntos o la geometría cuaterniónica de David Finkelstein,[5] la física octoniónica (§11.2 y §16.1) de Corinne Manogue y Tevian Dray,[6] etc.

También existe una interesante propuesta de una teoría de gravedad cuántica avanzada por Tullio Regge en 1959, según la cual se considera el espaciotiempo como un poliedro (o «politopo») 4-dimensional «tetraédrico» irregular, con su curvatura concentrada como funciones delta (§9.7) a lo largo de «aristas» 2-dimensionales, a las que Regge llamaba «huesos».[7] Véase la Fig. 33.3 (y la Fig. 32.3a, en §32.4). Se considera que el estado cuántico es una suma de tales espacios ponderada por números complejos, de acuerdo con la «suma sobre historias de Feynman» descrita en §26.5. La descripción de los propios espacios es completamente combinatoria, salvo por el hecho de que debe especificarse un «ángulo» en cada hueso para representar la intensidad de la curvatura. De hecho, la «cuerda cósmica» que se ha descrito en §28.2 (Fig. 28.4) es un ejemplo de este mismo tipo de geometría.

También se han hecho otras propuestas radicales, tales como la de Richard Jozsa[8] y Christopher Isham,[9] que emplean la *teoría de topos*. Esta es un tipo de teoría de conjuntos[10] que surge de la formalización de la «lógica intuicionista» (véase la nota 2.6) ¡que *niega* la validez del método de «demostración por contradicción» (§2.6 y §3.1)! No discu-

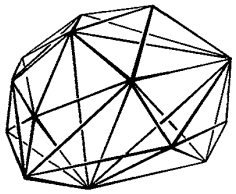


Fig. 33.3. En el «cálculo de Regge» el espaciotiempo se aproxima por un poliedro 4-dimensional (un «politopo»), normalmente construido a partir de «tetraedros» 4-dimensionales (5-simplexos). La curvatura reside (como funciones δ) a lo largo de aristas 2-dimensionales (normalmente triangulares) llamadas «huesos».

tiré ninguno de estos esquemas, y remitiré al lector interesado en el tema a la literatura.

Otra idea que quizá algún día encuentre un papel importante que desempeñar en la teoría física es la teoría de *categorías* y su generalización a la teoría de n-categorías. La teoría de categorías, introducida en 1945 por Samuel Eilenberg y Saunders MacLane,[11] es un formalismo (o armazón) algebraico extraordinariamente general basado en nociones abstractas muy primitivas (aunque confusas), estimulado en sus inicios por ideas de la topología algebraica. (Sus procedimientos se suelen denominar coloquialmente «sinsentidos abstractos».) Su gran potencia es engañosa, dado el carácter muy elemental de sus ingredientes básicos, que son tan solo «flechas» que conectan «objetos», y tiene una apariencia muy «combinatoria», en línea con otras ideas mencionadas en esta sección. La extensión de la teoría de categorías a la de n-categorías refleja el modo en que la «homotopía» refina la noción de «homología», tal como se ha discutido brevemente en §7.2. La teoría de categorías ya ha aportado un ingrediente en la teoría de twistores (en relación con §33.9), y la teoría de n-categorías tiene relación con lazos, enlaces, espumas de espín (§32.7) y estructuras q-deformadas (§32.5).[12] No me sorprendería que estas nociones llegasen a tener un papel importante para reemplazar las nociones espaciotemporales convencionales en la física del siglo XXI.

Bastante más en línea con las ideas de la corriente principal, está la noción de *geometría no conmutativa*, desarrollada en particular por el matemático Alain Connes, que ganó la Medalla Fields. ¿Qué es una «geometría no conmutativa»? Para apreciar la idea, pensemos primero en una variedad ordinaria, real y suave $\mathcal{M}$. A continuación, consideremos la familia de funciones (escalares) suaves de valor real definidas sobre $\mathcal{M}$ (que podemos considerar que son C^∞-suaves; véase §6.3). Tales fun-

ciones pueden sumarse o multiplicarse entre sí, y multiplicarse por números reales ordinarios (constantes). De hecho, constituyen un sistema algebraico $\mathcal{A}$, denominado un *álgebra conmutativa sobre los reales* $\mathbb{R}$. (Compárese con §12.2 y la nota 12.5) Resulta ahora[13] que si solo conocemos $\mathcal{A}$ como un álgebra, y no se ofrece ninguna información de su procedencia, podemos, pese a todo, reconstruir la variedad $\mathcal{M}$ simplemente a partir del álgebra $\mathcal{A}$. Vemos así que cada uno de los $\mathcal{M}$ y $\mathcal{A}$ pueden construirse a partir del otro, de donde se sigue que estas dos estructuras matemáticas son mutuamente *equivalentes* en un sentido preciso.

En mecánica cuántica, a menudo se encuentran álgebras que son, por el contrario, *no* conmutativas. Un ejemplo sería el álgebra de x^a y p_a que satisface las reglas de conmutación canónicas estándar $p_b x^a - x^a p_b = \mathrm{i}\hbar\delta^a_b$, de §21.2. Si tratamos de reconstruir una «variedad» a partir de esta álgebra, de un modo semejante a como $\mathcal{M}$ se obtendría a partir de $\mathcal{A}$, entonces obtenemos lo que se conoce como una geometría *no conmutativa*. A modo de ejemplo, consideremos otro caso particular y empecemos con las componentes $\boldsymbol{L}_1, \boldsymbol{L}_2, \boldsymbol{L}_3$ del *momento angular* mecanocuántico de §22.8 (recordemos que el álgebra de $\boldsymbol{L}_1, \boldsymbol{L}_2, \boldsymbol{L}_3$ está definida por las leyes no conmutativas $\boldsymbol{L}_1\boldsymbol{L}_2 - \boldsymbol{L}_2\boldsymbol{L}_1 = \mathrm{i}\hbar\boldsymbol{L}_3$, $\boldsymbol{L}_2\boldsymbol{L}_3 - \boldsymbol{L}_3\boldsymbol{L}_2 = \mathrm{i}\hbar\boldsymbol{L}_1$, $\boldsymbol{L}_3\boldsymbol{L}_1 - \boldsymbol{L}_1\boldsymbol{L}_3 = \mathrm{i}\hbar\boldsymbol{L}_2$). Podemos considerar que estos operadores generan las rotaciones de una esfera ordinaria S^2. Resulta que, a partir del álgebra generada por $\boldsymbol{L}_1, \boldsymbol{L}_2, \boldsymbol{L}_3$ podemos obtener una geometría no conmutativa que podemos llamar la «esfera no conmutativa». Hay muchas sutilezas matemáticas, bellas estructuras e inesperadas aplicaciones de esta idea, pero aquí no entraré en detalles. Volveré al tema en §33.7, en relación con la cuantización de twistores.

Connes y sus colegas han desarrollado la noción de geometría no conmutativa con la idea de producir una teoría física que incluya el modelo estándar de la física de partículas.[14] Su modelo utiliza un álgebra $\mathcal{A}$ que es un producto $\mathcal{A}_1 \times \mathcal{A}_2$, donde $\mathcal{A}_1$ es el álgebra (conmutativa) de funciones en el espaciotiempo (pero tomada con una métrica definida positiva) y $\mathcal{A}_2$ es un álgebra no conmutativa que procede de los grupos de simetría interna del modelo estándar de la física de partículas y proporciona «dos copias» del espaciotiempo. Tal como está, este modelo no incorpora las ideas lorentzianas de la relatividad especial, ni por supuesto las de la relatividad general. Más aún, no me parece que hasta ahora se haya utilizado toda la riqueza potencial de la idea de la geometría no conmutativa. Pese a ello, el modelo está en marcha y tiene características intrigantes que entrañan predicciones con respecto a la masa del bosón de Higgs.[15]

Todas estas ideas se concentran en la construcción de nociones de «espaciotiempo» que adoptan aspectos de discretización o características «cuánticas» de algún tipo. En el resto de este capítulo describiré una idea muy diferente, a saber, la de la *teoría de twistores* (¡a la que personalmente he dedicado más de cuarenta años!), en la que no hay discretización impuesta específicamente sobre el espaciotiempo. En lugar de ello, los puntos espaciotemporales son privados de su papel primario en la teoría física. El espaciotiempo se considera una construcción (secundaria) a partir de las nociones twistoriales más primitivas. La teoría de twistores guarda ciertas relaciones con la teoría de la red de espín y con las variables de Asthekar, y posiblemente con la geometría no conmutativa, pero no lleva directamente a una noción de un «espaciotiempo discreto». En su lugar, su alejamiento de la continuidad de los números reales toma la dirección opuesta, pues invoca la *magia de los números complejos* como un principio guía primario para la física. Según la teoría de twistores, hay un papel fundamental subyacente para los números complejos en la definición de la estructura espaciotemporal, además del bien establecido papel básico de estos números en la mecánica cuántica. De este modo, se percibe un importante hilo conector entre la física de lo grande y la física de lo pequeño.

33.2. Los twistores como rayos de luz

Como hemos visto en los capítulos 21 y 22, la estructura de los números complejos es fundamental para la mecánica cuántica. Las «amplitudes» que aparecen como coeficientes en la ley de superposición básica de la mecánica cuántica son números complejos, lo que conduce a los espacios de Hilbert complejos de la teoría. Mientras que estas amplitudes se suelen considerar cantidades abstractas, y desempeñan el papel básico de proporcionar probabilidades cuando tiene lugar una

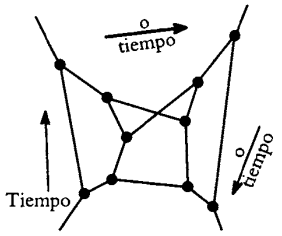


Fig. 33.4. En una red de espín del tipo original, las líneas se pueden leer como enlaces de cuanlazamiento (véanse 23.9 y la Fig. 23.7). Es posible escoger cualquier dirección temporal —hacia delante, de lado, o hacia atrás—, obteniéndose una interpretación de la red de espín igualmente válida.

medida, hemos visto (en §22.9) que hay una fuerte interconexión entre estos números complejos y la geometría espacial. Esto es más manifiesto en la mecánica cuántica de una partícula de espín 1/2, donde los posibles estados de espín corresponden a las diferentes direcciones espaciales, mediante la noción de una *esfera de Riemann*; además, en §22.10 hemos visto que los estados de espín para un espín más alto pueden describirse también en términos de la geometría espacial de la esfera de Riemann mediante la representación de Majorana. Pero no es solo en la mecánica cuántica donde vemos un papel geométrico fundamental para la esfera de Riemann. Recordemos (de §18.5) que dicha esfera desempeña un importante papel espaciotemporal en la teoría de la relatividad, puesto que el campo de visión de un observador puede considerarse también como una esfera de Riemann. Este hecho tiene una importancia seminal en la teoría de twistores, como veremos enseguida.

Otro principio guía tras la teoría de twistores es la *no localidad* cuántica. Recordemos de los extraños efectos EPR expuestos en §§23.3-6, y más en concreto del papel del «cuanlazamiento», tal como se manifiesta especialmente en el fenómeno de la teleportación cuántica, descrito en §23.10, que el comportamiento cuántico no puede entenderse por completo en términos de influencias enteramente locales de carácter causal «normal». Esto sugiere que se necesita una teoría en la que tales aspectos no locales estén incorporados en un nivel básico.

Una guía para conseguir esto puede obtenerse de la teoría de *redes de espín*. Recordemos de §32.6 que hemos de considerar que todas las redes de espín están construidas inicialmente a partir de pares EPR. Las líneas de una red de espín que aparecen como consecuencia pueden

considerarse legítimamente como enlaces de cuanlazamiento. La «información cuántica» que representa el cuanlazamiento puede «viajar» en una dirección u otra a lo largo de una línea de cuanlazamiento, o línea de red de espín. No hay especificación de un *tiempo* en la teoría de la red de espín estricta (y, de hecho, una red de espín puede leerse igualmente utilizando sentidos temporales diferentes —hacia delante, hacia atrás, lateral, etc.; véase la Fig 33.4). Por eso, los curiosos aspectos «retroactivos» del cuanlazamiento son solo reflejos de esta indiferencia a una dirección del flujo del tiempo que es una característica de las redes de espín.

Es posible considerar la teoría de twistores como una continuación del programa de la red de espín para obtener un esquema *relativista*, en el que rayos de luz idealizados (o sus generalizaciones, con espín) parecen ser, en cierto sentido, los portadores del cuanlazamiento. Las nociones espaciotemporales ordinarias no están inicialmente entre los ingredientes de la teoría de twistores, sino que van a *construirse* a partir de ellos. Esto tiene mucho en común con la filosofía subyacente tras mis redes de espín *originales*, en la que las nociones espaciales deben construirse a partir de las redes de espín, antes que pensar en las redes de espín como algo que habita en una geometría espacial previamente asignada.

La descripción twistorial del espaciotiempo resulta ser una descripción no local; además, hay un carácter fundamentalmente «holístico» en la descripción twistorial de los campos físicos que es resultado de una extraordinaria característica de la magia compleja (a saber, la cohomología de haces holomorfa) que todavía no hemos visto en este libro —aunque llegaremos a ella en §33.9 (y ya había un indicio de ello en la teoría de hiperfunciones de §9.7)— y que se mezcla con otro aspecto de la magia de los números complejos, a saber, el carácter holomorfo subyacente de la condición esencial de *frecuencia positiva* de la teoría cuántica de campos (§24.3 y §33.10). Vemos así que los aspectos no locales de la teoría de twistores están íntimamente ligados con la más importante de las motivaciones subyacentes, a saber, el deseo de explotar la magia de los números complejos, en la creencia de que la propia naturaleza puede depender muy bien de tales cosas en un nivel profundo. En esta sección, y en las próximas, veremos cómo todos es-

tos aspectos de la magia de los números complejos empiezan a reunirse en el armazón de la teoría de twistores. También veremos que la teoría de twistores tiene una profunda relación, notable e inesperada, con la relatividad *general*, y que eso proporciona una perspectiva intrigante sobre la QFT, la física de partículas y la posible generalización no lineal de la mecánica cuántica.

¿Cómo empiezan a unirse estas ideas en la teoría de twistores? Como primer paso hacia la comprensión de la teoría de twistores podemos considerar que un twistor representa un *rayo de luz* en el espaciotiempo (de Minkowski) ordinario $\mathbb{M}$. Podemos considerar que dicho rayo de luz proporciona el primitivo «enlace causal» entre un par de *sucesos* (i.e., de puntos espaciotemporales). Pero los propios sucesos deben considerarse como constructos secundarios, obtenidos a partir de sus papeles como intersecciones de rayos de luz. De hecho, es posible caracterizar un suceso R (punto espaciotemporal R) por medio de la familia de rayos de luz que pasan por R; véase la Fig. 33.5. Así pues, mientras que en la imagen espaciotemporal normal un rayo de luz Z es un lugar geométrico y un suceso R es un punto, hay una chocante inversión en el espacio twistorial, puesto que ahora el rayo de luz se describe como un *punto* $\mathbf{Z}$ y un suceso se describe como un *lugar geométrico* $\mathbf{R}$.

El *espacio twistorial* al que nos referimos aquí, cuyos puntos indivi-

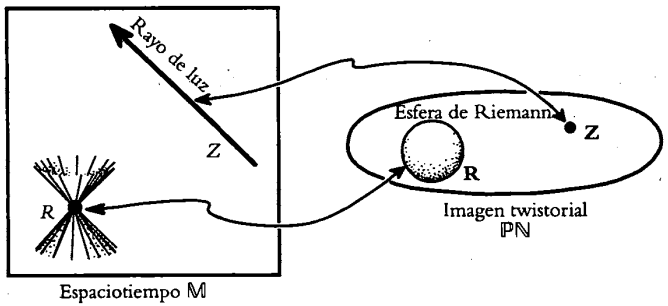


Fig. 33.5. Un rayo de luz Z en el espacio de Minkowski $\mathbb{M}$ se representa como un único punto $\mathbf{Z}$ en el espacio twistorial $\mathbb{PN}$ (espacio twistorial nulo proyectivo); un único punto $\mathbf{R}$ en $\mathbb{M}$ está representado por una esfera de Riemann $\mathbf{R}$ en $\mathbb{PN}$ (esfera que representa la «esfera celeste» de rayos de luz en R). Para la correspondencia completa, esto requiere el espacio de Minkowski compactificado $\mathbb{M}^{\#}$ descrito en la Fig. 33.9.

duales representan rayos de luz en $\mathbb{M}$, se denota[16] por $\mathbb{PN}$. (Se adopta esta notación para encajar con la terminología de §33.5). Así, el punto $\mathbf{Z}$ en $\mathbb{PN}$ corresponde al lugar geométrico Z (un rayo de luz) dentro de $\mathbb{M}$ y el punto R en $\mathbb{M}$ corresponde al lugar geométrico $\mathbf{R}$ (una esfera de Riemann; véase §18.5) dentro de $\mathbb{PN}$. Ahora bien, una parte esencial de la filosofía de la teoría de twistores es que las nociones físicas ordinarias, que normalmente se describen en términos espaciotemporales, deben traducirse en una descripción equivalente (aunque relacionada no localmente) en el espacio twistorial. Vemos que la relación entre $\mathbb{M}$ y $\mathbb{PN}$ es en realidad una correspondencia no local, en lugar de una transformación punto a punto. Sin embargo, el espacio $\mathbb{PN}$ nos ofrece los comienzos de una traducción semejante. La riqueza completa de la geometría twistorial —que resulta ser muy notable— solo se revela poco a poco, a medida que se desarrolla en todos sus detalles la correspondencia entre conceptos espaciotemporales y geometría en el espacio twistorial

Este lugar geométrico $\mathbf{R}$ dentro de $\mathbb{PN}$ describe la «esfera celeste» (el campo de visión total) de un observador en R, donde la esfera celeste de R se considera como la familia de rayos de luz que pasan por R. Como se ha señalado antes, esta esfera es naturalmente una *esfera de Riemann* que es un *espacio 1-dimensional complejo* (una curva compleja; véase el capítulo 8). Así, consideramos los puntos espaciotemporales como *objetos holomorfos* en el espacio twistorial $\mathbb{PN}$, de acuerdo con la filosofía de números complejos que subyace en la teoría de twistores. En §§33.5,6 veremos explícitamente cómo puede extenderse esta «filosofía holomorfa» a la geometría de un espacio twistorial más completo $\mathbb{T}$, y en §§33.8-12 cómo nos permite codificar de una forma notable la información de los campos sin masa lineales y no lineales.

No obstante, el propio espacio $\mathbb{PN}$, de los rayos de luz, no encaja inmediatamente en la «filosofía holomorfa» porque no es un espacio complejo. $\mathbb{PN}$ no puede ser una variedad compleja porque tiene *cinco* dimensiones reales[33.1] y cinco es un número impar, mientras que cualquier n-variedad compleja debe tener un número *par*, $2n$, de dimensiones reales (véase §12.9). En §33.6 veremos que si hacemos nuestros

[33.1] ¿Por qué los rayos de luz tienen cinco grados de libertad?

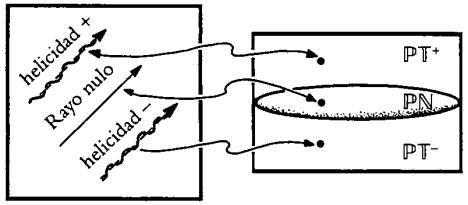


Fig. 33.6. La 5-variedad real $\mathbb{PN}$ divide el espacio twistorial proyectivo $\mathbb{PT}$ en dos 3-variedades complejas, $\mathbb{PT}^+$ y $\mathbb{PT}^-$, que representan partículas sin masa de helicidad positiva y negativa, respectivamente.

«rayos de luz» un poco más parecidos a partículas físicas sin masa, asignándoles espín (en realidad, helicidad; véase §22.7) y energía, entonces obtenemos un espacio $\mathbb{PT}$ *seis*-dimensional que realmente *puede* interpretarse como un espacio complejo, de *tres* dimensiones complejas. El espacio $\mathbb{PN}$ se acomoda dentro de $\mathbb{PT}$, dividiéndolo en dos variedades complejas $\mathbb{PT}^+$ y $\mathbb{PT}^-$, donde puede considerarse que $\mathbb{PT}^+$ representa partículas sin masa de helicidad positiva y $\mathbb{PT}^-$ representa partículas sin masa de helicidad negativa; véase la Fig. 33.6. Sin embargo, no sería correcto pensar en los twistores como partículas sin masa. En su lugar, los twistores proporcionan las variables en cuyos términos deben expresarse las partículas sin masa. (Esto es comparable al uso ordinario de un 3-vector de posición $\mathbf{x}$ para etiquetar un punto en el espacio. Aunque una partícula podría ocupar el punto etiquetado por $\mathbf{x}$, no sería correcto identificar la partícula con el vector $\mathbf{x}$.)

La perspectiva twistorial nos lleva a una visión del «espaciotiempo cuantizado» muy diferente de la que se suele proponer. Un punto de vista bastante «convencional» sostiene que los procedimientos de la teoría (cuántica) de campos deben aplicarse al tensor métrico g_{ab}, considerado este como un campo tensorial en (la variedad de) el espaciotiempo. Se expresa la idea de que la *métrica cuantizada* mostrará aspectos de «borrosidad» debido al principio de incertidumbre de Heisenberg. Se nos presenta la imagen de cierto tipo de espacio tetradimensional que posee una «métrica borrosa», de modo que, en particular, los conos nulos —y en consecuencia, la noción de *causalidad*— quedan sujetos a «incertidumbres cuánticas» (véase la Fig. 33.7a). Por consiguiente, no hay una noción clásicamente bien definida acerca de si un vector del espaciotiempo es de género espacio, de género tiempo o nulo. Esta cuestión ha planteado dificultades fundamentales para cualquier «teoría cuántica de la gravedad» demasiado convencional, pues una característica básica de la QFT es que los requisitos de causalidad exigen la conmutación de los operadores de campo definidos en sucesos con separación de género espacio (§26.11). Si la propia noción de «de género espacio» está sujeta a incertidumbres cuánticas (o se ha convertido en una noción cuántica), entonces los métodos estándar de la QFT —que implican la especificación de relaciones de conmutación para operadores de campo (§§26.2,3)— no pueden aplicarse directamente. La teoría de twistores sugiere una imagen muy diferente. Ahora, cualesquiera que sean los procedimientos de «cuantización» adecuados, deben ser aplicados dentro de un espacio twistorial y no dentro del espaciotiempo (como hubiera sido el punto de vista «convencional»). De forma análoga a como en la aproximación convencional los «sucesos» quedan intactos mientras que los «conos de luz» se hacen borrosos, en una aproximación basada en twistores son ahora los «rayos de luz» los que quedan intactos mientras que los «sucesos» se hacen borrosos (véase la Fig. 33.7b).

Fig. 33.7. (a) Con respecto a la posible naturaleza de un «espaciotiempo cuantizado», ha sido un punto de vista común que debería ser un tipo de espaciotiempo con una métrica «borrosa», que lleva a una especie de cono de luz «borroso», donde la noción de dirección en un punto, ya sea nula, de género tiempo o de género espacio, estaría sujeta a incertidumbres cuánticas. (b) Una perspectiva más «twistorial» sería tomar el espacio twistorial (en este caso $\mathbb{PN}$) para retener un tipo de existencia (de modo que seguiría habiendo rayos de luz), pero la condición de su intersección estaría ahora sometida a incertidumbres cuánticas. Por consiguiente, la que ahora se haría borrosa es la noción de «punto espaciotemporal».

Como acabamos de ver, la teoría de twistores, explota inicialmente una manifestación de la magia de los números complejos diferente de las que se encuentran en la teoría cuántica, a saber, la característica *clasica* de la geometría espaciotemporal por la que la esfera celeste puede considerarse como una esfera de Riemann, que es una variedad

compleja 1-dimensional. La idea es que esto nos ofrece indicios respecto al esquema *real* de cosas de la naturaleza, que en definitiva debe unificar la estructura espaciotemporal con los procedimientos de la mecánica cuántica. Hay que decir que esta característica de la geometría espaciotemporal es específica de la dimensión y la signatura concretas que nuestro espaciotiempo posee realmente. En realidad, el hecho de que la esfera de Riemann desempeñe un papel importante como esfera celeste en la teoría de la relatividad (§18.5) requiere que el espaciotiempo sea 4-dimensional y lorentziano, en abierto contraste con las ideas subyacentes en la teoría de supercuerdas y otros esquemas de tipo Kaluza-Klein. La magia compleja de la teoría de twistores es muy específica de la geometría espaciotemporal 4-dimensional de la teoría de la relatividad (especial) ordinaria, y no tiene la misma relación estrecha con la «geometría espaciotemporal» de dimensiones más altas (véase §33.4 *infra*).

Para ir más lejos, volvamos a la imagen de la red de espín pura original, y recordemos que su carencia principal era que no había ninguna referencia al desplazamiento espacial. En dicha teoría, los ángulos euclídeos surgen como una especie de «límite geométrico» de la teoría de red de espín pura; pese a todo, en esta teoría no aparecen distancias. En el esquema de variables de lazo, el aspecto de «distancia» de las cosas es abordado por los números ($n = 2j$) en las líneas que se refieren al *área* antes que al espín. Pero esto es diferente de la interpretación de la teoría de redes de espín original, donde no hay medida de distancia porque el espín es momento *angular*, que tiene que ver meramente con rotaciones y ángulos. Para poder incorporar desplazamientos traslacionales y distancias reales necesitaríamos un papel correspondiente para el momento *lineal* en la teoría. Por consiguiente, parece que necesitamos pasar del grupo de rotación al grupo completo de movimientos euclídeos y, para un esquema propiamente relativista, al grupo de Poincaré (§18.12).[17]

A finales de la década de 1950 y principios de la de 1960, cuando me dedicaba activamente a este tema, todavía no se había desarrollado la teoría de las variables de lazo, y contemplaba una generalización de las redes de espín en la que el grupo de Poincaré participa directamente. Sin embargo, me preocupaba un aspecto incómodo del grupo de Poincaré —que no es semisimple (véase §13.7)— que tiene desagradables implicaciones con respecto a sus representaciones. En aquella época, tenía la idea de que una extensión del grupo de Poincaré a lo que se conoce como el grupo *conforme* (que *sí* es semisimple) podría constituir el análogo relevante de la teoría de redes de espín en una estructura más satisfactoria matemáticamente. El grupo conforme amplía el grupo de Poincaré exigiendo meramente que se conserven los conos de luz, antes que la métrica del espacio de Minkowski. De hecho, resulta que el grupo conforme tiene un papel importante en la teoría de twistores, pues es también el grupo de simetría del espacio $\mathbb{PN}$ de los rayos de luz (idealizados). (La parte no reflexiva del grupo conforme es también el grupo de simetría de cada uno de los espacios $\mathbb{PT}^+$ y $\mathbb{PT}^-$, antes mencionados, que describen partículas sin masa con helicidad y energía.) En las dos secciones siguientes veremos de forma más explícita cuál es el papel de este grupo.

33.3. El grupo conforme; el espacio de Minkowski compactificado

Antes me he referido al grupo conforme del espaciotiempo. Tratemos de explorar un poco más en detalle el papel de este grupo. Tiene una importancia especial en física en relación con los campos sin masa (por ejemplo, el campo de Maxwell), pues resulta que las ecuaciones de campo para los campos sin masa son invariantes bajo este grupo más grande, y no meramente bajo el grupo de Poincaré.[18] Se puede adoptar la postura de que en el nivel más fundamental las partículas/campos sin masa son los ingredientes básicos, siendo la masa algo que entra en una etapa posterior. De hecho, este parece ser el punto de vista implícito en el modelo estándar, descrito en el capítulo 25, en el que la masa se introduce por medio del bosón de Higgs y se considera que surge solo a través de un mecanismo de ruptura de simetría (§25.5). Sea como fuere, una de las importantes motivaciones subyacentes tras la teoría de twistores fue en realidad una creencia en la importancia básica de los campos sin masa y el grupo conforme. Observaremos (§33.8) que las partículas y los campos sin masa tienen una descripción extra-

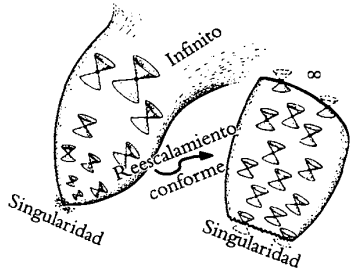


Fig. 33.8. La estructura de conos nulos de una variedad lorentziana $\mathcal{M}$ es equivalente a su estructura conforme. Un reescalamiento conforme de $\mathcal{M}$ afecta a su métrica, pero no a sus propiedades de causalidad. (Una métrica $\boldsymbol{g}$ reescala conformemente a $\boldsymbol{g}'$ si $\boldsymbol{g}' = \Omega^2\boldsymbol{g}$, siendo el campo escalar Ω positivo en todas partes.) En circunstancias favorables, dichos reescalamientos pueden ser útiles para «traer a la vista» singularidades y regiones infinitas.

ordinariamente concisa en la teoría de twistores, y este hecho constituye una de las piedras angulares de dicha teoría.

¿Qué es exactamente el grupo conforme? Estrictamente hablando, este grupo no actúa sobre el espacio de Minkowski $\mathbb{M}$, sino sobre una ligera extensión de $\mathbb{M}$ conocida como espacio de Minkowski *compactificado* $\mathbb{M}^\#$. El espacio $\mathbb{M}^\#$ es una variedad cerrada bellamente simétrica, que en muchos aspectos tiene una geometría más elegante que la del propio espacio de Minkowski. No obstante, no hay que pensar en él como «espaciotiempo real», sino como una conveniencia matemática. Es un útil intermediario para la comprensión de la geometría twistorial y su relación con la geometría espaciotemporal física.

Una imagen que no hay que olvidar es la esfera de Riemann y su relación con el plano complejo. Recordemos de §8.3 que la esfera de Riemann se obtiene a partir del plano complejo añadiendo a este un «elemento infinito», a saber, el punto etiquetado ∞, y cuando lo hemos hecho obtenemos una estructura geométrica con una simetría mayor incluso que la del plano del que hemos partido. De manera similar, el «espacio de Minkowski compactificado» $\mathbb{M}^\#$ se obtiene a partir del espacio de Minkowski ordinario $\mathbb{M}$ añadiéndole un «elemento infinito» que, esta vez, resulta ser un *cono de luz en el infinito* completo. El espacio resultante tiene una simetría mayor (a saber, el grupo conforme) que el propio espacio de Minkowski.

Veamos cómo funciona esto. El espacio $\mathbb{M}^\#$ resulta ser una variedad compacta real 4-dimensional con una métrica conforme lorentziana. Recordemos de §27.12 que una métrica conforme lorentziana es precisamente la familia de conos nulos especificados en el espacio. Esta es-

tructura se expresa más comúnmente en términos de una clase de equivalencia de métricas, donde una métrica $\boldsymbol{g}$ se considera equivalente a una métrica $\boldsymbol{g}'$ si $\boldsymbol{g}' = \Omega^2\boldsymbol{g}$ para algún campo escalar Ω que es positivo en todas partes. Este reescalamiento conserva, de hecho, los conos nulos (Fig. 33.8). Ahora, para pasar de $\mathbb{M}$ (considerada como una variedad conforme) a la variedad conforme compacta $\mathbb{M}^\#$, añadimos la 3-superficie $\mathscr{I}$ a la que antes nos hemos referido como el «cono de luz en el infinito». Recordemos las 3-superficies $\mathscr{I}^-$ y $\mathscr{I}^+$ (llamadas «scri menos» y «scri más», respectivamente) que representan los infinitos nulos pasado y futuro del espacio de Minkowski (véase la Fig. 27.16b). Podemos construir $\mathbb{M}^\#$ *identificando* $\mathscr{I}^-$ con $\mathscr{I}^+$ de la manera indicada en la Fig. 33.9. Un punto de $\mathscr{I}^-$ se considera el mismo que un punto correspondiente de $\mathscr{I}^+$ que es espacialmente *antipodal* a él (sobre la 2-esfera que representan la mayoría de los puntos del diagrama). El cono de luz de un punto a^- sobre $\mathscr{I}^-$ se concentra de nuevo en el punto a^+ sobre $\mathscr{I}^+$, y son los puntos a^- y a^+ los que deben identificarse. Además, los tres puntos que representan infinitos espaciales y temporales i^-, i^0 e i^+ son también identificados como el punto único i.[33.2] La variedad conforme $\mathbb{M}^\#$ tiene más simetría que el espacio de Minkowski, pues tiene un grupo de simetría 15-dimensional —el *grupo conforme*— en lugar de meramente el grupo de Poincaré 10-dimensional.

Existe una forma elegante de describir el espacio $\mathbb{M}^\#$ y su grupo de transformaciones. Consideremos el «cono de luz» $\mathcal{K}$ del origen O en un 6-espacio pseudoeuclídeo $\mathbb{E}^{2,4}$ con signatura $++----$. Escojamos coordenadas estándar w, t, x, y, z, v para $\mathbb{E}^{2,4}$, de modo que $\mathcal{K}$ viene dado por la ecuación

$$w^2 + t^2 - x^2 - y^2 - z^2 - v^2 = 0,$$

siendo la métrica ds^2 de $\mathbb{E}^{2,4}$

$$ds^2 = dw^2 + dt^2 - dx^2 - dy^2 - dz^2 - dv^2.$$

Esto es un «cono» 5-dimensional con vértice O. He hecho lo me-

[33.2] Vea si puede describir con más detalle la geometría de $\mathbb{M}^\#$ explicando la identificación punto a punto de $\mathscr{I}^+$ con $\mathscr{I}^-$ en términos espaciotemporales ordinarios. ¿Puede ver por qué la topología de $\mathbb{M}^\#$ es $S^1 \times S^3$? ¿Puede pensar en una diferencia importante que ocurriría para un número impar de dimensiones espaciotemporales?

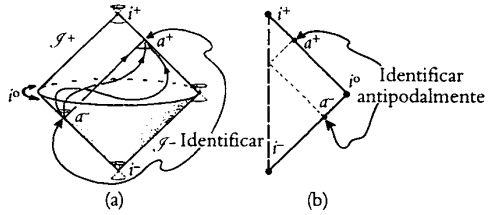


Fig. 33.9. El espacio de Minkowski compactificado $\mathbb{M}^\#$ se obtiene a partir del espacio de Minkowski ordinario $\mathbb{M}$ añadiendo sus infinitos nulos futuro $\mathscr{I}^+$ y pasado $\mathscr{I}^-$ e identificándolos luego adecuadamente. (a) El cono de luz futuro de cualquier punto a^- en $\mathscr{I}^-$ se concentra de nuevo en otro vértice a^+ en $\mathscr{I}^+$ (siendo este «cono de luz» en términos ordinarios simplemente la historia de un frente de onda plano que viaja a la velocidad de la luz), y a^- debe indentificarse con a^+. El infinito de género espacio i^0 y los infinitos pasado y futuro i^- e i^+ de género tiempo tienen que identificarse como un único punto. (b) Se muestra esta identificación $\mathscr{I}$, en términos del diagrama conforme estricto de la Fig. 27.16b, donde a^- es antipodal a a^+ sobre la «S^2 de rotación» para el diagrama entero.

jor que he podido con él en la Fig. 33.10, pero uno de los aspectos principales en que la imagen resulta equívoca es el hecho de que lo que parecen ser dos «piezas» distintas de $\mathcal{K}$ («pasado» y «futuro») están conectadas realmente en «una pieza».[33.3] Consideremos ahora la *sección* de $\mathcal{K}$ por el 5-plano nulo $w - v = 1$. Esta intersección es una 4-variedad («paraboloide») cuya métrica intrínseca, inducida a partir de la de $\mathbb{E}^{2,4}$ es[33.4]

$$ds^2 = dt^2 - dx^2 - dy^2 - dz^2.$$

Reconocemos esto como la forma métrica del 4-espacio de Minkowski plano ordinario (§18.1), de modo que podemos identificarla con $\mathbb{M}$, incluso si está insertado de una forma «doblada» en $\mathbb{E}^{2,4}$ (con la apariencia de una parábola en la Fig. 33.10). ¿Cómo encontramos $\mathbb{M}^\#$ en esta imagen? Es el espacio abstracto de los *generadores* completos de $\mathcal{K}$ (líneas rectas que pasan por O y yacen en $\mathcal{K}$, donde la recta compleja que pasa por O en ambas direcciones cuenta como un único gene-

[33.3] ¿Puede ver por qué?
[33.4] ¿Por qué?

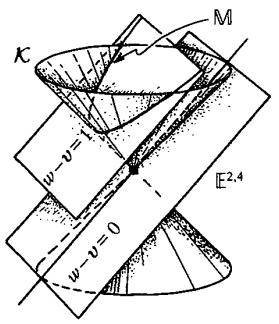


Fig. 33.10. El espacio de Minkowski compactificado $\mathbb{M}^\#$ puede identificarse como el espacio de generadores del «cono de luz» $\mathcal{K}$, en un $\mathbb{E}^{2,4}$ pseudoeuclídeo, dado por $w^2 + t^2 - x^2 - y^2 - z^2 - v^2 = 0$. La 4-variedad sección «paraboloide» $\mathbb{M}$ de $\mathcal{K}$ por el 5-plano nulo $w - v = 1$ tiene métrica intrínseca minkowskiana $ds^2 = dt^2 - dx^2 - dy^2 - dz^2$. La familia de generadores de $\mathcal{K}$ en $w - v = 0$ (no visible en este diagrama debido a la representación de solo una dimensión «temporal») son paralelas a $w - v = 1$ y no cortan a $\mathbb{M}$, y estos generadores proporcionan los puntos de $\mathscr{I}$

rador). Así, podemos considerar que cada punto de $\mathbb{M}^\#$ *es* simplemente un generador de $\mathcal{K}$ (Fig. 33.10), ¡de modo que $\mathbb{M}^\#$ es la «esfera celeste» para un «observador» situado en el origen de $\mathbb{E}^{2,4}$!

¿Por qué funciona esto? Cada generador que no yace en el 5-plano $w - v = 0$ corta a $\mathbb{M}$ en un único punto, de modo que esta familia de generadores está en una correspondencia continua 1-1 con $\mathbb{M}$. Pero, además, están los generadores que *yacen* en este 5-plano. Estos suministran a $\mathbb{M}$ los puntos adicionales que constituyen $\mathscr{I}$. El espacio $\mathbb{M}^\#$ definido de esta manera tiene una métrica lorentziana conforme que está proporcionada localmente por la de cualquier sección transversal local de $\mathcal{K}$.[33.5]

El grupo pseudoortogonal O(2, 4) que actúa sobre $\mathbb{E}^{2,4}$ (véanse §13.8 y §§18.1,2) consta de las «rotaciones» que conservan la métrica ds^2. Esto envía generadores de $\mathcal{K}$ a otros generadores de $\mathcal{K}$, de modo que envía $\mathbb{M}^\#$ sobre sí mismo. Además, conserva la estructura conforme de $\mathbb{M}^\#$.[33.6] Hay exactamente dos elementos de O(2, 4) que actúan como la identidad sobre $\mathbb{M}^\#$, a saber, el propio elemento identidad de O(2, 4) y el elemento identidad negativo de O(2, 4), que simplemente invierte la dirección de cada generador. Aparte de la naturaleza dos a uno de la correspondencia que surge de esta reversibilidad de las di-

[33.5] ¿Por qué la métrica conforme que proporciona una sección transversal cualquiera es la misma que la de cualquier otra? ¿Por qué los puntos de $\mathscr{I}$, definidos de esta manera, coinciden con la definición dada antes? *Sugerencia*:Véanse §18.4 y la Fig. 18.9.
[33.6] ¿Por qué? Puede suponer el resultado del ejercicio [33.5].

recciones de los generadores, O(2, 4) es el grupo conforme. Incluye un subgrupo 10-dimensional que conserva el 5-plano $w - v = 0$, y esto da el *grupo de Poincaré* de $\mathbb{M}$.[33.7] De hecho, este argumento es tan solo una versión en dimensiones más altas del que hemos dado en §18.5 cuando mostrábamos que las transformaciones conformes de una esfera ordinaria (que es el plano euclídeo compactificado) proporcionan una realización del grupo de Lorentz O(1, 3); véase la Fig. 18.9.

33.4. Los twistores como espinores de dimensión superior

¿Cómo encajan los twistores en todo esto? La forma más fácil (pero en absoluto la más transparente) de describir un twistor (en un espacio de Minkowski) es decir que es un *espinor reducido* (o medio espinor) para O(2, 4). (No hay que alarmarse por el laconismo matemático de esta descripción; ¡pronto daré una imagen mucho más física!) Véase §11.5 para una breve mención de la noción de un espinor reducido. Para un espacio $2n$-dimensional, sobre el que actúa un grupo pseudoortogonal $O(n - r, n + r)$, el espacio de espinores reducidos es 2^{n-1}-dimensional. En el caso presente, $n = 3$ (y $r = 1$), así que tenemos un espacio 4-dimensional de espinores reducidos, conocido como *espacio twistorial*.[19]

Por desgracia, con una definición como esta, no nos hacemos una imagen física o geométrica clara de cómo es un twistor. Además, vemos que debería existir una teoría de twistores para cualquier número par $2(n - 1)$ de dimensiones espaciotemporales, pese a lo que se ha dicho hacia el final de §33.2. Generalizamos la construcción anterior para $\mathcal{K}$ (tomándolo ahora de $2n - 1$ dimensiones) y la compactificación del espaciotiempo de Minkowski $2(n - 1)$-dimensional funciona de un modo exactamente análogo al que se ha dado antes, donde simplemente introducimos las dos nuevas coordenadas v y w, como antes, una con un signo menos en la métrica, y la otra con un signo más. El «espacio twistorial» es ahora 2^{2n-1}-dimensional. Esto también funcionará para un número impar $2n - 1$ de dimensiones espaciotemporales, salvo

[33.7] ¿Cuál es la condición explícita sobre una matriz (6×6) para que represente un elemento infinitesimal de O(2, 4)? ¿Cuál de estas matrices da transformaciones de Poincaré infinitesimales?

que ahora no tenemos la noción de espinores reducidos y es todo el espacio de espín 2^n-dimensional el que tendría que contar como nuestros «twistores». Sin embargo, en el caso de dimensión impar se pierde una característica importante de los twistores, a saber, su naturaleza quiral (que discutiremos más en detalle en §§33.7,12,14). Solo cuando pasamos a los espacios de espín reducido conseguimos un formalismo esencialmente quiral (de modo que las entidades dextrógiras y levógiras reciben descripciones twistoriales diferentes; véase §33.7), y podemos mantener la esperanza de que, en última instancia, puedan incorporarse los aspectos quirales de las interacciones débiles (§25.3). También veremos más adelante por qué esta definición general n-dimensional de un twistor carece de muchas de las propiedades físicas (y holomorfas) clave que hacen tan efectiva la teoría de twistores.

Puesto que los twistores se refieren a un grupo activo de transformaciones espaciotemporales (el grupo conforme) donde puntos del espaciotiempo son enviados a otros puntos del espaciotiempo bajo la acción del grupo, los twistores se ven como entidades que se refieren globalmente al espaciotiempo total, más que a puntos individuales en el espaciotiempo. Las cantidades locales tales como vectores, tensores o espinores ordinarios se refieren al grupo de simetría que actúa en un punto; véase §14.1 (por ejemplo, el grupo de rotaciones o el grupo de Lorentz; §13.8). Aunque esto hace de los twistores algo más difícil de manejar que los vectores, tensores o espinores ordinarios, esta globalidad tiene una ventaja cuando buscamos un formalismo que pretenda reemplazar el espaciotiempo y no solo definirlo en referencia a una variedad espaciotemporal previamente dada. De hecho, como se ha mencionado en §33.2, uno de los objetivos principales de la teoría de twistores es obtener un formalismo semejante. El inconveniente principal de esta aproximación es que resulta difícil ver cómo un formalismo semejante puede aplicarse a un espaciotiempo curvo general $\mathcal{M}$, cuando cosas tales como el grupo conforme no aparecen como una simetría de $\mathcal{M}$. En §§33.11,12 veremos cómo la teoría de twistores resuelve esta dificultad de una forma notable.

La definición anterior de un twistor, como un espinor reducido para O(2, 4), nos ofrece solo una perspectiva muy limitada sobre las ideas y la motivación de la teoría de twistores. No hay nada específica-

mente 4-dimensional en esta aproximación, y no da ninguna indicación clara de por qué habría que interesarse en la teoría de twistores con el fin de que nos proporcione una guía para avanzar en nuestra búsqueda de una teoría más profunda de la naturaleza. Para apreciar con más detalle lo que la teoría de twistores está tratando de decirnos en este aspecto, recordemos el mensaje de los capítulos 29 y 30. Mientras que normalmente se acepta que la unión apropiada cuántico/gravedad debe ser un objetivo principal en la búsqueda de una perspectiva fundamentalmente nueva en física, el mensaje de estos capítulos es que deberíamos buscar un desarrollo en el que las propias reglas de la teoría cuántica (de campos) no sigan siendo sacrosantas, sino que deberían doblegarse, igual que deberían hacerlo nuestras imágenes espaciotemporales convencionales. De todas formas, hay evidentemente mucha verdad y belleza en los principios de la mecánica cuántica, y estos no deberían ser abandonados sin más. En la teoría de twistores, en lugar de imponer reglas QFT, se examinan dichas reglas y se trata de extraer características que se mezclen con las de las ideas de Einstein, buscando armonías ocultas entre la relatividad y la mecánica cuántica. Como se ha dicho antes, un elemento clave para guiarse es la magia de los números complejos que ha intervenido en tantos lugares de este libro. Otro es una armonía especial con la teoría de Einstein del 4-espacio lorentziano antes que con sus generalizaciones a dimensiones más altas o a otras signaturas.

¿Qué hay de especial a este respecto en el 4-espacio lorentziano? Como se ha señalado en §18.5 y en §33.2, la esfera celeste de un observador tiene una estructura conforme natural y puede interpretarse como una esfera de Riemann. Habría que tener en cuenta que algo de esta naturaleza ocurre realmente en *cualquier* número (no nulo) de dimensiones de espacio y tiempo, donde la esfera celeste tiene siempre la estructura de una variedad conforme.[33.8] Lo que hay de especial en el caso 4-dimensional lorentziano es que dicha variedad conforme puede interpretarse de modo natural como una variedad *compleja* (la esfera de Riemann), una propiedad que no aparece en ningún otro número de dimensiones de espacio y tiempo. ¿Cuál es la importancia de esto? En la

[33.8] Explique por qué.

teoría de twistores, la magia de los números complejos se explota al máximo. No solo el espacio twistorial resulta ser una variedad compleja, sino que dicha variedad compleja tiene una interpretación física directa. De hecho, resultados generales nos dicen que los únicos casos en los que el «espacio twistorial» es un espacio complejo de cualquier tipo[20] son aquellos en que la diferencia entre el número de dimensiones espaciales y el número de dimensiones temporales deja resto 2 cuando se divide por 4. Hay que señalar que este no es el caso para la teoría de Kaluza-Klein original, ni para la teoría de la supergravedad 10 u 11-dimensional, ni para la teoría de cuerdas 26-dimensional original, ni para la teoría de supercuerdas 10-dimensional, ni para la supergravedad o la teoría M 11-dimensional, ni siquiera para la teoría-F 12-dimensional (puesto que en este caso hay 2 dimensiones temporales).

33.5. Geometría twistorial básica y coordenadas

¿Cuál es la interpretación física o geométrica de un twistor general para el 4-espacio minkowskiano ordinario? Es más fácil describirlo si utilizamos coordenadas minkowskianas estándar t, x, y, z para un punto R de $\mathbb{M}$, donde tomamos la velocidad de la luz como la unidad, $c = 1$. El espacio twistorial completo $\mathbb{T}$ es un *espacio vectorial complejo 4-dimensional*, para el que se utilizarán las coordenadas complejas Z^0, Z^1, Z^2, Z^3. Decimos que un twistor $\mathbf{Z}$ con estas coordenadas es *incidente* con el punto R del espaciotiempo (o que R es incidente con $\mathbf{Z}$) si se satisface la relación matricial clave (véase §13.3 para la notación matricial)

$$\begin{pmatrix} Z^0 \\ Z^1 \end{pmatrix} = \frac{\mathrm{i}}{\sqrt{2}} \begin{pmatrix} t+z & x+\mathrm{i}y \\ x-\mathrm{i}y & t-z \end{pmatrix} \begin{pmatrix} Z^2 \\ Z^3 \end{pmatrix},$$

¡de la que se siguen todas las propiedades básicas de la geometría twistorial en un espacio plano![33.9]

De acuerdo con la notación de §12.8, la notación de índices (abstactos) Z^α se utilizará a veces para representar el twistor $\mathbf{Z}$ (donde las componentes de $\mathbf{Z}$ en un sistema estándar serían Z^0, Z^1, Z^2, Z^3). Cada twistor $\mathbf{Z}$, o Z^α (un elemento de $\mathbb{T}$) tiene un complejo conjugado $\bar{\mathbf{Z}}$,

[33.9] Escriba esta ecuación en notación algebraica ordinaria.

que es un twistor *dual* (elemento del espacio twistorial dual $\mathbb{T}^*$). En la forma de índices, $\bar{\mathbf{Z}}$ se escribe $\bar{Z}_\alpha$, con un subíndice, y sus componentes (en el sistema estándar) serían

$$(\bar{Z}_0, \bar{Z}_1, \bar{Z}_2, \bar{Z}_3) = (\bar{Z}^2, \bar{Z}^3, \bar{Z}^0, \bar{Z}^1).$$

Esta notación es probablemente algo confusa. Las cuatro cantidades (números complejos) de la izquierda son simplemente las cuatro componentes del twistor dual $\bar{\mathbf{Z}}$. Las cuatro de la derecha son los respectivos complejos conjugados de los números complejos Z^2, Z^3, Z^0, Z^1. Así pues, la componente $\bar{Z}_0$ de $\bar{\mathbf{Z}}$ es el complejo conjugado de la componente Z^2 de $\mathbf{Z}$, etc. Nótese el intercambio de las dos primeras con las dos segundas cuando se forma la conjugación compleja. Puesto que $\bar{\mathbf{Z}}$ es un tensor dual, podemos formar su producto escalar (hermítico) (véanse §13.9 y §22.3) con el twistor original $\mathbf{Z}$ para obtener la *norma twistorial* (al cuadrado)

$$\begin{aligned}\bar{\mathbf{Z}} \bullet \mathbf{Z} = \bar{Z}_\alpha Z^\alpha &= \bar{Z}_0 Z^0 + \bar{Z}_1 Z^1 + \bar{Z}_2 Z^2 + \bar{Z}_3 Z^3 = \\ &= \overline{Z^2} Z^0 + \overline{Z^3} Z^1 + \overline{Z^0} Z^2 + \overline{Z^1} Z^3 = \\ &= \tfrac{1}{2}(|Z^0 + Z^2|^2 + |Z^1 + Z^3|^2 - |Z^0 - Z^2|^2 - |Z^1 - Z^3|^2),\end{aligned}$$

donde esta última fórmula muestra que la expresión hermítica $\bar{Z}_\alpha Z^\alpha$ tiene signatura $(+ + - -)$, de acuerdo con §13.9.[33.10] (La simetría del espacio twistorial exhibe la equivalencia local, mencionada en §3.10, del grupo SU(2, 2) con el O(2, 4) de §33.3.) A partir de la relación de incidencia clave dada arriba, observamos que un twistor Z^α puede ser incidente con un suceso en el espacio de Minkowski real $\mathbb{M}$ solo si su norma *se anula*: $\bar{Z}_\alpha Z^\alpha = 0$.[33.11] Cuando $\bar{Z}_\alpha Z^\alpha = 0$, decimos que el twistor $\mathbf{Z}$ es *nulo*.

Para enlazar con la exposición de §33.2, deberíamos primero familiarizarnos con el espacio twistorial *proyectivo* $\mathbb{PT}$, que es el 3-espacio proyectivo complejo ($\mathbb{CP}^3$) construido a partir del espacio vectorial complejo $\mathbb{T}$. (Véase §15.6 para una discusión general de los espacios proyectivos.) Buena parte de la geometría twistorial se expresa muy fácilmente en términos de $\mathbb{PT}$ en lugar de $\mathbb{T}$. Los números Z^0, Z^1, Z^2, Z^3 proporcionan ahora *coordenadas homogéneas* para $\mathbb{PT}$, de modo que las tres razones independientes

$$Z^0 : Z^1 : Z^2 : Z^3,$$

sirven para etiquetar puntos de $\mathbb{PT}$. Los twistores proyectivos *nulos* constituyen el espacio $\mathbb{PN}$, que es el subespacio 5-real dimensional del espacio 6-real dimensional $\mathbb{PT}$ para el que la norma twistorial se anula:

$$\bar{Z}_\alpha Z^\alpha = 0.$$

Esta ecuación define también el subespacio 7-real dimensional $\mathbb{N}$ de los twistores *no* proyectivos nulos, en el espacio vectorial $\mathbb{T}$. Cuando $\bar{Z}_\alpha Z^\alpha > 0$, obtenemos el espacio $\mathbb{T}^+$ de los twistores *positivos*, y cuando $\bar{Z}_\alpha Z^\alpha < 0$, obtenemos el espacio $\mathbb{T}^-$ de los twistores *negativos*. Los espacios proyectivos $\mathbb{PT}^+$ y $\mathbb{PT}^-$ se definen en correspondencia con esto. Véase la Fig. 33.11 (y compárese con la Fig. 33.6).

Exploremos la relación geométrica entre $\mathbb{PN}$ y $\mathbb{M}$ que se muestra en la Fig. 33.5 como una consecuencia de la relación de incidencia cla-

[33.10] Verifique esta expresión final; explique por qué esto nos dice la signatura.
[33.11] Demuéstrelo; demuestre, recíprocamente, que un suceso semejante existe siempre si $\bar{Z}_\alpha Z^\alpha = 0$, con tal de que Z^2 y Z^3 no se anulen simultáneamente.

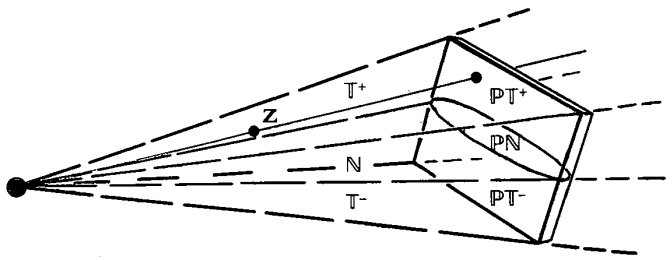


Fig. 33.11. El espacio twistorial $\mathbb{T}$ es un espacio vectorial complejo con una métrica pseudohermítica. El espacio twistorial proyectivo $\mathbb{PT}$ es el espacio de rayos (subespacios 1-dimensionales) en $\mathbb{T}$. Así pues, si un twistor $\mathbf{Z}$ tiene coordenadas (Z^0, Z^1, Z^2, Z^3), las razones $Z^0 : Z^1 : Z^2 : Z^3$ determinan el punto correspondiente en $\mathbb{PT}$. El subespacio 7-real dimensional $\mathbb{N}$ (de twistores nulos $\bar{Z}_\alpha Z^\alpha = 0$) divide el espacio twistorial $\mathbb{T}$ en los 4-espacios complejos $\mathbb{T}^+$ (de twistores positivos $\bar{Z}_\alpha Z^\alpha > 0$) y $\mathbb{T}^-$ (de twistores negativos $\bar{Z}_\alpha Z^\alpha < 0$). Las versiones proyectivas respectivas de estos espacios son los 5-real dimensionales $\mathbb{PN}$ (que representan rayos de luz en $\mathbb{M}^\#$) y las dos 3-variedades complejas $\mathbb{PT}^+$ (que representan partículas sin masa de helicidad positiva) y $\mathbb{PT}^-$ (que representan partículas sin masa de helicidad negativa).

ve dada al principio de esta sección. Puede verse directamente a partir de esta relación que dos puntos P, R de $\mathbb{M}$ (sucesos) que son incidentes con el mismo twistor $\mathbf{Z}$ diferente de cero (necesariamente un twistor nulo) deben tener una *separación nula* entre sí (i.e., cada uno de ellos yace en el cono de luz del otro). Se sigue que $\mathbf{Z}$ define un *rayo de luz* —línea recta nula en $\mathbb{M}$—, puesto que todos los puntos de $\mathbb{M}$ que son incidentes con $\mathbf{Z}$ deben tener *separación nula* entre sí. Véase la Fig. 33.12. Además, el twistor $\mathbf{Z}$ representa el mismo rayo de luz si reemplazamos Z_α por λZ^α, donde λ es cualquier número complejo no nulo. El lugar geométrico de los sucesos incidentes con un twistor proyectivo nulo (diferente de cero) es, de hecho, un rayo de luz; pero en la situación particular en la que $Z^2 = Z^3 = 0$ debemos interpretar esto de manera adecuada, pues no obtenemos puntos reales en $\mathbb{M}$ incidentes con Z^α, pese a lo cual podemos seguir considerando que tal twistor nulo describe un *rayo de luz en el infinito* (un generador de $\mathscr{I}$, que yace en $\mathbb{M}^\#$, en lugar de $\mathbb{M}$).[33.12]

Examinemos ahora el recíproco. Fijando el suceso R, con coordenadas reales t, x, y, z, observamos que el espacio de twistores $\mathbf{Z}$ incidentes con R está definido por dos relaciones homogéneas lineales en las componentes Z^0, Z^1, Z^2, Z^3. Cada una de estas relaciones lineales define un *plano* en $\mathbb{PT}$ y su *intersección* (el conjunto de puntos en $\mathbb{PT}$ que satisfacen ambas relaciones) nos da una *línea proyectiva* $\mathbf{R}$ en $\mathbb{PT}$ (un $\mathbb{CP}^1$) —que yace en $\mathbb{PN}$, de hecho—, que es por consiguiente una esfera de Riemann, como se requiere (§§15.4,6). Así pues, los puntos de $\mathbb{M}$ (sucesos) se representan, en el espacio twistorial, por líneas proyectivas en $\mathbb{PN}$. En la situación particular $Z^2 = 0 = Z^3$, obtenemos una línea proyectiva particular en $\mathbb{PN}$ a la que llamaremos $\mathbf{I}$. Esta línea especial representa el punto i que es el vértice del cono de luz $\mathscr{I}$ en el infinito. Cualquier otro punto Q de $\mathscr{I}$ está representado, en $\mathbb{PN}$, por una línea proyectivas que corta a $\mathbf{I}$.[33.13] La situación se ilustra en la Fig. 33.12.

La forma en que estas estructuras complejas representan la geometría del espacio de Minkowski (en el número de dimensiones estándar de espacio y tiempo) es muy notable. Podemos reinterpretar el espa-

[33.12] Demuestre explícitamente las afirmaciones de este párrafo.
[33.13] ¿Por qué?

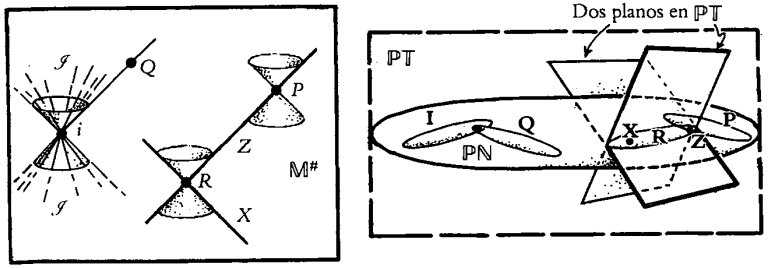


Fig. 33.12. La geometría de los lugares geométricos básicos en $\mathbb{M}^\#$ y $\mathbb{PN}$, dada por la relación de incidencia de la correspondencia twistorial. (a) Fijemos un punto (twistor nulo proyectivo) $\mathbf{Z}$ en $\mathbb{PN}$. Los puntos de $\mathbb{M}^\#$ (por ejemplo, P, R) que son incidentes con $\mathbf{Z}$ constituyen un rayo de luz, puesto que todos los puntos semejantes tienen una separación nula entre sí. (b) Fijemos un punto R en $\mathbb{M}^\#$. Los puntos de $\mathbb{PN}$ (por ejemplo, $\mathbf{Z}$, $\mathbf{X}$) que son incidentes con R (al yacer en la intersección de dos planos complejos en $\mathbb{PT}$) constituyen una línea proyectiva compleja, que es una esfera de Riemann. Puntos P y R en $\mathbb{M}^\#$ que tienen una separación nula a lo largo del rayo de luz Z tienen esferas de Riemann correspondientes $\mathbf{P}$ y $\mathbf{R}$, que se cortan en el punto $\mathbf{Z}$. (He dibujado estas esferas de Riemann muy alargadas, como un compromiso con el hecho de que son también líneas rectas proyectivas en la geometría proyectiva de $\mathbb{PT}$.) Una de estas esferas de Riemann particular es $\mathbf{I}$, que representa el punto i y en $\mathbb{M}^\#$. El punto i especifica infinito de género espacio/de género tiempo; es el vértice del cono de luz $\mathscr{I}$ en el infinito. Cualquier otro punto Q de $\mathscr{I}$ está representado en $\mathbb{PN}$ por una línea proyectiva $\mathbf{Q}$ que corta a $\mathbf{I}$.

cio de Minkowski como el espacio de líneas complejas que yacen en $\mathbb{PN}$ (o en $\mathbb{PN} - \mathscr{I}$, si queremos solo los puntos espaciotemporales finitos), tomando $\mathbb{PN}$ como estructura primaria y $\mathbb{M}$ como secundaria. Esto equivale a tomar los rayos de luz como algo más primitivo que los propios puntos espaciotemporales. La *intersección* de rayos de luz Z y X se representa por la existencia de una línea proyectiva sobre $\mathbb{PN}$ que contiene los puntos correspondientes $\mathbf{Z}$ y $\mathbf{X}$ de $\mathbb{PN}$ y, como hemos visto, la condición de que dos puntos espaciotemporales P y R tengan *separación nula* viene representada por la condición de que las correspondientes líneas proyectivas $\mathbf{P}$ y $\mathbf{R}$, en $\mathbb{PN}$, se corten (Fig. 33.12). Vemos así que un espacio twistorial proporciona una perspectiva completamente diferente, en geometría física, de la imagen espaciotemporal normal. Los puntos espaciotemporales ordinarios están representados

como esferas de Riemann en $\mathbb{PN}$. Los puntos de $\mathbb{PN}$ están representados como rayos de luz en el espaciotiempo. En cualquiera de las formas la correspondencia es no local. Pese a todo, podemos pasar de una imagen a la otra mediante reglas geométricas precisas.

33.6. Geometría de twistores como partículas sin masa con espín

Recordemos que la más fundamental de las ideas motivadoras tras la teoría de twistores es que habría que explotar al máximo la magia de los números complejos. A pesar de contener un sistema grande (de 4 parámetros reales) de líneas proyectivas complejas, el propio $\mathbb{PN}$ no es una variedad compleja (lo que difícilmente podía ser, como se ha señalado en §33.2, puesto que es de dimensión impar). Sin embargo, se convierte en una, a saber $\mathbb{PT}$ (que es una $\mathbb{CP}^3$), cuando tan solo se le añade una dimensión real más. ¿Podemos interpretar estos puntos extra de una forma físicamente significativa? En realidad sí (como se ha sugerido en §33.2). Recordemos que los fotones libres reales tienen más estructura que la de ser simplemente rayos de luz en $\mathbb{M}$. Un rayo de luz describe una partícula puntual que viaja a la velocidad de la luz en una dirección determinada, pero los fotones reales tienen energía y espín. Por ahora podemos pensar en ellos clásicamente. Las dos formas básicas en que puede girar un fotón son en sentido dextrógiro y en sentido levógiro respecto a la dirección de movimiento (helicidad positiva y negativa, respectivamente, definidas por polarización circular y dextrógira levógira; véase §22.7). En cualquiera de los casos, la magnitud de esta helicidad es $\hbar$. Resulta que los fotones clásicos de helicidad positiva pueden representarse como puntos de $\mathbb{PT}^+$, y los de helicidad negativa, como puntos de $\mathbb{PT}^-$, donde las dimensiones extra proceden de la energía del fotón. Esta descripción es también válida para cualquier otra partícula sin masa con espín $\frac{1}{2}n\hbar$ no nulo.

¿Cómo funciona esto? No es este el lugar para entrar en detalles, pero los rasgos esenciales se pueden esbozar de la siguiente forma. Como primer paso, hay que señalar que las dos primeras componentes

Z^0 y Z^1 del twistor $\mathbf{Z}$ son realmente las dos componentes de un 2-espinor $\boldsymbol{\omega}$, con forma de índices ω^A, donde $\omega^0 = Z^0$, $\omega^1 = Z^1$ (véanse §22.8 y §25.2). Las dos componentes restantes Z^2 y Z^3 de $\mathbf{Z}$ son las componentes de un espinor (dual) *primado* $\boldsymbol{\pi}$, con forma de índices $\pi_{A'}$, donde $\pi_{0'} = Z^2$, $\pi_{1'} = Z^3$. A veces escribimos

$$\mathbf{Z} = (\boldsymbol{\omega}, \boldsymbol{\pi}),$$

y nos referimos a $\boldsymbol{\omega}$ y $\boldsymbol{\pi}$ como las *partes espinoriales* del twistor $\mathbf{Z}$. El twistor conjugado complejo $\bar{\mathbf{Z}}$ tiene sus partes espinoriales en orden inverso

$$\bar{\mathbf{Z}} = (\bar{\boldsymbol{\pi}}, \bar{\boldsymbol{\omega}}),$$

de modo que la norma del twistor puede expresarse

$$\bar{Z}_\alpha Z^\alpha = \bar{\mathbf{Z}} \bullet \mathbf{Z} = \bar{\boldsymbol{\pi}} \bullet \boldsymbol{\omega} + \bar{\boldsymbol{\omega}} \bullet \boldsymbol{\pi} = \bar{\pi}_A \omega^A + \bar{\omega}^{A'} \pi_{A'}.$$

La relación de *incidencia* entre el twistor $\mathbf{Z}$ y el punto espaciotemporal R, con coordenadas de Minkowski t, x, y, z, se escribe ahora

$$\boldsymbol{\omega} = \mathrm{i}r\boldsymbol{\pi},$$

que representa $\omega^A = \mathrm{i}r^{AA'}\pi_{A'}$, donde $\boldsymbol{r}$ (o $r^{AA'}$) tiene la matriz de componentes

$$\begin{pmatrix} r^{00'} & r^{01'} \\ r^{10'} & r^{11'} \end{pmatrix} = \frac{1}{\sqrt{2}} \begin{pmatrix} t+z & x+\mathrm{i}y \\ x-\mathrm{i}y & t-z \end{pmatrix}.$$

El espinor $\boldsymbol{\pi}$ está asociado con el *momento* de la partícula sin masa, en el sentido de que el producto exterior $\bar{\boldsymbol{\pi}}\boldsymbol{\pi}$ (sin contracciones; véase §14.3) describe su 4-momento. El espinor $\boldsymbol{\omega}$ está asociado con el *momento angular* de la partícula, en el sentido de que el producto simetrizado de $\boldsymbol{\omega}$ con $\bar{\boldsymbol{\pi}}$ describe la parte anti-autodual del 6-momento angular de la partícula (§18.7, §19.2, §22.12 y §32.2) y el producto simetrizado de $\bar{\boldsymbol{\omega}}$ con $\boldsymbol{\pi}$ describe su parte autodual.[21] A diferencia del caso del momento, el momento *angular* depende de una elección del origen espaciotemporal O, y se suele hablar de momento angular *con respecto a* O. Esta independencia/dependencia del origen se refleja en el comportamiento traslacional de las dos partes espinoriales $\boldsymbol{\pi}$ y $\boldsymbol{\omega}$ de un twistor $\mathbf{Z}$. Bajo desplazamiento del origen O a un nuevo punto espaciotemporal Q, con vector de posición $\boldsymbol{q}$ con respecto a O, encontra-

mos (con $\boldsymbol{q}$ en forma matricial, como antes) que las partes espinoriales experimentan las transformaciones[33.14]

$$\boldsymbol{\pi} \mapsto \boldsymbol{\pi} \quad \text{y} \quad \boldsymbol{\omega} \mapsto \boldsymbol{\omega} - i\boldsymbol{q}\boldsymbol{\pi}.$$

Hay también una cantidad escalar independiente del origen que puede construirse a partir del momento y del momento angular, a saber, la *helicidad* s. Resulta que la helicidad es la mitad de la norma del twistor:

$$s = \frac{1}{2}\bar{Z}_\alpha Z^\alpha = \frac{1}{2}\bar{\mathbf{Z}} \bullet \mathbf{Z}$$

(y notamos de lo anterior que esto es tan solo la parte real de $\bar{\boldsymbol{\omega}} \bullet \bar{\boldsymbol{\pi}}$). De hecho, los twistores ofrecen un formalismo considerablemente más conciso para manejar partículas sin masa que la aproximación convencional 4-vector/tensorial de §22.12. Ahora tenemos una imagen física clara de un twistor no nulo (salvo el reescalamiento de fase $\mathbf{Z} \mapsto e^{i\theta}\mathbf{Z}$, con θ real) como una partícula clásica giratoria y sin masa;[33.15] compárese con la Fig. 33.6.

Aún no tenemos una imagen *geométrica* muy clara de un twistor no nulo. Podemos obtenerla si estamos dispuestos a considerar un espacio de Minkowski *complexificado* $\mathbb{CM}$ (o su compactificación $\mathbb{CM}^\#$), donde las coordenadas espaciotemporales t, x, y, z se consideran ahora números complejos. En efecto, siempre hay un lugar geométrico no trivial 2-complejo dimensional de puntos de $\mathbb{CM}^\#$ incidentes con un twistor Z^α (distinto de cero), llamado un *α-plano*, que es *autodual* en el sentido de que una 2-forma tangente al mismo es autodual. Este α-plano representa a Z^α salvo proporcionalidad; véase la Fig. 33.13. Análogamente, un twistor dual W_α define un *β-plano* que es un 2-plano complejo *anti*-autodual en $\mathbb{CM}^\#$.[33.16]

Hasta ahora, esto solo nos da una imagen espaciotemporal *compleja* de un twistor. ¿Podemos obtener una imagen «real» que podamos visualizar realmente? La estructura de realidad de $\mathbb{CM}^\#$ está contenida en

[33.14] Demuestre que la relación de incidencia entre un twistor y un punto espaciotemporal se conserva en esta transformación; demuestre que la norma del twistor se conserva.

[33.15] Explique por qué existe esta libertad de fase y por qué para una partícula de helicidad dada $s > 0$ la *energía* de la partícula está codificada en la localización del punto en $\mathbb{PT}^+$.

[33.16] Demuéstrelo.

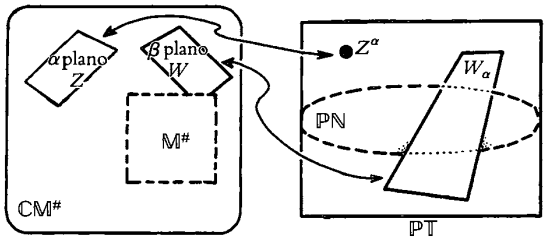


Fig. 33.13. Descripción espaciotemporal compleja de twistores (en general no nulos) y twistores duales. Para cualquier twistor Z^α distinto de cero, hay siempre un lugar geométrico 2-complejo dimensional en $\mathbb{CM}^\#$ de puntos incidentes con él, llamado un α-plano, que es autodual en todo lugar. Para cualquier twistor dual W_α distinto de cero, los puntos en $\mathbb{CM}^\#$ incidentes con él constituyen siempre un plano 2-complejo dimensional que es anti-autodual, llamado un plano. Solo para twistores nulos o twistores duales nulos hay puntos reales en estos lugares geométricos, y entonces los puntos reales constituyen un rayo de luz, de acuerdo con la Fig. 33.12.

su noción de *conjugación compleja* (§18.1); esta intercambia α-planos con β-planos, de acuerdo con el hecho de que intercambia superíndices twistoriales con subíndices (i.e., twistores con twistores duales) e intercambia «autodual» con «anti-autodual». En términos de la geometría proyectiva de $\mathbb{PT}$, la conjugación compleja intercambia puntos con planos, puesto que un twistor dual determina un plano en $\mathbb{PT}$.[33.17] Este hecho nos permite obtener una imagen de un twistor proyectivo no nulo Z^α en términos de una geometría espaciotemporal real. Lo primero que tenemos que hacer es representar Z^α por su conjugado complejo $\bar{Z}_\alpha$, que, siendo un twistor real, está asociado con un plano complejo en $\mathbb{PT}$. Este plano está determinado por su intersección con $\mathbb{PN}$, que es un lugar geométrico 3-dimensional real. Podemos interpretar este lugar geométrico como una familia 3-paramétrica de rayos de luz en $\mathbb{M}$. Así, esta familia de rayos de luz representa geométricamente al twistor Z^α (salvo proporcionalidad). Véase la Fig. 33.14.

Los rayos de luz se retuercen de una forma complicada, pero es posible obtener una imagen sorprendente de la configuración. Consideremos un instante de tiempo $\mathbb{E}^3$ (i.e., una sección 3-espacial euclídea

[33.17] ¿Por qué?

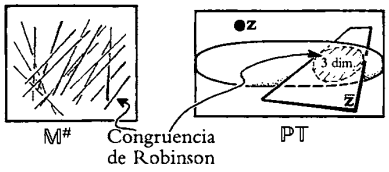


Fig. 33.14. Podemos obtener una imagen «real» de un twistor Z^α no nulo pasando primero a su complejo conjugado $\bar{Z}_\alpha$, definiendo así un plano proyectivo complejo en $\mathbb{PT}$. Este plano está determinado por su intersección con $\mathbb{PN}$, que es un lugar geométrico 3-real dimensional. Este lugar geométrico define una familia 3-paramétrica de rayos de luz en $\mathbb{M}^\#$ llamada una congruencia de Robinson.

ordinaria —«ahora»— a través del espaciotiempo de Minkowski $\mathbb{M}$). Cualquier rayo de luz en $\mathbb{M}$ —una partícula puntual que se mueve en una dirección concreta con la velocidad de la luz— viene representado en $\mathbb{E}^3$ por un punto con una «flecha» unida a él, flecha que determina la dirección de movimiento. Tenemos que imaginar una *familia* 3-paramétrica de rayos de luz semejantes —llamada una *congruencia de Robinson*— para representar nuestro único twistor $\mathbf{Z}$. En la Fig. 33.15, vemos un sistema de círculos orientados (y una recta) que llena la totalidad del 3-espacio ordinario $\mathbb{E}^3$. Habrá una partícula de nuestra familia en cada punto de $\mathbb{E}^3$, que se mueve (a la velocidad de la luz) en la dirección indicada por la tangente orientada al círculo que pasa por dicho punto. Es bastante significativo que, a medida que avanza el tiempo, toda esta configuración se propaga simplemente como un todo con la velocidad de la luz en la dirección (negativa) de la línea recta en la imagen, y esa propagación representa el movimiento de la partícula giratoria y sin masa descrita por el twistor. Esta configuración de círculos es, de hecho, la proyección estereográfica (§8.3, Fig. 8.7a), en el 3-espacio euclídeo ordinario, de la configuración de paralelos de Clifford en S^3 (§15.4).

No debemos pensar en estos «rayos de luz» como entidades físicas; solo nos proporcionan una realización geométrica de un twistor (proyectivo). Esta configuración codifica en realidad la estructura del momento angular de la partícula (clásica) giratoria y sin masa.[22] Es ciertamente una imagen no local. En la Fig. 33.15 hay un círculo mínimo cuyo radio es el espín dividido por la energía de la partícula. El centro

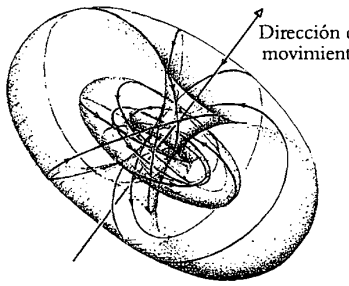


Fig. 33.15. Una imagen espacial que da una «instantánea» de una congruencia de Robinson. (Paralelos de Clifford proyectados estereográficamente sobre S^3; véanse las Figs. 8.7a y 15.8, que es una familia 3-paramétrica de círculos —y una línea recta— que llena la totalidad de $\mathbb{E}^3$.) Imaginamos una partícula en cada punto de $\mathbb{E}^3$, que se mueve en línea recta con la velocidad de la luz (un rayo de luz) en la dirección del círculo (orientado) en el que yace. La configuración completa se propaga con la velocidad de la luz en la dirección (negativa) de la línea recta en la figura. Esto representa el movimiento y el momento angular de la partícula sin masa giratoria descrita por la Z^α.

de este círculo, a grandes rasgos, representa la «localización» de la partícula giratoria (pero no puede considerarse que la historia de este centro sea un rayo de luz que representa la historia de la partícula sin masa, porque no se comporta adecuadamente bajo transformaciones de Lorentz).[33.18] Fue esta configuración la que provocó originalmente el nombre de «twistor».[23]

33.7. Teoría cuántica twistorial

Esto esboza la geometría básica de la teoría de twistores en el espacio plano. Pero quizá algunos lectores estén comprensiblemente impacientes y se pregunten cómo una imagen semejante, con toda su elegancia geométrica, va a servirnos para avanzar en *física*. ¿Qué tiene que decir la teoría de twistores acerca de la unificación de la estructura del espaciotiempo con los principios mecanocuánticos? Hasta ahora, hemos visto algunas «bonitas» maneras geométricas y algebraicas de

[33.18] Encuentre el centro (en las coordenadas de §33.5) y muestre cómo se transforma bajo una transformación general de Lorentz para la velocidad.

describir partículas sin masa, pero ni las ideas mecanocuánticas ni las de la relatividad general han desempeñado papel alguno. ¡Yo tendría que saberlo!

Volvamos a la idea más básica de la teoría de twistores. Esta consiste en considerar que todas las nociones espaciotemporales son *subsidiarias* de las del espacio twistorial $\mathbb{T}$. Al ser un espacio completamente complejo, $\mathbb{T}$ ofrece la potencialidad de explotar la magia de los números complejos de formas que no están inmediatamente presentes en el marco espaciotemporal estándar. Por consiguiente, antes que utilizar descripciones en términos de coordenadas espaciotemporales reales, en su lugar se utilizan las variables twistoriales complejas Z^α. Ahora bien, las variables twistoriales son mezclas de variables de posición y de momento, y debemos preguntar: ¿qué es lo que toma el lugar de la regla de cuantización estándar (§21.2)

$$p_a \mapsto \mathrm{i}\hbar\frac{\partial}{\partial x^a}$$

(o alternativamente $x^a \mapsto -\mathrm{i}\hbar\partial/\partial p_a$)? La respuesta es que, en analogía con el hecho de que x^a y p_a son variables *canónicas conjugadas*, como se expresa en la ley de conmutación $p_b x^a - x^a p_b = \mathrm{i}\hbar\delta_b^a$ de §21.2, las variables twistoriales Z^α y $\bar{Z}_\alpha$ van a tomarse como operadores canónicos conjugados

$$Z^\alpha\bar{Z}_\beta - \bar{Z}_\beta Z^\alpha = \hbar\delta^\alpha_\beta,$$

donde, como la posición y el momento por separado, las Z^α y $\bar{Z}_\alpha$ conmutan entre sí: $Z^\alpha Z^\beta - Z^\beta Z^\alpha = 0$ y $\bar{Z}_\alpha\bar{Z}_\beta - \bar{Z}_\beta\bar{Z}_\alpha = 0$.[33.19]

Como un aparte, puede comentarse que esta no conmutación *cuántica* de $\bar{Z}_\alpha$ con Z^α plantea algunas cuestiones intrigantes con respecto al tipo de «geometría» que podría aparecer si nos tomamos más en serio el hecho de que las «coordenadas» fundamentales para un espacio twistorial cuántico podrían ser tales entidades no conmutativas. Clásicamente, cuando consideramos la estructura de 8-variedad real del espacio twistorial $\mathbb{T}$, podemos utilizar Z^α y $\bar{Z}_\alpha$ como variables conmutativas independientes (véase §10.1). Pero en esta imagen cuántica,

[33.19] A partir de consideraciones generales sobre operadores en un espacio de Hilbert, ¿puede ver por qué es apropiado que no hubiera ninguna «i» en el conmutador de twistores?

Z^α y $\bar{Z}_\alpha$ no conmutan. Intentar utilizar un par «cuántico» semejante Z^α y $\bar{Z}_\alpha$ como coordenadas independientes nos llevaría al área de la geometría no conmutativa, que se ha examinado brevemente en §33.1. Podría resultar interesante proseguir esta línea, pero hasta ahora nadie lo ha hecho.

Recordemos que en una función de onda ordinaria $\psi(\mathbf{x})$ en el espacio de posición para una partícula no aparecen las variables de momento $\mathbf{p}$, sino que, en su lugar, el momento se representa en términos del *operador* $\partial/\partial x^a$ (como antes). ¿Cuál es el análogo twistorial de esto? Parece que necesitamos que nuestra «función de onda twistorial» $f(Z^\alpha)$ fuera «independiente de $\bar{Z}_\alpha$» y que $\bar{Z}_\alpha$ fuera representado en su lugar en términos del operador $\partial/\partial Z^\alpha$. De hecho, esto es correcto, pero ¿qué *significa* realmente que f sea «independiente de $\bar{Z}_\alpha$»? Formalmente, esta «independencia» se expresaría como $\partial f/\partial\bar{Z}_\alpha = 0$, que (como recordamos de §10.5) son simplemente las ecuaciones de Cauchy-Riemann que afirman que $f(Z^\alpha)$ es una función *holomorfa* de Z^α.

Este es un hecho muy sorprendente y bastante notable. Las funciones de onda twistoriales son entidades holomorfas, de modo que pueden entrar en contacto adecuado con el mundo mágico de los números complejos. El papel cuántico de las variables complejo conjugadas $\bar{Z}_\alpha$ consiste en que aparecen como diferenciación

$$\bar{Z}_\alpha \mapsto -\hbar\frac{\partial}{\partial Z^\alpha},$$

que es un operador holomorfo, de modo que, en el nivel de descripción cuántico, la holomorficidad se conserva. Resulta tranquilizador que la interpretación de los twistores en términos de momento y de momento angular para una partícula sin masa sea compatible con las reglas de conmutación de twistores, que dan correctamente los conmutadores de momento y momento angular (§22.12) y los engloban en los conmutadores de twistores que se han dado antes.[24]

Una cantidad de particular interés es la *helicidad*, considerada ahora como un operador, cuyos valores propios son los diversos posibles valores semienteros $(\ldots, -2\hbar, -\frac{3}{2}\hbar, -\hbar, -\frac{1}{2}\hbar, 0, \frac{1}{2}\hbar, \frac{3}{2}\hbar, 2\hbar, \ldots)$ que están permitidos para una partícula sin masa. Es especialmente digno

de mención que, teniendo en cuenta correctamente la no conmutación, el operador helicidad se convierte en[25,[33.20]]

$$s = \frac{1}{4}(Z^\alpha \bar{Z}_\alpha + \bar{Z}_\alpha Z^\alpha) \mapsto \frac{1}{2}\hbar\left(2 + Z^\alpha \frac{\partial}{\partial Z^\alpha}\right).$$

El operador

$$\boldsymbol{Y} = Z^\alpha \frac{\partial}{\partial Z^\alpha}$$

se denomina *operador homogeneidad de Euler*. (Recordemos a nuestro viejo amigo Leonhard Euler, de quien he hablado en los capítulos 5, 6, 7 y 9, en particular.) Como demostró Euler, $\boldsymbol{Y}$ tiene la notable propiedad de que sus funciones propias son *homogéneas*, siendo el valor propio el grado de homogeneidad. Es decir, la ecuación

$$\boldsymbol{Y}f = uf,$$

donde u es cierto número, es la condición para que se satisfaga la propiedad de homogeneidad[33.21]

$$f(\lambda Z^\alpha) = \lambda^u f(Z^\alpha).$$

Se sigue que una función de onda twistorial para una partícula sin masa con una helicidad definida de valor S (de modo que $sf = \hbar Sf$, donde s es el operador y S el valor propio) debe ser homogénea de grado $-2S - 2$ además de ser holomorfa.[33.22]

Así pues, en particular, una función de onda twistorial de un fotón ($S = \pm 1$) sería suma de dos partes, una homogénea de grado 0, que describe la componente levógira ($S = -1$), y una de grado -4, que describe la componente dextrógira ($S = 1$). Un neutrino, tomado como partícula sin masa, tendría una función de onda homogénea de grado -1 (puesto que la helicidad es $-1/2$), mientras que la función de onda de un antineutrino (sin masa) sería de grado -3. La función de onda de una partícula escalar sin masa es homogénea de grado -2. Lo que importa para nuestra discusión aquí es el caso de un *gravitón*, que consideraremos (provisionalmente) como una partícula

[33.20] Verifique la igualdad entre estas dos expresiones para s.
[33.21] Vea si puede demostrarlo.
[33.22] ¿Por qué este valor?

sin masa de espín 2 en un fondo de Minkowski plano ($S = \pm 2$). Su parte levógira ($S = -2$) tiene una función twistorial homogénea de grado 2 y su parte dextrógira ($S = 2$), una función de onda twistorial de grado -6.

Este sesgo es sorprendente e ilustra la naturaleza esencialmente quiral de la teoría de twistores. Veremos enseguida que este sesgo se hace especialmente fuerte cuando tratamos de llevar la relatividad general bajo el paraguas twistorial. Por el momento, tratemos de entender cómo hay que interpretar las funciones de onda twistoriales. Para estas, el sesgo no plantea problemas, y todo funciona con gran suavidad. No obstante, hay una sutileza importante acerca de cómo se va a interpretar nuestra función de onda $f(Z^\alpha)$, llamada normalmente *función twistorial*. Vayamos a eso ahora.

33.8. Descripción twistorial de campos sin masa

En el caso de la representación espaciotemporal de la función de onda de una partícula libre sin masa y de espín general, la ecuación de Schrödinger se traduce en una ecuación conocida como la *ecuación de campo libre y sin masa*.[26] Hemos visto un ejemplo de esto en el caso de espín 1/2 en la ecuación del neutrino (Dirac-Weyl) sin masa (§25.3). No es oportuno entrar ahora en detalles, pero la ecuación misma es suficientemente simple de escribir una vez que tenemos a mano el formalismo 2-espinorial, como se ha utilizado en §22.8 y §25.2. Para la helicidad negativa $S = -\frac{1}{2}n$, tenemos una cantidad $\psi_{AB\ldots D}$, y para la helicidad positiva $S = \frac{1}{2}n$, una cantidad con índices primados $\psi_{A'B'\ldots D'}$. Cada una de estas es completamente simétrica en todos sus n índices, y cada una de ellas tiene frecuencia positiva, y satisfacen las respectivas ecuaciones

$$\nabla^{AA'}\psi_{AB\ldots D} = 0, \quad \nabla^{AA'}\psi_{A'B'\ldots D'} = 0,$$

donde $\nabla^{AA'}$ es simplemente la traducción 2-espinorial del operador gradiente ordinario ∇^a (escrito en forma de índices elevados; véase

§14.3).[33.23] Para espín 0, tenemos simplemente la ecuación de ondas $\Box\psi = 0$, donde $\Box$ es el d'alembertiano ordinario introducido en §24.5. De hecho, la notación 2-espinorial conveniente para estas ecuaciones ilumina algunas sutilezas. Cuando $n = 2$ (espín 1), estas dos ecuaciones se convierten simplemente en las ecuaciones de Maxwell para campo libre en los casos anti-autodual y autodual, respectivamente.[33.24] Cuando $n = 4$, se convierten en las ecuaciones de Einstein para el campo débil, donde la curvatura se considera como una perturbación infinitesimal del espacio plano $\mathbb{M}$.[27]

¿Qué tienen que ver estas ecuaciones con las funciones twistoriales? Curiosamente resulta[28] que hay una expresión explícita como *integral de contorno* (§7.2) que da automáticamente la solución general de frecuencia positiva del campo libre sin masa, partiendo simplemente de la función twistorial $f(Z^\alpha)$. De hecho, la expresión también funciona perfectamente sin este requisito de frecuencia positiva, aunque el requisito está fácilmente garantizado en el formalismo twistorial, como veremos en §33.10. No es oportuno dar aquí todos los detalles, pero la idea básica es que, en el caso de la helicidad positiva, $f(Z^\alpha)$ se multiplica primero por $\boldsymbol{\pi}$ (§33.6) tomado n veces (lo que proporciona n índices primados), o, en el caso de la helicidad negativa, sobre él actúa primero $\partial/\partial\boldsymbol{\omega}$ tomado n veces (lo que suministra n índices espinoriales no primados); luego se multiplica por la 2-forma $\boldsymbol{\tau} = \mathrm{d}\pi_{0'} \wedge \mathrm{d}\pi_{1'}$ y se integra sobre un contorno 2-dimensional apropiado, donde se incorpora primero la relación de incidencia $\boldsymbol{\omega} = \mathrm{i}\boldsymbol{r}\boldsymbol{\pi}$ para eliminar $\boldsymbol{\omega}$ en favor de $\boldsymbol{\pi}$ y $\boldsymbol{r}$. Esta integración elimina $\boldsymbol{\pi}$, de modo que terminamos con una cantidad indexada $\psi_{\dots}$ en cualquier punto espacio temporal R escogido (de modo que $\psi_{\dots}$ es función solo de $\boldsymbol{r}$). El contorno debe yacer dentro del lugar geométrico $\boldsymbol{\omega} = \mathrm{i}\boldsymbol{r}\boldsymbol{\pi}$ (para cada $\boldsymbol{r}$ fijo), i.e., dentro de (la versión no proyectiva[29] de) la línea R, en $\mathbb{N}$, que representa el suceso R; véase la Fig. 33.16.

[33.23] Escriba explícitamente estas ecuaciones, para helicidad $-\frac{1}{2}n$, utilizando la notación $\psi_r = \psi_{00\dots011\dots1}$, donde hay $n - r$ 0s y r 1s, y traduzca $\nabla^{AA'}$ a partir de ∇^a de la misma forma que la cantidad $r^{AA'}$ se traduce a partir de las coordenadas de Minkowski ordinarias t, x, y, z, como se ha descrito antes.

[33.24] Intente demostrar esto, donde $\psi_{00} = C_1 - \mathrm{i}C_2$, $\psi_{01} = -\mathrm{i}C_3$, $\psi_{11} = -C_1 - \mathrm{i}C_2$, donde $\mathbf{C} = 2\mathbf{E} - 2\mathrm{i}\mathbf{B}$ (véase §19.2), y donde son válidas expresiones correspondientes para $\psi_{A'B'}$.

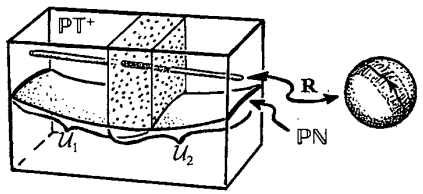


Fig. 33.16. La integral de contorno twistorial básica. Una función twistorial f (para helicidad $\frac{1}{2}n$), homogénea de grado $-n - 2$, se multiplica n veces por $\boldsymbol{\pi}$, o sobre ella actúa n veces $\partial/\partial\boldsymbol{\omega}$ (n negativo), lo que suministra los índices, y luego se multiplica por la 2-forma $\boldsymbol{\tau} = \mathrm{d}\pi_{0'} \wedge \mathrm{d}\pi_{1'}$. Para cualquier elección particular del punto espaciotemporal R, con vector de posición $\boldsymbol{r}$, se realiza entonces una integral de contorno en la región $\mathbf{R}$ del espacio twistorial definido por la relación de incidencia $\boldsymbol{\omega} = \mathrm{i}\boldsymbol{r}\boldsymbol{\pi}$. Esto integra la $\boldsymbol{\pi}$-dependencia, y nos deja con una solución de las ecuaciones de campo sin masa. En el caso ilustrado, $\mathbf{R}$ se toma en la mitad superior del espacio twistorial $\mathbb{PT}$ y f es holomorfa en la intersección de $\mathcal{U}_1$ con $\mathcal{U}_2$, donde los conjuntos abiertos $\mathcal{U}_1$ y $\mathcal{U}_2$, cubren juntos la totalidad de $\mathbb{PT}$.

La condición de frecuencia positiva está asegurada exigiendo que la integral de contorno siga siendo válida también cuando se permite que la línea $\mathbf{R}$ se aventure dentro de la región twistorial $\mathbb{PT}^+$. (Las líneas en $\mathbb{PT}$ corresponden a «puntos espaciotemporales complejos», como hemos visto en §33.6, y las que yacen enteramente en la subregión $\mathbb{PT}^+$ corresponden a puntos de la subregión $\mathbb{M}^+$ de $\mathbb{CM}$ que se conoce como el *tubo avanzado*.[30] Volveremos a esta cuestión en §33.10.) También pueden describirse dentro de este marco campos sin masa de helicidad mixta —tales como un fotón plano polarizado, que es una suma de una parte levógira y una dextrógira— donde las funciones twistoriales para las dos helicidades diferentes simplemente se suman.

La existencia misma de una expresión semejante me resulta algo mágico. Para el campo sin masa, las ecuaciones parecen evaporarse en el formalismo twistorial, al convertirse, de hecho, en «holomorficidad pura». Cuando examinamos esta expresión con más cuidado, observamos que existe una sutileza importante acerca de cómo debe interpretarse una función twistorial, y esto se relaciona de forma sorprendente con la separación en frecuencia positiva/negativa de campos sin masa (§33.10). Esta sutileza resulta también crucial para la forma en que se manifiestan

las funciones twistoriales, y nos proporciona espacios twistoriales curvos. ¿Cuál es esta sutileza? Se trata de que las funciones twistoriales no deben verse realmente como «funciones» en el sentido ordinario, sino como lo que se denominan *elementos de cohomología de haces holomorfa.*[31]

33.9. Cohomología de haces twistorial

¿Qué es la cohomología de haces? Las ideas son bastante sofisticadas matemáticamente, pero muy naturales en realidad. Aquí nos interesaremos solo en lo que se denomina cohomología de *primer* haz. Quizá la manera más fácil de representar esta noción sea pensar primero en la forma en que puede construirse una variedad en términos de varias cartas de coordenadas, como se ha discutido en §10.2 y §12.2 y se ha ilustrado en la Fig. 12.5a. En cada solapamiento entre un *par* de cartas se define una *función de transición* (que proporciona el *pegamento* de las cartas). Recordemos de §12.2, Fig. 12.5a, que estas funciones de transición están sujetas a ciertas condiciones de consistencia en solapamientos triples entre cartas.

Consideremos ahora una variedad construida así pero donde las funciones de transición difieren de la identidad en solo una cantidad infinitesimal. Véase la Fig. 33.17. Este cambio infinitesimal entre una carta $\mathcal{U}_i$ y otra carta $\mathcal{U}_j$ estaría descrito por un campo vectorial $\boldsymbol{F}_{ij}$ en la parte de $\mathcal{U}_i$ que se solapa con $\mathcal{U}_j$, que describe cómo debe «deslizarse» infinitesimalmente $\mathcal{U}_i$ con respecto a $\mathcal{U}_j$. De forma equivalente, podemos pensar que $\mathcal{U}_j$ está deslizado con respecto a $\mathcal{U}_i$ pero en dirección contraria. Esto está descrito por el campo vectorial $\boldsymbol{F}_{ji}$ en la parte de $\mathcal{U}_j$ que se solapa con $\mathcal{U}_i$, por lo que en este solapamiento tenemos

$$\boldsymbol{F}_{ji} = -\boldsymbol{F}_{ij}$$

(véase la Fig. 33.18a). En un triple solapamiento entre cartas $\mathcal{U}_i, \mathcal{U}_j$ y $\mathcal{U}_k$, encontramos (Fig. 33.18b) que debe cumplirse la relación de compatibilidad[33.25]

$$\boldsymbol{F}_{ij} + \boldsymbol{F}_{jk} = \boldsymbol{F}_{ik}.$$

[33.25] Demuestre que la antisimetría en F_{ij} es un requisito de consistencia de la condición de solapamiento triple.

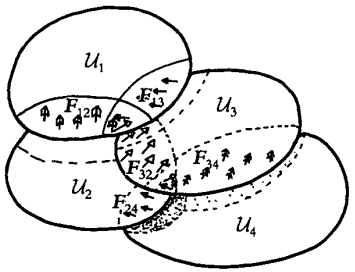


Fig. 33.17. Recordemos (de la Fig. 12.5a) cómo se construye una variedad a partir de varias cartas de coordenadas. (Definida en cada solapamiento entre un par de cartas hay una «función de transición» que proporciona el «pegamento» de las cartas.) Aquí consideramos funciones de transición que difieren solo infinitesimalmente de la identidad, y que por consiguiente están dadas por un campo vectorial $\boldsymbol{F}_{ij}$ en cada solapamiento de cartas, lo que nos dice cómo está «desplazada» cada carta en relación con las otras con las que se solapa. (Las «cartas» son conjuntos abiertos $\mathcal{U}_1, \mathcal{U}_2, \mathcal{U}_3, \ldots$ en el espacio de coordenadas plano.)

Existen también deformaciones infinitesimales «triviales» que surgen simplemente de cambiar (infinitesimalmente) el sistema de coordenadas en cada carta. Podemos pensar que estas vienen dadas por un campo vectorial $\boldsymbol{H}_i$ en cada carta particular $\mathcal{U}_i$, que simplemente «desliza» la carta entera a lo largo de sí misma. Esto nos daría una familia de $\boldsymbol{F}_{ij}$ «triviales» de la forma

$$\boldsymbol{F}_{ij} = \boldsymbol{H}_i - \boldsymbol{H}_j$$

en solapamientos entre pares de cartas, que no cambian la variedad (Fig. 33.18c).

Estas ideas nos dicen en esencia las reglas de la cohomología de primer haz.[32] Sin embargo, no tenemos por qué interesarnos solo en campos vectoriales. Las funciones ordinarias f_{ij} serán tan buenas como los campos vectoriales $\boldsymbol{F}_{ij}$ que hemos estado considerando. Simplemente exigimos que cada f_{ij} esté definida en la intersección de $\mathcal{U}_i$ con $\mathcal{U}_j$, que $f_{ij} = -f_{ji}$, que $f_{ij} + f_{jk} + f_{ki} = 0$ en cada solapamiento triple, y que la colección completa $\{f_{ij}\}$ sea considerada equivalente a otra colección semejante $\{g_{ij}\}$ si cada miembro de la colección de las correspondientes *diferencias* $\{f_{ij} - g_{ij}\}$ tiene la forma «trivial» $\{h_i - h_j\}$. Decimos que las $\{f_{ij}\}$ están reducidas *módulo* cantidades de la forma $\{h_i - h_j\}$, que es en esencia el mismo sentido en que se ha utilizado el término «módulo» en §16.1 (véase también la noción de «clase de equivalencia» mencionada en el prefacio).

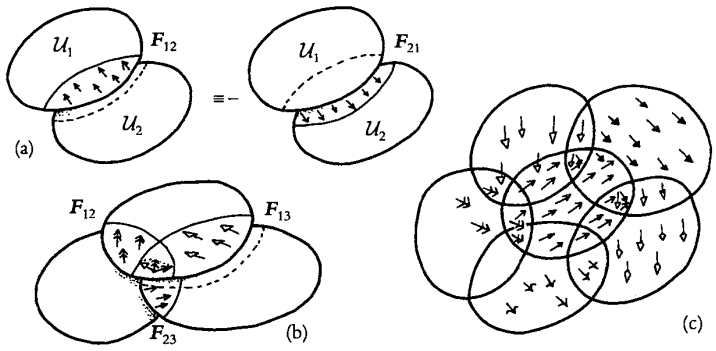


Fig. 33.18. Los campos vectoriales $\boldsymbol{F}_{ij}$ están sujetos a ciertos requisitos. (a) En la intersección de $\mathcal{U}_i$ con $\mathcal{U}_j$, tenemos $\boldsymbol{F}_{ij} = -\boldsymbol{F}_{ji}$. Esto se ilustra como el movimiento de $\mathcal{U}_i$ sobre $\mathcal{U}_j$, como se muestra por el vector $\boldsymbol{F}_{12}$ en $\mathcal{U}_1$. Pero el mismo movimiento relativo se consigue por el negativo de esto sobre $\mathcal{U}_2$. (b) Se ilustra la condición de solapamiento triple $\boldsymbol{F}_{ij} + \boldsymbol{F}_{jk} = \boldsymbol{F}_{ik}$. En la intersección triple de $\mathcal{U}_1$, $\mathcal{U}_2$ y $\mathcal{U}_3$, el movimiento $\boldsymbol{F}_{12}$ de $\mathcal{U}_1$ sobre $\mathcal{U}_2$ es la suma del movimiento $\boldsymbol{F}_{13}$ de $\mathcal{U}_1$ sobre $\mathcal{U}_3$ y el movimiento de $\boldsymbol{F}_{32}$ de $\mathcal{U}_3$ sobre $\mathcal{U}_2$. (c) Si todas las cartas se toman individualmente en su totalidad, esto no tiene ningún efecto (excepto por un cambio de coordenadas en cada carta). Esto ilustra que el desplazamiento global $\boldsymbol{F}_{ij} = \boldsymbol{H}_i - \boldsymbol{H}_j$ no cuenta para nada y debe ser «factorizado».

De hecho, la clase de funciones (f_{ij} o h_i) en las que podemos estar interesados en la teoría de la cohomología puede ser extraordinariamente general. En la teoría de twistores solemos trabajar con funciones holomorfas. Esto nos da la noción de «cohomología de haces holomorfa».

Específicamente, esta idea de cohomología se aplica a funciones twistoriales. De hecho, tenemos que considerar que una «función twistorial» *no* es simplemente una única función holomorfa f, sino que está proporcionada por una *colección* de funciones holomorfas $\{f_{ij}\}$, donde cada f_{ij} está definida en el solapamiento entre un par de conjuntos abiertos $\mathcal{U}_i$, $\mathcal{U}_j$, donde $f_{ji} = -f_{ij}$, donde en solapamientos triples $f_{ij} + f_{jk} + f_{ki} = 0$, y donde la colección completa de estos conjuntos abiertos $\{\mathcal{U}_i\}$ cubre la región entera $\mathcal{Q}$ del espacio twistorial bajo consideración. Un *primer elemento de cohomología* en $\mathcal{Q}$ (con respecto al recubrimiento $\{\mathcal{U}_i\}$ viene representado como esta colección $\{f_{ij}\}$, reducida

módulo las cantidades de la forma $\{h_i - h_j\}$, con h_i definida en $\mathcal{U}_i$. La colección de funciones $\{f_{ij}\}$ no debe considerarse *como* el elemento de cohomología, sino como algo que proporciona una forma de representar ese misterioso «elemento». Llamamos a las $\{f_{ij}\}$ *representantes* de este primer elemento de cohomología.

No obstante, para la definición estricta de cohomología, tendríamos que considerar también el tomar el límite de recubrimientos cada vez más finos de la región $\mathcal{Q}$. Afortunadamente, hay teoremas que nos dicen que para la cohomología de haces *holomorfa* podemos parar el refinamiento cuando las $\mathcal{U}_i$ son tipos de conjuntos suficientemente simples conocidos como conjuntos de *Stein*.[33] (La cohomología de primer haz holomorfa se anula siempre en cualquier conjunto de Stein.) Así pues, si restringimos nuestra atención a recubrimientos para los que cada $\mathcal{U}_i$ es un conjunto de Stein, no necesitamos decir «con respecto al recubrimiento $\{\mathcal{U}_i\}$» cuando nos referimos a un elemento de cohomología definido en $\mathcal{Q}$. La noción de cohomología no depende de la elección concreta del recubrimiento de Stein. Un elemento de cohomología es un «objeto» definido en $\mathcal{Q}$, que resulta ser el mismo cualquiera que sea el recubrimiento utilizado.[34] ¡Este hecho notable es parte de la magia de la cohomología de haces (holomorfa)!

¿Cómo se aplica todo esto a las funciones twistoriales y las integrales de contorno que hemos considerado en §33.8? La situación más simple surge cuando hay solo dos cartas $\mathcal{U}_i$ y $\mathcal{U}_j$, que, juntas, cubren la región del espacio twistorial bajo consideración. Allí hay solo una función por considerar, y esta es la función twistorial de §33.8: $f(Z^\alpha) = f_{12} = -f_{21}$. De acuerdo con las reglas anteriores de la cohomología de haces, decimos que $f(Z^\alpha)$ es *equivalente* a $g(Z^\alpha)$ si la diferencia es *trivial* en el sentido anterior, i.e., si

$$f - g = h_1 - h_2,$$

donde la función holomorfa h_1 está definida globalmente en $\mathcal{U}_1$ y h_2 globalmente en $\mathcal{U}_2$. Es sencillo demostrar que la integral de contorno apropiada aplicada a f es, de hecho, la misma que la aplicada a g cuando quiera que estas funciones son equivalentes en este sentido. Sin embargo, a veces es necesario considerar «carteados» más complicados. En esencia, las anteriores «reglas de cohomología» para la equivalencia en-

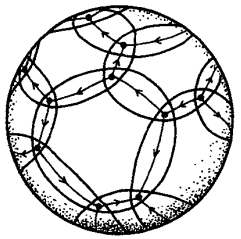


Fig. 33.19. Un «contorno ramificado» (sobre la esfera de Riemann), aplicable a la evaluación espaciotemporal de funciones twistoriales para las que el recubrimiento consiste en más de dos conjuntos.

tre funciones twistoriales están adaptadas para preservar las respuestas que proporcionan las expresiones con integrales de contorno, pero donde la noción de una integral de contorno debe ahora ser generalizada a la de una «integral de contorno ramificada», con una rama en cada región del solapamiento. Esto se indica en la Fig. 33.19.[35]

Una característica importante de la cohomología es que es *esencialmente no local*. Podríamos tener un elemento de cohomología definido en una región $\mathcal{Q}$. Entonces tiene sentido considerar la restricción de dicho elemento a una región más pequeña $\mathcal{Q}'$, contenida en $\mathcal{Q}$. La característica no local de la cohomología se manifiesta en el hecho de que para cualquier subregión (abierta) suficientemente pequeña $\mathcal{Q}'$, en $\mathcal{Q}$, el elemento se *anula* necesariamente cuando se restringe a $\mathcal{Q}'$, en el sentido de que, dada f_{ij} en $\mathcal{Q}'$, entonces siempre pueden encontrarse h, en $\mathcal{Q}'$, para las que $f_{ij} = h_i - h_j$.

Esta no localidad de las funciones twistoriales nos dice que no hay que atribuir significado al valor alcanzado por f_{ij} en un punto particular. De hecho, podemos restringirnos a una región suficientemente pequeña que rodea a dicho punto y encontrar que el elemento de cohomología desaparece por completo.Véase la Fig. 33.20. Esta no localidad que muestran las funciones twistoriales (consideradas como elementos de la primera cohomología) recuerda inevitablemente los aspectos no locales de los efectos EPR y el cuanlazamiento (§23.10). En mi opinión, algo importante está sucediendo entre bastidores que algún día puede dar sentido a la misteriosa naturaleza no local de los fenómenos EPR, aunque, si es así, todavía no ha sido completamente revelada.

Tenemos que pensar que este «elemento de cohomología» es un «objeto» definido en el espacio $\mathcal{Q}$, que es un poco como una función definida en $\mathcal{Q}$ pero que es fundamentalmente no local. Un ejemplo de

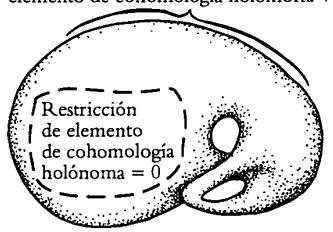


Fig. 33.20. Un elemento de cohomología siempre puede restringirse a una región más pequeña. Pero si esta región es suficientemente pequeña, la cohomología desaparece siempre. Esto ilustra la no localidad de la cohomología.

un «objeto» de este tipo es realmente un *fibrado vectorial* (complejo) entero sobre $\mathcal{Q}$, como se ha descrito en §§15.2,5. Recordemos que en la definición de un fibrado la parte que yace por encima de una región suficientemente pequeña del espacio base (aquí $\mathcal{Q}$) es «trivial» en el sentido de que esta parte es solo un producto topológico (§15.2). (Véase la Fig. 15.3.) Este es un ejemplo del hecho de que si restringimos nuestro primer elemento de cohomología a una región suficientemente pequeña, también se hace «trivial», i.e., se anula. Así pues, la «información» que se está expresando en un elemento de cohomología es algo de carácter fundamentalmente no local.

Quizá valga la pena ofrecer un ejemplo elemental que ilustra la noción de cohomología, aunque solo en un caso sencillo, de una manera particularmente gráfica. La Fig. 33.21 es un dibujo de un «objeto imposible» a veces conocido como una «tribarra».[36] Es evidente que el «objeto 3-dimensional» que el dibujo parece mostrar no puede existir realmente en el espacio euclídeo ordinario. Pese a todo, *localmente* no hay nada imposible en el dibujo. La imposibilidad es algo no local, que desaparece si consideramos una región suficientemente pequeña en el dibujo. De hecho, esta noción de «imposibilidad» en un dibujo semejante puede expresarse como un elemento de cohomología específico.[37][33.26] No obstante, es un tipo de cohomología relativamente simple, en la que las funciones $\{f_{ij}\}$ se toman constantes.

Solo he tocado aquí algunas ideas básicas de la cohomología de haces. Existen muchas aplicaciones de estas ideas en matemáticas, no todas

[33.26] Vea si puede hacer esto rompiendo el dibujo en varios dibujos que se solapan (las $\{\mathcal{U}_i\}$), cada uno de los cuales representa individualmente una estructura 3-espacial consistente, y utilizando el logaritmo de la distancia de esta estructura 3-espacial al ojo del observador para calcular las f_{ij}.

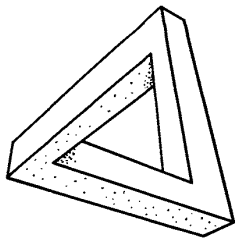


Fig. 33.21. Dibujo de un «objeto imposible» (una «tribarra»). Localmente, no hay nada imposible en lo que el dibujo representa. La «imposibilidad» viene medida por un elemento de cohomología, que desaparece en cualquier región suficientemente pequeña en el diagrama.

ellas relacionadas con la holomorficidad. Los «haces» en los que estamos principalmente interesados en la teoría de twistores son los expresados en términos de funciones holomorfas, y hay una magia especial en la teoría de la cohomología en este contexto particular. (A grandes rasgos, el término «haz» se refiere al tipo de función que nos interesa, pero la noción de haz es en realidad mucho más general que la de una función ordinaria.)[38] Hay muchos otros usos de la cohomología, entre ellos algunos que tienen importancia en el estudio de los espacios de Calabi-Yau que aparecen en la teoría de cuerdas (§31.14), por ejemplo. Además, hay otras formas completamente diferentes de definir elementos de cohomología de haces, y puede demostrarse que todas ellas son matemáticamente equivalentes, pese a sus diferentes aspectos.[39] En mi opinión, la cohomología (de haces) es un ejemplo excelente de una noción platónica (§1.3), que, al igual que el propio sistema $\mathbb{C}$ de los números complejos, parece tener una «vida propia» que está más allá de cualquier forma concreta que pudiéramos escoger para representarla.

33.10. Los twistores y la separación en frecuencia positiva/negativa

¿Cómo incorporamos la condición de *frecuencia positiva*, tan fundamental para la QFT, en la teoría de twistores? Recordemos (§§9.2,3) de qué modo la división de la esfera de Riemann S^2 en los hemisferios sur y norte S^- y S^+ nos proporciona la separación de una función, definida en el ecuador S^1, en sus partes de frecuencia positiva y negativa. La parte de frecuencia positiva se extiende en S^-, y la parte de frecuencia negativa, en S^+ (Fig. 33.22a) El espacio twistorial proyectivo hace algo

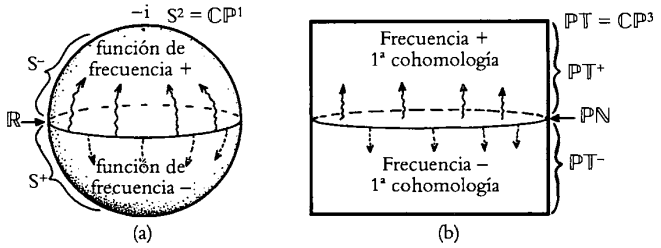


Fig. 33.22. Una analogía entre la esfera de Riemann $S^2(= \mathbb{CP}^1)$ y el espacio twistorial proyectivo $\mathbb{PT}$. (a) Función compleja (i.e., un «elemento de 0ª cohomología»), definida sobre el eje real $\mathbb{R}$ de S^2, se divide en su parte de frecuencia positiva, que se extiende de manera holomorfa en lo que aquí se representa como el hemisferio norte S^-, y su parte de frecuencia negativa, que se extiende al hemisferio sur S^+. (La esfera de Riemann está dibujada aquí de modo que $\mathbb{R}$ es su ecuador, pero tal que $-i$ está en el polo norte e i en el polo sur; compárese con las Figs. 8.7, 9.10 y §9.5.) (b) Un elemento de cohomología, definido sobre $\mathbb{PN}$ (y que representa un campo sin masa), se divide en su parte de frecuencia positiva, que se extiende de manera holomorfa a la mitad superior $\mathbb{PT}^+$ del espacio twistorial proyectivo, y su parte de frecuencia negativa, que se extiende a la mitad inferior $\mathbb{PT}^-$.

correspondiente, pero de una forma global que se aplica directamente a campos sin masa en su totalidad. Lo hace según una analogía directa entre la esfera de Riemann y el espacio twistorial proyectivo $\mathbb{PT}$, donde el análogo de una función en la esfera de Riemann S^2 es un primer elemento de cohomología en $\mathbb{PT}$. El análogo del ecuador S^1 va a ser el espacio $\mathbb{PN}$, y notamos que $\mathbb{PN}$ divide a $\mathbb{PT}$ (que es un $\mathbb{CP}^3$) en dos mitades $\mathbb{PT}^+$ y $\mathbb{PT}^-$ precisamente de la misma forma que S^1 divide a S^2 (que es un $\mathbb{CP}^1$) correspondientemente[40] en dos hemisferios S^- y S^+ (Fig. 33.22b).

Más explícitamente, el análogo de una función (compleja) ordinaria definida en S^1, o en S^-, o en S^+ es, respectivamente, un primer elemento de cohomología definido en $\mathbb{PN}$, o en $\mathbb{PT}^+$, o en $\mathbb{PT}^-$, respectivamente. Campos sin masa en $\mathbb{M}$ (estrictamente, en $\mathbb{M}^\#$) se representan como elementos de primera cohomología en $\mathbb{PN}$. Cada uno de estos puede expresarse (esencialmente de modo unívoco) como una suma de un elemento que se extiende en $\mathbb{PT}^+$ y un elemento que se extiende en $\mathbb{PT}^-$. El primero describe un campo sin masa de frecuencia positiva, y el segundo, uno de frecuencia negativa.[41] En términos espaciotem-

porales, esta parte de frecuencia positiva del campo se extiende para estar definida en el *tubo avanzado*, que recordemos de §33.8 es la región $\mathbb{M}^+$ de $\mathbb{CM}^\#$ que consiste en puntos que están representados en el espacio twistorial por líneas proyectivas en $\mathbb{PT}^+$. En $\mathbb{CM}^+$, estos son los puntos (complejos) cuyos vectores de posición tienen partes imaginarias que son de género tiempo y apuntan al pasado.[33.27]

Esta analogía entre $\mathbb{PT}$ y la esfera de Riemann lleva a un posible modo en que las ideas de la teoría de twistores podrían encontrar una analogía con algunas ideas de la teoría de cuerdas. Recordemos de §§31.5,13 que las superficies de Riemann se utilizan para representar «historias de cuerdas» en dicha teoría. La esfera de Riemann ($\mathbb{CP}^1$) es la más simple de tales superficies, pero se introducen superficies con diferentes números de «asas» (superficies de Riemman de género superior; véase §8.4) para representar tipos más generales de historias de cuerdas. Estas superficies de Riemann también pueden tener «agujeros» (con fronteras S^1), además de asas (véase la Fig. 31.5). Por analogía,[42] podemos considerar generalizaciones del espacio $\mathbb{PT}$, que adquieren asas de un modo correspondiente, y también «agujeros» (con fronteras que son copias de $\mathbb{PN}$). Estas se conocen como «espacios twistoriales pretzel», y es posible desarrollar una forma de QFT basada en estos espacios (véase la Fig. 33.23). Por el momento, el estatus de estas ideas no ha sido completamente establecido.

Históricamente, el requisito de frecuencia positiva —y esta propiedad de que $\mathbb{PN}$ divide a $\mathbb{PT}$ en dos mitales tales— ofreció una motivación clave en la formulación original de la teoría de twistores; fue en 1963, más de doce años antes del descubrimiento de que los campos sin masa tienen una descripción twistorial como cohomología holomorfa de primer haz.[43] Resulta sorprendente que aquí tengamos una vez más una propiedad que es específica del hecho de que el espaciotiempo tiene cuatro dimensiones con signatura lorentziana. Hay también algo muy concreto en el hecho de que tenemos primeros elementos de cohomología que desempeñan un papel en la teoría de

[33.27] Pruebe esto: Demuestre, a partir de la ecuación de incidencia, que un vector de posición complejo r^a para un punto R de $\mathbb{CM}$ representado por una línea proyectiva en $\mathbb{PT}^-$, si y solo si la parte imaginaria de r^a es de género tiempo y apunta al pasado.

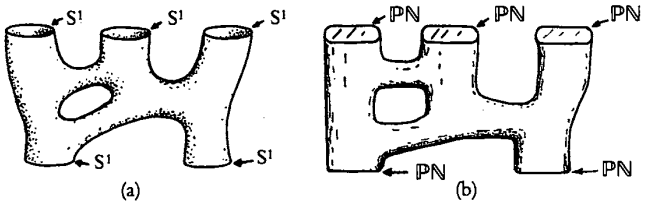


Fig. 33.23. (a) Teoría de campo conforme (modelo tipo teoría de cuerdas), basada en generalizaciones de la esfera de Riemann a superficies de Riemann de género más alto que pueden tener «agujeros» de tamaño finito, así como asas (véase la Fig. 31.5: los agujeros representan lugares donde se introduce información externa). (b) Versión twistorial que utiliza generalizaciones de $\mathbb{PT}$, que adquiere «asas» de una manera que corresponde a superficies de Riemann, y también «agujeros» cuyas fronteras son copias de $\mathbb{PN}$ («espacios twistoriales pretzel»).

twistores, en lugar de funciones ordinarias —que son elementos de cohomología «cero»—, o elementos de segunda cohomología y superiores. También existen nociones de orden superior de cohomología (y tienen un papel que desempeñar en la teoría de twistores), pero hay algo único en la primera cohomología que es fundamental para la teoría de twistores. Pues solo entonces estas cantidades encuentran un papel directo para generar *deformaciones* del espacio twistorial. Vayamos a ello.

33.11. El gravitón no lineal

Los elementos de cohomología (funciones twistoriales) que hemos considerado deberían pensarse, de momento, como enteramente «pasivos», en el sentido de que están simplemente «pintados» en el espacio (twistorial). Esto corresponde al hecho de que describen campos espaciotemporales que simplemente residen en el espaciotiempo y no influyen en otros campos. Para ver cómo pueden proporcionar una influencia activa, pensemos que la «pintura» en el espacio twistorial «se seca», de modo que ahora el espacio se distorsiona (Fig. 33.24). Para ver cómo puede suceder esto, consideremos que nuestra función twistorial previamente pasiva f_{ij} está asociada de la forma apropiada a un campo

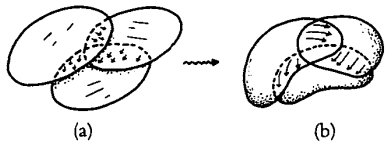


Fig. 33.24. Un elemento de campo vectorial de primera cohomología es «pasivo» (i.e., simplemente «pintado en» el espacio). Para que tenga una influencia activa, consideremos el «secado de la pintura» como un resultado de una exponenciación del campo vectorial en cada solapamiento. El resultado de esto es un «deslizamiento» finito de una carta sobre otra, lo que da una distorsión finita, o un «espacio curvo».

vectorial $\boldsymbol{F}_{ij}$. «Deslizando una carta sobre otra» una cantidad infinitesimal en la dirección de estos campos vectoriales, empezamos a «secar la pintura» y construir un espacio twistorial infinitesimalmente «curvado». Podemos imaginar que esta deformación es «exponenciada» (§14.6) hasta que se obtiene una deformación finita del espacio twistorial (¡pintura completamente seca!).

La primera situación en la que se aplicó con éxito este procedimiento fue en el caso de la gravedad anti-autodual.[44] En el caso infinitesimal (campo débil) tenemos un campo sin masa de helicidad $S = -2$, de modo que, utilizando la fórmula anterior $-2S - 2$ para el grado de homogeneidad, tenemos una función twistorial $f(= f_{ij})$ de homogeneidad 2. Aquí estamos suponiendo por simplicidad que hay solo dos cartas $\mathcal{U}_1$ y $\mathcal{U}_2$, cada una de las cuales se toma como una porción de un espacio twistorial plano $\mathbb{T}$ con las coordenadas estándar de §33.5. El campo vectorial $\boldsymbol{F}$ requerido, construido a partir de f, resulta ser

$$\boldsymbol{F} = \frac{\partial f}{\partial \omega^0}\frac{\partial}{\partial \omega^1} - \frac{\partial f}{\partial \omega^1}\frac{\partial}{\partial \omega^0}.$$

Nótese que el grado de homogeneidad 2 de f compensa exactamente el de los dos operadores diferenciales para dar un operador que es homogéneo de grado cero, de modo que actúa sobre el espacio twistorial proyectivo.[33.28]

Imaginemos ahora que exponenciamos este desplazamiento infinitesimal de una carta sobre la otra (véase la Fig. 33.25). Entonces obte-

[33.28] ¿Por qué el grado cero implica que esto da un campo vectorial en una región en $\mathbb{PT}$? *Sugerencia*: ¿Cuál es el conmutador de $\boldsymbol{F}$ con $\boldsymbol{Y}$ (§33.7)?

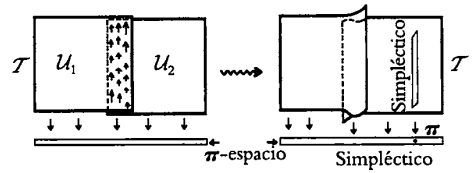


Fig 33.25. Aplicamos la idea de la Fig. 33.24 en el caso de la descripción twistorial de la gravedad anti-autodual (con dos cartas). El campo vectorial es $(\partial f/\partial\omega^0)\partial/\omega^1 - (\partial f/\partial\omega^1)\partial/\omega^0$, con f homogénea de grado 2. Obtenemos un espacio twistorial curvo $\mathcal{T}$. Existe una proyección global de $\mathcal{T}$ en el $\boldsymbol{\pi}$-espacio. Cada fibra de esta proyección es un 2-espacio simpléctico complejo, como lo es el propio $\boldsymbol{\pi}$-espacio.

nemos un espacio twistorial *curvo* $\mathcal{T}$. La ausencia de $\boldsymbol{\pi}$-derivadas en nuestra relación de carteado infinitesimal implica que el twistor en una carta debe tener la misma parte $\boldsymbol{\pi}$ que el twistor con el que empalma en la carta contigua. Se sigue de ello que la operación que «destaca» el espinor $\boldsymbol{\pi}$ del espacio carteado entero $\mathcal{T}$ es consistente sobre la totalidad de $\mathcal{T}$. Es decir, existe una proyección global de $\mathcal{T}$ en el espacio de espinores $\boldsymbol{\pi}$. Ignoremos (o preferiblemente eliminemos) los «elementos cero» tanto de $\mathcal{T}$ como del espacio $\boldsymbol{\pi}$. Entonces encontramos que $\mathcal{T}$ es un tipo de *fibrado* sobre el π-espacio (véase §15.2).[45] Cada *fibra* (imagen inversa de cualquier $\boldsymbol{\pi}$ concreta, i.e., la parte de $\mathcal{T}$ que yace «sobre» $\boldsymbol{\pi}$) resulta ser una 2-variedad completa con una estructura simpléctica, como la tiene el propio espacio $\boldsymbol{\pi}$ (véase §14.8; aquí esto solo significa que una medida de área está definida en esta 2-variedad), un hecho que está asegurado por la forma específica del carteado que se ha dado antes.

¿Cómo vamos a retomar contacto desde este espacio twistorial curvo con alguna noción de «espaciotiempo»? La respuesta es que cada «punto espaciotemporal» corresponde unívocamente a una sección transversal holomorfa del fibrado $\mathcal{T}$. (La noción de una sección transversal holomorfa se ha visto en §15.5; aquí es una aplicación del $\boldsymbol{\pi}$-espacio de nuevo en $\mathcal{T}$.) ¿Por qué es esta una definición razonable? En el caso plano $\mathbb{T}$, esto equivale a representar el punto espaciotemporal R (posiblemente complejo) por la aplicación que lleva $\boldsymbol{\pi}$ a $\mathbf{Z} = (i\boldsymbol{r}\boldsymbol{\pi}, \boldsymbol{\pi})$. En términos del espacio twistorial *proyectivo* plano $\mathbb{PT}$, esta sección transversal es simplemente la línea recta $\mathbf{R}$ (una esfera de Riemann, $\mathbb{CP}^1$) en

$\mathbb{PT}$ que hemos utilizado en §33.5 para representar a R.[33.29] Es muy notable que esta definición de un «punto espaciotemporal» funcione igualmente bien para el espacio twistorial curvo $\mathcal{T}$. Vemos[46] que existe una familia de secciones transversales holomorfas con 4 parámetros complejos, igual que en el caso plano. (En el espacio proyectivo $\mathbb{PT}$, esta es una familia de líneas $\mathbb{CP}^1$ con 4 parámetros complejos.) Tenemos así una variedad compleja 4-dimensional $\mathcal{M}$ para representar esta familia. La 4-dimensionalidad es un hecho notable —un ejemplo de la magia compleja en dimensiones más altas— que se sigue de los teoremas del matemático japonés Kunihiko Kodaira.[47] (La experiencia solo con variedades *reales* podría haber llevado a esperar que hubiera una familia de *infinitos* parámetros de tales objetos. Pero ya hemos advertido en §15.5 que las secciones transversales holomorfas pueden ser muy restringidas.)

En la Fig. 33.26 se ilustra gráficamente este procedimiento (en la descripción proyectiva). Partamos de una región apropiada $\mathcal{R}$ del espaciotiempo de Minkowski complejo $\mathbb{CM}$. Por simplicidad, tomemos $\mathcal{R}$ como algún entorno (abierto) apropiado de un punto $\mathcal{R}$ en $\mathbb{CM}$. La correspondiente región $\mathcal{Q}$ del espacio proyectivo $\mathbb{PT}$ es la barrida por la familia de líneas, cada una de las cuales representa un punto de $\mathcal{R}$. Este será un entorno (conocido como *entorno tubular*) de la línea $\mathbf{R}$, en $\mathbb{PT}$, que representa a $\mathcal{R}$ (Figs. 33.26b,c). Podemos considerar que la topología de $\mathcal{Q}$ es $S^2 \times \mathbb{R}^4$, donde S^2 proviene de la topología de la línea $\mathbf{R}$ —o, de forma equivalente, del $\boldsymbol{\pi}$-espacio proyectivo— y el $\mathbb{R}^4$ describe la parte transversal del entorno inmediato de cada punto de $\mathbf{R}$. Pensemos ahora en S^2 (aquí el $\boldsymbol{\pi}$-espacio proyectivo) separado en dos hemisferios, ligeramente ampliados, de modo que hay un «collar» de solapamiento, y entonces consideramos $\mathcal{Q}$ construido a partir de las dos piezas (conjuntos abiertos) $\mathcal{U}_1$ y $\mathcal{U}_2$ solapadas que yacen por encima de cada uno de estos hemisferios ligeramente ampliados (Fig. 33.26d). Ahora «desplazamos» $\mathcal{U}_2$ con relación a $\mathcal{U}_1$, según el campo vectorial anterior, para obtener nuestra región espacio twistorial proyectiva deformada $\mathbb{PT}$ (Figs. 33.26e,f).

Sigue habiendo una proyección global en el $\boldsymbol{\pi}$-espacio (Fig. 33.26f) que da la estructura fibrada. Sin embargo, las «líneas rectas» originales en

[33.29] Explique en qué sentido es esta línea una «sección transversal» de $\mathbb{PT} - \mathbf{I}$.

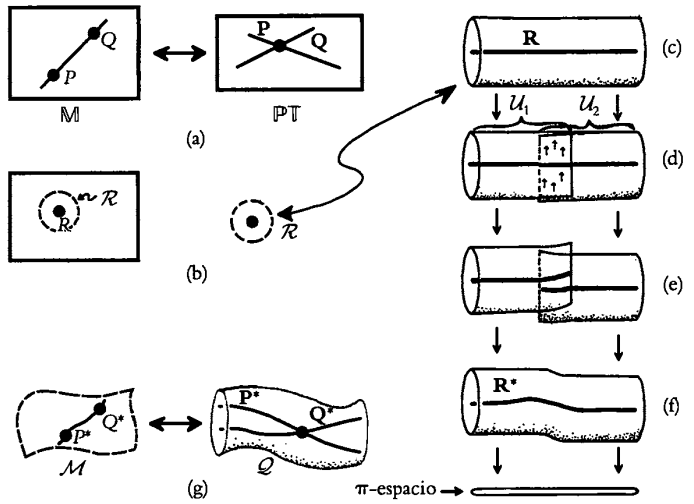


Fig 33.26. Construcción de un gravitón no lineal levógiro. (a) En la correspondencia twistorial estándar en el espacio plano, puntos P y Q de $\mathbb{CM}$ tienen una separación nula cuando quiera que las líneas correspondientes $\mathbf{P}$ y $\mathbf{Q}$ en $\mathbb{PT}$ se cortan. (b) Queremos deformar $\mathbb{PT}$ de algún modo en un espacio twistorial curvo, pero algunos teoremas matemáticos nos dicen que esto no puede hacerse globalmente. Por consiguiente, tomamos solo una vecindad adecuada (abierta) $\mathcal{R}$, de un punto R en $\mathbb{CM}$ como nuestro «espaciotiempo» de partida. (c) Esto corresponde a una vecindad tubular $\mathcal{Q}$, en $\mathbb{PT}$, de la línea $\mathbf{R}$. (d) Podemos aplicar ahora el procedimiento de la Fig. 33.25 para deformar $\mathcal{Q}$ (considerada como la unión de dos conjuntos abiertos $\mathcal{U}_1$ y $\mathcal{U}_2$. (e) Sin embargo, observamos que la línea original $\mathbf{R}$ está ahora rota, y no se puede utilizar como una definición razonable de un «punto espaciotemporal». (f) Un teorema de Kodaira viene en nuestra ayuda, para decirnos que existe una familia 4-paramétrica de «líneas» $\mathbf{R}^*$ (curvas holomorfas compactas, que pertenecen a la misma clase topológica que nuestras líneas originales), que servirán para este propósito. (g) Los puntos de nuestro espacio «gravitón no lineal» buscado $\mathcal{M}$ (un 4-espacio complejo) están dados por curvas de Kodaira $\mathbf{R}^*$. La métrica (conforme compleja) de $\mathcal{M}$ está definida (como en (a)) por la condición de que P^* y Q^* tienen separación nula cuando quiera que las líneas correspondientes $\mathbf{P}^*$ y $\mathbf{Q}^*$ se cortan. La curvatura de Weyl de $\mathcal{M}$ resulta ser automáticamente anti-autodual, y es también Ricci-plana en virtud de los detalles de la construcción.

$\mathcal{U}_1$ y $\mathcal{U}_2$ están ahora rotas, de modo que no dan secciones transversales, pero, según el teorema de Kodaira, hay una nueva familia de 4 parámetros de curvas holomorfas en $\mathbb{PT}$, siendo estas las secciones transversales holomorfas de la estructura fibrada. El espacio requerido $\mathcal{M}$ está construido de modo que cada uno de sus puntos corresponde a una de estas secciones transversales (Figs. 33.26f,g). Resulta que a $\mathcal{M}$ se le puede asignar una métrica $\boldsymbol{g}$ de una manera natural y que su curvatura de Weyl es anti-autodual, y es Ricci-plana. Podemos encontrar fácilmente los conos nulos de $\boldsymbol{g}$ (estructura conforme) utilizando el hecho de que los dos puntos P^* y Q^* de $\mathcal{M}$ tienen separación nula si y solo si se cortan las líneas correspondientes $\mathbf{P}^*$ y $\mathbf{Q}^*$ en $\mathbb{PT}$ (Fig. 33.26).

Quizá el lector se preocupe por lo que en realidad significa físicamente este «espaciotiempo» $\mathcal{M}$. Resulta *complejo* (y, por consiguiente, 8-dimensional, en lugar de 4-dimensional, cuando se considera como variedad real). En el caso plano, podemos discriminar los puntos espaciotemporales *reales* (sucesos en $\mathbb{M}$) tomando secciones transversales de $\mathbb{T}$ que yacen en $\mathbb{N}$, y considerar entonces que nuestro $\mathcal{M}$ es simplemente la complexificación $\mathbb{CM}$ del espacio de Minkowski $\mathbb{M}$. Pero en el caso curvo no se nos permite tal lujo. De hecho, en el caso curvo, el «espaciotiempo» que obtenemos mediante esta construcción es necesariamente una variedad compleja por sí misma, y no puede aparecer como una complexificación de un espaciotiempo real lorentziano.

¿Por qué es así? Una 4-variedad lorentziana con una curvatura de Weyl anti-autodual es necesariamente Weyl-plana (puesto que la conjugada compleja de la parte autodual cero es la parte anti-autodual, que por consiguiente es también cero). Si es también Ricci-plana, entonces es simplemente plana en general. En el caso complejo,[48] por el contrario, hay una gran familia de 4-variedades Ricci-planas anti-autoduales y no triviales. ¡Es un hecho sorprendente que *todas* estas pueden obtenerse (al menos localmente) por medio del ya mencionado procedimiento twistorial!

¿Qué haremos con este espacio complejo $\mathcal{M}$? *Físicamente*, la interpretación de un 4-espacio complejo Ricci-plano y anti-autodual (si puede decirse que sea de «frecuencia positiva» en un sentido apropiado) es que representa un *gravitón levógiro*. De hecho, es un gravitón *no lineal*, en el sentido de que es un tipo de «función de onda», pero ahora es una

solución de la ecuación no lineal de Einstein para el vacío (Ricci-planitud), en lugar de su aproximación lineal. Lo último habría sido el caso si solo hubiéramos tomado la función twistorial f como un elemento de cohomología, en lugar de permitir que «la pintura se seque» y deforme así el propio espacio twistorial. Vemos que la teoría de twistores nos ha llevado en una dirección curiosa y previamente inesperada en la unificación de ideas de la teoría cuántica con la estructura espaciotemporal. Nuestras funciones de onda twistoriales son ahora entidades no lineales, de modo que empiezan a aparecer desviaciones respecto a las reglas estándar de la mecánica cuántica lineal (§§22.2-4).

Hay un aspecto de esta construcción que es particularmente digno de mención. Si tomamos cualquier punto $\mathbf{Z}$ del espacio twistorial curvo $\mathcal{T}$, observamos que cualquier entorno suficientemente pequeño de $\mathbf{Z}$ tiene una estructura idéntica a la de un entorno de cualquier punto escogido $\mathbf{Z}'$ del espacio twistorial *plano* $\mathbb{T}$ (que no esté en la región «infinita» $\mathbf{I}$; véase §33.5). Por consiguiente, la estructura local que posee el espacio twistorial es «blanda», en el sentido en que se ha utilizado esta palabra en §14.8. Así, toda la información relativa a curvatura, etc., del espacio $\mathcal{M}$ está almacenada en $\mathcal{T}$ *globalmente*, no localmente. Esto refleja el hecho mencionado antes de que un elemento de cohomología definido por una función twistorial desaparece por completo cuando se restringe a una región suficientemente pequeña. No hay «ecuaciones de campo» en el espacio twistorial. El tipo de información que está normalmente almacenada en las soluciones de las ecuaciones de campo en el espaciotiempo (en este caso, la ecuación de Einstein anti-autodual) parece estar almacenada solo no localmente en una construcción espacio twistorial.[49]

33.12. Twistores y relatividad general

Esta «construcción del gravitón no lineal» ha sido fundamental para el desarrollo de la teoría de twistores desde la década de 1970. En su forma inicial, pedía a gritos avances en dos direcciones diferentes. La más obvia de estas apuntaba a una construcción correspondiente para el gravitón no lineal *dextrógiro*, y para que este se combinara con el levó-

giro de modo que pudieran formarse estados de polarización mixta (tales como gravitones no lineales plano polarizados). Esta podría ser una parte clave del programa twistorial. Como se ha señalado antes, la noción de un «gravitón no lineal» está muy en el espíritu de búsqueda de una teoría, como la que se ha defendido en el capítulo 30, en la que las reglas lineales estándar de la teoría **U**-cuántica ordinaria necesitan doblegarse para que se pueda obtener la unión correcta con la relatividad general de Einstein. Sin embargo, el «gravitón» que ha aparecido en la construcción anterior es solo «medio gravitón», en cuanto que solo ha sido incorporado uno de los dos posibles estados de helicidad.

Algunos lectores avisados podrían aventurar la sugerencia de que, si pasamos a una descripción en términos de twistores *duales* W_α, en lugar de twistores Z^α, entonces se obtendría una función de onda no lineal para un gravitón *dextrógiro*, repitiendo la construcción anterior en términos de twistores duales.[33.30] De esta forma, sería el gravitón dextrógiro el que corresponde a la homogeneidad de grado 2 (en W_α) y el levógiro el que corresponde a la homogeneidad –6. No obstante, esto no nos saca de dificultades, porque ahora perdemos los estados de helicidad levógira, y no tendría sentido utilizar las variables W_α para los estados dextrógiros y las variables Z^α para los levógiros, sobre todo porque también necesitamos describir estados de helicidad mixta.[50]

El problema de «exponenciar» de algún modo las funciones twistoriales $f(Z^\alpha)$ de homogeneidad –6 para obtener un gravitón no lineal dextrógiro ha sido calificado como el problema *googly* (gravitacional). (El término «googly» se utiliza en el críquet para describir una bola que gira en sentido dextrógiro con relación a su dirección de movimiento, aunque su lanzamiento se parece al que normalmente impartiría un giro levógiro.) Se ha tardado veinte años en encontrar una solución plausible, pero estudios recientes parecen ofrecer una construcción apropiada para esto.[51] Pese a todo, en el momento de escribir este libro muchos aspectos de los procedimientos siguen siendo conjeturas. No voy a intentar describir estos desarrollos aquí, excepto para decir que la

[33.30] ¿Por qué? *Sugerencia:* ¿Por qué una reflexión espacial convierte twistores en twistores duales?

nueva característica esencial es que las *fibras* de la proyección de nuestro espacio twistorial curvo $\mathcal{T}$ en el espacio proyectivo $\mathbb{P}\mathcal{T}$ se «retuercen» de una forma que está definida por una función twistorial de homogeneidad –6. (El «giro» se efectúa exponenciando un campo vectorial, sobre un par de cartas que se solapan, de la forma engañosamente simple $Cf_{-6}Z^\alpha\partial/\partial Z^\alpha$, donde C es una constante apropiada y f_{-6} es una función twistorial de grado de homogeneidad –6.) Esto permite que las dos partes levógira y dextrógira del gravitón estén incorporadas simultáneamente.

Al menos en el caso de un espaciotiempo asintóticamente plano $\mathcal{M}$ de la forma adecuada, existe una construcción explícita directa para $\mathcal{T}$ en términos de $\mathcal{M}$. Además, hay una propuesta tentativa para obtener $\mathcal{M}$ a partir de un $\mathcal{T}$ dado, i.e., para construir puntos espaciotemporales a partir de la estructura puramente twistorial de $\mathcal{T}$, que se conjetura para asegurar que la Ricci-planitud (ecuación del vacío de Einstein) se incorpora correctamente. De forma significativa, la propuesta está relacionada con un proyecto de investigación a largo plazo, debido a Ezra T. Newman y sus colegas, para interpretar puntos espaciotemporales en términos de lo que se denominan «cortes de conos de luz», que son las intersecciones de los conos de luz en $\mathcal{M}$ con el infinito nulo futuro $\mathscr{I}^+$.[52] Sin embargo, aunque parece prometedora, algunos aspectos importantes de esta construcción twistorial siguen sin estar resueltos en el momento de escribir esto.[53]

La otra dirección en la que la construcción del gravitón no lineal levógiro (de 1975-1976) reclamaba avances era en generalizaciones desde la teoría gravitacional a otros campos gauge. Muy pronto, en 1976-1977, Richard Ward demostró cómo los campos gauge anti-autoduales podían obtenerse también utilizando una construcción twistorial algo similar a la gravitatoria. De hecho, la construcción de Ward ha generado un considerable interés matemático y desarrollos posteriores por parte de Ward y otros, sobre todo en el área de los *sistemas integrables* (ecuaciones no lineales que pueden resolverse, en un sentido apropiado, en el caso general). Aquí la teoría de twistores ha proporcionado una potente visión de conjunto de este tema.[54] Parece probable que los avances mencionados hacia una completa solución al problema de la helicidad mixta gravitacional marcarán el camino para tratar los

campos gauge generales (con helicidad mixta) también dentro del formalismo twistorial.

33.13. Hacia una teoría twistorial de la física de partículas

Esto nos lleva a la cuestión de cómo podría desarrollarse la teoría de twistores para dar una teoría física generalizada, lo que, por el momento, no es. Para ello es importante que se desarrollen dos áreas adicionales de estudio en la teoría de twistores. La primera trata de proporcionar un tratamiento global de la QFT. De hecho, ha habido una considerable actividad, llevada a cabo principalmente por Andrew Hodges y sus estudiantes en Oxford (con alguna aportación inicial mía a comienzos de la década de 1970), que proporciona un enfoque perturbativo de la QFT en el que los diagramas de Feynman se reemplazan por construcciones conocidas como *diagramas twistoriales*. Estos implican una integración de contorno en altas dimensiones, y el formalismo consigue algunos éxitos sorprendentes al evitar muchos de los infinitos que se encuentran en los procedimientos de Feynman convencionales.[55] No obstante, el enfoque es todavía algo más complicado de lo que sería deseable, y carece de un principio guía independiente que nos diga exactamente qué integral de contorno hay que realizar, sin que tengamos que apelar como intermediarias a las expresiones convencionales de Feynman.

La segunda de estas áreas es la *teoría twistorial de partículas*, que fue desarrollada básicamente por Zoltan Perjés, George Sparling, Lane Hughston, Paul Tod y Florence Tsou (Tsou Seung Tsun) a partir de ideas que introduje desde mediados de la década de 1970 hasta principios de la de 1980, pero que desde entonces se ha estancado. La idea básica aquí es que mientras que las partículas sin masa pueden describirse mediante funciones de onda twistoriales de una sola variable twistorial $f(Z^\alpha)$, las partículas masivas requieren más variables, por ejemplo, $X^\alpha, \ldots, Z^\alpha$. Existe una expresión para el momento lineal y el momento angular de una partícula masiva que supone sumar las contribuciones individuales de todos estos twistores, pero ahora hay un *grupo* de simetría interno que surge de las transformaciones entre estas variables twistoriales y sus conjugadas complejas que no afectan al momento total ni al momento angular. Quizá valga la pena mencionar que se obtienen grupos que incluyen, pero que generalizan ligeramente, el U(2) de las interacciones electrodébiles y el SU(3) de las interacciones fuertes. Se advirtieron varias relaciones sorprendentes con la clasificación estándar de las partículas según el modelo estándar, pero el esquema llegó a estancarse por ciertas razones técnicas. Parece haber una perspectiva razonable de que los recientes desarrollos en el «problema googly» —especialmente si pueden aplicarse a campos gauge— pudieran reabrir el tema.

Hay también, en mi opinión, una posibilidad significativa de que la propuesta de Chan-Tsou para un modelo de física de partículas, brevemente descrito en §25.8, pudiera ligarse de forma importante con estos desarrollos. Dicha propuesta requiere que exista un grupo dual para cada grupo de simetría (no abeliano) para partículas, además del grupo gauge original. La teoría de twistores sugiere que, de acuerdo con la construcción de Ward que se ha mencionado antes —junto con su conjeturada versión «googly»—, cada gupo debería intervenir en las versiones anti-autodual y autodual, y esto parece exigir que la forma dual del grupo gauge debería desempeñar un papel significativo además del papel del grupo gauge original. Así, utilizando ideas de la propuesta de Chan-Tsou, el programa twistorial de partículas muy bien podría tomar parte en una futura física de partículas. Hay que anticipar, además, que un progreso exitoso en esta área debería tener también un impacto importante en el programa QFT de la teoría de diagramas twistoriales.

33.14. ¿El futuro de la teoría de twistores?

En mis descripciones de la teoría de twistores no he advertido al lector de que mis ideas sobre el tema quizá no reflejen las de la comunidad de físicos en general. De hecho, puesto que he dedicado más de la mitad de mi vida (de manera intermitente) a la teoría de twistores, apenas es probable que mi propia perspectiva se corresponda estrechamente con la de muchos otros que no han estado tan implicados. Ade-

más, debería dejar claro que la comunidad de físicos que dominan el tema es relativamente pequeña, y desde luego muy pequeña en relación con la de aquellos que saben de teoría de cuerdas o de supersimetría. La teoría de twistores no podría calificarse ni mucho menos como una actividad en la «corriente principal» de los físicos teóricos actuales.

Pese a todo, la teoría de twistores, como la teoría de cuerdas, ha ejercido una influencia significativa en las matemáticas puras, y esto se ha considerado una de sus mayores virtudes. La teoría de twistores ha tenido un impacto importante sobre la teoría de los sistemas integrables (como se ha mencionado antes brevemente), en la teoría de representaciones[56] y en la geometría diferencial. (En esta última área debería mencionar el trabajo de Sergei A. Merkulov y L. J. Schwachhöfer, que fueron capaces de encontrar una solución al que se conoce como el «problema de la holonomía», utilizando métodos desarrollados a partir de los de la construcción del gravitón no lineal original.[57] En trabajos relacionados, la teoría de twistores tiene un valor importante en la construcción de lo que se denominan «variedades hyperkhaler», «espacios de Zoll», etc.)[58] La teoría de twistores se ha visto guiada en gran medida por consideraciones de elegancia e interés matemático, y buena parte de su fuerza reside en su estructura matemática rigurosa y fructífera.

Todo esto está muy bien, podría estar inclinado a decir el lector ingenuo con cierta justificación, pero ¿no me quejé yo, en el capítulo 31, de que una debilidad de la teoría de *cuerdas* es que su impulso era básicamente matemático y tenía poca guía procedente de la naturaleza del mundo físico? En algunos aspectos, esta es también una crítica válida para la teoría de twistores. Ciertamente no hay una razón de peso procedente de los datos observacionales modernos que nos obligue a creer que la teoría de twistores ofrece el camino a seguir por la física moderna. Además, muchos podrían tener la sensación de que la naturaleza fuertemente quiral de la teoría lleva las cosas demasiado lejos en la dirección de la asimetría espacial. Después de todo, no hay evidencia física de que una asimetría izquierda/derecha tenga que desempeñar un papel en la física gravitacional. En los capítulos 27, 28 y 30 he señalado la necesidad de una asimetría temporal en la unión cuántico/gravi-

tacional apropiada, pero no hay un requisito físico aparente de una asimetría *espacial* (salvo quizá indirectamente, vía el teorema CPT de la QFT; véanse §25.4 y §30.2).

Por supuesto, podría darse el caso de que la asimetría espacial en el formalismo no se tradujera en una asimetría en los efectos físicos. La mejor razón para esperar que esto sea cierto reside en el hecho de que las álgebras generadas por el par $(Z^\alpha, -\hbar\partial/\partial Z^\alpha)$, por una parte, y $(\hbar\partial/\partial\bar{Z}_\alpha, \bar{Z}_\alpha)$, por otra, son formalmente idénticas. Esto sugiere que, cualquiera que sea la conclusión a la que llegáramos a alcanzar utilizando una descripción twistorial (variables Z^α), podría obtenerse igualmente utilizando una descripción dual twistorial (utilizando variables $W_\alpha = \bar{Z}^\alpha$), y que esta similaridad es tan completa que ninguna simetría izquierda/derecha en gravitación emergería en la teoría resultante. Por otra parte, si el formalismo tiene que reflejar la naturaleza, entonces exigiremos una asimetría izquierda/derecha cuando la teoría llegue a describir las interacciones débiles (§25.3). Pero tal como está la teoría de twistores, en su relativamente primitivo estado actual, no hay ninguna razón clara para esta diferencia.

Por ahora la principal crítica que se puede formular contra la teoría de twistores es que no se trata realmente de una teoría *física*. Ciertamente no hace ninguna predicción física inequívoca. Mi propia perspectiva (super)optimista sería considerar que la teoría de twistores es vagamente comparable al formalismo hamiltoniano de la física clásica. La teoría hamiltoniana no introdujo cambios físicos, pero ofreció una perspectiva diferente acerca de la física clásica que más tarde resultó ser la que se requería para la nueva teoría cuántica según las recetas de Schrödinger, como se ha descrito en los capítulos 21-23. Análogamente, la teoría de twistores es meramente una reformulación que no introduce cambios físicos. La esperanza optimista está en que su armazón pudiera proporcionar también un trampolín para algunos importantes desarrollos físicos en el futuro.

Por supuesto, el escéptico no está obligado a creer que tales desarrollos vayan a tener lugar, y el argumento básico para la teoría de twistores reside, de hecho, como en la teoría de cuerdas (o en la teoría M), en la fuerza de su atractivo estético y matemático. Sin embargo, ambas teorías son incompatibles matemáticamente tal como están, puesto que

operan con diferente número de dimensiones espaciotemporales. Se podría decir justamente (pero con demasiada dureza) que una predicción de la teoría de twistores es que las aspiraciones de la teoría de cuerdas son erróneas. Esta incompatibilidad no se extiende a las variantes o reinterpretaciones de la teoría de cuerdas (o la teoría M) donde las dimensiones extra no se consideran como dimensiones espaciotemporales en absoluto, sino que se consideran como dimensiones «internas» de algún tipo. Aunque una reinterpretación semejante parece ofrecer un punto de vista consistente, está en desacuerdo con la fuerza impulsora que hay tras la teoría de cuerdas tal como se acepta habitualmente.

En relación con esto, debería recordar al lector cierto trabajo muy reciente, mencionado en §31.18, básicamente de Edward Witten.[59] Este apunta a algunas posibilidades fascinantes para una nueva perspectiva sobre las amplitudes de dispersión de Yang-Mills. Combina algunas ideas de la teoría de twistores con otras de la teoría de cuerdas, ¡pero ahora en un contexto 4-dimensional!

En cualquier caso, la teoría de twistores requiere algún aporte nuevo. Entre los ingredientes más importantes de otras teorías físicas de éxito destacan los lagrangianos y las integrales de camino de Feynman, que proporcionan la forma adecuada en QFT de tratar las ecuaciones de campo (véase §26.6). Sin embargo, la teoría de twistores se jacta de la evaporación de las ecuaciones de campo (§§33.9,11), así que, al parecer, se necesitan nuevas ideas para el desarrollo de una QFT twistorial completa.[60]

¿Hace alguna otra «predicción» precisa la teoría de twistores? Lo más próximo a una predicción que puedo considerar es que las motivaciones subyacentes en la teoría parecen implicar que el universo debería tener una curvatura espacial negativa, i.e., $K < 0$. Para ver la razón de esta expectativa, recordemos primero de los capítulos 27 y 28 (especialmente, §27.13) que el big bang parece haber tenido una naturaleza extraordinariamente uniforme, con un estrecho parecido a uno de los modelos de FLRW. Estos modelos son conformemente planos (anulación de la curvatura de Weyl) y pueden describirse de forma muy simple en términos de un espacio twistorial plano ($\mathbb{CP}^3$).[61] En cada uno de los casos $K > 0$, $K = 0$, $K < 0$, hay un grupo de simetría

exacta, pero solo en el caso $K < 0$ es este un grupo holomorfo. De hecho, en tal caso el grupo es precisamente aquel del que hemos partido con la «magia compleja» de la teoría de twistores, a saber, el grupo de Lorentz O(1,3), que (ignorando las reflexiones) es el grupo de transformaciones holomorfas de la esfera de Riemann. ¿Dónde está esta esfera de Riemann? Es el «infinito» del 3-espacio hiperbólico —como el círculo limitador de la imagen de Escher, reproducida en la Fig. 2.11— análogo a la esfera celeste de §18.5, como una frontera para el 3-espacio hiperbólico de §18.4; véase la Fig. 18.10.

Vemos que $K < 0$ no es tanto una predicción de la teoría de twistores como de la filosofía holomorfa subyacente. ¿Podemos ir más lejos y decir algo sobre la constante cosmológica Λ? Las construcciones twistoriales propuestas en la actualidad (véase §33.12) parecen poder acomodar la ecuación del vacío de Einstein solo en el caso $\Lambda = 0$, y es difícil ver cómo podría modificarse el tipo actual de procedimiento para acomodar $\Lambda \neq 0$. ¿Nos dice esto que $\Lambda = 0$ es una predicción de la teoría de twistores? ¡Más valdría que no (a pesar de mi preferencia personal por $\Lambda = 0$)! Pues datos observacionales muy recientes (véase §28.10) indican con fuerza $\Lambda > 0$. Esto simplemente plantea a la teoría de twistores nuevos desafíos. ¡Está claro que la teoría de twistores tendrá que hacerlo mucho mejor que esto si tiene que llegar a hacerse respetable como teoría física!

¿Qué pasa con las reglas de la teoría cuántica? ¿Señala la teoría de twistores alguna dirección concreta para el cambio, de acuerdo con las aspiraciones del capítulo 30? El gravitón no lineal de §33.11 empieza a indicar que la aproximación twistorial implicará finalmente una modificación (no lineal) de las reglas de la mecánica cuántica. Sin embargo, todavía no hay mucho, dentro del formalismo twistorial, que indique la presencia de una asimetría temporal fundamental en estas modificaciones, como se requeriría de acuerdo con las discusiones de §§30.2,3,9. Sin embargo, una posible característica sugestiva de los desarrollos «googly» concretos que se han discutido brevemente en §33.12 es que realmente dependen de una descripción con asimetría temporal. La fuerza de esta posibilidad tendrá que esperar a futuros desarrollos, y deberían tenerse en cuenta los comentarios de los párrafos precedentes. Por consiguiente, la teoría de twistores no dice hasta ahora nada útil

sobre la reducción del estado cuántico, pese a que este fenómeno ha proporcionado una parte significativa de los impulsos motivadores iniciales tras la teoría.

Para concluir, abordemos la cuestión del estatus de la filosofía holomorfa subyacente, que constituye uno de los impulsos principales que hay tras la teoría de twistores. Creo que es justo decir que se ha mantenido esta filosofía y que ha ofrecido una poderosa fuerza impulsora, que en algunos aspectos ha superado las expectativas (como sucede con las representaciones twistoriales de campos sin masa, tanto lineales (§§33.8-10) como no lineales (§§33.11,12)). Pese a todo, en algún momento la teoría tendrá que decir algo sobre la presencia de números reales en la física y el comportamiento no holomorfo, tales como la emergencia de valores de probabilidad (de acuerdo con la regla no holomorfa del módulo al cuadrado $z \mapsto |z|^2$) y los puntos espaciotemporales *reales*, donde esperaríamos ser capaces de acomodar el comportamiento no analítico (no digamos ya el no holomorfo). Con respecto a esta última cuestión, podría extraerse cierto aliento de la extraordinaria noción de las hiperfunciones, introducidas al final del capítulo 9 (véase §9.7), según las cuales el comportamiento no analítico puede representarse de forma muy elegante dentro del contexto de operaciones holomorfas. Queda para el futuro saber en qué medida una teoría de twistores futura será realmente capaz de abordar estas cuestiones.

Notas

Sección 33.1

33.1. Véase Ahmavaara (1965).

33.2. Véanse Schild (1949), 't Hooft (1984) y Snyder (1947).

33.3. Véanse Sorkin (1991), Rideout y Sorkin (1999) y Markopoulos y Smolin (1997); uno de los desarrollos más importantes en este campo fue Markopoulos (1998).

33.4. Véanse Kronheimer y Penrose (1967), Geroch *et al.* (1972), Hawking *et al.* (1976), Myrheim (1978) y 't Hooft (1978).

33.5. Véase Finkelstein (1969).

33.6. Véanse Smolin (2001), Gürsey y Tze (1996), Dixon (1994), Manogue y Schray (1993) y Manogue y Dray (1999).

33.7. Véase Regge (1962) para la referencia original. Immirzi (1997) ha escrito una revisión informal (e informativa).

33.8. Jozsa desarrolló estas ideas en su tesis doctoral. Véase Jozsa (1981).

33.9. Véase Isham y Butterfield (2000).

33.10. Véase Goldblatt (1979).

33.11. Véanse Eilenberg y Mac Lane (1945), Mac Lane (1988) y Lawvere y Schanuel (1997).

33.12. Véanse Baez y Dolan (1998), Baez (2000), Baez (2001) y Chari y Pressley (1994).

33.13. Véase Connes y Berberian (1995).

33.14. Hay también otros muchos usos de la geometría no conmutativa, tanto en matemáticas puras como aplicados a la física. Véase Connes (1990, 1998). Como ejemplo de los últimos, se ha desarrollado un elegante formalismo para el tratamiento global de la renormalización con la ayuda de la geometría no conmutativa; véanse 26.9 y Kreimer (2000).

33.15. Véase Connes y Berberian (1995).

Sección 33.2

33.16. Para una exactitud estricta, necesitamos incluir un equivalente a una esfera (de Riemann) de «rayos de luz en el infinito» para completar la definición de $\mathbb{PN}$; véase 33.3.

33.17. Esto necesitaría las apropiadas cantidades (escalares) no direccionales para el grupo de Poincaré (a saber, sus operadores de Casimir; véase §22.12). Estas son el espín total y la masa en reposo (al cuadrado). Sin embargo, no se sabe que la masa en reposo se construya como múltiplos enteros de algo, de modo que los aspectos combinatorios de un esquema semejante no son tan claros. Este enfoque fue desarrollado, de todas formas, por John Moussouris en su tesis doctoral en Oxford en 1983 (véase Moussouris, 1983). Requería una etiqueta adicional unida a las líneas de la red además de masa y espín.

Sección 33.3

33.18. Véanse McLennan (1965) y Penrose (1963, 1964, 1965a, 1986).

Sección 33.4

33.19. Véase Penrose y Rindler (1986), en particular el apéndice.

33.20. Véanse Harvey (1990), Penrose y Rindler (1986) y Budinich y Trautman (1988).

Sección 33.6

33.21. Véanse Penrose y Rindler (1986) y Huggett y Tod (2001).

33.22. En cualquier suceso x en el espaciotiempo hay dos direcciones nulas especificadas. Está la dirección del «rayo de luz» de esta familia que pasa por x y está la dirección del 4-momento de la partícula giratoria que representa el twistor. Estas dos direcciones nulas son las «direcciones nulas principales», i.e., las direcciones definidas por la representación de Majorana (véase §22.10), del (la parte autodual o anti-autodual) momento angular que posee dicha partícula, tomado respecto a x. Véanse Wald (1984) y Huggett y Tod (2001).

33.23. Véanse Penrose (1967, 1975, 1987b) y Penrose y Rindler (1986).

Sección 33.7

33.24. Véanse Penrose (1968b) y Huggett y Todd (2001).

33.25. Véanse Huggett y Todd (2001), Penrose y Rindler (1986) y Hughston (1979).

Sección 33.8

33.26. Véanse Dirac (1936), Fierz (1938, 1940) y Penrose (1965a).

33.27. Véanse Fierz y Pauli (1939), Penrose y Rindler (1986), Penrose (1965a) y Penrose y MacCallum (1972).

33.28. Véanse Penrose (1968b, 1969c, 1987b); Huggett y Todd (2001); Hughston (1979); Whittaker (1903); Bateman (1904, 1944).

33.29. Este es el $\mathbb{C}^2$ que representa la recta $\mathcal{R}$ en $\mathbb{PN}$ (Fig. 33.11). Más familiar para la mayoría de los teóricos de twistores es la versión completamente proyectiva de esta integral de contorno, para la que se utiliza la 1-forma $\iota = \pi_{0'}d\pi_{1'} - \pi_{1'}d\pi_{0'}$ en lugar de la 2-forma $\tau = d\pi_{0'} \times d\pi_{1'}$. La integral de contorno es ahora 1-dimensional, y su relación con la receta 2-dimensional dada aquí es que una de estas dimensiones del contorno (dada por un círculo S^1) reduce la versión no proyectiva dada aquí a la versión proyectiva más familiar. Una ventaja de la presente versión es que permite describir estados de helicidad mixta.

33.30. Véanse Huggett y Todd (2001), Hughston (1979) y Penrose y Rindler (1986).

33.31. La teoría de twistores tiene una gran deuda con sir Michael Atiyah por

un importante ingrediente inicial para esta consecución. Véanse Penrose (1979b) para los argumentos originales para la cohomología twistorial, y Eastwood *et al.* (1981) para un tratamiento completo.

Sección 33.9

33.32. Esto es lo que se conoce como cohomología de Cech. Hay también muchas otras maneras de llegar al concepto de cohomología. Véanse Wells (1991), Ward y Wells (1989) y Griffiths y Harris (1978).

33.33. Véanse Gunning y Rossi (1965), Ward y Wells (1989), Wells (1991), así como también Penrose y Rindler (1986).

33.34. Véanse Gunning y Rossi (1965) y Penrose y Rindler (1986).

33.35. Penrose y Rindler (1986).

33.36. Véase Penrose y Penrose (1958).

33.37. Véase Penrose (1991).

33.38. Véanse Gunning y Rossi (1965), Griffiths y Harris (1978), Chern (1979) y Wells (1991).

33.39. Véanse las referencias en la nota 33.38, así como Eastwood *et al.* (1981).

Sección 33.10

33.40. No tiene importancia la inversión de los + y los − aquí. Se trata tan solo de un molesto accidente notacional. Véase §9.2.

33.41. Hay aquí algunas cuestiones técnicas importantes. Si el campo original no es analítico (no es C^ω), entonces estos campos (en $\mathbb{M}$) pueden resultar hiperfuncionales en el sentido de §9.7; véase Bailey *et al.* (1982).

33.42. Véase Hodges *et al.* (1989).

33.43. Véase Penrose (1987b).

Sección 33.11

33.44. Véanse Penrose (1976a, 1976b), Ward (1977), Penrose y Ward (1980) y Penrose y Rindler (1986).

33.45. Hay una sutileza técnica en cuanto que no es un fibrado *holomorfo* (§15.5) pese a que todas las operaciones en la construcción son holomorfas, puesto que, localmente en el $\mathcal{PT}$-espacio, no es estrictamente un espacio producto holomorfo. $\mathcal{T}$ se conoce como una *fibración holomorfa*. Véase Penrose (1976b).

33.46. En circunstancias normales; véanse Huggett y Tod (2001), Ward y Wells (1989) y Penrose y Ward (1980).

33.47. Véase Kodaira (1962).

33.48. O en el caso real definido positivo (++++) o en el caso de signatura dividida (+ + − −). Véanse Penrose (1976b), Hansen *et al.* (1978), Atiyah *et al.* (1978) y Dunajski (2002).

33.49. Parece haber así una relación significativa con la noción de una *QFT topológica*, como se ha mencionado en §32.5; véase la nota 32.17.

Sección 33.12

33.50. Hay una aproximación con simetría de reflexión a estos problemas que utiliza lo que se conoce como *ambitwistores*, y esta ha disfrutado de algunos éxitos parciales significativos en esta dirección; véanse Penrose (1975), LeBrun (1985, 1990), Isenberg *et al.* (1978) y Witten (1978). Véase también Penrose y Rindler (1986). Un ambitwistor en un espacio plano es básicamente un par (W_α, Z^α), donde $W_\alpha Z^\alpha = 0$, y describe un rayo de luz complejo. Sin embargo, esto no encaja con la filosofía antes adoptada de que «las funciones twistoriales son funciones de onda», puesto que una descripción ambitwistorial es más parecida a una clásica, en la que aparecen una variable y su variable conjugada —aquí Z^α y W_α— en lugar de solo una o la otra, como sería apropiado para una función de onda. La aproximación ambitwistorial también encuentra algunas dificultades matemáticas en su descripción de campos no lineales.

33.51. Véase Penrose (2001).

33.52. Véanse, por ejemplo, Frittelli *et al.* (1997), Bramson (1975) y Penrose (1992).

33.53. Véase Penrose (2001).

33.54. Véase Mason y Woodhouse (1996).

Sección 33.13

33.55. Para algunas de las referencias más antiguas, véanse Penrose y MacCallum (1972) y Penrose (1975). Para trabajos más recientes, véase Hodges (1982, 1985, 1990, 1998).

Sección 33.14

33.56. Véanse Bailey y Baston (1990), Baston y Eastwood (1989) y Mason y Woodhouse (1996) para el uso de twistores en matemáticas.

33.57. Véase Merkulov y Schwachhöfer (1998).

33.58. Véanse Gindikin (1986, 1990) y Lebrun y Mason (2002).

33.59. Véase la nota 31.81.

33.60. Aunque los lagrangianos han desempeñado un papel periférico en la teoría de twistores para entender las interacciones físicas, todavía no han encontrado una formulación general adecuada dentro de dicha teoría. Hay quizá cierta ironía en que el éxito que ha tenido la teoría de twistores al presentar los campos físicos de una manera que implícitamente resuelve sus ecuaciones de campo (por medio de funciones twistoriales homogéneas, en el caso de campos libres sin masa) es lo que lleva a una dificultad con su formulación lagrangiana. En la formulación cuántica convencional, las ecuaciones de campo proceden de una «suma sobre historias» (§26.6), donde debe ser explícitamente posible *violar* las ecuaciones de campo en la formulación, para que esta idea tenga sentido, y entonces las correcciones cuánticas a la teoría clásica proceden del examen más detallado de esta integral de camino. ¡Todo esto se pierde si la formulación no permite la violación de las ecuaciones de campo! Creo que es necesaria una reevaluación de lo que debe ser la auténtica «esencia» de los lagrangianos en la teoría de twistores y, de hecho, en la teoría física en general. Quizá esto tiene relación con las preocupaciones que he expresado al final de §26.6 y con las cuestiones auténticamente profundas presentadas por el prácticamente ubicuo problema de la divergencia que aparece con las integrales de camino (véase §26.6). Pero véanse la parte final de 32.5 y la nota 32.17.

33.61. Penrose y Rindler (1986), §9.5.

34

¿Dónde está el camino a la realidad?

34.1. Las grandes teorías de la física del siglo xx... ¿y más allá?

Tratemos de recapitular lo que nuestras teorías físicas nos han enseñado —cuando empezamos a explorar el tercer milenio a.C.— acerca de la naturaleza fundamental de este notable mundo en el que nos encontramos. No hay duda de que se han producido extraordinarios avances en el conocimiento, y de que estos han llegado a través de una cuidadosa observación física y una soberbia experimentación, a través de razonamientos físicos de gran profundidad e intuición, y a través de razonamientos matemáticos que van desde los complicados pero rutinarios hasta los más inspirados. Estos nos ha llevado desde las ideas de los antiguos griegos respecto a la geometría del espacio hasta las magníficas estructuras de la mecánica clásica, pasando por la mecánica newtoniana; y luego a la teoría electromagnética de Maxwell y a la termodinámica. Más recientemente, el siglo xx nos ha dado la relatividad especial, que llevó a la extraordinaria y bien verificada teoría de la relatividad general de Einstein, y también tenemos la profundamente misteriosa aunque profundamente exacta y muy general mecánica cuántica, y su desarrollo en la teoría cuántica de campos (QFT); en concreto, disponemos de los extraordinarios y exitosos modelos estándar de la física de partículas y de la cosmología.

No es extraño que teóricos confiados crean que quizá estemos «casi al final», y que pueda haber una «teoría de todo» no mucho más allá de los desarrollos posteriores de finales del siglo xx. Con frecuencia estos comentarios se han hecho con un ojo puesto en la «teoría de

cuerdas» que existiera en cada momento. Es más difícil mantener este punto de vista ahora que la teoría de cuerdas se ha metamorfoseado en algo (teoría **F** o teoría **M**) de cuya naturaleza se admite que es esencialmente desconocida en el momento presente.

Personalmente, creo que todavía estamos mucho más lejos que esto de una «teoría final». No tengo ninguna fe en que los desarrollos esbozados en el capítulo 31 se acerquen siquiera a las líneas correctas. Es cierto que varios desarrollos *matemáticos* extraordinarios son producto de ideas de la teoría de cuerdas (y otras relacionadas). Sin embargo, sigo siendo profundamente escéptico respecto a que sean otra cosa que fragmentos asombrosos de matemáticas, aunque con cierto aporte de profundas ideas físicas. En el caso de las teorías cuya dimensionalidad espaciotemporal supera la que directamente observamos (a saber, 1 + 3), no veo ninguna razón para creer (§§31.11,12) que por sí mismas nos lleven mucho más lejos en la dirección del conocimiento *físico*. Con respecto a los otros esquemas que se han propuesto, tales como los esbozados en los capítulos 32 y 33, con los que estoy mucho más de acuerdo, no hay duda de que también a estos les faltan algunas ideas importantes. Sería imprudente predecir con mucha confianza que estas teorías estén a punto de dar los pasos adicionales necesarios que nos pudieran guiar hacia el verdadero camino a la comprensión de la realidad física.

Pese a todo, ya en el siglo xx la especie humana ha hecho sin duda progresos extraordinarios hacia tal comprensión, y en este libro he intentado transmitir algo de lo que se ha logrado. La relatividad general de Einstein sobresale, en mi opinión, como el mayor logro individual de este siglo. Muchos científicos podrían muy bien considerar la teoría cuántica (y la QFT) como un logro incluso mayor. Desde mi perspectiva concreta sobre la cuestión, no me siento capaz de compartir esta opinión. Aunque es indudablemente cierto que la teoría cuántica ha explicado un número de cosas incomparablemente mayor que la relatividad, y en un abanico muchísimo mayor de fenómenos diferentes, no considero que la teoría haya alcanzado todavía la coherencia necesaria *como* tal. El problema reside, por supuesto, en la paradoja de la medida, considerada en el capítulo 29. En mi opinión, la teoría cuántica es incompleta. Cuando esté completa —lo que diría que sucederá

en algún momento del siglo XXI—, representará sin duda un logro mayor incluso que la relatividad general de Einstein. En realidad, como sugieren firmemente las afirmaciones del capítulo 30, semejante mecánica cuántica completa debería incluir la teoría de Einstein como caso límite para grandes masas y distancias. (Y espero que mis comentarios en §31.8 hayan dejado claro al lector que no creo que la teoría de cuerdas ya lo haya conseguido, pese a las numerosas afirmaciones que se hacen en sentido contrario.)

En mi opinión, la relatividad general va a permanecer probablemente como una descripción del espaciotiempo en el límite de gran escala (donde se permite la presencia de una constante cosmológica Λ como parte de la teoría de Einstein), aunque debemos esperar serias modificaciones en sus descripciones a la distancia absurdamente minúscula de Planck de 10^{-35} m, o allí donde las densidades puedan aproximarse al valor de Planck de aproximadamente 5×10^{93} veces la del agua, en la vecindad de ciertas singularidades espaciotemporales. Esta posición sobre el estatus de la relatividad general debe considerarse ahora como la convencional. El estatus observacional de la teoría, al menos cerca del extremo más grande de la escala de distancias donde se sitúan las estrellas de neutrones en órbita y los efectos de lente gravitacional, e incluso de agujeros negros, debe considerarse excelente. Y aquí me refiero a la teoría de Einstein estándar, sin una constante cosmológica.

Pero ¿qué pasa con la constante cosmológica? Las observaciones de los últimos años parecen favorecer un valor positivo para ella. Si existe realmente Λ, es ciertamente muy pequeña. Si se considera Λ como una curvatura, entonces es el recíproco del cuadrado de una distancia que es al menos comparable al radio del universo observable, de modo que Λ es ciertamente despreciable salvo en escalas cosmológicas. Cuando se interpreta Λ como una densidad efectiva Ω_Λ, entonces dicha densidad no puede ser mayor que 2 o 3 veces la minúscula densidad media que tiene ahora nuestro universo, que es aproximadamente 10^{-27} kg m^{-3}, considerablemente menor que el mejor vacío producido artificialmente en la Tierra. Una vez más, Λ solo podría tener relevancia en escalas cosmológicas. Pese a todo, y sobre la base del punto de vista expresado a menudo por los teóricos de los campos cuánticos, Λ es realmente una

medida de la densidad efectiva del vacío, generada por «fluctuaciones mecanocuánticas del vacío» (una característica de la incertidumbre de Heisenberg en QFT; véase §21.11, así como también §29.6 y §30.14) y, por consiguiente, «debería» tener un valor (comparable al valor de Planck) ¡que es algo del orden de 10^{120} veces mayor (o, según algunas propuestas, posiblemente solo 10^{60} veces mayor) que el límite superior de lo que se observa! Esto se considera un enigma fundamental en QFT,[1] no resuelto por ninguno de los enfoques convencionales de la gravedad cuántica ni por la teoría de cuerdas. Mi propia actitud se ve menos perturbada por esto que lo que parece estar para muchos teóricos. Mi conjetura (véase §30.14) es que la cuestión general de las «fluctuaciones del vacío» tendrá que ser radicalmente revisada cuando tengamos una mejor teoría cuántica de la gravitación y, realmente, una mejor QFT.

Debemos reconocer, por supuesto, el extraordinario abanico de fenómenos que apoyan la mecánica cuántica y la QFT existentes. Pero debería dejar claro que esto no se contradice con el punto de vista que he defendido en el capítulo 30, en el que se prevén cambios en los fundamentos de la teoría cuántica. Ningún experimento hasta la fecha parece haberse acercado mucho a la exploración del nivel de «gravedad cuántica» en el que espero que se pongan de manifiesto los cambios y ocurra objetivamente la reducción del vector de estado (**RO** gravitatoria). Los entrelazamientos cuánticos observados en distancias de hasta 15 kilómetros[2] son perfectamente compatibles con estos cambios esperados, puesto que estos entrelazamientos implican solo pares de fotones con energías del orden de 10^{-19} J, y no cabe esperar la reducción de estado espontánea de acuerdo con la **RO** gravitatoria hasta que los fotones se midan realmente (en cuyo instante tendría lugar **RO** en el aparato de medida). La situación experimental actual con respecto a la validez de la mecánica cuántica cuando están involucrados movimientos de masa importantes queda bien manifiesta en recientes experimentos realizados por Anton Zeilinger y colegas en Viena.[3] Han llevado a cabo lo que en esencia es un experimento de doble rendija con «bucky bolas» de C_{60} (y también C_{70}). Estas son *fullerenos*, en los que cada molécula tiene 60 átomos de carbono, en una disposición geométrica muy bella que recuerda la forma de las costuras en un balón de

fútbol moderno (o, en el otro caso, una disposición menos simétrica que incluye 70 átomos de carbono). Estas moléculas de fullereno tienen un diámetro aproximado de un nanometro e interfieren consigo mismas después de haber estado en una superposición de dos localizaciones separadas por aproximadamente 10^{-7} m, que es unas 100 veces el diámetro de la bucky-bola. Según el esquema sugerido en §30.11, una superposición semejante duraría al menos unos cien mil años antes de reducirse de manera espontánea según la **RO** gravitatoria, de modo que claramente no hay contradicción aquí con el experimento de Zeilinger.

Por supuesto, la situación podría muy bien ser diferente con algunos experimentos futuros. Algo del tipo de la propuesta FELIX en el espacio de §30.13, o más probablemente algún experimento relacionado tal como podría resultar del trabajo de Dik Bouwmeester en Santa Bárbara, pondría a prueba directamente el esquema **RO** gravitatorio, y muy bien podría suceder que existan otras posibilidades para experimentos que pudieran realizarse a principios del siglo XXI. Considero particularmente excitante esta perspectiva, y existe la posibilidad de que tales experimentos pudieran cambiar de forma significativa nuestra visión actual de la mecánica cuántica. Como mínimo, podrían limitar seriamente las especulaciones acerca de las posibles modificaciones de la mecánica cuántica de acuerdo con alguna teoría futura.

Esto contrasta con la situación experimental actual (o plausiblemente extrapolada) con respecto a otros intentos de combinar la teoría cuántica y la gravitación, tales como los esbozados en los capítulos 31-33. La mayor parte de los experimentos diseñados para abordar tales propuestas de gravitación cuántica implican partículas lanzadas a energías extraordinariamente altas, absurdamente alejadas de las capacidades de cualquier acelerador de partículas existente (o en proyecto). (Las únicas excepciones a esto que conozco son experimentos diseñados para comprobar la posibilidad de existencia de «grandes» dimensiones extra —§31.4— que podrían afectar, por ejemplo, a la ley inversa del cuadrado para la gravedad a pequeñas distancias, o a algunas otras propuestas vagamente relacionadas dirigidas a ver si la covariancia de Lorentz podría violarse a alta energía, debido a efectos cuántico gravitatorios sugeridos.)[4] Hay, de hecho, una seria dificultad al afrontar la

puesta a prueba de cualquier esquema convencional de «gravitación cuántica», pues los efectos de esta unión modificarían únicamente la estructura espaciotemporal (en los extraordinariamente minúsculos tiempos o distancias de Planck), dejando intactos los procedimientos estándar de la mecánica cuántica. Estamos en mejores condiciones, experimentalmente, si las reglas de la mecánica cuántica son modificadas por efectos de la relatividad general, tal como se ha sugerido en el capítulo 30, puesto que estos efectos propuestos están dentro del alcance de la tecnología actual. Si tales experimentos se realizan con éxito e indican la necesidad de un cambio en las reglas de la mecánica cuántica, entonces habría al fin una buena guía física para complementar los desiderata básicamente matemáticos que dirigen la investigación actual en la gravitación cuántica.

La ausencia de datos experimentales relativos a las propuestas normales de gravitación cuántica ha conducido a una curiosa situación en la investigación en la física teórica fundamental. Al parecer, existe un consenso general en que para hacer progresos reales más allá de los modelos estándar de la física de partículas y la cosmología, y obtener un conocimiento más profundo de los ingredientes básicos del universo, será necesario tener una teoría que englobe la gravedad además de las fuerzas fuerte, débil y electromagnética. Parece ser que la razón para esto es en parte la idea (sin duda físicamente justificada) de que una QFT *finita* (en oposición a meramente renormalizable) exigirá que las divergencias se «corten» a la minúscula distancia de Planck, donde la gravedad debe formar necesariamente parte de la imagen (véase §31.1). Pero puesto que no hay experimentos en esta área, los esfuerzos de los teóricos han estado muy dirigidos al mundo interno de los desiderata matemáticos.

34.2. Física fundamental matemáticamente dirigida

La relación entre ideas matemáticas y comportamiento físico ha sido un tema constante en este libro. A lo largo de la historia de la ciencia física, el progreso se ha conseguido al encontrar el equilibrio correcto entre las restricciones, tentaciones y revelaciones de la teoría matemá-

tica, por una parte, y la observación precisa de las acciones del mundo físico, por otra, normalmente mediante experimentos cuidadosamente controlados. Cuando la guía experimental está ausente, como es el caso con la investigación fundamental más actual, este equilibrio se rompe. La coherencia matemática[5] está lejos de ser un criterio suficiente para decirnos si es probable que estemos «en la pista correcta» (y en muchos casos incluso estos requisitos que parecen esenciales son arrojados al viento). Vemos que las virtudes matemáticas estéticas empiezan a desempeñar un papel mucho más importante que antes. Los investigadores suelen señalar los éxitos de Dirac, Schrödinger, Einstein, Feynman y muchos otros que estuvieron guiados en un grado considerable por los atractivos estéticos de las ideas teóricas concretas que ellos proponían. Creo que no hay que negar el valor de tales consideraciones estéticas, que desempeñan un papel de crucial importancia en la selección de propuestas plausibles de nuevas teorías de la física fundamental.

Algunos de estos juicios estéticos pueden expresar a veces simplemente una clara necesidad de un esquema matemático coherente; pues, en efecto, belleza y consistencia matemáticas están íntimamente relacionadas. Creo que la necesidad de tal coherencia, en cualquier modelo físico propuesto, es indiscutible. Más aún, a diferencia de muchos criterios estéticos, la coherencia matemática tiene la ventaja de que es claramente algo objetivo. La dificultad con los juicios estéticos reside, en general, en que suelen ser muy subjetivos.

Pese a todo, la coherencia matemática no tiene por qué ser algo que se aprecie de inmediato. Quienes han trabajado larga y arduamente con un conjunto de ideas matemáticas, quizá están en una mejor posición para apreciar la sutil, y con frecuencia inesperada, unidad que puede haber dentro de algún esquema concreto. Quienes, por el contrario, llegan a tal esquema desde fuera, quizá lo vean con más perplejidad, y tal vez les cueste apreciar por qué tal y cual propiedad debería tener un mérito particular, o incluso por qué algunas cosas en la teoría deberían considerarse más sorprendentes —y, quizá por ello, más bellas— que otras. Pero, una vez más, puede haber circunstancias en que son quienes vienen de fuera los que mejor pueden hacer juicios objetivos, ¡pues quizá pasar muchos años tratando con un conjunto muy limitado de problemas matemáticos que surgen dentro de un enfoque concreto produce juicios distorsionados!

Pero, a pesar de su indudable valor, la elegancia y coherencia en las matemáticas de una teoría física están muy lejos de ser suficientes. Normalmente, las condiciones físicas tienen una importancia mucho mayor. Pero en las situaciones donde se carece de una guía física, las cualidades matemáticas cobran una importancia mucho mayor. Yo no argumentaría, por supuesto, que haya respuestas simples a estas cuestiones. Creo que los investigadores individuales tienen razón en seguir sus impulsos estéticos. Sin embargo, no deberían sorprenderse si encuentran que algún colega se muestra indiferente ante la supuesta trascendencia de las conclusiones a las que puedan llevar tales impulsos. Considero que tales motivaciones estéticas son una parte esencial del desarrollo de cualquier idea nueva importante en la ciencia teórica. Pero sin las restricciones del experimento y la observación, tales motivaciones llevan con frecuencia la teoría mucho más allá de lo que está justificado físicamente.

A lo largo de la historia, podemos ver muchos ejemplos de un bello esquema matemático que al principio parecía proporcionar una forma nueva y revolucionaria de desvelar los secretos de la naturaleza, pero en donde estas esperanzas iniciales no se han visto satisfechas, al menos, no en la forma inicialmente prevista. Un buen ejemplo puede ser el sistema de los cuaterniones, por lo que se refiere a su bella propiedad de constituir un álgebra de división. Como hemos visto en §11.2, una vez que Hamilton los descubrió en 1843, se vio tentado a dedicar los veintidós años restantes de su vida a intentar representar las leyes de la naturaleza enteramente dentro de ese marco. Sin embargo, este trabajo en «cuaterniones puros» (por los que entiendo sus cuaterniones reales originales, con su propiedad de álgebra de división) tuvo muy poco efecto directo sobre el desarrollo posterior de la ciencia física básica. Otras influencias de Hamilton en la teoría física han sido ciertamente enormes, y muy directas. En efecto, fueron sus propias investigaciones anteriores en lo que ahora llamamos «hamiltonianos», «principio de Hamilton», y la «ecuación de Hamilton-Jacobi», etc. —que forman parte de una exploración de la analogía entre partículas newtonianas y ondas—, las que proporcionaron el

punto de despegue para el desarrollo en el siglo XX de la mecánica cuántica y la QFT (véanse §20.2 y §§21.1,2). Pero la influencia de los cuaterniones en la física fue lejana, solo a través de generalizaciones en las que la propiedad del álgebra de división tuvo que ser desestimada.

A mediados del siglo XIX, debió de ser demasiado fácil quedar hipnotizado por esa bella característica matemática de los cuaterniones que hace posible dividir por ellos (§11.1). Esa maravillosa propiedad, de la que disfrutan los cuaterniones, y relativamente pocas álgebras más, ha ejercido una influencia importante en matemáticas puras, pero no directamente en la corriente principal de la física. Fue la generalización que hizo Clifford de los cuaterniones a dimensiones más altas, junto con las ideas posteriores de Pauli y en especial de Dirac, en las que se adopta la signatura lorentziana relevante para el espaciotiempo, la que permitió finalmente que fuera posible dar un gran paso en la teoría física (§11.5, §§24.6,7). Pero en estos desarrollos posteriores, que fueron muy importantes para la física, la bella propiedad de división ¡tuvo que ser necesariamente abandonada!

Volveré en §34.9 a esta cuestión algo misteriosa del éxito que puede tener en física la belleza en matemáticas. Pero estos puntos afectan a la importante y fascinante cuestión complementaria de los «retornos» matemáticos. Ya desde la época de los antiguos griegos, las teorías que empezaron próximas al comportamiento del mundo han generado siempre vastas áreas de bellas matemáticas, inicialmente estudiadas por dichas razones pero que a menudo encontraron aplicaciones muy alejadas de aquellas consideraciones físicas que las originaron. A veces se necesitaron varios siglos antes de que se hallaran esas aplicaciones (como fue el caso del estudio de las secciones cónicas por parte de Apolonio, hacia el 200 a.C., que desempeñó un papel fundamental en la comprensión de los movimientos planetarios que ofrecieron Kepler y Newton en los siglos XVI y XVII, o el caso del «pequeño teorema» de Fermat de 1640, que encontró importantes aplicaciones en la criptografía a finales del siglo XX). Las matemáticas —en especial, las buenas matemáticas— tienen la costumbre de encontrar sus aplicaciones en campos muy dispares, y esta es una de las razones de su fuerza y solidez. Las obras de la naturaleza

han proporcionado con frecuencia una fuente maravillosa de tales ideas matemáticas. Quizá no sea sorprendente que hubiera precisión y fiabilidad en estas ideas estimuladas por la naturaleza si aceptamos que esta opera exactamente de acuerdo con leyes matemáticas. Más notable es la *sutileza* de las matemáticas que parecen estar involucradas en las leyes de la naturaleza, y la propensión que parecen tener aquellas a encontrar aplicaciones en áreas muy alejadas de su propósito original (como fue el caso, en especial, del cálculo infinitesimal de Newton y Leibniz; véase el capítulo 6).

Pero ¿podemos argumentar recíprocamente que una presunta teoría física que estimula mucha investigación productiva en áreas matemáticas muy dispersas gana credibilidad en virtud de esto? Esta cuestión tiene especial importancia para los esquemas físicos de los capítulos 31,32 y 33. Creo que no se puede responder de una forma sencilla, sino que debe recomendarse una especial cautela.

La teoría de cuerdas, en particular, ha estimulado mucha y muy bella investigación matemática, y buena parte de su fuerza le viene de este atractivo matemático. (Lo mismo podría decirse de gran parte de la teoría de twistores y de los enfoques de Ashtekar y Hawking). Pero no está claro en qué medida esto es indicativo de un acuerdo subyacente con la realidad física. Pese a todo, a menudo he oído a matemáticos puros expresar con satisfacción que cierto resultado que han podido encontrar tiene aplicaciones en física ¡simplemente porque tiene una relevancia matemática para la teoría de cuerdas! Puedo comprender muy bien que muchos matemáticos puros deseen que aspectos de su bella disciplina encuentren una aplicación importante para el funcionamiento del mundo físico. Pero debería quedar claro que no existe (por el momento) ninguna razón observacional para creer que la teoría de cuerdas (en particular) *sea* física, aunque está ciertamente motivada por poderosas aspiraciones físicas. La teoría de cuerdas es también un tema que han estudiado muchos físicos, pero ¿eso la hace física? Esto plantea la cuestión de las modas en la investigación física fundamental, un tema que me gustaría abordar a continuación.

34.3. El papel de las modas en la teoría física

Permítanme empezar citando un informe realizado por Carlo Rovelli, y hecho público en su alocución durante el Congreso Internacional sobre Relatividad General y Gravitación, celebrado en Pune, India, en diciembre de 1997.[6] Rovelli es uno de los creadores del enfoque de las variables de lazo en la gravedad cuántica, como se ha descrito en §32.4, y no pretendía ningún corporativismo en la realización de su iforme. Pese a todo, los resultados que encontró reflejan las que hubieran sido mis propias expectativas (no contrastadas). Hizo un recuento de los artículos sobre la gravedad cuántica publicados el año anterior, tal como estaban registrados en los archivos de Los Ángeles. El promedio aproximado de artículos por mes, en los diferentes enfoques de esta disciplina, era el siguiente

Teoría de cuerdas	69
Gravedad cuántica de lazo	25
QFT en espacios curvos	8
Aproximaciones en retículos	7
Gravedad cuántica euclídea	3
Geometría no conmutativa	3
Cosmología cuántica	1
Twistores	1
Otros	6

El lector advertirá que en el espacio que he dedicado en este libro a las exposiciones de estas respectivas teorías, solo he seguido muy vagamente las demandas de la moda. (He tocado brevemente la QFT en espacios curvos en relación con el efecto Hawking, como se ha examinado en §30.4. Las aproximaciones en retículos adoptan modelos espaciotemporales discretos en lugar de continuos —véase §33.1—. La gravedad cuántica euclídea interviene en el enfoque de Hawking, como se ha examinado en §28.9. La cosmología cuántica utiliza espaciotiempos simplificados en los que se ignoran la mayoría de los grados de libertad gravitatorios. Los otros enfoques citados se han discutido en los capítulos 31-33). Se advertirá que había más artículos en el área de la

teoría de cuerdas que en todas las demás áreas juntas. Parece que hay una opinión general de que si hoy se volviera a hacer un informe semejante, la preponderancia de los artículos de la teoría de cuerdas sería aún mayor. Si pensáramos que la investigación física debe regirse por los principios del gobierno democrático, entonces veríamos que, debido a una mayoría absoluta por parte de los teóricos de cuerdas, ¡todas las decisiones respecto a la investigación que deberían hacerse estarían dictadas por ellos![7]

Afortunadamente, los criterios de la ciencia no son los del gobierno democrático. Es correcto y oportuno que las actividades de la minoría no deban sufrir en virtud del simple hecho de que están en minoría. La coherencia matemática y el acuerdo con la observación son mucho más importantes. Pero ¿podemos ignorar por completo los caprichos de la moda? Ciertamente no podemos. Además de algunas ideas mucho menos creíbles, muy de moda en su día (tales como la noción de la supergravedad 11-dimensional de siete dimensiones extra que constituyen una «7-esfera aplastada»),[8] recuerdo muchas modas del pasado que me parecieron —y aún me parecen— contener verdades muy significativas (tales como las trayectorias de Regge —véase §31.5— y la matriz S analítica[9] de Geoffrey Chew), pero que hoy hace décadas que han pasado de moda. Hasta cierto punto, la popularidad de una teoría proporciona una medida de su plausibilidad científica, pero solo hasta *cierto* punto.

También es cierto que, como sucede en el mundo empresarial, son las grandes compañías las que tienen una tendencia natural a hacerse más grandes a expensas de las más pequeñas. No es difícil ver por qué esto debería suceder también con las modas científicas, sobre todo en el mundo moderno de los vuelos a reacción e internet, donde las ideas nuevas científicas se difunden a gran velocidad por todo el planeta, propagadas de viva voz en congresos científicos o transmitidas casi instantáneamente por correo electrónico o en artículos científicos (con frecuencia no sometidos a revisión) en internet. La a menudo frenética competitividad que genera esta facilidad de comunicación conduce a efectos «caravana», donde los investigadores temen quedar rezagados si no se suben al carro. La moda no tiene mucho que decir sobre ideas teóricas que continuamente caen bajo el escrutinio experimental. Pero

con ideas que están tan lejos de la posibilidad de confirmación o refutación experimental como lo están las de la gravedad cuántica, debemos ser especialmente cautos en tomar la popularidad de un enfoque como índice real de su validez.

La moda tiene también desempeña un papel en otras áreas tales como la notación o el formalismo matemático específico. Esta es quizá una cuestión menos importante que las que se han planteado antes, pero continúa siendo importante para el desarrollo de la investigación. Permítanme describir un ejemplo concreto, a saber, el uso tan dominante del formalismo espinorial de 4 componentes de Dirac en lugar del formalismo posterior de 2 componentes de Van der Waerden (véanse §22.8, §24.7 y §25.2). Hay cierta ironía en esto, como veremos. De hecho, en electrodinámica cuántica el formalismo 4-espinorial es utilizado casi universalmente, mientras que, como ha demostrado Robert Geroch,[10] es en realidad mucho más sencillo utilizar 2-espinores (brevemente descritos en §22.8). Cuando Dirac descubrió su ecuación, en 1928, utilizó 4-espinores. La ecuación de Dirac estimuló mucho interés en la importancia de los espinores, y un año después el destacado matemático holandés B. L. van der Waerden formuló el potente cálculo 2-espinorial.[11] Sin embargo, la expectación causada por entonces a raíz del descubrimiento de la ecuación del electrón de Dirac significó que la mayoría de los físicos siguieron el enfoque original de Dirac, pues no había muchos que estuvieran siquiera familiarizados con el formalismo más flexible y refinado de Van der Waerden. De todas formas, parece que el propio Dirac apreció finalmente la potencia de lo que Van der Waerden había realizado. De hecho, a comienzos de la década de 1950 asistí a un curso de conferencias de Dirac donde expuso una bella introducción al cálculo 2-espinorial, que clarificó para mí el tema general cuando me había parecido casi desconcertante en las exposiciones que había leído anteriormente.

El propio Dirac había utilizado en realidad el enfoque 2-espinorial, en 1936, para encontrar generalizaciones de su ecuación para el electrón a partículas de espín más alto.[12] Pero no sintiéndose cómodos con el formalismo 2-espinorial, parece que varios investigadores redescubrieron casos especiales de las ecuaciones de Dirac para el espín más alto, que ahora suelen recibir nombres tales como la «ecuación de Duffing-Kemmer-Petiau» (1936-1939, para espín 0 y 1), la «ecuación de Proca» (1930) y la «ecuación de Rarita-Schwinger» (1941, para espín 3/2). Son los trabajos de estos otros investigadores en esta área (utilizando métodos tensor/4-espinor), más que el trabajo anterior de Dirac, los que cita la gente.

Dirac no era un seguidor de la moda, y parece que ¡ni siquiera siguió siempre la moda que él mismo había establecido! De todas formas, otros se ven a veces arrastrados a ella incluso cuando no lo pretenden. Conocí un ejemplo de esto cuando visité el CERN a mediados de la década de 1970 para hablar con Bruno Zumino, una de las personas a quien se deben las ideas básicas de la supersimetría. (Su trabajo de 1974 con Julius Weiss[13] había tenido una conexión precisa con la teoría de twistores, y yo quería explorarla.) Me dijo que había apreciado la fuerza del formalismo 2-espinorial, y había escrito en cierta ocasión un artículo en el que utilizaba 2-espinores para formular una idea suya. Sin embargo, unos meses más tarde, según me comentó, el reputado físico Abdus Salam propuso la misma idea aunque utilizando 4-espinores. Todo el mundo se refirió entonces al artículo de Salam y nadie al suyo. ¡Zumino llegó a la conclusión de que no cometería de nuevo el error de utilizar el formalismo 2-espinorial (técnicamente superior)!

Hay una cuestión relacionada que dificulta a los investigadores, en especial a los jóvenes, romper con las líneas de moda en investigación, aunque quieran hacerlo. Es la extraordinaria cantidad de ideas matemáticas difíciles y dispares a las que deben enfrentarse en la física matemática moderna. Resulta muy difícil discriminar una pequeña parte de una línea de trabajo concreta e intentar dominarla. Hacer un estudio comparativo autorizado de los méritos globales de varias líneas diferentes estaría ciertamente más allá de las capacidades de la mayoría de los investigadores jóvenes. Si tienen que elegir, deben fiarse de las preferencias de aquellos que son investigadores ya consagrados, y esto solo puede sumarse a la propagación de las líneas de trabajo que ya están de moda, a costa de aquellas que son peor conocidas.

Aunque mis comentarios anteriores estaban dirigidos al tipo de investigación teórica que no está limitada por resultados experimentales, el elemento de la moda no deja de ser importante también en relación

con el experimento, si bien por una razón algo diferente. Esta deriva en esencia del enorme coste que normalmente conlleva el montaje de los experimentos en la frontera de la física fundamental. Puesto que la mayoría de los experimentos son muy caros, se suele requerir un apoyo gubernamental, o de grandes empresas, y serán necesarios numerosos comités para decidir si seguir adelante con un experimento, o si este u otro experimento utilizará mejor los limitados fondos. Es natural que los expertos científicos que forman parte de dichos comités sean aquellos que han destacado por su contribución al desarrollo de ideas que han llevado con éxito a las perspectivas actuales. Así, ellos tenderán a favorecer experimentos que directamente aborden cuestiones que parecen naturales desde estas perspectivas concretas. Hay, por lo tanto, una fuerte tendencia a que la teoría se quede «bloqueada» en direcciones concretas. Por razones como estas, será intrínsecamente muy difícil hacer un cambio importante de dirección.

34.4. ¿Puede refutarse experimentalmente una teoría errónea?

Se podría pensar que aquí no hay ningún peligro real, porque, si la dirección es errónea, entonces el experimento la refutará, de modo que nos veremos obligados a seguir una nueva dirección. Esta es la imagen tradicional de cómo progresa la ciencia. De hecho, el bien conocido filósofo de la ciencia Karl Popper ofreció un criterio que parecía razonable[14] para la admisibilidad científica de una teoría propuesta, a saber, que sea *refutable observacionalmente*. Pero me temo que esta es una visión demasiado estricta de la ciencia en este mundo moderno de la «gran ciencia».

Tomemos el ejemplo de la supersimetría en la moderna física de partículas. Es una idea teórica con una evidente elegancia matemática, y que hace más fácil la vida del teórico a la hora de construir QFT renormalizables (§31.2). Y lo más importante, es un ingrediente fundamental en la teoría de cuerdas. Su estatus entre los teóricos es tan fuerte actualmente que casi se considera una parte del modelo «estándar» en la física de partículas actual. Pese a todo, y tal como están las cosas,

no tiene ningún apoyo experimental (serio). La teoría predice «supercompañeras» para todas las partículas observadas en la naturaleza, pero ninguna de aquellas ha sido observada hasta ahora. La razón de que no lo hayan sido, según los teóricos de la supersimetría, es que un mecanismo de ruptura de simetría (de naturaleza desconocida) hace que las supercompañeras sean tan masivas que las energías necesarias para crearlas están más allá de las posibilidades de los aceleradores actuales. Con posibilidades de energía aumentadas, las supercompañeras podrían encontrarse, y con ello se habría conseguido un hito en la teoría física con implicaciones importantes para el futuro. Pero supongamos que las supercompañeras siguen sin encontrarse. ¿Refutaría esto la idea de supersimetría? En absoluto. Podría argumentarse (y probablemente se haría) que había habido demasiado optimismo sobre el grado de la ruptura de simetría, y que se necesitarían energías aún mayores para encontrar las supercompañeras perdidas.

Vemos que no es tan fácil desalojar una idea teórica popular mediante el tradicional método científico del experimento crucial, incluso si resulta que la idea es realmente errónea. Además, el enorme coste económico de los experimentos a altas energías hace considerablemente más difícil poner a prueba una teoría de lo que podría haber sido de otra forma. Existen otras muchas propuestas teóricas, en la física de partículas, en las que las partículas predichas tienen masas-energías demasiado altas para cualquier posibilidad seria de refutación. Varias versiones concretas de GUT o de la teoría de cuerdas hacen muchas «predicciones» de este tipo que son completamente inmunes a la refutación por razones de este tipo.

El carácter «apopperiano» de tales modelos, ¿las hace inaceptables como teorías científicas? Pienso que un juicio popperiano tan estricto sería decididamente demasiado severo. Como ejemplo intrigante, recordemos el argumento de Dirac (§28.2) de que la mera existencia de un único monopolo magnético en alguna parte del cosmos proporcionaría una explicación al hecho de que cada partícula del universo tiene una carga eléctrica que es un múltiplo entero de un valor dado (como se observa en realidad). La teoría que afirma que tal monopolo existe *en alguna parte* es típicamente apopperiana. Dicha teoría podría ser establecida mediante el descubrimiento de tal partícula, pero pare-

ce que no es *refutable*, como requeriría el criterio de Popper; pues si la teoría es falsa, ¡por mucho que los experimentadores busquen en vano, su incapacidad para encontrar un monopolo no refutaría la teoría![15] Pese a todo, se trata de una teoría científica, perfectamente digna de una consideración seria.

Podría hacerse un comentario similar en relación con la cosmología. La región del universo que se encuentra fuera de nuestro horizonte de partículas (§27.12) está más allá de la observación directa. Pese a todo, parece una propuesta científica razonable la de que dicha región debería asemejarse, en una amplia escala, a la región que es accesible a la observación directa. La teoría de que la región inobservable se parece a la región observacionalmente accesible, que, de hecho, es parte del modelo estándar de la cosmología (§27.11), aunque no de la mayoría de los esquemas inflacionarios (§28.4), no es refutable observacionalmente.

Más aún, incluso si restringimos la atención a la parte directamente observable del universo, podemos preguntar si la geometría espacial, supuesta homogénea e isótropa a gran escala, tiene curvatura positiva, negativa o nula (los casos respectivos $K > 0$, $K < 0$ o $K = 0$; véase §27.11). Si nuestra teoría afirma que $K = 0$, entonces esto tiene el carácter de ser refutable observacionalmente, porque, para cualquier desviación finita de la planitud espacial, una observación suficientemente precisa podría discernir (en principio, aunque quizá no en la práctica) esta separación de la planitud, por muy pequeña que pudiera ser esa curvatura. Pero si nuestra teoría afirma que $K \neq 0$, y la realidad fuera $K = 0$, entonces dicha teoría no podría ser refutada porque siempre habría algún margen de incertidumbre en las observaciones que no descartaría una muy ligera curvatura espacial positiva o negativa. Notemos que el caso $K > 0$ podría en principio ser refutado si en realidad $K < 0$, y $K < 0$ podría ser refutado si en realidad $K > 0$. Por el contrario, $K = 0$ no puede ser *confirmado* (directamente),[16] mientras que $K \neq 0$ podría ser confirmado observacionalmente (si el universo resultase ser de esta forma). Por ello, las dos afirmaciones $K > 0$ y $K < 0$ son popperianas en el sentido restringido de que son refutables en ciertas circunstancias —aunque no son susceptibles de ser refutadas si, en realidad, $K = 0$— y también son confirmables individualmente. ¡Notemos

que $K = 0$ es completamente popperiana, en principio, pero no confirmable!

A la vista de estas diversas posibilidades, no estoy seguro de dónde nos deja la perspectiva popperiana. Me parece evidente que cada una de las teorías $K > 0$, $K < 0$ o $K = 0$ es igualmente «científica» como afirmación, a pesar de estas sutiles diferencias con respecto al criterio de Popper. Y, en cualquier caso, la mayoría de los cosmólogos no adoptarían la línea pedante que he adoptado aquí, a saber, que «$K = 0$» significa realmente que esto se va a cumplir *exactamente*. De todas formas, una teoría *correcta* estaría en mejores condiciones si resulta que predice $K > 0$ o $K < 0$, puesto que entonces tiene la oportunidad de ser confirmada observacionalmente (y la confirmación es lo que busca una teoría científica, pese a la perspectiva más negativa de Popper sobre la cuestión de la aceptabilidad científica). Una teoría que predice $K = 0$ tendría que depender de otras justificaciones para ganar aceptación.

Una justificación semejante podría ser que $K = 0$ sea una consecuencia de una teoría concreta que encuentra confirmación observacional de alguna otra manera. De hecho, esta es la situación afirmada para la cosmología inflacionaria, ahora tan de moda, que se ha discutido en §28.4. Recordemos que, igual que la supersimetría es «casi» parte del modelo estándar de la física de partículas, ¡la cosmología inflacionaria es con frecuencia *casi* considerada una parte del modelo «estándar» de la cosmología! Tratemos de examinar el estatus de la inflación con respecto a la cuestión de la refutabilidad de Popper.

Podría pensarse que la situación es clara, y que la inflación es, de hecho, una teoría popperiana. Durante una década se había afirmado sistemáticamente que una de las implicaciones de la idea de inflación es que $K = 0$,[17] y recuerdo haber asistido a numerosas conferencias, dadas por defensores de la inflación, en las que se hacían tales predicciones.[18] Así, si las observaciones nos dicen concluyentemente que $K \neq 0$, ¡entonces la inflación está descartada! Esto parece una adhesión al principio de Popper tan clara como uno pudiera pedir. Más aún, existen muchas predicciones detalladas acerca del fondo de microondas que proceden de la inflación (junto con algunas otras hipótesis), y estas parecen disfrutar de algún apoyo experimental, de un modo gene-

ral, y en particular con respecto a la observada *invariancia de escala* de las fluctuaciones, en acuerdo con la mayoría de las predicciones inflacionarias. Sin embargo, a mediados de la década de 1990 empezó a acumularse evidencias, procedentes de varias observaciones de diferente tipo, de que la densidad promedio de materia Ω_d del universo (barión ica más oscura) está muy lejos de la que sería necesaria para una planitud espacial global, pues es como máximo de aproximadamente un tercio de dicho valor. (Las densidades Ω_d y Ω_Λ se toman como una fracción de la *densidad crítica*, que es la densidad que hubiera dado $K = 0$ en la teoría de Einstein sin término cosmológico; véase §28.10.) En concreto, Ω_d es aproximadamente 0,3. De acuerdo con esta tendencia en las observaciones, los teóricos de la inflación empezaron a ofrecer modelos inflacionarios que ahora permitían $K \neq 0$, y en concreto $K < 0$.[19] Por si fuera poco, podemos señalar que la escuela de Hawking, en la que parecía que $K > 0$ era una predicción definida (en relación con la propuesta de «ausencia de frontera» de Hartle-Hawking; véase §28.9), también empezó a concebir modos de acomodar $K < 0$ dentro de su esquema.[20]

Esta situación duró hasta 1998, cuando observaciones de supernovas lejanas (§28.10) parecían decirnos que es necesario incorporar una constante cosmológica positiva en la ecuación de Einstein, i.e., $\Lambda > 0$. Esto proporciona una densidad efectiva adicional Ω_Λ que, cuando se suma a la densidad de materia Ω_d, podría proporcionar el total crítico percibido $\Omega_d + \Omega_\Lambda = 1$ (o, en su lugar, $\Omega_d + \Omega_\Lambda > 1$, que sería necesario para la propuesta original de Hartle-Hawking). De esta forma, la planitud espacial global ($K = 0$) puede ser compatible con la observación (como podría serlo la curvatura espacial global positiva), con $\Omega_\Lambda \approx 0{,}7$. En vista de ello, la mayoría de los inflacionistas parecen haber vuelto a que $K = 0$ es una predicción de la cosmología inflacionaria. ¡No estoy seguro de lo que hubiera dicho Popper sobre todo esto!

De hecho, ahora hay una propuesta inflacionaria exótica en la que se introduce un nuevo ingrediente (un nuevo campo) para el universo, conocido como «quintaesencia», que proporcionaría una constante cosmológica *efectiva* a través de una «energía oscura» dinámica de presión negativa. Véase Steinhardt *et al.* (1999). Se ha argumentado que

esto podría indicar que se nos viene una nueva fase de inflación (véase §28.10). Cabe esperar que sugerencias que suenan tan fantásticas como estas encontrarán modos de ser concluyentemente establecidas observacionalmente, aunque, en la práctica, las cosas pocas veces parecen tan claras.

En mi opinión, debemos ser extraordinariamente cautos respecto a este tipo de afirmaciones, incluso si están aparentemente apoyadas por resultados experimentales de gran calidad. Estos son a menudo analizados desde la perspectiva de alguna teoría científica de moda. Por ejemplo, las soberbias observaciones de BOOMERANG[21] del fondo cósmico de microondas fueron originalmente interpretadas desde la perspectiva inflacionaria, y se ha proclamado con firmeza que las observaciones muestran realmente que $K = 0$ (y, por consiguiente, $\Lambda = 0$). Más aún, en el caso de algunos experimentos (como en el de BOOMERANG), en los que se ha recogido una gran cantidad de datos y hay mucho lugar para diferentes tipos de análisis, es posible que los datos en bruto no se distribuyan libremente durante un período de varios años para que las personas involucradas puedan tener (muy razonablemente) «prioridad» en ello. Durante el período intermedio, hay poco lugar para que los datos sean analizados desde otro punto de vista. De hecho, en el caso de BOOMERANG, Vahe Gurzadyan, con algunos miembros del equipo, pudieron acceder realmente a los datos y aplicar su análisis de elipticidad (§28.10), y encontraron un importante indicio directo de que $K < 0$ (más tarde apoyado por su correspondiente análisis de los datos de WMAP). Como sucede con la «anómala» medida WMAP $\ell = 2$ (extrañamente oculta por el eje vertical en la Fig. 28.19), esto no es demasiado favorable para la posición inflacionaria. ¡Tendremos que esperar hasta que se asiente el polvo por completo antes de llegar a una conclusión clara sobre todo esto!

Vemos así con qué fuerza pueden influir las cuestiones de moda científica en las direcciones de la investigación teórica en ciencia, pese a la tradicional insistencia de los científicos en la objetividad de su disciplina. De todas formas, debería dejar absolutamente claro que la aparente falta de objetividad no es culpa de la propia naturaleza. Existe un mundo físico objetivo ahí fuera, y los físicos consideran correctamente que su trabajo es descubrir su naturaleza y entender su comporta-

miento. La aparente subjetividad que vemos en las acusadas influencias de la moda mencionadas antes son solo características de nuestros escarceos en esta comprensión, donde las presiones sociales, las presiones de financiación y (comprensiblemente) las debilidades y limitaciones humanas desempeñan papeles importantes en las algo caóticas y a menudo mutuamente incompatibles imágenes que se nos presentan en la actualidad.

34.5. ¿Dónde podemos esperar nuestra próxima revolución en física?

Creo que en mis descripciones en este capítulo quizá he presentado una imagen del progreso actual hacia una comprensión fundamental de la física bastante más pesimista que la que se suele encontrar en las exposiciones divulgativas. Pero creo también que es una imagen mucho más realista. Por otra parte, no quiero sugerir que hayamos alcanzado una etapa en la que sea prácticamente imposible un progreso fundamental, como algunos divulgadores han tratado de mantener.[22] Existe una enorme cantidad de datos observacionales a los que todavía hay que dar sentido, y esto sería así incluso si no se hicieran más experimentos.

Los datos resultantes de los experimentos modernos se suelen almacenar de forma automática, y solo un aspecto particular de la información almacenada puede ser de interés para los teóricos y los experimentadores que están directamente involucrados. Sería así probable que el conjunto de datos fuera analizado solo en el aspecto concreto que aborda las cuestiones en las que aquellos están interesados. Es ciertamente posible que en tales datos haya ocultas muchas claves sobre la naturaleza, incluso si todavía no las hemos leído de manera adecuada. Recordemos que la relatividad general de Einstein estaba basada fundamentalmente en una idea (el principio de equivalencia; véase §17.4) que había estado implícita en datos observacionales que existían desde (e incluso antes de) la época de Galileo, aunque no del todo apreciados. Muy bien puede haber otras claves ocultas en las inconmensurablemente más extensas observaciones modernas. Quizá haya incluso

claves «obvias» ante nuestros propios ojos que necesitan que se les dé la vuelta para contemplarlas desde una perspectiva diferente, de modo que pueda obtenerse una perspectiva fundamentalmente nueva respecto a la naturaleza de la realidad física.

De hecho, creo que es necesaria una nueva perspectiva de esta naturaleza, y que este cambio de punto de vista tendrá que abordar las cuestiones profundas planteadas por la paradoja de la medida en mecánica cuántica y la no localidad relacionada que es inherente a los efectos EPR y la cuestión del «entrelazamiento» (capítulos 23 y 29). He discutido en el capítulo 30 que la paradoja de la medida debe estar profundamente interconectada con los principios de la relatividad general (y, específicamente, con el principio de equivalencia de Galileo-Einstein que se ha mencionado). Quizá nuevos experimentos (tales como el de FELIX, o una alternativa más realista basada en tierra; §30.13) puedan señalar el camino hacia una mejor comprensión de la teoría cuántica. Quizá habrá otros tipos de experimentos que arrojen luz sobre la naturaleza de la gravedad cuántica (tales como los diseñados para poner a prueba la posibilidad de una dimensionalidad más alta para el espaciotiempo). Tal vez, por el contrario, serán las consideraciones teóricas las que nos hagan avanzar.

¿Van a encontrarse las semillas de tales supuestos desarrollos teóricos en las ideas descritas en los capítulos previos de este libro? Evidentemente habría numerosos puntos de vista diferentes sobre esta cuestión, y la opinión personal debe desempeñar un papel importante en cualquier respuesta. Mi esperanza (durante más de cuarenta años) ha sido —y lo sigue siendo— que la herramienta de la teoría de twistores pueda aportar ideas que puedan conducir a un cambio semejante en el punto de vista físico. Pero a pesar de los avances que se han hecho (véase el capítulo 33), no puede decirse que la teoría de twistores nos haya llevado significativamente, tal como están las cosas, en alguna dirección que nos ayude a resolver la paradoja de la medida.

Cualquiera que pueda ser la posición propia concerniente a los méritos relativos de las teorías que he descrito, se necesitan decididamente nuevas ideas y nuevas perspectivas. ¿Cómo van a llegar estas? ¿Podemos esperar un «nuevo Einstein» trabajando en solitario que dé con ideas tan revolucionarias a partir fundamentalmente de elucubra-

ciones interiores? ¿O nos veremos conducidos de nuevo por hallazgos experimentales enormemente enigmáticos? En el caso de Albert Einstein, sus intuiciones internas le llevaron finalmente a la relatividad general, que es en gran medida la «teoría de una persona» (pese a la aportación esencial que Einstein recibió de Lorentz, Poincaré, Mach, Minkowski, Grossmann y otros). La teoría cuántica, por el contrario, fue en gran medida la «teoría de muchas personas», pues estaba impulsada externamente por los extraordinarios resultados de cuidadosos experimentos. En el clima actual de la investigación fundamental, parece que resulta mucho más difícil que *individuos* aislados hagan progresos sustanciales de lo que lo era en la época de Einstein. El trabajo en equipo, los cálculos de computador en masa, el seguimiento de ideas en boga... Estas son las actividades que solemos ver en la investigación actual. ¿Podemos confiar en que las nuevas perspectivas fundamentalmente necesarias salgan de tales actividades? Queda por verlo, pero tengo mis dudas sobre ello. Quizá si las nuevas direcciones tuvieran un impulso más experimental, como fue el caso de la mecánica cuántica en el primer tercio del siglo XX, podría funcionar el enfoque de «muchas personas». Pero veo que en el área de la gravedad cuántica esto solo puede suceder si existen experimentos que revelen una influencia de los principios de la relatividad general en la propia estructura de la mecánica cuántica (como he planteado en el capítulo 30). A falta de esto, tengo la sensación de que se necesitará algo más parecido a la aproximación einsteniana de «una persona». Y para eso hay pocas dudas, en mi opinión, de que la estética matemática debe ser una fuerza impulsora importante además de la intuición física.

La razón para esta creencia es que cuanto más profundamente sondeamos los fundamentos del comportamiento físico, más encontramos que está controlado de forma precisa por las matemáticas. Más aún, las matemáticas que encontramos no son solo del tipo que lleva a un cálculo directo; son de un carácter profundamente sofisticado, donde hay una belleza y una sutileza que no se ven en otras matemáticas que son relevantes en un nivel menos fundamental. De acuerdo con esto, el progreso hacia una comprensión física más profunda, cuando no puede ser guiado en detalle por el experimento, debe basarse cada vez con más fuerza en una capacidad para apreciar la relevancia física y la profundidad de las matemáticas, y «olfatear» las ideas apropiadas mediante el uso de una apreciación matemática estética y profundamente sensible.

Por la propia naturaleza del problema, es extraordinariamente difícil establecer cualquier tipo de criterio fiable para conseguir esto. En el contraste entre los enfoques que se han descrito en los últimos capítulos de este libro, ya hemos visto cuántos y qué diferentes desarrollos matemáticos, cada uno de ellos guiado por su propio conjunto de criterios estéticos físicos y matemáticos, pueden desarrollarse en direcciones mutuamente contradictorias. Algunos han argumentado que quizá deberíamos buscar la forma de reunir *todos* estos enfoques en algún tipo de síntesis, quizá destilando lo que es apropiado del conjunto de todos ellos. Por el contrario, podría argumentarse razonablemente que las contradicciones entre los diferentes enfoques son demasiado grandes, y que a lo sumo puede sobrevivir uno de ellos, teniendo que descartarse todos los demás. Particularmente, sospecho que la verdad está en algún lugar entre estos extremos, y que todavía puede hallarse algo de importancia incluso en muchas de las teorías cuyas ideas principales tendrán que ser abandonadas al final.

Algunas de las teorías que he descrito, aunque no completamente compatibles, tienen una base común apreciable. En particular, la aproximación de variables de lazo del capítulo 32 tiene características importantes en común con la teoría de twistores (capítulo 33) y puedo imaginar perfectamente que una combinación adecuada de las ideas de cada una de ellas (que quizá incluya redes de espín, espumas de espín, teoría de n-categorías, o incluso geometría no conmutativa) pudiera llevar a un avance. Pero la teoría de cuerdas, tal como está hoy día, con su dependencia de dimensiones espaciales extra, está para mí demasiado lejos de la teoría de twistores o de variables de lazo para que surja una unión previsible. Las cuerdas propiamente dichas no son una razón para la incompatibilidad (§31.5). Incluso la supersimetría ha sido unida a ideas twistoriales.[23] Pero la insistencia de la teoría de cuerdas en dimensiones más altas (especialmente en aquellas dimensiones/signaturas concretas que violan la filosofía holomorfa de la teoría de twistores; véase el párrafo final de §33.4) representa un conflicto fundamental tanto con la teoría de twistores como con la de las variables de lazo. Hasta muy recientemente los teóricos de cuerdas no han mostrado

ninguna inclinación a ofrecer una teoría (1 + 3)-dimensional consistente. Sin embargo, como se ha mencionado en §31.18 y §33.14, se ha producido un cambio reciente, y ahora parecen tomarse en serio aplicaciones de ideas de la teoría de cuerdas a un espaciotiempo (1 + 3)-dimensional ordinario.

34.6. ¿Qué es la realidad?

Como el lector habrá podido deducir, no creo que hayamos encontrado todavía el verdadero «camino hacia la realidad», pese a los extraordinarios progresos que se han hecho durante dos mil quinientos años, sobre todo en los últimos siglos. Ciertamente, se necesitan ideas fundamentalmente nuevas. Pese a todo, algunos lectores podrían muy bien adoptar el punto de vista de que el propio camino puede ser un espejismo. Es cierto, podrían argumentar, que hemos tenido la suerte de dar con esquemas matemáticos que presentan una notable coincidencia con la naturaleza, pero la unidad de la naturaleza en su conjunto con un esquema matemático puede no ser más que una «quimera». Otros podrían adoptar el punto de vista de que la propia idea de una «realidad física» verdaderamente objetiva, con independencia de cómo decidamos mirarla, es en sí misma una quimera.

De hecho, podemos muy bien preguntar:¿qué *es* la realidad física? Esta pregunta se ha planteado durante miles de años, y a lo largo de los tiempos los filósofos han intentado todo tipo de respuestas. Hoy miramos en retrospectiva desde nuestro punto de vista de la ciencia moderna y decimos adoptar una posición más moderada. Más que intentar responder a la pregunta «¿qué?», la mayoría de los científicos modernos tratarán de evitarla. Intentarán argumentar que la pregunta ha sido mal planteada: no deberíamos preguntar *qué* es la realidad, sino meramente *cómo* se comporta. «¿Cómo?» es, de hecho, una pregunta fundamental que podemos considerar que ha sido uno de los intereses principales de este libro. ¿Cómo describimos las leyes que rigen nuestro universo y sus contenidos?

Pese a todo, muchos lectores sentirán sin duda que esta respuesta es algo decepcionante: un perfecto «escaqueo». Saber cómo se comportan

los contenidos del universo no parece decirnos mucho acerca de *qué* es lo que se está comportando así. Esta pregunta «¿qué?» está íntimamente relacionada con otra pregunta antigua y profunda, a saber, «¿por qué?». ¿Por qué los objetos de nuestro universo se comportan de la forma particular que lo hacen? Pero sin saber *qué* son estos objetos, es difícil ver *por qué* deberían hacer una cosa antes que otra.

La ciencia moderna debería ser prudente al intentar dar respuestas a preguntas «¿por qué?» tanto como a preguntas «¿qué?». Pese a todo, preguntas como «¿qué?» y «¿por qué?» reciben frecuentemente respuestas. Se cree que está bien hacerlo así siempre que las preguntas no estén preguntando algo sobre la realidad en sus niveles más profundos. Se puede esperar una respuesta a preguntas como las siguientes: «¿de qué está compuesta una molécula de colesterol?», «¿por qué se enciende una cerilla cuando se frota rápidamente sobre una superficie rugosa apropiada?», «¿qué es una aurora?», «¿por qué brilla el Sol?», «¿cuáles son las fuerzas que mantienen unidos a un átomo de hidrógeno o a una molécula de hidrógeno?», «¿por qué es inestable el núcleo de uranio?». Pero otras preguntas posibles podrían causar más embarazo, tales como: «¿qué es un electrón?» o «¿por qué el espacio tiene solo tres dimensiones?». Sin embargo, estas preguntas pueden encontrar significado dentro de alguna imagen más fundamental de la realidad física.

Se verá, sobre todo por las discusiones de los capítulos 31-33, que los físicos modernos describen las cosas invariablemente en términos de modelos matemáticos. Esto es independiente de la familia concreta de propuestas que puedan sostener. Es como si trataran de encontrar la «realidad» dentro del mundo platónico de las ideas matemáticas. Semejante visión parecería ser una consecuencia de alguna «teoría de todo» propuesta, pues entonces la realidad física aparecería como un reflejo de leyes puramente matemáticas. Como he argumentado en este capítulo, sin duda estamos a mucha distancia de una teoría semejante, y es materia de controversia si se encontrará alguna vez algo parecido a una «teoría de todo». Sea como fuere, se da el caso de que cuanto más profundamente sondeemos los secretos de la naturaleza, más profundamente nos vemos dirigidos hacia el mundo platónico de las ideas matemáticas a medida que buscamos el conocimiento. ¿Por qué es así? Por el momento, solo podemos verlo como un misterio. Es el primero de

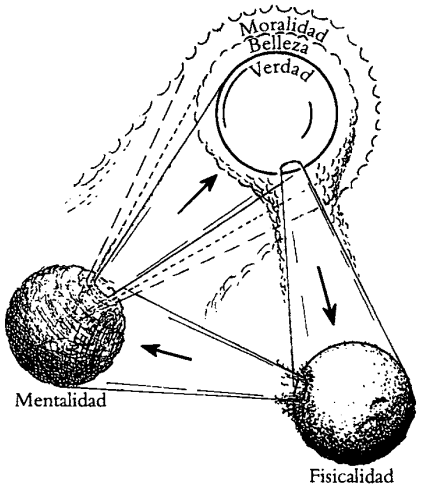


Fig. 34.1. Una repetición del diagrama (Fig. 1.3) que muestra «tres mundos y tres misterios», pero embellecido con los otros «absolutos platónicos» de belleza y moralidad, además de la verdad absoluta que debe encontrarse en las matemáticas. Belleza y verdad están entretejidas, la belleza de una teoría física actúa como guía para su corrección en relación con el mundo físico, mientras que la cuestión general de la moralidad depende en última instancia del mundo de la mentalidad.

los tres profundos misterios mencionados en §1.4 e ilustrados en la Fig. 1.3, y aquí vueltos a dibujar y algo adornados como en la Fig. 34.1.

Pero ¿son las nociones matemáticas cosas que habitan realmente en un «mundo» propio? Si es así, parece que hemos encontrado que nuestra realidad última tiene su hogar dentro de este mundo completamente abstracto. Algunas personas tienen dificultades para aceptar que el mundo matemático de Platón es en cierto sentido «real», y no obtendrán ningún consuelo de una visión en la que la propia realidad física está construida a partir de nociones abstractas. Mi postura en esta cuestión es que ciertamente deberíamos considerar que el mundo de Platón proporciona un tipo de «realidad» para las nociones matemáticas (y he tratado de argumentar con fuerza esta postura en §1.3), pero me resistiría a intentar *identificar* la realidad física dentro de la realidad abstracta del mundo de Platón. Creo que mi posición sobre esta cuestión está bien expresada en la Fig. 1.3, donde cada uno de los tres mundos —matemático-platónico, físico y mental— tiene su propio tipo de realidad, y donde cada uno está basado (profundamente y misteriosamente) en el que le precede (tomando los mundos cíclicamente). Me gusta pensar que, en cierto sentido, quizá sea el mundo platónico el más fundamental de los tres, puesto que las matemáticas son una necesidad que prácticamente se conjura a sí misma mediante la sola lógica.

Sea como fuere, existe el misterio, o paradoja, adicional del aspecto cíclico de estos mundos, donde cada uno de ellos parece capaz de englobar en su totalidad al que le sigue, aunque en sí mismo parezca depender solo de una pequeña parte de su predecesor.

34.7. Los papeles de la mentalidad en la teoría física

Debemos tener en cuenta que cada «mundo» posee su propio tipo de existencia, diferente del de los otros dos. En cualquier caso, no creo que, en última instancia, seamos capaces de considerar cualquiera de estos «mundos» por sí mismo, separado de los otros dos. Puesto que uno de estos es el mundo de la mentalidad, esto plantea la cuestión del papel de la mente en la teoría física, y también de cómo interviene la mentalidad en las estructuras físicas con las que está relacionada (tales como, al menos, los cerebros humanos vivos, despiertos y sanos). Me he abstenido de manera deliberada de abordar extensamente en este libro la cuestión de la mentalidad consciente, pese al hecho de que esta cuestión debe ser en definitiva importante en nuestro intento de entender la realidad física. (He discutido en detalle estas cuestiones en otra obra, y aquí no plantearé muchos de los puntos controvertidos que aparecen.)[24] Pero no sería apropiado que tratara de evitar por completo la cuestión de la mentalidad. Aparte de que el mundo de la mentalidad deba considerarse en unión de los otros dos mundos, de acuerdo con la Fig. 34.1, hay varios apartados en este libro donde la cuestión de la consciencia ha desempeñado ya un papel importante en la teoría física, ya sea implícita o explícitamente.

Uno de estos está relacionado con el principio antrópico, mencionado en §27.3 y discutido con cierto detalle en §§28.6,7. Cualquier universo que pueda «ser observado» debe ser capaz, como necesidad lógica, de soportar la mentalidad consciente, puesto que la consciencia es precisamente la que desempeña el papel final como «observador». Este requisito fundamental podría muy bien proporcionar ligaduras sobre las leyes físicas o los parámetros físicos del universo, para que pueda existir (y exista) mentalidad consciente. Por consiguiente, el principio antrópico afirma que el universo que nosotros, como observadores

conscientes, realmente observamos debe operar con leyes y valores apropiados de los parámetros que sean compatibles con dichas ligaduras. Tales ligaduras podrían manifestarse en valores concretos para las constantes (adimensionales) fundamentales de la naturaleza, discutidas en §31.1. De hecho, se ha convertido en un tópico el considerar que los valores que en realidad encontramos son el resultado de algún tipo de aplicación del principio antrópico.

Por desgracia, incluso si los valores de estas constantes están determinados por el principio antrópico —antes que, digamos, por consideraciones matemáticas—, el principio es casi inutilizable, pues sabemos muy poco sobre las condiciones que son necesarias para que exista la consciencia y realmente se produzca. Es casi del todo inutilizable en un universo espacialmente infinito y esencialmente uniforme ($K \leq 0$), porque en un universo semejante cualquier configuración de materia que pudiera ocurrir por azar, ocurriría en algún lugar, de modo que incluso condiciones muy desfavorables para la vida consciente estarían permitidas por el principio; véase §28.6. En mi opinión, una posibilidad mucho más optimista es que estas constantes fundamentales sean en realidad números matemáticamente determinables. En un universo espacialmente infinito, esta necesidad no plantea problemas importantes con el principio antrópico.

Hay un papel importante e independiente que desempeña la consciencia en muchas interpretaciones de la parte **R** de la mecánica cuántica, como se ha expuesto en el capítulo 29 (en particular, en §§29.7,8). De hecho, casi todas las interpretaciones «convencionales» dependen en definitiva de la presencia de un «ser perceptivo», y por ello parecen exigir que sepamos qué es realmente un ser perceptivo. Recordemos que la interpretación de Copenhague (punto de vista (a) en §29.1) considera que la función de onda no es una entidad física objetivamente real, sino que, en efecto, es algo cuya existencia está «en la mente del observador». Además, al menos en una de sus manifestaciones, esta interpretación requiere que una medida sea una «observación», lo que presumiblemente significa algo finalmente observado por un ser consciente, aunque, en un nivel de aplicabilidad más práctico, la medida es algo realizado por un aparato de medida «clásico». No obstante, esta dependencia de un aparato clásico es solo una «medida provisional», puesto que cualquier aparato real sigue estando hecho de componentes cuánticos, y en realidad no se comportaría de forma clásica —ni siquiera aproximadamente— si se atuviera a la evolución **U** cuántica estándar. (Esta es simplemente la cuestión del gato de Schrödinger; véanse §§29.7-9 y §§30.10-13.) La cuestión de la decoherencia por el entorno (punto de vista (c) en §29.1) también nos ofrece una posición provisional, puesto que la inaccesibilidad de la información «perdida en el entorno» no significa que esté perdida *realmente*, en un sentido objetivo. Pero para que la pérdida sea subjetiva, nos vemos abocados de nuevo a la cuestión de lo «subjetivamente percibido... ¿por quién?» que nos devuelve a la cuestión del observador consciente.

En cualquier caso, incluso con la decoherencia por el entorno, si mantenemos la adhesión rigurosa a la evolución **U** para la «verdadera» descripción cuántica del universo, entonces nos vemos dirigidos a la descripción de los muchos universos para la realidad (punto de vista (b) en §29.1). El punto de vista de los muchos universos es manifiestamente dependiente de tener una comprensión adecuada de lo que constituye un «observador consciente», puesto que cada «realidad» percibida está asociada con un «estado de observador», de modo que no sabemos qué estados de realidad (i.e., «mundos») están permitidos hasta que sepamos qué estados de observador están permitidos. Dicho de otra forma, el comportamiento del mundo aparentemente objetivo que es realmente percibido depende del camino que siga la consciencia de uno a través de las miríadas de alternativas cuánticamente superpuestas. En ausencia de una teoría adecuada de los observadores conscientes, la interpretación de los muchos universos debe necesariamente seguir estando fundamentalmente incompleta (véase §29.8).[25]

El enfoque de las historias consistentes (punto de vista (d) en §29.2) también depende explícitamente de cierta noción de lo que pudiera ser un «observador» (la noción mencionada como un IGUS en el esquema de Gell-Mann/Hartle).[26] El punto de vista sugerido por Wigner (una versión del punto de vista (f) en §29.2), según el cual la consciencia (o quizá los sistemas vivos en general) podría *violar* la evolución **U** también es otro punto de vista que hace referencia explícitamente al papel de la mente (o de lo que quiera que sea lo que constituye un «observador») en la interpretación de la mecánica cuántica. Por lo que

puedo deducir, las únicas interpretaciones que *no* dependen necesariamente de una noción de «observador consciente» son las de De Broglie-Bohm (punto de vista (e) en §29.2)[27] y la mayoría de aquellas (puntos de vista (f) en §29.2) que requieren algún cambio fundamental en las reglas de la mecánica cuántica, según las cuales tanto **U** como **R** deben tomarse como aproximaciones a cierta evolución física objetivamente real.

Como he afirmado en muchos apartados de este libro (en especial, en el capítulo 30), me adhiero a este último punto de vista, según el cual es en los fenómenos gravitatorios donde una **R** objetiva (i.e., **RO**) domina sobre **U**. Esta **RO** gravitatoria tendría lugar espontáneamente, y no requiere que un observador consciente forme parte del proceso. En circunstancias normales, habría frecuentes manifestaciones **RO** que ocurren continuamente, y estas conducirían a que a gran escala emerja un mundo clásico como excelente aproximación. Por consiguiente, no hay necesidad de invocar ningún observador consciente para conseguir la reducción del estado cuántico (**R**) cuando tiene lugar una medida.

Por otra parte, imagino que el fenómeno de la consciencia —que considero que es un proceso físico *real*, que aparece «ahí fuera» en el mundo físico— hace uso fundamentalmente del proceso **RO** real. Así pues, mi posición es básicamente la inversa de las antes mencionadas, en las que, de un modo u otro, se imagina que la consciencia es responsable del proceso **R**. En mi opinión, es un proceso **R** físicamente real el que es (parcialmente) responsable de la propia consciencia.[28]

¿Es esta afirmación verificable experimentalmente? Creo que lo es. En primer lugar, hay sugerencias concretas acerca de estructuras en el cerebro que podrían ser relevantes —muy en particular, los microtúbulos neuronales A-reticulares, como sugirió originalmente Stuart Hameroff[29] (pero quizá también otras estructuras, como las clatrinas sinápticas,[30] cuya estructura se asemeja mucho a las moléculas de C^{60})— y hay un espacio considerable para confirmar/refutar estas ideas concretas. El esquema requeriría un tipo de coherencia cuántica a gran escala que actuara en regiones considerables del cerebro (con características en común con la superconductividad a alta temperatura;[31] véase §28.1), y se supone que los microtúbulos neuronales A-reticulares desempeñarían un papel importante en ello. Un suceso consciente estaría asociado con una reducción parcial (**RO** orquestada) del estado de dicho sistema cuántico. Esto implicaría normalmente a muchas partes del cerebro juntas, para conseguir suficiente movimiento coherente en moléculas de proteína para producir la **RO** gravitatoria de acuerdo con la propuesta descrita en §§30.11,12.

Hay una ingeniosa sugerencia propuesta por Andrew Duggins según la cual estas conjeturas —aunque no específicas de la hipótesis de los microtúbulos— podrían ser puestas a prueba. Esta sugerencia depende del hecho de que regiones completamente diferentes del cerebro son responsables de diferentes aspectos de la percepción (tales como la percepción visual del movimiento, del color o de la forma), y, pese a todo, en la imagen que resulta en la consciencia, todos estos aspectos diferentes se unen para formar una única imagen. Esto se conoce a veces como el *problema de la encuadernación*. La idea de Duggins consiste en tratar de ver si hay violaciones significativas de las desigualdades de Bell implicadas en la formación de una imagen consciente, lo que indicaría la presencia de efectos tipo EPR no locales (§§23.3-5), que sugerirían con fuerza que efectos cuánticos a gran escala forman parte de la percepción consciente. Hasta ahora, los resultados preliminares no son concluyentes, aunque sí algo alentadores.[32]

Cualquiera que sea el estatus de estas ideas, creo que una teoría física «fundamental» que proclame cualquier tipo de compleción en los niveles más profundos de los fenómenos físicos debe tener también la capacidad de acomodar la mentalidad consciente. Algunas personas tratarían de evitar (o empequeñecer) este problema, argumentando que la consciencia «emerge» simplemente como una especie de «epifenómeno». Por consiguiente, se afirmaría, para la emergencia de la consciencia no importa el tipo exacto de física que pueda subyacer en los procesos físicos (no necesariamente biológicos) relevantes. Una postura estándar es la del *funcionalismo computacional*, según el cual es meramente la actividad computacional (de cierta naturaleza apropiada pero aún no especificada) la que da lugar a la mentalidad consciente. He argumentado con vehemencia contra esta idea (en parte utilizando razonamientos basados en el teorema de Gödel y la noción de computabilidad de Turing; véase §16.6), y he sugerido, de hecho, que la consciencia depen-

de realmente de la teoría **RO** (gravitatoria) de la que carecemos.[33] Mis argumentos exigen que esta teoría debe ser una teoría *no computacional* (i.e., sus acciones caen fuera del ámbito de la simulación por una máquina de Turing; §16.6). Las ideas teóricas para producir un modelo **RO** de este tipo están hoy día en una fase muy preliminar, pero quizá haya aquí algunas claves.

34.8. Nuestro largo camino matemático a la realidad

Espero que haya quedado claro de las discusiones ofrecidas en las secciones precedentes que nuestro camino hacia la comprensión de la naturaleza del mundo real está todavía muy lejos de su objetivo. Quizá este objetivo no se alcance nunca, o tal vez surja con el tiempo una teoría final en cuyos términos pueda entenderse, en principio, lo que llamamos «realidad». Si es así, la naturaleza de dicha teoría debe diferir enormemente de lo que hasta ahora hemos visto en las teorías físicas. La idea simple más importante que ha surgido de nuestro viaje, de unos tres mil quinientos años de duración, es que existe una profunda unidad[34] entre ciertas áreas de las matemáticas y el funcionamiento del mundo físico, y este es el «primer misterio» mostrado en las Fig. 1.3 y 34.1. En mi opinión, si el «camino a la realidad» alcanza por fin su objetivo, entonces tendría que haber una simplicidad profundamente subyacente en dicho punto final. Pero no veo esto en ninguna de las propuestas existentes.

Esta idea de los antiguos griegos de que son las *matemáticas* las que subyacen en el funcionamiento de la realidad física nos ha servido extraordinariamente bien, y espero haber dejado claro que, pese a la distancia que queda respecto a nuestro pretendido objetivo, hemos llegado a una impresionante comprensión de las operaciones del universo en los niveles más profundos que conocemos. Algunos conceptos matemáticos destacan por haber sido particularmente exitosos en el pasado. Entre ellos están el sistema de los números reales y las ideas de la geometría. Inicialmente fue la geometría euclídea, estudiada de manera sistemática por primera vez por los antiguos griegos, pero luego las ideas se desarrollaron alejándose de la geometría de Euclides para llegar a las de Lambert, Gauss, Lobachevski, Bolyai, Riemann, Beltrami y otros. Luego Minkowski nos habló de incorporar el tiempo al espacio, y Einstein nos presentó su magnífica geometría espaciotemporal curva de la relatividad general. El cálculo diferencial e integral de Arquímedes, Fermat, Newton, Leibniz, Euler, Cartan y muchos otros, y también las ideas relacionadas de las ecuaciones diferenciales, ecuaciones integrales y derivadas variacionales, se han mostrado absolutamente vitales para las teorías exitosas que describen el funcionamiento del mundo, uniéndose con la geometría de maneras que tienen un significado profundo. También han sido fundamentales las ideas estadísticas que nos permiten manejar grandes y complicados sistemas físicos con un número enorme de ingredientes individuales, tal como nos han enseñado Maxwell, Boltzmann, Gibbs, Einstein y otros. Las matemáticas subyacen profundamente en la teoría cuántica, desde las ideas de la teoría de matrices de Heisenberg hasta los espacios de Hilbert complejos, las álgebras de Clifford, la teoría de la representación, el análisis funcional en infinitas dimensiones, etc., de Dirac, Von Neumann, y muchos otros.

Quisiera destacar tan solo dos aspectos particulares de las matemáticas que subyacen en el funcionamiento del mundo, discutiendo cada uno de ellos por turno, pues creo que pueden llevar a cuestiones de principio importantes aunque no abordadas en nuestra teoría física. El primero es el papel del sistema de los *números complejos*, que encontramos tan fundamental para las operaciones de la mecánica cuántica, en contraste con el sistema de los números reales, que había proporcionado la base de todas las teorías previas de éxito. El segundo es el papel de la *simetría*, que tiene una importancia capital en prácticamente todas las teorías del siglo xx, en particular en relación con la formulación de las teorías gauge de las interacciones físicas.

En primer lugar, consideremos los números complejos. Ha sido un tema recurrente de este libro que no solo hay una magia especial en las matemáticas de estos números, sino que la propia naturaleza parece aprovecharse de ella cuando teje su universo en sus niveles más profundos. Pese a todo, podemos muy bien preguntarnos si esto es realmente una característica verdadera de nuestro mundo, o si es meramente la utilidad matemática de dichos números la que ha llevado a su uso extendido en la teoría física. Creo que muchos físicos se inclinarían

hacia la segunda opinión. Pero para ellos sigue habiendo algo de misterio —que necesita alguna explicación— en por qué el papel de dichos números debería aparecer de forma tan universal en el armazón de la teoría cuántica, puesto que subyacen en el fundamental principio de la superposición cuántica y, con un disfraz algo diferente, en la ecuación de Schrödinger, la condición de frecuencia positiva y la «estructura compleja» de dimensión infinita (§26.3) que interviene en la teoría cuántica de campos. Para estos físicos, los números reales parecen «naturales» y los números complejos «misteriosos». Pero desde una posición puramente matemática, no hay nada especialmente más «natural» en los números reales que en los números complejos. De hecho, en vista del estatus matemático algo mágico de los números complejos, se podría adoptar muy bien el punto de vista contrario y considerarlos característicamente más «naturales» o «infusos» que los reales.

Desde mi peculiar punto de vista, la importancia de los números complejos —o, más específicamente, la importancia de la holomorficidad (o analiticidad compleja)— en la base de la física debe verse como algo «natural», y quizá el enigma haya que verlo al revés.[35] ¿Cómo es posible que estructuras *reales* parezcan desempeñar un papel tan importante en física? Debería dejar claro que incluso el formalismo estándar de la mecánica cuántica, aunque basado en los números complejos, no es una teoría completamente holomorfa. Lo vemos en el requisito habitual de que los observables cuánticos sean descritos por operadores hermíticos (o incluso normales, como se ha descrito en §22.5) y en la naturaleza unitaria (y no simplemente lineal compleja) de la evolución cuántica, que dependen de la noción de conjugación compleja ($z \mapsto \bar{z}$). En relación con esto, la importante propiedad de ortogonalidad entre estados es una operación no holomorfa. La hermiticidad tiene que ver con la exigencia habitual (pero no necesaria) de que los resultados de las medidas sean números reales, y la unitariedad tiene que ver con que «la probabilidad se conserve», i.e., que la regla del módulo al cuadrado (que también tiene que ver con las medidas) se mantenga, por lo que una amplitud compleja z se convierte en una probabilidad, de acuerdo con la operación no holomorfa

$$z \mapsto z\bar{z}.$$

Vemos que es básicamente en la conversión de «información cuántica» (i.e., entrelazamiento; véase §23.10) en «información clásica» donde se rompe la holomorficidad cuántica. La ortogonalidad de alternativas es de nuevo una característica crucial de la medida. Así pues, la no holomorficidad parece intervenir solo en el momento en que se introducen las medidas en la teoría cuántica.

Por supuesto, también vemos el papel de los números reales en el espaciotiempo de fondo dentro del cual se sitúa el formalismo de la teoría cuántica. Si la **RO** gravitatoria resulta ser la verdadera base de la reducción del estado cuántico, entonces veremos que la estructura (no holomorfa) de números reales del espaciotiempo real se relaciona con la de la operación $z \mapsto z\bar{z}$. ¿Hay aquí quizá una lección para la teoría de twistores, dada la dependencia particular de la teoría respecto de las operaciones holomorfas? ¿Quizá, por el contrario, deberíamos buscar un papel para principios combinatorios discretos que emergen de alguna manera a partir de la magia compleja, de modo que el «espaciotiempo» debería tener una estructura discreta subyacente en lugar de una basada en números reales (como se ha examinado en §§3.3,5, §32.6 y §33.1)? En cualquier caso, creo que hay aquí cuestiones importantes y profundas concernientes a la misma base matemática de la realidad física.

Vayamos ahora al papel fundamental de la *simetría* en la teoría física moderna. No hay duda de la utilidad de esta noción. Tanto la teoría de la relatividad (en relación con el grupo de Lorentz) como la teoría cuántica hacen un uso exhaustivo de ella. Pero ¿debemos considerar la simetría como algo fundamental para los modos de la naturaleza, o una característica accidental o aproximada?

Parece que un principio en muchas de las modernas aproximaciones a la física de partículas es tomar la simetría como algo realmente fundamental, y considerar las desviaciones de la simetría observadas en la actualidad como un resultado de la ruptura de simetría en el universo primitivo. De hecho, como se ha señalado en §13.1 y en §§15.2,4, la simetría exacta es una característica necesaria de la idea de conexión fibrada. Además, recordemos de §§25.5,8 que la actitud estándar hacia la teoría electrodébil es considerar la simetría SU(2) como fundamentalmente exacta, de modo que pueda desempeñar un papel como sime-

tría gauge de las fuerzas electrodébiles (§§15.1,8), pero donde normalmente se concibe que la simetría se rompe de manera espontánea (10^{-12} segundos después del big bang). Recordemos de §28.3 que existen ciertas dificultades en invocar el universo primitivo para proporcionar la ruptura de simetría necesaria. Esto se aplica a la simetría SU(2) de la teoría electrodébil y también a las simetrías mucho mayores que se utilizan en las teorías GUT.

¿Realmente simplifican los grupos grandes de simetría de las teorías GUT nuestra imagen de la física de partículas? ¿O sería esta más simple si muchas de estas simetrías aparentes estuvieran fundamentalmente rotas desde el principio? Desde esta segunda perspectiva alternativa (que es, de hecho, una perspectiva consistente incluso para la teoría electrodébil; véase §28.3), muchas de las simetrías que percibimos en nuestras teorías fundamentales serían realmente solo aproximadas en el nivel fundamental, y debemos buscar con más profundidad para entender de dónde proceden esas simetrías aparentes.

En la teoría cuántica ordinaria tenemos ejemplos de ambos tipos de simetrías rotas. Existen situaciones bien comprendidas en las que manifiestamente ocurre ruptura espontánea de simetría, tal como sucede en la superconductividad (ruptura U(1)) y otros fenómenos. Por otro lado, hay ejemplos en los que pueden utilizarse ideas de simetría para ofrecer una excelente comprensión de un fenómeno, pero donde se sabe que la simetría es solo una aproximación que surge de una teoría más profundamente subyacente que es más exacta pero menos simétrica, tal como sucede en la clasificación de los espectros atómicos.[36] Queda por ver cuál de estos dos tipos de situaciones tendrá mayor importancia en una teoría más profunda de la física de partículas.[37]

Como punto de interés relacionado, hay circunstancias en las que un grupo de simetría exacta puede intervenir incluso en estructuras en las que inicialmente no se impone ninguna simetría. Lo vemos con la propia esfera de Riemann, que podemos imaginar compuesta de cartas del plano complejo de una forma específica que carece de cualquier simetría. Pero con tal de que la topología de la variedad compleja resultante sea, de hecho, S^2, encontramos que es *equivalente* a la esfera de Riemann, en tanto que variedad compleja (por un teorema de Riemann), de modo que su grupo de simetría es exactamente SL(2, $\mathbb{C}$),

i.e., el grupo de Lorentz no reflexivo (§18.5), por muy irregularmente que esté compuesta.[38]

Hay una cuestión, en parte relacionada, concerniente a las misteriosas constantes de la naturaleza (§31.1). ¿Están estos números determinados en el universo extremadamente primitivo (tal como sucede en la propuesta de tipo Wheeler/Smolin mencionada en §28.6), en analogía con la idea de ruptura de simetría? De hecho, se suele considerar que algunas de estas constantes, como los ángulos de Cabbibo y Weinberg (§25.7), surgen de esta manera. ¿O en realidad podrían estos números determinarse matemáticamente a partir de alguna teoría más profunda subyacente? Personalmente, preferiría esto último, pero no parece que estemos cerca de tener una teoría creíble de este tipo.[39]

Una cuestión interesante en relación con esto es la simetría *quiral* de las interacciones débiles (§25.3). En la aproximación normal al modelo estándar, esta simetría quiral se incorpora en el armazón de la teoría. Pero ahora se ha observado que los neutrinos (al menos la mayoría de ellos) son partículas masivas (i.e., con masa en reposo no nula), y este hecho ya representa una desviación del modelo electrodébil estándar original. No podemos simplemente «culpar» a un neutrino levógiro de toda la asimetría quiral de las interacciones débiles. Un neutrino masivo no es completamente una partícula «zig» levógira, puesto que con masa tendría también una parte «zag» dextrógira (véase §25.2). Podríamos imaginar que en algunas versiones del modelo estándar (ligeramente ampliadas de modo que se incorporen neutrinos masivos) había una ruptura espontánea de simetría a partir de un modelo previo con simetría izquierda/derecha. Pero, en este caso, la «perspectiva» convencional es que la asimetría esta ahí desde el principio, en lugar de aparecer de una ruptura espontánea de simetría en el universo primitivo.

También podríamos contemplar si la asimetría temporal (exigida por la discusión de los capítulos 27 y 28) es un punto que debería ser reexaminado desde esta perspectiva. Sin embargo, ciertamente no puede aparecer de una convencional «ruptura espontánea de simetría en el universo primitivo». La imagen convencional utiliza la segunda ley de la termodinámica; no puede utilizarse para deducirla.

34.9. Belleza y milagros

Volvamos ahora a algunos aspectos más generales y misteriosos de las matemáticas que se ha encontrado que subyacen en la teoría física en sus niveles más profundos, al menos en un nivel tan profundo como se nos ha revelado hasta ahora. Dos poderosas fuerzas impulsoras internas han tenido una gran influencia en la dirección de la investigación teórica, pese a que normalmente pasan desapercibidas en escritos teóricos serios, por miedo, sin duda, a que pueda parecer que dichas influencias se hayan desviado demasiado de las reglas estrictas del método científico correcto. La primera de estas es la belleza, o elegancia, y ya he tocado el tema en muchos apartados de este libro. En cuanto a la segunda, a saber, el irresistible atractivo de lo que frecuentemente se califica de «milagros», tan solo la he sugerido (en §19.8, §21.5 y §31.14); pero, como puedo dar fe por mi experiencia personal, estos pueden ejercer realmente una poderosa influencia en la dirección de la propia investigación de cada uno.

Antes de entrar en la cuestión de los milagros, que son el interés principal de esta sección, volvamos primero a la cuestión de la belleza, puesto que las dos están relacionadas. Como se ha indicado antes, muchas de las ideas que se cree que han impulsado un avance importante en la teoría física también se consideran de una belleza cautivadora. Está la indudable belleza de la geometría euclídea, que formó la base de la primera teoría física profundamente precisa, a saber, la teoría del espacio formulada por los antiguos griegos. Mil quinientos años más tarde llegó la extraordinaria elegancia de la dinámica newtoniana, con su profunda y bella estructura subyacente de geometría simpléctica, como se reveló posteriormente por los formalismos lagrangiano y hamiltoniano (§20.4). La forma matemática del electromagnetismo de Maxwell también es exquisita, y no hay duda de la suprema belleza matemática de la relatividad general de Einstein. Lo mismo puede decirse de la estructura de la mecánica cuántica y muchas de sus características específicas. Yo destacaría la extraordinaria elegancia matemática del espín mecanocuántico, de la ecuación de ondas relativista de Dirac, y del formalismo de integrales de camino para la QFT, tal como lo desarrolló Feynman.

Pese a todo, podemos preguntarnos si la indudable belleza matemática de estos esquemas es algo que brillaría independientemente, tan solo como puras matemáticas, si no hubiera sido por el extraordinario hecho de que están en tan buen acuerdo con el funcionamiento de nuestro universo. Consideradas solo como estructuras matemáticas, ¿cómo soportarían la comparación con algunas de las gemas o faros de las matemáticas puras? Creo que la soportarían bastante bien, aunque no de forma abrumadora. Existen muchos dominios de las matemáticas puras, sin relaciones discernibles con el mundo físico, cuya belleza iguala o incluso supera a la de las teorías físicas por las que hemos pasado. (Véase también §16.3.)

Consideremos algunos bellos y profundos desarrollos en matemáticas, cuya influencia sobre la física —al menos hasta ahora— ha sido mínima. La teoría de Cantor del infinito es un ejemplo digno de mención. En mi opinión, es una de las contribuciones más profundamente bellas en el conjunto de la historia de las matemáticas. Sin embargo, parece que muy poco de ella tiene relevancia para el funcionamiento del mundo físico tal como lo conocemos (véanse §§16.3,4,7). La misma cuestión surge en relación con otro de los monumentales logros del conocimiento matemático, un descendiente muy próximo de la teoría de Cantor del infinito, a saber, el famoso teorema de incompletitud de Gödel (§16.6). Están también las interesantes y amplias ideas de la teoría de categorías (§33.1) en las que por el momento se ha visto muy poca relación con la física.

En estas dos últimas áreas hay algún indicio plausible de que podría haber una relación importante con una física que pudiera desarrollarse en el siglo xxi (§34.7 y §33.1), pero esto es muy especulativo. Parece mucho menos plausible que se revele alguna conexión importante con la física en el caso de la inmensa mayoría de las otras teorías matemáticas bellas y profundas que se han desarrollado. Consideremos, por ejemplo, el extraordinario logro de Andrew Wiles en el siglo xx al establecer la verdad de la afirmación de 350 años de antigüedad conocida como «último teorema de Fermat». Esto parece estar muy alejado de las leyes físicas, tal como las entendemos hoy, pese a las magníficas ideas matemáticas implicadas. Muchos otros maravillosos desarrollos matemáticos se produjeron en el siglo xx, tales como la clasificación de los grupos

simples, tanto continuos como discretos. Aquí ha habido ciertamente aplicaciones a la física, pero esto no quiere decir ni mucho menos que la teoría de los grupos simples nos proporcione una «teoría física». Se trata simplemente de que las clasificaciones matemáticas son útiles para los físicos, pues les permiten ver qué posibilidades hay. Consideremos otro ejemplo. El siglo XIX vio la exquisita teoría de Riemann de la función ζ y su relación con la distribución de los números primos. Esto parece tan igualmente alejado de la física como lo anterior, pese a la indudable belleza y gran importancia matemática de la aún no demostrada hipótesis de Riemann (§7.4). De hecho, *hay* aquí algunas relaciones intrigantes con la física,[40] pero sería difícil mantener que la teoría de Riemann nos ofrece algo parecido a un modelo del mundo físico.

¿Cabe esperar que haya una estrecha relación con un cuerpo mayor de bellas y profundas matemáticas en la física del futuro, o vamos a ser engañados por los éxitos que hemos visto hasta ahora en la teoría física para hacernos creer que la relación entre matemáticas y física es más estrecha de lo que realmente es? La pregunta puede parafrasearse de forma sucinta en términos de la Fig. 1.3. ¿Cuánto del mundo matemático de Platón yace en la base de la flecha que muestra el «primer misterio»?

También podemos preguntar si puede haber alguna forma de percibir qué tipo de matemáticas son las que encuentran un papel profundo en el gobierno del comportamiento del mundo físico. Esta es una pregunta intrigante. Quizá los factores cruciales subyacentes que gobiernan la relación misteriosa entre matemáticas y física serán mejor entendidos en alguna fecha futura.

Espero que consideraciones como las anteriores aclaren al lector que la belleza matemática es, en sí misma, una guía ambigua en el mejor de los casos. Pese a todo, como he comentado en numerosos apartados en este libro, es la difícil dudar del extraordinario papel que los juicios estéticos, tanto matemáticos como físicos, desempeñan en la toma de decisiones respecto a las líneas más fructíferas que se deben en la investigación en la física teórica. La mayoría de estas son de un carácter sutil, y resulta fácil creer que la valoración del atractivo de las distintas alternativas que se pueden seguir es una cuestión muy personal. En ocasiones, sin embargo, puede surgir algo en la investigación de

las teorías matemáticas del mundo físico que tenga un impacto mucho más poderoso sobre tales elecciones que la mera elegancia matemática, y esto es a lo que yo llamo «milagro».

Puedo pensar en muchos ejemplos de esto en la historia reciente de las ideas de la «gravedad cuántica». Uno de ellos se dio en la teoría de la supergravedad (§31.2), donde se observó que, mientras que el enfoque perturbativo de la QFT de la teoría estándar de la relatividad general de Einstein conducía a divergencias no renormalizables en segundo orden, los términos divergentes se cancelaban milagrosamente cuando se introducía la supersimetría.[41] Esta cancelación implicaba grandes números de términos, y durante algún tiempo muchos investigadores en supergravedad pensaron que este «milagro» aparente de la cancelación era una señal de que la teoría estaba en el camino correcto, de modo que la renormalizabilidad debía esperarse en todos los órdenes; y que, por consiguiente, ¡la verdadera teoría de la gravedad cuántica sería pronto revelada! Por desgracia para los investigadores en supergravedad, cuando fueron capaces de completar el cálculo a tercer orden volvieron las divergencias no renormalizables. Esto llevó a consideraciones de dimensionalidad superior, pero las cosas se estancaron durante un tiempo. Luego la supergravedad fue reavivada como parte de la ruta que lleva a la teoría M, como se describe en §§31.4,14.

Estoy seguro de que la propia teoría de cuerdas y la teoría M han sido guiadas por muchos de estos milagros. Uno de los más importantes fue seguramente el descubrimiento de las simetrías espejo, que ofrecía un indicio de que la enigmática colección de teorías de cuerdas, en apariencia completamente diferentes (descritas en §31.14), podrían unificarse en un gran esquema conocido como «teoría M». Dichas simetrías espejo actuaban como magia, y se descubrió que números que previamente había parecido que tenían poco que ver entre sí, eran los mismos, como fue el caso de los cálculos realizados por Candelas y sus colegas, descritos en §31.14. Esto ciertamente lo califica como un milagro, en el sentido en que utilizo aquí el término. Estoy seguro de que cuando el número 31 7206 375, obtenido en el cálculo de Candelas utilizando simetría espejo, fue confirmado finalmente por los geómetras algebraicos, esto fue jaleado como un milagro que proporcionaba la evidencia concluyente de que la nueva teoría de cuerdas M estaba en

el camino correcto. Sea o no este el caso, el «milagro» proporcionaba un excelente apoyo a los aspectos matemáticos de la idea de la simetría espejo. De hecho, posteriormente se ha suscitado mucho interés puramente matemático por esa cuestión, y ahora se ha obtenido una buena comprensión puramente matemática de mucho de lo que está implicado.[42]

¿Realmente son estos aparentes milagros buenas guías del carácter correcto de una aproximación a una teoría física? Esta es una pregunta profunda y difícil. Puedo imaginar que a veces lo son, pero hay que ser extraordinariamente cautos sobre esto. Puede ser muy bien que el descubrimiento de Dirac de que su ecuación de ondas relativista incorporaba automáticamente el espín del electrón pareciera uno de estos milagros, como lo había sido el uso que hizo Bohr de la cuantización del momento angular para obtener los espectros atómicos correctos, y análogamente la comprensión de Einstein de que su aproximación a la gravedad mediante el espacio curvo de la relatividad general diera realmente la respuesta correcta para la posición del perihelio de Mercurio —que previamente había intrigado a los astrónomos durante setenta años—. Pero estas eran claramente consecuencias físicas apropiadas de las teorías que se estaban proponiendo, y los milagros ofrecían una confirmación impresionante de las respectivas teorías. Menos claro está cuál es la fuerza de los milagros puramente matemáticos, como es el caso de la supergravedad o de la simetría espejo. Cuando por fin se obtiene una comprensión matemática de un milagro de este tipo matemático, existe la posibilidad de que pueda producir, en alguna medida, una «devaluación» del milagro en cuestión. Incluso así, quizá esto no elimine por completo la fuerza psicológica del propio milagro, que siempre debe verse en su contexto histórico apropiado.

Sin embargo, una cosa es cierta, y es que tales milagros matemáticos no siempre pueden ser una guía segura. En el curso de mis propios estudios de la teoría de twistores, he encontrado varios elementos alentadores que parecían caer bajo el encabezamiento de «milagros» en el sentido en que estoy utilizando aquí el término. El descubrimiento (§33.8) de que funciones homogéneas de twistores simples generan soluciones generales de las ecuaciones de campos sin masa fue uno de estos, y la construcción del gravitón no lineal de §33.11 fue otro. ¿En

qué medida son estos indicios fuertes de que la teoría de twistores está «en la línea correcta»? Una vez más, hay que ser prudentes. No pretendo hacer una comparación entre los milagros de la teoría de twistores y los de la teoría de cuerdas. Pero ambos no pueden ser indicadores inequívocos, porque, como se ha señalado en §33.14, ¡las dos teorías, tal como están, son mutuamente incompatibles!

Pese a todo, estos comentarios se aplican solamente a lo que entiendo que es el estado de estas teorías «en este momento». Algunos desarrollos que parecen interesantes y que se han llevado a cabo solo en los últimos meses podrían dar la vuelta por completo a la conclusión a la que parezco haber llegado en mis comentarios finales en el párrafo anterior. Se trata de algunas aplicaciones muy innovadoras de las ideas de la teoría de cuerdas en el contexto de la teoría de twistores; y se deben, principalmente, a Edward Witten[43] (las he mencionado brevemente en §31.18 y §33.14). En estos desarrollos, la teoría de cuerdas se aplica a la física espaciotemporal *tetra*dimensional estándar y se interesa en interacciones de tipo Yang-Mills que probablemente tienen relevancia directa para las interacciones entre partículas reales, con lo que representan una ruptura sustancial con la forma de la teoría de cuerdas a la que me he referido en los párrafos anteriores. ¿Cómo se consigue esto? En esencia, se hace considerando que el «espacio imagen» en el que deben aplicarse las superficies de Riemann de la teoría de cuerdas no es una 3-variedad compleja de Calabi-Yau (§31.14), que había sido invocada para suministrar las «dimensiones espaciales extra» de la teoría de cuerdas «estándar», sino la 3-variedad compleja que es el *espacio twistorial proyectivo* $\mathbb{PT}$ (un $\mathbb{CP}^3$; véase §33.5). Como hemos visto, la geometría twistorial remite explícitamente a un espacio 4-dimensional ordinario, y *no* hay «dimensiones espaciales extra». Tal como se han descrito estas nuevas ideas, hay todavía implicada alguna supersimetría, y esta versión supersimétrica de $\mathbb{PT}$ puede considerarse realmente como un tipo de «espacio de Calabi-Yau». (Esto es así para que se cancele cierta «anomalía», aunque creo que la necesidad de cancelación de la anomalía puede estar sobrevalorada, y quizá la supersimetría no sea realmente necesaria.) Tal como están estas nuevas ideas, las superficies de Riemann se toman de género 0 (véase §8.4), i.e., son *esferas de Riemann*.[44] Esto permite establecer un contacto notable con buena parte

de la teoría de twistores anterior, en la que se habían incluido previamente ideas de «cuerda».[45]

Si la teoría de cuerdas puede llegar a cambiar de esta forma, que a mí me parece un cambio muy sustancial, ¿qué relevancia física tienen entonces los «milagros» de dicha teoría? Mi conjetura sería que pueden tener una relevancia importante (aunque indirecta), y que podría salir inmediatamente a la luz una comprensión mayor de lo que «está pasando entre bastidores» (como se sugiere en los comentarios de Richard Thomas citados en §31.18). ¿Podemos extraer lo que es potente en la teoría de cuerdas y despojarla de una dependencia necesaria de la supradimensionalidad espaciotemporal? Tal vez. Lo que parece cierto, en esencia, es que hay algo profundo en la idea de una teoría cuántica de campos basada en las aplicaciones de esferas de Riemann en variedades complejas[46] (o quizá también las aplicaciones de superficies de Riemann de género más alto), y este tipo de objeto seguiría teniendo relevancia en este contexto más nuevo, al ser la variedad compleja un espacio twistorial (proyectivo). Pero lo que está pasando *realmente* sigue pareciendo un misterio.

34.10. Preguntas profundas respondidas, preguntas más profundas planteadas

Cuestiones tales como las descritas en las secciones precedentes están lejos de tener respuestas dentro del conocimiento físico actual, y podemos confiar en que una física futura del siglo XXI arroje luz importante sobre ellas. Pero si volvemos la vista atrás para ver lo que ya hemos conseguido en nuestra comprensión a finales del siglo XX, la especie humana puede sentirse, y con razón, bastante orgullosa. Muchas cuestiones que eran profundamente enigmáticas —y a veces terroríficas— para los antiguos han encontrado respuestas, y con frecuencia es posible actuar de un modo positivo a la luz de dichas respuestas. Muchos de los terrores que antes causaban las enfermedades no provocan miedo ahora, no solo debido a los fármacos modernos (donde el método científico se ha mostrado de valor incalculable), sino a que puede utilizarse el diagnóstico precoz mediante el uso de la moderna tecnología

(rayos X, ultrasonidos, tomografía, etc.) y los tratamientos físicos sofisticados (radiación, láseres, etc.). A menudo esta tecnología depende de conocimientos profundos de física que no estaban a disposición de los antiguos. El mismo tipo de conocimiento nos ha dado muchas otras cosas, tales como la hidroelectricidad, la iluminación eléctrica, los modernos materiales que sirven de protección contra los elementos, las telecomunicaciones tales como la televisión y la telefonía móvil, la tecnología de ordenadores, internet, el transporte moderno en sus diversas formas y otros numerosos aspectos de nuestra vida moderna.

Muchos de estos desarrollos dependen directamente de la física de una forma u otra. Además, las reglas básicas de la química, tal como se entiende hoy día, son también básicamente físicas (en principio, si no en la práctica), pues proceden sobre todo de las reglas de la mecánica cuántica. La biología está mucho más lejos de ser reducible a leyes físicas, pero no tenemos ninguna razón para creer que (aparte de la consciencia) el comportamiento biológico no sea, en su raíz, puramente dependiente de acciones físicas que ahora comprendemos básicamente. En consecuencia, la biología también parece estar, en última instancia, controlada por las matemáticas.

Consideremos, por ejemplo, la forma milagrosa en que una semilla puede desarrollarse en una planta viva, donde la soberbia estructura de cada planta es similar en gran detalle a cada una de las demás que proceden del mismo tipo de semilla. Hay aquí física profundamente subyacente, puesto que el ADN que controla el crecimiento de la planta es una molécula, y la persistencia y fiabilidad de su estructura depende de forma crucial de las reglas de la mecánica cuántica (como señaló Schrödinger en 1944 en su influyente librito *¿Qué es la vida?*).[47] Además, el crecimiento de la planta está controlado en última instancia por las mismas fuerzas físicas que gobiernan las partículas individuales de las que está compuesta. Las relevantes son principalmente de origen electromagnético, pero la fuerza nuclear fuerte es vital para determinar qué nucleos son posibles, y por lo tanto qué tipos de átomos pueden existir.

También la fuerza débil tiene su papel en los fenómenos que vemos a gran escala, y es notable cómo, a pesar de su debilidad (solo unas 10^{-7} veces la intensidad de la fuerza fuerte y 10^{-5} veces la intensidad del

electromagnetismo), esta fuerza puede dar como resultado algunos de los sucesos más espectaculares que ha experimentado la humanidad. Pues es la fuerza débil la responsable, a través de las desintegraciones radiactivas en el interior de la Tierra, del calentamiento del magma terrestre. En particular, las erupciones volcánicas son su legado. Hubo un período de unos pocos años en la historia de la Tierra, a partir del 535 d.C. aproximadamente, durante el que se registraron hambrunas globales y un clima anormalmente frío, debido a una capa prácticamente continua de polvo que había sido arrojado en una enorme explosión volcánica. El volcán era probablemente el mismo objeto que el Krakatoa, cerca de Java, que al parecer entró en una violenta erupción en el año 535, y así lo hizo otra vez (pero con menos violencia) en 1883.

Posiblemente más dramática incluso para sus espectadores civilizados fue la explosión volcánica que destruyó la isla de Thera (Santorini), que habría sido visible desde Creta, a unos 160 kilómetros al sur, aproximadamente en 1628 a.C. Barrió la civilizada comunidad de la propia Thera y probablemente fue la responsable en definitiva del posterior ocaso de la pacífica y culta sociedad de Knossos en Creta, en cuyo Gran Palacio se dice que estaba localizado el famoso laberinto de Dédalo.[48] Se ha argumentado de manera convincente que la destrucción de Thera puede haber sido muy bien la fuente de la leyenda de la Atlántida.[49] Quizá encontremos consuelo en el hecho de que algunos de los cataclismos del pasado también pueden haber generado el crecimiento de nuevos avances que no se hubieran producido en otro caso. (El más espectacular de estos fue la aniquilación global de los dinosaurios, que permitió el desarrollo de los mamíferos que llevó finalmente a los seres humanos, aunque parece que la causa de esto fue la colisión con un asteroide antes que la actividad volcánica.) ¿Debe algo a este suceso volcánico catastrófico el extraordinario desarrollo de la cultura de la antigua Grecia en el milenio que siguió a la destrucción de Thera?

Quizá sea aún más sorprendente que las más violentas explosiones vistas en el universo son causadas por la fuerza más débil de todas —si es justo llamarla una fuerza—, a saber, la gravitación (solo aproximadamente 10^{-40} de la fuerza eléctrica, en un átomo de hidrógeno, y aproximadamente 10^{-38} de la intensidad de la fuerza débil), donde los agujeros negros alimentan las increíblemente poderosas fuentes de energía de los cuásares. Pero, pese a su extraordinaria potencia, su distancia a nosotros es tan grande que, visto desde la Tierra, el cuásar más brillante, 3C-273, tiene solo aproximadamente 10^{-6} del brillo de la cercana estrella Sirio. De hecho, cuando examinamos el cielo en una noche clara y tranquila, aunque podemos sentirnos sobrecogidos ante la inmensidad del universo, solo percibimos la más minúscula fracción de su enorme escala. Los objetos más lejanos visibles a simple vista (la galaxia Andrómeda) están a un ridículo 10^{-3} de la distancia a 3C-273 ¡y aproximadamente a 10^{-4} de la distancia al límite del universo observable!

Las singularidades espaciotemporales que residen en los núcleos de los agujeros negros están entre los objetos conocidos (o presuntos) del universo sobre los que permanecen los más profundos misterios, y que nuestras teorías actuales se ven impotentes para describir. Como hemos visto en §§34.5,7,8, en particular, hay otras cuestiones profundamente misteriosas de las que tenemos poca comprensión. Es muy probable que el siglo XXI revele ideas más maravillosas incluso que aquellas con las que nos ha bendecido el siglo XX. Pero para que esto suceda, necesitaremos nuevas y poderosas ideas que nos lleven en direcciones significativamente diferentes de las que en la actualidad se están siguiendo. Quizá lo que necesitamos fundamentalmente es algún cambio sutil de perspectiva... algo que todos hemos pasado por alto...

Notas

Sección 34.1

34.1. Véase, por ejemplo, Mukohyama y Randall (2003).
34.2. Véase Tittel *et al.* (1998).
34.3. Véase Arndt *et al.* (1999).
34.4. Véanse Amelino-Camelia *et al.* (1998), Gambini y Pullin (1999), Amelino-Camelia y Piran (2001), Sarkar (2002), o, para una perspectiva alternativa, Magueijo y Smolin (2002).

Sección 34.2

34.5. Utilizo aquí la palabra «coherencia» para transmitir algo que es solo un poco más fuerte que *consistencia*. La palabra quiere sugerir que hay cierta economía además de consistencia en una estructura matemática completamente *coherente*, donde diferentes aspectos de su formalismo funcionan en perfecto acuerdo.

Sección 34.3

34.6. Véase Rovelli (1998).

34.7. Algunos de mis colegas me han dicho que creen que este es realmente el caso ahora.

34.8. Recordemos de §15.4 que la S^7 está fibrada por esferas S^3 en analogía con los paralelos de Clifford en S^3. De hecho, S^7 es lo que se denomina «paralelizable», lo que significa que un 7-sistema de vectores tangentes puede ser continuamente asignado en todos sus puntos. El «aplastamiento» de S^7 es conseguido sistemáticamente a lo largo de estas direcciones «paralelas».Véase Jensen (1973).

34.9. Véanse Collins (1977) para las trayectorias de Regge, y Chew (1962) para la matriz S.

34.10. Véase Geroch (notas de clase en la Universidad de Chicago, inéditas).

34.11. Véanse Van der Waerden (1929), Infeld y Van der Waerden (1933), Penrose y Rindler (1984, 1986) y O'Donnell (2003).

34.12. Véase Dirac (1936); para otras versiones de las ecuaciones de espín más alto, véase Corson (1953).

34.13. Véase Weiss y Zumino (1974).

Sección 34.4

34.14. Véase Popper (1934).

34.15. Para un uso divertido de una idea de este tipo, véase Aldiss y Penrose (2000).

34.16. A pesar de recientes afirmaciones de que BOOMERanG ha proporcionado realmente dicha confirmación.Véase, por ejemplo, Bouchet *et al.* (2002) para una crítica (limitada) de tales afirmaciones.

34.17. El universo en estado estacionario, propuesto por Bondi, Gold y Hoyle a principios de la década de 1950, era acentuadamente popperiano, como fue bien establecido por Bondi, y entre sus vías de refutación hubiera estado también el establecimiento de $K \neq 0$. Sin embargo, cayó a causa de otros conflictos con la observación, muy especialmente la presencia de la radiación de fondo de 2,7 K, que era evidencia prácticamente directa del big bang; véanse §§27.7,10,11,13 y §§28.4,7,10.

34.18. Véase, por ejemplo, Linde (1993).

34.19. Véanse Bucher *et al.* (1995) y Linde (1995).

34.20. Véase Hawking y Turok (1998).

34.21. Véase Lange *et al.* (2001).

Sección 34.5

34.22. Véase, por ejemplo, *El fin de la ciencia*, de John Horgan.

34.23. Véanse Ferber (1978), Ward y Wells (1989), Delduc *et al.* (1993) e Ilyenko (1999).

Sección 34.7

34.24. Véase Penrose (1989, 1994, 1997a, 1997b).

34.25. Véanse Deutsch (2000) y Lockwood (1989).

34.26. Véanse Gell-Man (1994) y Hartle (2004).

34.27. No obstante, cualquiera que haya oído alguna vez las ideas de David Bohm sobre este tema se daría cuenta de que él no consideraría que las cuestiones de la consciencia carezcan de relación con las de la mecánica cuántica.

34.28. Véase Penrose (1989, 1994, 1997).

34.29. Véanse Hameroff y Watt (1982), Hameroff (1987, 1998) y Hameroff y Penrose (1996).

34.30. Véase Koruga *et al.* (1993).

34.31. Véase, por ejemplo, Anderson (1997).

34.32. Comunicación personal. Por lo que sé, las investigaciones de Duggins no están aún completas, y no hay ninguna publicación disponible hasta la fecha.

34.33. Véanse Penrose (1987a, 1994) y Hameroff y Penrose (1996).

Sección 34.8

34.34. Que esta unidad deba considerarse notable, y algo misteriosa, parece ser una cuestión de debate entre expertos. En una famosa conferencia, el físico matemático Eugene Wigner (1960) habló sobre «La irrazonable efectividad de las matemáticas en las ciencias físicas». Pero Andrew Gleason, un destacado matemático, dos de cuyos potentes teoremas han aparecido en otros lugares de este libro (véanse las notas 13.4 y 23.4), ha adoptado una postura contraria (1990), al considerar la con-

cordancia entre matemáticas y física como un mero reflejo del hecho de que «la matemática es la ciencia del orden». Mi propio punto de vista estaría más cerca del de Wigner que del de Gleason. No solo la extraordinaria precisión, sino también la sutileza y sofisticación que encontramos en las leyes matemáticas que operan en los fundamentos de la física, son para mí mucho más que la mera expresión de un «orden» subyacente en el funcionamiento del mundo.

34.35. Esta filosofía parece estar próxima a la de Geoffrey Chew, que valoraba mucho las propiedades holomorfas de la matriz S. Véase Chew (1962).

34.36. Véanse los comentarios introductorios por Freeman Dyson, en Dyson (1966).

34.37. Véase Penrose (1988b).

34.38. Este hecho tiene importancia en el estudio de las simetrías asintóticas de espaciotiempos asintóticamente planos en la relatividad general; véanse Sachs (1962a, 1962b) y Penrose y Rindler (1986).

34.39. Compárese con Eddington (1946).

Sección 34.9

34.40. Véanse Du Sautoy (2004) para una discusión de la conexión de la hipótesis de Riemann con la física, así como también Berry y Keating (1999). Otra relevancia de la función ζ de Euler para la física es la que se denomina «regularización de la función ζ». Una búsqueda rápida en los LANLarXiv mostró ¡142 ítems en el último recuento!

34.41. Véanse Wess y Bagger (1992) y la nota 31.13.

34.42. Véanse las notas 31.57 y 31.58.

34.43. Véase la nota 31.81.

34.44. Esto en sí mismo no restringe a dichas superficies de Riemann a ser las líneas del espacio twistorial que representan puntos espaciotemporales (véase §33.5), sino que pueden ser curvas de «orden superior». La «condición de género 0» nos dice básicamente que uno está interesado en procesos de árbol de Yang-Mills en los que no aparecen «lazos cerrados» (§26.8).

34.45. Véanse Shaw y Hughston (1990) y Hodges (1985, 1990a, 1990b).

34.46. Estos se conocen como «σ-modelos»; véase Ketov (2000).

Sección 34.10

34.47. Reimpreso en Schrödinger (1967).

34.48. Véase Davies (1997), pp. 89-94, para una descripción gráfica de este acontecimiento (lo he tomado prestado del prólogo). Para la leyenda, un viejo favorito está reimpreso en Hamilton (1999).

34.49. Véase Friedrich (2000) para un resumen reciente de estas ideas.

Epílogo

Antea, una estudiante posdoctoral de física, con un gran talento tanto artístico como matemático, procedía de una pequeña ciudad del sur de Italia. Observaba el cielo en una noche despejada desde una ventana oriental del Instituto Albert Einstein en Golm, cerca de Potsdam, Alemania. Este prestigioso instituto de investigación había sido fundado a finales del siglo XX cerca del lugar donde Einstein había tenido una pequeña casa de recreo. Una buena parte de la investigación estaba relacionada con la debatida cuestión de la «gravedad cuántica» que intenta unificar los principios subyacentes en la relatividad general de Einstein y los de la mecánica cuántica: un misterio en la base misma de las leyes del mundo.

En esta dirección se encaminaba la propia investigación de Antea, pero ella era una recién llegada y tenía algunas ideas poco ortodoxas y todavía no plenamente formadas acerca del camino a seguir, algunas de las cuales eran fundamentalmente opuestas a las de sus colegas. Esa noche había estado trabajando hasta la madrugada, en la biblioteca superior del instituto, cuando hacía rato que todos los demás se habían ido a dormir. Había estado estudiando una vieja investigación relativa a gigantescas emisiones de energía que tenían lugar en los centros de algunas galaxias. Afortunadamente, pensaba para sí, la Tierra y el sistema solar no están próximos a ninguno de estos lugares, pues si lo estuvieran se habrían vaporizado en su totalidad casi instantáneamente. La explicación establecida para estas explosiones increíblemente potentes consistía en que cada una de ellas estaba alimentada por un agujero negro de proporciones inmensas.

Antea sabía que un agujero negro es una región del espaciotiempo

en cuyo interior hay una estructura conocida como una «singularidad espaciotemporal», algo cuya descripción científica era aún profundamente esquiva y que dependía de la todavía carente teoría de la gravedad cuántica. Pero el interés real de Antea no estaba tanto en los agujeros negros galácticos como en una explosión aún más monstruosa, la explosión que termina con todas las explosiones —o, mejor, la que dio inicio a todas—, llamada el «big bang». Ella pensaba que era el origen de todas las cosas buenas tanto como de todas las cosas malas. Pero la singularidad espaciotemporal en el big bang planteaba misterios aún mayores que los de los agujeros negros. Antea sabía que en la raíz de estos misterios yace el secreto de cómo unificar la teoría de Einstein del espacio, el tiempo y la gravedad a gran escala con los principios mecanocuánticos de la física de lo pequeño.

Era una noche tranquila y las estrellas eran inequívocamente claras. Antea permaneció un rato pensativa, con los brazos cruzados reposando en la balaustrada de la escalera, observando las figuras de las estrellas a través del ventanal: no sabía durante cuánto tiempo. Siempre se sobrecogía al contemplar, en esa vasta cúpula aparentemente hemisférica, la gran distancia de esos minúsculos pinchazos de luz, aunque contara muy poco frente a la inmensidad de las escalas cosmológicas. Pese a todo, cavilaba, si alguna explosión cósmica llegara a hacerse visible para ella ahora, por muy lejana que fuese, sus pequeños fotones no habrían experimentado ningún tiempo en alcanzarla. Lo mismo se aplicaría a los minúsculos gravitones producidos en la explosión, algunos de los cuales podrían sentirse en el detector de ondas gravitatorias del Instituto cerca de Hannover a 250 km de distancia. Se estremeció con la idea de que, en efecto, estaría en inmediato contacto directo con ese suceso explosivo...

Mientras estaba allí, mirando al este, fue sorprendida por un fogonazo momentáneo e inesperado de luz verde, como si llegara el alba, antes de irrumpir el rojo vivo del sol. El fenómeno del «rayo verde» y su bien establecida explicación física eran muy conocidos para ella, pero nunca lo había presenciado antes, y le produjo un extraño efecto emocional. Esta experiencia se mezcló con varias ideas matemáticas que le habían dado problemas durante la noche.

Bibliografía

Además de los grandes avances en el conocimiento físico que se han logrado en el siglo XX —a partir de experimentos muy refinados y teoría matemática sofisticada—, la moderna tecnología y la innovación han mejorado enormemente la capacidad de distribuir y recoger información a escala mundial. En concreto, está la introducción de arXiv.org, un archivo online donde físicos y matemáticos, biólogos y científicos de la computación pueden publicar preprints (o e-prints) de su trabajo antes (¡o incluso en lugar de!) enviarlos a revistas. ArXiv.org ha hecho posible que los científicos comuniquen nuevas ideas a una velocidad increíblemente alta, y como consecuencia el ritmo de la actividad investigadora se ha acelerado en un grado sin precedentes (o, como algunos podrían considerar, alarmante).

He tratado de sacar provecho de esta importante nueva tendencia ofreciendo, donde fuera posible, los enlaces a arXiv.org para ítems en la bibliografía. Encontrar un artículo en arXiv.org es muy sencillo. Primero, utilice su buscador preferido para ir a www.arxiv.org. Luego, o bien busca el artículo, o bien entra www.arxiv.org/, seguido del código de identificación proporcionado entre paréntesis en la bibliografía. Por ejemplo, para llamar al artículo de 2003 de Lee Smolin «How far are we from the quantum theory of gravity?», entraríamos en la dirección web:

www.arxiv.org/hep-th/0303185.

Esto será especialmente útil para aquellos lectores que tengan acceso a la World Wide Web pero estén lejos de una biblioteca universitaria que pudiera guardar las revistas especializadas donde se suelen publicar los artículos. Espero que esta nueva característica en la bibliografía le anime a leer los muchos buenos artículos en los arXiv, ya estén o no referenciados aquí.

Abian, A. (1965), *The Theory of Sets and Transfinite Arithmetic*, Saunders, Filadelfia.

Abbott, B., *et al.* (2004), «Detector Description and Performance for the First Coincidence Observations between LIGO and GEO», *Nucl. Instrum. Meth.* **A517**, 154-179 [gr-qc/0308043].

Adams, C. C. (2000), *The Knot Book*, Owl Books, Nueva York.

Adams, F. J., y Atiyah, M. F. A. (1966), «On K-theory and Hopf invariant», *Quarterly J. Math.* **17,** 31-38.

Adler, S. L. (1995), *Quaternionic Quantum Mechanics and Quantum Fields*, Oxford University Press, Nueva York.

Afriat, A. (1999), «The Einstein, Podolsky, and Rosen Paradox», en *Atomic, Nuclear, and Particle Physics*, Plenum Publishing Corp.

Aharonov, Y., y Albert, D. Z. (1981), «Can we make sense out of the measurement process in relativistic quantum mechanics?», *Phys. Rev.* **D24**, 359-370.

Aharonov, Y., y Anandan, J. (1987), «Phase change during a cyclic quantum evolution», *Phys. Rev. Lett.* **58**, 1593-1596.

Aharonov, Y., y Bohm, D. J. (1959), «Significance of electromagnetic potentials in the quantum theory», *Phys. Rev.* **115**, 485-491.

Aharonov, Y., y Vaidman, L. (1990), «Properties of a quantum system during the time interval between two measurements», *Phys. Rev.* **A41**, 11.

Aharonov, Y., y Vaidman, L. (2001), «The Two-State Vector Formalism of Quantum Mechanics», en *Time in Quantum Mechanics* (eds. J. G. Muga *et al.*), Springer-Verlag.

Aharonov, Y., Bergmann, P., y Lebowitz, J. L. (1964), «Time symmetry in the quantum process of measurement», en *Quantum Theory and Measurement* (eds. J. A. Wheeler y W. H. Zurek), Princeton University Press, Princeton, New Jersey, 1983; originalmente en *Phys. Rev.* **134B**, 1410-1416.

Ahmavaara, Y. (1965), «The structure of space and the formalism of relativistic quantum theory», *I. J. Math. Phys.* **6,** 87-93.

Aitchison, I., y Hey, A. (2004), *Gauge Theories in Particle Physics: A Practical Introduction*, vols. 1 y 2, Institute of Physics Publishing, Bristol.

Albert, D., Aharanov, Y., y D'Amato, S. (1985), «Curious New Statistical Prediction of Quantum Mechanics», *Phys. Rev. Lett.* **54**, 5.

Aldiss, B. W., y Penrose, R. (2000), *White Mars*, St Martin's Press, Londres.

Alpher, B., y Gamow (1948), «The Origin of Chemical Elements», *Phys. Rev.* **73**, 803.

Ambjorn, J., Nielsen, J. L., Rolf, J., y Loll, R. (1999), «Euclidean and Lorentzian Quantum Gravity: Lessons from Two Dimensions», *Chaos Solitons Fractals* 10 [hep-th/9805108].

Amelino-Camelia, G., y Piran, T. (2001), «Planck-scale deformations of Lorentz symmetry as a solution to the UHECR and the TeV-MATH-gamma paradoxes», *Phys. Rev.* **D64**, 036005 [astro-ph/0008107].

Amelino-Camelia, G., *et al.* (1998), «Potential Sensitivity of Gamma-Ray Burster Observations to Wave Dispersion in Vacuo», *Nature,* **393**, 763-765 [astro-ph/9712103].

Anderson, P. W. (1997), *The Theory of Superconductivity in the High-Tc Cuprate Superconductors*, Princeton University Press, Princeton, New Jersey.

Anguige, K. (1999), «Isotropic cosmological singularities 3: The Cauchy problem for the inhomogeneous conformal Einstein-Vlasov equations», *Annals Phys.* **282**, 395-419.

Antoci, S. (2001), «The origin of the electromagnetic interaction in Einsteins unified field theory with sources» [gr-qc/018052].

Anton, H., y Busby, R. C. (2003), *Contemporary Linear Algebra*, John Wiley & Sons, Hoboken, NJ.

Apostol, T. M. (1976), *Introduction to Analytic Number Theory*, Springer-Verlag, Nueva York (hay traducción castellana: *Introducción a la teoría analítica de números*, Editorial Reverté, Barcelona, 1984).

Arfken, G., y Weber, H. (2000), *Mathematical Methods for Physicists*, Harcourt/Academic Press.

Arndt, M., *et al.* (1999), «Wave-particle duality of C60 molecules», *Nature* **401**, 680.

Arnol'd, V. I. (1978), *Mathematical Methods of Classical Mechanics*, Springer-Verlag, Nueva York (hay traducción castellana: *Mecánica clásica. Métodos matemáticos*, Thomson Paraninfo, Madrid, 1983].

Arnowitt, R., Deser, S., y Misner, C. W. (1962), en *Gravitation: An Introduction to Current Research* (ed. L. Witten), John Wiley & Sons, Inc., Nueva York.

Ashtekar, A. (1986), «New variables for classical and quantum gravity», *Phys. Rev. Lett.* **57**, 2244-2247.

Ashtekar, A. (1987), «New Hamiltonian formulation of general relativity», *Phys. Rev.* **D36**, 1587-1602.

Ashtekar, A. (1991), *Lectures on Non-Perturbative Canonical Gravity* (Apéndice), World Scientific, Singapur.

Ashtekar, A., y Das, S. (2000), «Asymptotically anti-de Sitter space-times», *Classical and Quantum Gravity*, **17**, L17-L30.

Ashtekar, A., y Lewandowski, J. (1994), «Representation theory of analytic holonomy algebras», en *Knots and Quantum Gravity* (ed. J. C. Baez), Oxford University Press, Oxford.

Ashtekar, A., y Lewandowski, J. (2001), «Relation between polymer and Fock excitations», *Class. Quant. Grav.* **18**, L117-L127.

Ashtekar, A., y Lewandowski, J. (2004), «Background Independent Quantum Gravity: A Status Report» [gr-qc/0404018].

Ashtekar, A., y Magnon, A. (1980), «A geometrical approach to external potential problems in quantum field theory», *General Relativity and Gravitation*, vol. 12, 205-223.

Ashtekar, A., y Schilling, T. A. (1998), en *On Einstein's Path* (ed. A. Harvey), Springer-Verlag, Berlín.

Ashtekar, A., Baez, J. C., y Krasnov, K. (2000), «Quantum geometry of isolated horizons and black hole entropy», *Adv. Theo. Math. Phys.* **4**, 1-95.

Ashtekar, A., Baez, J. C., Corichi, A., y Krasnov, K. (1998), «Quantum geometry and black hole entropy (1998)», *Phys. Rev. Lett.* **80**, 904-907.

Ashtekar, A., Bojowald, M., y Lewandowski, J. (2003), «Mathematical structure of loop quantum cosmology», *Adv. Theor. Math. Phys.* **7**, 233-268.

Atiyah, M. F. (1990), *The Geometry and Physics of Knots,* Cambridge University Press, Cambridge.

Atiyah, M. F., y Singer, I. M. (1963), «The Index of Elliptic Operators on Compact Manifolds», *Bull. Amer. Math. Soc.* **69**, 322-433.

Atiyah, M. F., Hitchin, N. J., y Singer, I. M. (1978), «Self-duality in four-dimensional Riemannian geometry», *Proc. Roy. Soc. Lond.* **A362**, 42561.

Baez, J. C. (1998), «Spin foam models», *Class. Quant. Grav.* **15**, 182758.

Baez, J. C. (2000), «An introduction to spin foam models of quantum gravity and BF theory», *Lect. Notes Phys.* **543**, 25-94

Baez, J. C. (2001), «Higher-dimensional algebra and Planck-scale physics», en *Physics Meets Philosophy at the Planck Scale* (eds. C. Callender y N. Huggett), Cambridge University Press, Cambridge [gr-qc/9902017].

Baez, J. C., y Dolan, D. (1998), «Categorification», en *Higher Category Theory* (eds. E. Getzler y M. Kapranov), Contemporary Mathematics vol. 230. AMS, Providence. RI (véase también http://xxx.lanl.gov/abs/math.QA/9802029).

Bahcall, N., Ostriker, J. P., Perlmutter, S., y Steinhardt., P. J. (1999), «The Cosmic Triangle: Revealing the State of the Universe», *Science* **284** [astro-ph/9906463].

Bailey, T. N., y Baston, R. J. eds. (1990), «Twistors in Mathematics and Physics», *LMS Lecture Note Series* **156**, Cambridge University Press, Cambridge.

Bailey, T. N., Ehrenpreis, L., y Wells, R. O., Jr. (1982), «Weak solutions of the massless field equations», *Proc. Roy. Soc. Lond.* **A384**, 40325.

Banchoff, T. (1990, 1996), *Beyond the Third Dimension*, Scientific American Library. Véase también http://www.faculty.fairfield.edu/jmac/cl/tb4d.htm.

Bar, I. (2000), «Survey of Two-Time Physics» [hep-th/0008164].

Barbour, J. B. (1989), *Absolute or Relative Motion*, vol. I: *The Discovery of Dynamics*, Cambridge University Press, Cambridge.

Barbour, J. B. (1992), *Time and the Interpretation of Quantum Gravity*, Syracuse University Preprint.

Barbour, J. B. (2001a), *The Discovery of Dynamics: A Study from a Machian Point of View of the Discovery and the Structure of Dynamical Theories*, Oxford University Press, Oxford.

Barbour, J. B. (2001b), *The End of Time*, Oxford University Press, Oxford.

Barbour, J. B. (2004), *Absolute or Relative Motion: The Deep Structure of General Relativity*, Oxford University Press, Oxford.

Barbour, J. B., Foster, B., y O Murchadha, N. (2002), «Relativity without relativity», *Class. Quant. Grav.* **19**, 3217-3248 [gr-qc/0012089].

Barrett, J. W., y Crane, L. (1998), «Relativistic spin networks and quantum gravity», *J. Math. Phys.* **39**, 3296-302.

Barrett, J. W., y Crane, L. (2000), «A Lorentzian signature model for quantum general relativity», *Class. Quant. Grav.* **17**, 3101-3118.

Barrow, J. D., y Tipler, F. J. (1986), *The Anthropic Cosmological Principle*, Oxford University Press, Oxford.

Barrow, J. D., y Tipler, F. J. (1988), *The Anthropic Cosmological Principle*, Oxford University Press, Oxford.

Baston, R. J., y Eastwood, M. G. (1989), *The Penrose Transform: Its Interaction with Representation Theory*, Oxford University Press, Oxford.

Bateman, H. (1904), «The solution of partial differential equations by means of definite integrals», *Proc. Lond. Math. Soc.* (2) 1, 451-459.

Bateman, H. (1944), *Partial Differential Equations of Mathematical Physics*, Dover, Nueva York.

Becker, R. (1982), *Electromagnetic Fields and Interactions*, Dover, Nueva York.

Begelman, M. C., Blandford, R. D., y Rees, M. J. (1984), «Theory of extragalactic radio sources», *Rev. Mod. Phys.* **56**, 255.

Bekenstein, J. (1972), «Black holes and the second law», *Lett. Nuovo Cim.* **4**, 737-740.

Belinskii, V. A., Khalatnikov, I. M., y Lifshitz, E. M. (1970), «Oscilliatory approach to a singular point in the relativistic cosmology», *Usp. Fiz. Nauk* **102**, 463-500 (traducción inglesa en *Adv. in Phys.* **19**, 525-573).

Bell, J. S. (1987), *Speakable and Unspeakable in Quantum Mechanics*, Cambridge University Press, Cambridge (hay traducción castellana: *Lo decible y lo indecible en mecánica cuántica*, Alianza, Madrid, 1990).

Beltrami, E. (1868), *Essay on the Interpretation of non-Euclidean Geometry*, traducido en Stillwell, J. C. (1996), «Sources of Hyperbolic Geometry», *Hist. Math.*, **10**, AMS Publications.

Bennet, C. L., *et al.* (2003), «First Year Wilkinson Microwave Anisotropy Probe (WMAP) Observations: Preliminary Maps and Basic Results», *Astrophys. J. Suppl.* 148, 1 [astro-ph/0302207].

Bergmann, P. G. (1956), *Helv. Phys. Acta Suppl.* **4**, 79.

Bergmann, P. G. (1957), «Two-component spinors in general relativity», *Phys. Rev.* **107**, 6249.

Bern, Z. (2002), «Perturbative Quantum Gravity and its relation to Gauge Theory», *Living Rev. Relativity*, **5** [www.livingreviews.org/Articles/Volume5/20025 bern/index.html].

Berry, M. V. (1984), «Quantal phase factors accompanying adiabatic changes», *Proc. Roy. Soc. Lond.* **A392**, 45-57.

Berry, M. V. (1985), «Classical adiabatic angles and quantal adiabatic phase», *J. Phys. A. Math. Gen.* **18**, 15-27.

Berry, M. V., y Keating, J. P. (1999), «The Riemann Zeros and Eigenvalue Asymptotics», *SIAM Review* **41**, 2, 236-266.

Berry, M. V., y Robbins, J. M. (1997), «Indistinguishability for quantum particles: spin, statistics and the geometric phase», *Proc. R. Soc. Lond.* A **453**, 1771-1790.

Biedenharn, L. C., y Louck, J. D. (1981), *Angular Momentum in Quantum Physics*, Addison-Wesley, Londres.

Bilaniuk, O.-M., y Sudarshan, G. (1969), «Particle beyond the light barrier», *Phys. Today* **22**, 43-51.

Birrell, N. D., y Davies, P. C. W. (1984), *Quantum Fields in Curves Space*, Cambridge University Press, Cambridge.

Bjorken, J. D., y Drell, S. D. (1965), *Relativistic Quantum Mechanics*, McGraw Hill, Nueva York y Londres.

Blanchard, A., Douspis, M., Rowan-Robinson, M., y Sarkar, S. (2003), «An alternative to the cosmological concordance model», *Astron. Astrophys.* **412**, 35-44.

Blanford, R. D., y Znajek, R. L. (1977), «Electromagnetic Extraction of Energy from Kerr Black Holes», *Monthly Notices of the Royal Astronomical Society* **179**, 433.

Bohm, D. (1951), *Quantum Theory* (Prentice-Hall, Englewood-Cliffs). Ch. 22,

sect. 15-19. Reimpreso como «The Paradox of Einstein, Rosen and Podolsky», en *Quantum Theory and Measurement* (eds. J. A. Wheeler y W. H. Zurek), Princeton University Press, Princeton, New Jersey, 1983.

Bohm, D., y Hiley, B. (1994), *The Undivided Universe*, Routledge, Londres.

Bojowald, M. (2001), «Absence of singularity in loop quantum cosmology», *Phys. Rev. Lett.* **86**, 5227-5230.

Bondi, H. (1957), «Negative mass in general relativity», *Rev. Mod. Phys.* **29**, 423-428; también en *Math. Rev.* **19**, 814.

Bondi, H. (1960), «Gravitational waves in general relativity», *Nature* (Londres) **186**, 535.

Bondi, H. (1961), *Cosmology*, Cambridge University Press, Cambridge.

Bondi, H. (1964), *Relativity and Common Sense*, Heinemann, Londres.

Bondi, H. (1967), *Assumption and Myth in Physical Theory*, Cambridge University Press, Cambridge.

Bondi, H., y Gold, T. (1948), «The Steady-State Theory of the Expanding Universe», *Mon. Not. Roy. Astron. Soc.* **108**, 252-270.

Bondi, H., van der Burg, M. G. J., y Metzner, A. W. K. (1962), «Gravitational waves in general relativity, VII. Waves from axisymmetric isolated systems», *Proc. Roy. Soc. Lond.* **A269**, 21-52.

Bonnor, W. B., y Rotenberg, M. A. (1966), «Gravitational waves from isolated sources», *Proc. Roy. Soc. Lond.* **A289**, 247-274.

Börner, G. (2003), *The Early Universe*, Springer-Verlag.

Bouchet, F. R., Peter, P., Riazuelo, A., y Sakellariadou, M. (2000), «Evidence against or for topological defects in the BOOMERanG data?», *Phys. Rev.* **D65** (2002), 021301 [astro-ph/0005022].

Boyer, C. B. (1968), *A History of Mathematics*, 2.ª ed. John Wiley & Sons, Nueva York (hay traducción castellana: *Historia de la matemática*, Alianza, Madrid, 2003).

Braginsky, V. (1977), «The Detection of Gravitational Waves and Quantum Non-Distributive Measurements», en *Topics in Theoretical and Experimental Gravitation Physics* (eds. V. De Sabbata y J. Weber), pp. 105-122. Plenum Press, Nueva York.

Bramson, B. D. (1975), «The alignment of frames of reference at null infinity for asymptotically flat Einstein-Maxwell manifold», *Proc. R. Soc. London, Ser A* 341, 451-461.

Brandhuber, A., Spence, B., y Ttavaglini, G. (2004), «One-Loop Gauge Theory Amplitudes in N = 4 Super Yang Mills from MHV Vertices» [hep-th/0407214].

Brauer, R., y Weyl, H. (1935), «Spinors in *n* dimensions», *Am. J. Math.* **57**, 425-449.

Brekke, L., y Freund, P. G. O. (1993), *P-adic Numbers in Physics*, North-Holland, Amsterdam.

Bremermann, H. (1965), *Distributions, Complex Variables and Fourier Transforms*, Addison Wesley, Reading, Massachusetts.

Brody, D. C., y Hughston, L. P. (1998a), «Geometric models for quantum statistical inference», en *The Geometric Universe; Science, Geometry, and the Work of Roger Penrose* (eds. S. A. Huggett, L. J. Mason, K. P. Tod, S. T. Tsou y N. M. J. Woodhouse), Oxford University Press, Oxford.

Brody, D. C., y Hughston, L. P. (1998b), «The quantum canonical ensemble», *J. Math. Phys.* **39** (12), 6502-6508.

Brody, D. C., y Hughston, L. P. (2001), «Geometric Quantum Mechanics», *J. Geom. and Phys.* **38** (1), 19-53.

Brown, J. W., y Churchill, R. V. (2004), *Complex Variables and Applications*. McGraw-Hill, Nueva York y Londres.

Bryant, R. L., Chern, S. -S., Gardner, R. B., Goldschmidt, H. L., y Griffiths, P. A. (1991), *Exterior Differential Systems*, MSRI Publications, 18. Springer-Verlag, Nueva York.

Bucher, M., Goldhaber, A., y Turok, N. (1995), «An open Universe from Inflation», *Phys. Rev.* **D52** [hep-ph/9411206].

Buckley, P., y Peat, F. D. (1996), *Glimpsing Reality*, University of Toronto Press, Toronto.

Budinich, P., y Trautman, A. (1988), *The Spinorial Chessboard. Trieste Notes in Physics*, Springer-Verlag, Berlín.

Burbidge, G. R., Burbidge, E. M., Fowler, W. A., y Hoyle, F. (1957), «Synthesis of the Elements in Stars», *Revs. Mod. Phys.* **29**, 547-650.

Burkert, W. (1972), *Lore and Science in Ancient Pythagoreanism*, Harvard University Press, Harvard.

Burkill, J. C. (1962), *A First Course in Mathematical Analysis*, Cambridge University Press, Cambridge.

Byerly, W. E. (2003), *An Elementary Treatise on Fouriers Series and Spherical, Cylindrical, and Ellipsoidal Harmonics, with Applications to Problems in Mathematical Physics*, Dover, Nueva York.

Cachazo, F., Svrcek, P., y Witten, E. (2004), «MHV Vertices and Tree Amplitudes in Gauge Theory» [hep-th/0403047].

Cachazo, F., Svrcek, P., y Witten, E. (2004b), «Twistor Space Structure of One-Loop Amplitudes in Gauge Theory» [hep-th/0406177].

Cachazo, F., Svrcek, P., y Witten, E. (2004c), «Gauge Theory Amplitudes in Twistor Space and Holomorphic Anomaly» [hep-th/0409245].

Candelas, P., de la Ossa, X. C., Green, P. S., y Parkes, L. (1991), «A pair of Calabi-Yau manifolds as an exactly soluble superconformal theory», *Nucl. Phys.* **B359**, 21.

Cartan, É. (1923), «Sur les varietés à connexion affine et la théorie de la relativité generalisée I», *Ann. École Norm. Sup.* **40**, 325-412.

Cartan, É. (1924), «Sur les varietés à connexion affine et la théorie de la relativité generalisée (suite)», *Ann. École Norm. Sup.* **41**, 1-45.

Cartan, É. (1925), «Sur les varietés à connexion affine et la théorie de la relativité generalisée II», *Ann. École Norm. Sup.* **42**, 17-88.

Cartan, É. (1945), *Les Systemes Differentiels Exterieurs et leurs Applications Geometriques*, Hermann, París.

Cartan, É. (1966), *The Theory of Spinors*, Hermann, París.

Carter, B. (1966), «Complete Analytic Extension of the Symmetry Axis of Kerr's Solution of Einstein's Equations», *Phys. Rev.* **141**, 4.

Carter, B. (1971), «Axisimmetric Black Holes Has Only Two Degrees of Freedom», *Phys. Rev. Lett.*, **26**, 331-332.

Carter, B. (1974), «Large Number Coincidences and the Anthropic Principle», en *Confrontation of Cosmological Theory with Astronomical Data* (ed. M. S. Longair), pp. 2918, Reidel, Dordrecht. (Reimpreso en Leslie, 1990.)

Cercignani, C. (1999), *Ludwig Boltzmann: The Man Who Trusted Atoms*, Oxford University Press, Oxford.

Chan, H-M., y Tsou, S. T. (1993), *Some Elementary Gauge Theory Concepts*, World Scientific Lecture Notes in Physics, vol. 47, Londres.

Chan, H-M., y Tsou, S. T. (2002), «Fermion Generations and Mixing from Dualized Standard Model», *Acta Physica Polonica B* **12**.

Chandrasekhar, S. (1981), «The maximum mass of ideal white dwarfs», *Astrophys. J.*, **74**, 81-82.

Chandrasekhar, S. (1983), *The Mathematical Theory of Black Holes*, Clarendon Press, Oxford.

Chari, V., y Pressley, A. (1994), *A Guide to Quantum Groups*, Cambridge University Press, Cambridge.

Chen, W. W. L. (2002), *Linear Functional Analysis*. [Disponible online: http://www.maths.mq.edu.au/~wchen/lnlfafolder/lnlfa.html.]

Cheng, K. S., y Wang, J. (1999), «The formation and merger of compact objects in central engine of active galactic nuclei and quasars: gamma-ray burst and gravitational radiation», *Astrophys., J.* **521**, 502.

Chern, S. S. (1979), *Complex Manifolds Without Potential Theory*, Springer-Verlag, Nueva York.

Chernoff, P. R., y Marsden, J. E. (1974), «Properties of infinite hamiltonian systems», *Lecture Notes in Mathematics*, vol. 425, Springer-Verlag, Berlín.

Chevalley, C. (1946), *Theory of Lie Groups*, Princeton University Press, Princeton.

Chevalley, C. (1954), *The Algebraic Theory of Spinors*, Columbia University Press, Nueva York.

Chew, G. F. (1962), *S-Matrix Theory of Strong Interactions*, Pearson Benjamin Cummings.

Choquet-Bruhat, Y., y DeWitt-Morette, C. (2000), *Analysis, Manifolds, and Physics*, partes I y II, North-Holland, Amsterdam.

Christian, J. (1995), «Definite events in NewtonCartan quantum gravity», Oxford preprint, enviado a *Phys. Rev. D.*

Christenson, J. H., Cronin, J. W., Fitch, V. L., y Turlay, R. (1964), «Evidence for the 2p decay of the K0 meson», *Phys. Rev. Lett.* **13**, 138-140.

Church, A. (1936), *The calculi of lambda-conversion*, Annals of Mathematics Studies, 6, Princeton University Press, Princeton, NJ.

Claudel, C. M., y Newman, K. P. (1998), «Isotropic Cosmological Singularities I. Polytropic Perfect Fluid Spacetimes», *Proc. R. Soc. Lond.* **454**, 1073-1107.

Clauser, J. F., Horne, M. A., Shimony, A., y Holt, R. A. (1969), «Proposed experiment to test local hidden-variable theories», *Phys. Rev. Lett.* **23**, 880.

Clifford, W. K. (1873), «Preliminary Sketch of Biquaternions», *Proc. London Math. Soc.* **4**, 381-395.

Clifford, W. K. (1878), «Applications of Grassmanns extensive algebra», *Am. J. Math.* **1**, 350-358.

Clifford, W. K. (1882), *Mathematical Papers by William Kingdon Clifford* (ed. R. Tucker), Londres.

Cohen, P. J. (1966), *Set Theory and the Continuum Hypothesis*, W. A. Benjamin, Nueva York.

Collins, P. D. B. (1977), *An Introduction to Regge Theory and High Energy Physics*, Cambridge University Press, Cambridge.

Colombeau, J. F. (1983), «A multiplication of distributions», *J. Math. Anal. Appl.* **94**, 96-115.

Colombeau, J. F. (1985), *Elementary Introduction to New Generalized Functions*, North-Holland, Amsterdam.

Connes, A. (1990), «Essay on physics and non-commutative geometry, en *The*

Interface of Mathematics and Particle Physics (eds. D. G. Quillen, G. B. Segal y Tsou S. T.), Clarendon Press, Oxford.

Connes, A. (1998), «Noncommutative differential geometry and the structure of space-time», en *The Geometric Universe; Science, Geometry, and the Work of Roger Penrose* (eds. S. A. Huggett, L. J. Mason, K. P. Tod, S. T. Tsou y N. M. J. Woodhouse), Oxford University Press, Oxford.

Connes, A., y Berberian, S. K. (1995), *Noncommutative Geometry*, Academic Press.

Connes, A., y Kreimer, D. (1998), «Hopf Algebras, Renormalization and Non-commutative Geometry» [hep-th/9808042].

Conway, J. H. (1976), *On Numbers and Games*, Academic Press, Londres.

Conway, J. H., y Kochen, S. (2002), «The geometry of the quantum paradoxes», en *Quantum [Un]speakables: From Bell to Quantum Information* (eds. R. A. Bertlmann y A. Zeilinger), Springer-Verlag, Berlín.

Conway, J. H., y Norton, S. P. (1979), «Monstrous Moonshine», *Bull. Lond. Math. Soc.* **11**, 308-339.

Conway, J. H., y Smith, D. A. (2003), *On Quaternions and Octonions*, A. K. Peters.

Corson, E. M. (1953), *Introduction to Tensors, Spinors, and Relativistic Wave-equations*, Blackie and Son Ltd., Londres.

Costa de Beauregard, O. (1995), «Macroscopic retrocausation», *Found. Phy. Lett.* **8** (3), 287-291.

Cotes, R. (1714), «Logometria», *Phil. Trans. Roy. Soc. Lond* (marzo).

Cox, D. A., y Katz, S. (1999), *Mirror Symmetry and Algebraic Geometry*, Mathematical Surveys and Monographs 68, American Mathematical Society, Providence, RI.

Cramer. J. G. (1988), «An overview of the transactional interpretation of quantum mechanics», *Int. J. Theor. Phys.* **27(2)**, 227-236.

Crowe, M. J. (1967), *A History of Vector Analysis: The Evolution of the Idea of a Vectorial System*, University of Notre Dame Press, Toronto (reimpreso con añadidos y correcciones por Dover, Nueva York, 1985).

Cvitanovič, P., y Kennedy, A. D. (1982), «Spinors in negative dimensions», *Phys. Scripta* **26**, 5-14.

Das, A., y Ferbel. T. (2004), *Introduction to Nuclear and Particle Physics*, World Scientific Publishing Company, Singapur.

Davenport, H. (1952), *The Higher Arithmetics: An Introduction to the Theory of Numbers*, Hutchinson's University Library.

Davies, M. (1997), *Europe: A History*, Oxford University Press, Oxford, pp. 89-94.

Davies, P. (2003), *How to Build a Time Machine*, Penguin, USA.

Davis, M. (1978), «What is a Computation?», en *Mathematics Today: Twelve Informal Essays* (ed. L. A. Steen), Springer-Verlag, Nueva York.

Davis, M. (1988), «Mathematical logic and the origin of modern computers», en *The Universal Turing Machine: A Half-Century Survey* (ed. R. Herken), Kammerer and Unverzagt, Hamburgo.

Davydov, A. S. (1976), *Quantum Mechanics*, Pergamon Press, Oxford.

De Bernardis, P., *et al.* (2000), «A Flat Universe from High-Resolution Maps of the Cosmic Microwave Background Radiation», *Nature* **404**, 955-959.

Delduc, F., Galperin, A., Howe, P., y Sokatchev, E. (1993), «A twistor formulation of the heterotic D = 10 superstring with manifest (8,0) worldsheet super-symmetry», *Phys. Rev.* **D47**, 578-593 [hep-th/9207050].

Derbyshire, J. (2003), *Prime Obsession: Bernhard Riemann and the Greatest Unsolved Problem in Mathematics*, Joseph Henry Press, Washington, DC.

Deser, S. (1999), «Nonrenormalizability of D = 11 supergravity» [hep-th/9905017].

Deser, S. (2000), «Infinities in quantum gravities», *Annalen Phys*, 9, 299-307 [gr-qc/9911073].

Deser, S., y Teitelboim, C. (1977), «Supergravity Has Positive Energy», *Phys. Rev. Lett.* **39**, 248-252.

Deser, S., y Zumino, B. (1976), «Consistent supergravity», *Phys. Lett.* **62B**, 335-337.

De Sitter, W. (1913), *Phys. Zeitz.*, **14**, 429 (en alemán).

Deutsch, D. (2000), *The Fabric of Reality*, Penguin, Londres (hay traducción castellana: *La estructura de la realidad*, Anagrama, Barcelona, 2002).

Devlin, K. (1988), *Mathematics: The New Golden Age*, Penguin Books, Londres.

Devlin, K. (2002), *The Millennium Problems: The Seven Greatest Unsolved Mathematical Puzzles of Our Time*, Basic Books, Londres/Perseus Books, Nueva York.

DeWitt, B. S. (1967), «Quantum Theory of Gravity. I. The Canonical Theory», *Phys. Rev.* **160**, 1113.

DeWitt, B. S. (1984), *Supermanifolds*, Cambridge University Press, Cambridge.

DeWitt, B. S., y Graham, R. D., eds. (1973), *The Many-Worlds Interpretation of Quantum Mechanics*, Princeton University Press, Princeton.

Dicke, R. H. (1961), «Dirac's Cosmology and Mach's Principle», *Nature* **192**, 440-441.

Dicke, R. H. (1981), «Interaction-free quantum measurements: A paradox?», *Am. J. Phys.* **49**, 925.

Dicke, R. H., Peebles, P. J. E., Roll, P. G., y Wilkinson, D. T. (1965), «Cosmic Black-Body Radiation», *Astrophys. J.* **142**, 414-419.

Dine, M. (2000), «Some reflections on Moduli, their Stabilization and Cosmology» [hep-th/0001157].

Diósi, L. (1984), «Gravitation and quantum mechanical localization of macro-objects», *Phys. Lett.* **105A**, 199-202.

Diósi, L. (1989), «Models for universal reduction of macroscopic quantum fluctuations», *Phys. Rev.* **A40**, 1165-1174.

Dirac, P. A. M. (1928), «The quantum theory of the electron», *Proc. Roy. Soc. Lond.* **A117**, 610-24; *ibid*, parte II, **A118**, 351-361.

Dirac, P. A. M. (1932), *Proc. Roy. Soc.* **A136**, 453.

Dirac, P. A. M. (1933), «The Lagrangian in Quantum Mechanics», *Physicalische Zeitschrift der Sowjetunion*, Band 3, Heft 1.

Dirac, P. A. M. (1936), «Relativistic Wave Equations», *Proc. R. Soc. Lond.* **A155**, 447-459.

Dirac, P. A. M. (1937), «The Cosmological Constants», *Nature* **139**, 323.

Dirac, P. A. M. (1938), «A new basis for cosmology», *Proc. R. Soc. Lond.* **A165**, 199.

Dirac, P. A. M. (1950), «Generalized Hamiltonian dynamics», *Can. J. Math.* **2**, 129.

Dirac, P. A. M. (1964), *Lectures on Quantum Mechanics*, Yeshiva University, Nueva York.

Dirac, P. A. M. (1966), *Lectures in Quantum Field Theory*, Academic Press, Nueva York.

Dirac, P. A. M. (1982a), *The Principles of Quantum Mechanics*, 4.ª ed., Clarendon Press, Oxford (hay una traducción castellana de la tercera edición inglesa: *Principios de mecánica cuántica*, Ariel, Barcelona, 1958, agotada).

Dirac, P. A. M. (1982b), «Pretty mathematics», *Int. J. Theor. Phys.* **21**, 603-605.

Dirac, P. A. M. (1983), «The Origin of Quantum Field Theory», en *The Birth of Particle Physics* (eds. Brown y Hoddeson), Cambridge University Press, Nueva York.

Dixon, G. (1994), *Division Algebras, Quaternions, Complex Numbers and the Algebraic Design of Physics*, Kluwer Academic Publishers, Boston.

Dodelson, S. (2003), *Modern Cosmology*, Academic Press, Londres.

Dolan, L. (1996), «Superstring twisted conformal field theory: Moonshine, the Monster, and related topics» (South Hadley, MA, 1994), *Contemp. Math.* **193**, 9-24.

Domagala, M., y Lewandowski, J. (2004), «Black Hole Entropy from Quantum Geometry» [gr-qc/0407041].

Donaldson, S. K., y Kronheimer, P. B. (1990), *The Geometry of Four-Manifolds*, Oxford University Press, Oxford.

Douady, A., y Hubbard, J. (1985), «On the dynamics of polynomial-like map-pings», *Ann. Sci. Ecole Norm. Sup.* **18**, 287-343.

Dowker, F., y Kent, A. (1996), «On the consistent histories approach to quantum mechanics», *J. Stat. Phys.* **82** [gr-qc/9412067].

Drake, S. (1957), *Discoveries and Opinions of Galileo*, Doubleday, Nueva York.

Drake, S., trad. (1953), *Galileo Galilei: Dialogue Concerning the Two Chief World SystemsPtolemaic and Copernican*, University of California, Berkeley (hay una edición castellana de los diálogos de Galileo: *Diálogo sobre los dos máximos sistemas del mundo*, edición y traducción de Antonio Beltrán Marí, Alianza, Madrid, 1995).

Dray, T., y Manogue, C. A. (1999), «The Exceptional Jordan Eigenvalue Problem», *Int. J. Theor. Phys.* **38**, 2901-16 [math-ph/99110004].

Dreyer, O., Markopoulou, F, y Smolin, L. (2004), «Symmetry and entropy of black hole horizons» [hep-th/0409056].

Duffin, R. J. (1938), «On the characteristic matrices of covariant systems», *Phys. Rev.* **54**, 1114.

Dunajski, M. (2002), «Anti-self-dual four-manifolds with a parallel real spinor», *R. Soc. Lond. Proc. Ser. A Math. Phys. Eng. Sci.* **458(2021)**, 1205-1222.

Du Sautoy, M. (2004), *The Music of the Primes*, Perennial, Nueva York.

Dunham, W. (1999), *Euler: The Master of Us All*, Math. Assoc. Amer., Washington, DC.

Dyson, F. J. (1966), *Symmetry Groups in Nuclear and Particle Physics: A Lecture-note and Reprint Volume*, W. A. Benjamin, Nueva York.

Eastwood, M. G., Penrose, R., y Wells, R. O., Jr. (1981), «Cohomology and massless fields», *Comm. Math. Phys.* **78**, 305-351.

Eddington, A. S. (1929a), *The Nature of the Physical World*, Cambridge University Press, Cambridge (hay traducción castellana: *La naturaleza del mundo físico*, Editorial Sudamericana, Buenos Aires, 1945, agotada).

Eddington, A. S. (1929b), «A Symmetrical Treatment of the Wave Equation», *Proc. R. Soc. Lond.* **A121**, 524-542.

Eddington, A. S. (1946), *Fundamental Theory*, Cambridge University Press, Cambridge.

Eden, R. J., Landshoff, P.V., Olive, D. I., y Polkinghorne, J. C. (2002), *The Analityc S-Matrix*, Cambridge University Press.

Edwards, C. H., y Penney, D. E. (2002), *Calculus with Analytic Geometry*, Prentice Hall, 6.ª ed.

Ehrenberg, W., y Siday, R. E. (1949), «The refractive index in electron optics and the principles of dynamics», *Proc. Phys. Soc.* **LXIIB**, 8-21.

Ehrenfest, P., y Ehrenfest, T. (1959), *The Conceptual Foundations of the Statistical Approach in Mechanics*, Cornell University Press, Ithaca, NY.

Eilenberg, S., y Mac Lane, S. (1945), «General theory of natural equivalences, *Trans. Am. Math. Soc.* **58**, 231-294.

Einstein, A. (1914), en Lorentz, *et al.* (1952).

Einstein, A. (1917), «Kosmologische Betrachtungen zur allgemeinen Relativitätstheorie», *Sitzungsberichte der Preussischen Akademie der Wissenschaften*, 142-152.

Einstein, A. (1925), «S. B. Preuss», *Akad. Wiss.* 22, 414.

Einstein, A. (1945), «A generalization of the relativistic theory of gravitation», *Ann. Math.* **46**, 578.

Einstein, A. (1948), «A generalized theory of gravitation», *Rev. Mod. Phys.* **20**, 35.

Einstein, A. (1955), «Relativistic theory of the non-symmetric field», en *Appendix II: The Meaning of Relativity*, 5.ª ed., pp. 133-66, Princeton University Press, Princeton, NJ.

Einstein, A., y Kaufman, B. (1955), «A new form of the general relativistic field equations», *Ann. Math.* **62**, 128.

Einstein, A., y Straus, E. G. (1946), «A generalization of the relativistic theory of gravitation II», *Ann. Math.* **47**, 731.

Einstein, A., Podolsky, P., y Rosen, N. (1935), «Can quantum-mechanical description of physical reality be considered complete?», en *Quantum Theory and Measurement* (eds. J. A. Wheeler y W. H. Zurek), Princeton University Press, Princeton, New Jersey, 1983; originalmente en *Phys. Rev.* **47**, 777-780.

Elitzur, A. C., y Vaidman, L. (1993), «Quantum mechanical interaction-free measurements», *Found. Phys.* **23**, 987-997.

Elliott, J. P., y Dawber, P. G. (1984), *Symmetry in Physics*, vol. 1. Macmillan, Londres.

Ellis, J., Mavromatos, N. E., y Nanopoulos, D.V. (1997a), «Vacuum fluctuations and decoherence in mesoscopic and microscopic systems», en *Symposium on Flavour-Changing Neutral Currents: Present and Future Studies*, UCLA.

Ellis, J., Mavromatos, N. E., y Nanopoulos, D.V. (1997b), «Quantum decoherence in a D-foam background», *Mod. Phys. Lett.* **A12**, 2029-2036.

Engelking, E. (1968), *Outline of General Topology*, North-Holland & PWN, Amsterdam.

Euler, L. (1748), *Introductio in Analysis Infinitorum*.

Everett, H. (1957), «Relative State formulation of quantum mechanics», en *Quantum Theory and Measurement* (eds. J. A. Wheeler y W. H. Zurek), Princeton University Press, Princeton, New Jersey, 1983; originalmente en *Rev. Mod. Phys.* **29**, 454-262.

Fauvel, J., y Gray, J. (1987), *The History of Mathematics: A Reader*, Macmillan, Londres.

Ferber, A. (1978), «Supertwistors and conformal supersymmetry», *Nucl. Phys.* **B132**, 55-64.

Fernow, R. C. (1989), *Introduction to Experimental Particle Physics*, Cambridge University Press, Cambridge.

Feynman, R. P. (1948), «Space-time approach to nonrelativistic quantum mechanics», *Rev. Modern Phys.* **20**, 367-387.

Feynman, R. P. (1949), «The theory of positrons», *Phys. Rev.* **76**, 749.

Feynman, R. P. (1987), *Elementary Particles and the Laws of Physics: The 1986 Dirac Memorial Lectures*, Cambridge University Press, Cambridge.

Feynman, R. P., y A. Hibbs. (1965), *Quantum Mechanics and Path Integrals*, McGraw-Hill, Nueva York.

Fierz, M. (1938), «Uber die Relativitische Theorie kräftefreier Teichlen mit belie-bigem Spin», *Helv. Phys. Acta.* **12**, 3-37.

Fierz, M. (1940), «Uber den Drehimpuls von Teichlen mit Ruhemasse null und beliebigem Spin», *Helv. Phys. Acta.* **13**.,45-60.

Fierz, M., y Pauli, W. (1939), «On relativistic wave equations for particles of arbitrary spin in an electromagnetic field», *Proc. Roy. Soc. Lond.* **A173**, 211-232.

Finkelstein, D. (1969), «Space-time code», *Phys. Rev.* **184**, 1261-1279.

Finkelstein, D., y J. Rubinstein (1968), «Connection between spin, statistics, and kinks», *J. Math. Phys.* **9**, 1972.

Flanders, H. (1963), *Differential Forms*, Academic Press (reeditado por Dorf, 1989).

Floyd, R. M., y Penrose, R. (1971), «Extraction of Rotational Energy from a Black Hole», *Nature Phys. Sci.* **229**, 177.

Fortney, L. R. (1997), *Principles of Electronics, Analogic and Digital*, Harcourt Brace Jovanovich.

Frankel, T. (2001), *The Geometry of Physics*, Cambridge University Press, Cambridge.

Frenkel, A. (2000), «A Tentative Expresion of the Károlyházy Uncertainty of the Space-time Structure through Vacuum Spreads in Quantum Gravity» [quant-ph/0002087].

Friedlander, F. G. (1982), *Introduction to the Theory of Distributions*», Cambridge University Press, Cambridge.

Friedrich, W. L. (2000), *Fire in the Sea: The Santorini Volcano: Natural History and the Legend of Atlantis* (trad. A. R. McBirney), Cambridge University Press, Cambridge.

Frittelli, S., Kozameh, C., y Newman, E. T. (1997), «Dynamics of light cone cuts at null infinity», *Phys. Rev.* **D56**, 8.

Fröhlich, J., y Pedrini, B. (2000), «New applications of the chiral anomaly», en *Mathematical Physics 2000* (eds. A. Fokas, A. Grigoryan, T. Kibble, y B. Zegarlinski), pp. 947, Imperial College Press, Londres.

Gambini, R., y Pullin, J. (1999), «Nonstandard optics from quantum spacetime», *Phys. Rev.* **D59** 124021.

Gandy, R. (1988), «The confluence of ideas in 1936», en *The Universal Turing Machine: A Half-Century Survey* (ed. R. Herken), Kammerer and Unverzagt, Hamburgo.

Gangui, A. (2003), «Cosmology from Topological Defects», *AIP Conf. Proc.* **668** [astro-ph/0303504].

Gardner, M. (1990), *The New Ambidextrous Universe*, W. H. Freeman, Nueva York (hay traducción castellana: *El nuevo universo ambidiestro*, RBA Coleccionables, Barcelona, 1994).

Gauss, C. F. (1900), *Werke*, vol. VIII, pp. 357-362, Leipzig.

Gelfand, I., y Shilov, G. (1964), *Generalized Functions*, vol. 1, Academic Press, Nueva York.

Gell-Mann, M. (1994), *The Quark and the Jaguar: Adventures in the Simple and the Complex*, W. H. Freeman, Nueva York (hay traducción castellana: *El Quark y el Jaguar: aventuras en lo simple y lo complicado*, Tusquets, Barcelona, 1995).

Gell-Mann, M., y Hartle, J. B. (1995), «Strong Decoherence», en *Proceedings of the 4th Drexel Conference on Quantum Non-Integrability: The Quantum-Classical Correspondence* (eds. D.-H. Feng y B.-H. Lu), International Press of Boston, Hong Kong, 1988 [gr-qc/9509054].

Gell-Mann, M., y Neeman, Y. (2000), *Eightfold Way*, Perseus Publishing.

Geroch, R. (1968), «Spinor structure of space-times in general relativity I», *J. Math. Phys.* **9**, 1739-1744.

Geroch, R. (1970), «Spinor structure of space-times in general relativity II», *J. Math. Phys.* **11**, 343-348.

Geroch, R. (1984), «The Everett Interpretation, *Nous*, 18, 617-633.

Geroch, R. (inédito), *Geometrical Quantum Mechanics*, Lecture notes given at University of Chicago.

Geroch, R. y Hartle, J. (1986), «Computability and physical theories», *Found. Phys.* **16**, 533.

Geroch, R., Kronheimer, E. H., y Penrose, R. (1972), «Ideal points for space-times», *Proc. Roy. Soc. Lond.* **A347**, 545-567.

Ghirardi, G. C., Grassi, R., y Rimini, A. (1990), «Continuous spontaneous reduction model involving gravity», *Phys. Rev.* **A42**, 1057-1064.

Ghirardi, G. C., Rimini, A., y Weber, T. (1986), «Unified dynamics for microscopic and macroscopic systems», *Phys. Rev.* **D34**, 470.

Gibbons, G. W. (1984), «The isoperimetric and Bogomolny inequalities for black holes», en *Global Riemannian Geometry* (eds. T. Willmore y N. J. Hitchin), Ellis Horwood, Chichester.

Gibbons, G. W. (1997), «Collapsing Shells and the Isoperimetric Inequality for Black Holes», *Class. Quant. Grav.* **14**, 2905-2915 [hep-th/9701049].

Gibbons, G. W., y Hartnoll, S. A. (2002), «Gravitational instability in higher dimensions» [hep-th/0206202].

Gibbons, G. W., y Perry, M. J. (1978), «Black Holes and Thermal Green's Function», *Proc. Roy. Soc. Lond.* **A358**, 467-494.

Gibbs, J. (1960), *Elementary Principles in Statistical Mechanics*, Dover, Nueva York.

Gindikin, S. G. (1986), «On one construction of hyperkähler metrics», *Funct. Anal. Appl.* **20**, 82-83 (en ruso).

Gindikin, S. G. (1990), «Between integral geometry and twistors», en *Twistors in Mathematics and Physics* (eds. T. N. Bailey y R. J. Baston), LMS Lecture Note Series 156, Cambridge University Press, Cambridge.

Gisin, N. (1989), «Stochastic quantum dynamics and relativity», *Helv. Phys. Acta.* **62**, 363.

Gisin, N. (1990), *Phys. Lett.* **143A**, 1.

Gisin, N., de Riedmatten, H., Scarani, V., Marcikic, I., Acin, A., Tittel, W., y Zbinden, H. (2004), «Two independent photon pairs versus four-photon entangled states in parametric down conversion», *J. Mod. Opt.* **51**, 1637 [quant-ph/0310167].

Glashow, S. (1959), «The renormalizability of vector meson interactions», *Nucl. Phys.* **10**, 107.

Gleason, A. M. (1957), «Measures on the Closed Subspaces of a Hilbert Space», *J. Math. and Mech.* **6**, 885-893.

Gleason, A. M. (1990), en *More Mathematical People* (eds. D. J. Albers, G. L. Alexanderson, y C. Reid), Harcourt Brace Jovanovich, Boston, p. 94.

Goddard, P., *et al.* (1973), «Quantum dynamics of a massless, relativistic string», *Nucl. Phys.* **B56**, 109.

Gold, T. (1962), «The Arrow of Time», *Am. J. Phys.* **30**, 403.

Goldberg, J. N., Macfarlane, A. J., Newman, E. T., Rohrlich, F., y Sudarshan, E. C. G. (1967), «Spin-s spherical harmonics and eth», *J. Math. Phys.* **8**, 2155-2161.

Goldblatt, R. (1979), *Topoi: The Categorial Analysis of Logic*, North-Holland Publishing Company, Oxford y Nueva York.

Goldstein, S. (1987), «Stochastic mechanics and quantum theory», *J. Stat. Phys.* **47**.

Gottesman, D., y Preskill, J. (2003), «Comment on The black hole final state» [hep-th/0311269].

Gouvea, F. Q. (1993), *P-Adic Numbers: An Introduction*, Springer-Verlag; 2nd edition (2000), Berlín y Nueva York.

Grassmann, H. G. (1844), *Die lineare Ausdehnungslehre*, 4.ª ed., Springer-Verlag.

Grassmann, H. G. (1862), *Die lineare Ausdehnungslehre Vollständig und in strenger Form bearbeitit.*

Gray, J. (1979), *Ideas of Space: Euclidean, Non-Euclidean, and Relativistic*, Oxford University Press, Oxford (hay traducción castellana: *Ideas de espacio*, Mondadori, Barcelona, 1992).

Green, M. B. (2000), «Superstrings and the unification of physical forces», en *Mathematical Physics 2000* (eds. A. Fokas, T. W. B. Kibble, A. Grigouriou, y B. Zegarlinski), pp. 59-86, Imperial College Press, Londres.

Green, M. B., Schwarz, J. H., y Witten, E. (1987), *Superstring Theory*, vols. I y II, Cambridge University Press, Cambridge.

Green, M. B., Schwarz, J. H., y Witten, E. (1988), *Superstring Theory*, Cambridge University Press, Cambridge.

Greene, B. (1999), *The Elegant Universe; Superstrings, Hidden Dimensions, and the Quest for the Ultimate Theory*, Random House, Londres (hay traducción castellana: *El universo elegante: supercuerdas, dimensiones ocultas y la búsqueda de una teoría definitiva*, Crítica, Barcelona, 2005).

Griffiths, P., y Harris, J. (1978), *Principles of Algebraic Geometry*, John Wiley & Sons, Nueva York.

Grishchuk, L. P., *et al.* (2001), «Gravitational Wave Astronomy: in Anticipation of First Sources to be Detected», *Phys. Usp.* **44**, 1-51 [astro-ph/0008481].

Groemer, H. (1996), *Geometric Applications of Fourier Series and Spherical Harmonics*, Cambridge University Press, Cambridge.

Gross, D. J., y Periwal, V. (1988), «String Perturbation Theory Diverges», *Phys. Rev. Lett.* **60**, 2105.

Gross, M. W., Huybrechts, D., Joyce, D., y Winkler, G. D. (2003), *Calabi-Yau Manifolds and Related Geometries*, Springer-Verlag.

Grosser, M., Kunzinger, M., Oberguggenberger, M., y Steinbauer, R. (2001), *Geometric Theory of Generalized Functions with Applications to General Relativity*, Kluwer Academic Publishers, Boston y Dordrecht, Países Bajos.

Guenther, D. B., Krauss, L. M., y Demarque, P. (1998), «Testing the Constancy of the Gravitational Constant Using Helioseismology», *Astrophys. J.* **498**, 871-876.

Gunning, R. C., y Rossi, H. (1965), *Analytic Functions of Several Complex Variables*, Prentice-Hall, Englewood Cliffs, New Jersey.

Gürsey, F. (1983), «Quaternionic and octonionic structures in physics: episodes in the relation between physics and mathematics», *Symm. Phys. (1600-1980)*, pp. 557-592, Sant Feliu de Guíxols. Univ. Autónoma Barcelona, Barcelona, 1987.

Gürsey, F., y Tze, C.-H. (1996), *On the Role of Division, Jordan, and Related Algebras in Particle Physics*, World Scientific, Singapur.

Gurzadyan, V. G., *et al.* (2002), «Ellipticity analysis of the BOOMERANG CMB maps», *Int. J. Mod. Phys.* **D12**, 1859-1874 [astro-ph/0210021].

Gurzadyan, V. G., *et al.* (2003), «Is there a common origin for the WMAP low multipole and for the ellipticity in BOOMERANG CMB maps?» [astro-ph/0312305].

Gurzadyan, V. G., *et al.* (2004), «WMAP confirming the ellipticity in BOOMERANG and COBE CMB maps» [astro-ph/0402399].

Gurzadyan, V. G., y Kocharyan, A. A. (1992), «On the problem of isotropization of cosmic background radiation», *Astron. Astrophys.* **260**, 14.

Gurzadyan, V. G., y Kocharyan, A. A. (1994), *Paradigms of the Large-Scale Universe*, Gordon and Breach, Lausana, Suiza.

Gurzadyan, V. G., y Torres, S. (1997), «Testing the effect of geodesic mixing with COBE data to reveal the curvature of the universe», *Astron. and Astrophys.* **321**, 19-23 [astro-ph/9610152].

Guth, A. (1997), *The Inflationary Universe*, Jonathan Cape, Londres (hay traducción castellana: *El universo inflacionario: la búsqueda de una nueva teoría sobre los orígenes del cosmos*, Debate, Barcelona, 1999).

Haag, R. (1992), *Local Quantum Physics: Fields, Particles, Algebras*, Springer-Verlag, Berlín.

Haehnelt, M. G. (2003), «Joint Formation of Supermassive Black Holes and Galaxies», en *Carnegie Observatories Astrophysics Series*, vol. 1: *Coevolution of*

Black Holes and Galaxies (ed. L. C. Ho), Cambridge University Press, Cambridge [astro-ph/0307378].

Halverson, N. W. (2001), «DASI First Results: A Measurement of the Cosmic Microwave Background Angular Power Spectrum» [astro-ph/0104489].

Halzen, F., y Martin, A. D. (1984), *Quarks and Leptons: An Introductory Course in Modern Particle Physics*, John Wiley & Sons, Nueva York.

Hameroff, S. R. (1998), «Fundamental geometry: the Penrose Hameroff Orch OR model of consciousness», en *The Geometric Universe; Science, Geometry, and the Work of Roger Penrose* (eds. S. A. Huggett, L. J. Mason, K. P. Tod, S. T. Tsou, y N. M. J. Woodhouse), Oxford University Press, Oxford.

Hameroff, S. R. (1987), *Ultimate Computing. Biomolecular Consciousness and Nano-Technology*, North-Holland, Amsterdam.

Hameroff, S. R., y Penrose, R. (1996), «Conscious events as orchestrated space-time selections», *J. Consc. Stud.* **3**, 36-63.

Hameroff, S. R., y Watt, R. C. (1982), «Information processing in microtubules», *J. Theor. Biol.* **98,** 549-561.

Hamilton, E. (1999), *Mythology: Timeless Tales of Gods and Heroes*, Warner Books, Nueva York.

Han, M. Y., y Nambu, Y. (1965), «Three-Triplet Model with Double SU(3) Symmetry», *Phys. Rev.* **139**, B1006-1010.

Hanany, S., *et al.* (2000), «MAXIMA-1: A Measurement of the Cosmic Microwave Background Anisotropy on angular scales of 10 arcminutes to 5 degrees», *Astrophys. J.* **545**, L5.

Hanbury Brown, R., y Twiss, R. Q. (1954), «A new type of interferometer for use in radio astronomy», *Phil. Mag.* **45**, 663-682.

Hanbury Brown, R., y Twiss, R. Q. (1956), «Correlation between photons in 2 coherent beams of light», *Nature* **177**.

Hannabuss, K. (1997), *An Introduction to Quantum Theory*, Oxford University Press, Oxford.

Hansen, B. M. S., y Murali, C. (1998), «Gamma Ray Bursts from Stellar Collisions» [astro-ph/9806256].

Hansen, R.O., Newman, E.T., Penrose, R., y Tod, K. P. (1978), «The metric and curvature properties of H-space», *Proc. Roy. Soc. Lond.* **A363**, 445-468.

Hardy, G. H. (1914), *A Course of Pure Mathematics*, 2.ª ed., Cambridge University Press, Cambridge.

Hardy, G. H. (1940), *A Mathematician's Apology*, Cambridge University Press, Cambridge (hay traducción castellana: *Apología de un matemático*, Nivola Libros y Ediciones, Madrid, 1999).

Hardy, G. H. (1949), *Divergent Series*, Oxford University Press, Oxford.

Hardy, G. H., y Wright, E. M. (1945), *An Introduction to the Theory of Numbers*, 2.ª ed. Clarendon Press, Oxford.

Hardy, L. (1992), «Quantum mechanics, local realistic theories, and Lorentz-invariant realistic theories», *Phys. Rev. Lett.* **68**, 2981 [/astract/PRL/v68/i20/p2981_1].

Hardy, L. (1993), «Nonlocality for two particles without inequalities for almost all entangled states», *Phys. Rev. Lett.* **71(11)**, 1665.

Hartle, J. B. (2002), *Gravity: An Introduction to Einsteins General Relativity*. Addison-Wesley, San Francisco, CA y Londres.

Hartle, J. B. (2004), «The Physics of "Now"» [gr-qc/0403001].

Hartle, J. B., y Hawking, S. W. (1983), «The wave function of the Universe», *Phys. Rev.* **D28**, 2960.

Harvey, F. R. (1966), «Hyperfunctions and linear differential equations», *Proc. Nat. Acad. Sci.* **5**, 1042-1046.

Harvey, F. R. (1990), *Spinors and Calibrations*, Academic Press, San Diego, CA.

Haslehurst, L., y Penrose, R. (2001), «The most general (2,2) self-dual vacuum: a googly approach», en *Further Advances in Twistor Theory*, vol. III: *Curved Twistor Spaces* (eds. L. J. Mason, L. P. Hughston, P. Z. Kobak y K. Pulvere), pp. 34-59.

Hawking, S. W. (1972), «Black holes in general relativity», *Commun. Math. Phys.* **25**, 152-166.

Hawking, S. W. (1974), «Black hole explosions», *Nature* **248**, 30.

Hawking, S. W. (1975), «Particle creation by black holes», *Commun. Math. Phys.* **43**.

Hawking, S. W. (1976a), «Black holes and thermodynamics», *Phys. Rev.* **D13(2)**, 191.

Hawking, S. W. (1976b), «Breakdown of predictability in gravitational collapse», *Phys. Rev.* **D14**, 2460.

Hawking, S. W., King, A. R., y McCarthy, P. J. (1976), «A new topology for curved space-time which incorporates the causal, differential, and conformal structures», *J. Math. Phys.* **17**, 174-181.

Hawking, S. W., y Ellis, G. F. R. (1973), *The Large-Scale Structure of Space-Time*, Cambridge University Press, Cambridge.

Hawking, S. W., e Israel, W., ed. (1987), *300 Years of Gravitation*, Cambridge University Press, Cambridge.

Hawking, S. W., y Penrose, R. (1970), «The singularities of gravitational collapse and cosmology», *Proc. Roy. Soc. Lond.* **A314**, 529-548.

Hawking, S. W., y Penrose, R. (1996), *The Nature of Space and Time*, Princeton University Press, Princeton, New Jersey (hay traducción castellana: *La naturaleza del espacio y el tiempo*, Debate, Barcelona, 1996).

Hawking, S. W., y Turok, N. (1998), «Open Inflation Without False Vacua», *Phys. Lett.* **B425** [hep-th/9802030].

Hawkins, T. (1977), «Weierstrass and the theory of matrices», *Arch. Hist. Exact Sci.* **17**, 119-163.

Hawkins, T. (2000), *Emergence of the Theory of Lie Groups*, Springer-Verlag, Nueva York.

Heisenberg, W. (1971), *Physics and Beyond*, Addison Wesley, Londres (hay traducción castellana: *La parte y el todo: conversando en torno a la física atómica*, Ellago Ediciones, Castellón, 2004).

Heisenberg, W. (1989), «What is an elementary particle?», en *Encounters with Einstein*, Princeton University Press, Princeton.

Helgason, S. (2001), *Differential Geometry and Symmetric Spaces*, AMS Chelsea Publishing, Providence, RI.

Hestenes, D. (1990), «The Zitterwebegung Interpretation of Quantum Mechanics», *Found. Physics.* **20** (10), 1213-1232.

Hestenes, D., y Sobczyk, G. (1999), *Clifford Algebra to Geometric Calculus: A Unified Language for Mathematics and Physics*, Reidel, Dordrecht, Holanda.

Heyting, A. (1956), *Intuitionism, Studies in Logic and the Foundations of Mathematics*, North-Holland, Amsterdam.

Heywood, P., y Redhead, M. I. G. (1983), «Non-locality and the Kochen-Specker paradox», *Found. Phys.* **13** (5), 481-499.

Hicks, N. J. (1965), *Notes on Differential Geometry*, Van Nostrand, Princeton.

Hirschfeld, J. W. P. (1998), *Projective Geometries over Finite Fields*, 2.ª ed., Clarendon Press, Oxford.

Hodges, A. P. (1982), «Twistor diagrams», *Physica*, **114A**, 157-175.

Hodges, A. P. (1985), «A twistor approach to the regularization of divergences», *Proc. Roy. Soc. Lond.* **A397**, 341-374. «Mass eigenstatates in twistor theory», *ibid*, 375-396.

Hodges, A. P. (1990a), «String Amplitudes and Twistor Diagrams: An Analogy», en *The Interface of Mathematics and Particle Physics* (eds. D. G. Quillen, G. B. Segal y S. T. Tsou), Oxford University Press, Oxford.

Hodges, A. P. (1990b), «Twistor diagrams and Feynman diagrams», en *Twistors in Mathematics and Physics, LMS Lect. Note Ser.* 156 (eds. T. N. Bailey y R. J. Baston), Cambridge University Press, Cambridge.

Hodges, A. P. (1998), «The twistor diagram programme», en *The Geometric Universe; Science, Geometry, and the Work of Roger Penrose* (eds. S. A. Huggett, L. J. Mason, K. P. Tod, S. T. Tsou y N. M. J. Woodhouse), Oxford University Press, Oxford.

Hodges, A. P., Penrose, R., y Singer, M. A. (1989), «A twistor conformal field theory for four space-time dimensions», *Phys. Lett.* **B216**, 48-52.

Hollands, S., y Wald, R. M. (2001), «Local Wick Polynomials and Time Ordered Products of Quantum Fields in Curved Spacetime», *Commun. Math. Phys.* **223**, 289-326 [gr-qc/0103074].

Home, D. (1997), *Conceptual Foundations of Quantum Physics: An Overview from Modern Perspectives*, Plenum Press, Nueva York y Londres.

Hopf, H. (1931), «Über die Abbildungen der dreidimensionalen Sphäre auf die Kugelfläche», *Math. Ann.* **104**, 637.

Horgan, J. (1996), *The End of Science*, Perseus Publishing, Nueva York (hay traducción castellana: *El fin de la ciencia*, Paidós Ibérica, Barcelona, 1998).

Horowitz, G. T. (1998), «Quantum states of black holes», en *Black Holes and Relativistic Stars* (ed. R. M. Wald), pp. 241-266, University of Chicago Press, Chicago.

Horowitz, G. T, y Maldacena, J. (2003), «The black hole final state» [hep-th/0310281].

Horowitz, G. T., yPerry, M. J. (1982), «Gravitational energy cannot become negative», *Phys. Rev. Lett.* **48**, 371-374.

Howie, J. (1989), «On the SQ-universality of T(6)-groups», *Forum Math.* **1**, 251-272.

Hoyle, C. D., *et al.* (2001), «Submillimeter Test of the Gravitational Inverse-Square Law: A Search for Large Extra Dimensions», *Phys. Rev. Lett.* **86** (8), 1418-1421.

Hoyle, F. (1948), «A New Model for the Expanding Universe», *Mon. Not. Roy. Astron. Soc.* **108**, 372.

Hoyle, F., Fowler, W. A., Burbidge, G. R., y Burbidge, E. M. (1956), «Origin of the elements in stars», *Science* **124**, 611-614.

Huang (1949), «On the zitterbewegung of the electron», *Am. J. Phys.* **47**, 797.

Huggett, S. A., y Jordon, D. (2001), *A Topological Aperitif*, Springer-Verlag, Londres.

Huggett, S. A., y Tod, K. P. (2001), *An Introduction to Twistor Theory*, Cambridge University Press, Cambridge.

Hughston, L. P. (1979), «Twistors and Particles», *Lecture Notes in Physics* 97, Springer-Verlag, Berlín.

Hughston, L. P. (1995), «Geometric Aspects of Quantum Mechanics», en *Twistor Theory* (ed. S. A. Huggett), pp. 59-79, Marcel Dekker, Nueva York.

Hughston, L. P., Jozsa, R., y Wooters, W. K. (1993), «A complete classification of quantum ensembles having a given density matrix», *Phys. Letts.* **A183**, 14-18.

Ilyenko, K. (1999), *Twistor Description of Null Strings*, Oxford, Tesis Doctoral no publicada.

Immirzi, G. (1997), «Quantum Gravity and Regge Calculus» [gr-qc/9701052].

Infeld, L., y van der Waerden, B. L. (1933), «Die Wellengleichung des Elektrons in der allgemeinen Relativitätstheorie», *Sitz. Ber. Preuss. Akad. Wiss. Phisik. Math. Kl.* 9, 380-401.

Isenberg, J., Yasskin, P. B., y Green, P. S. (1978), «Non-self-dual gauge fields», *Phys. Lett.* **78B**, 462-464.

Isham, C. J. (1975), *Quantum Gravity: An Oxford Symposium*, Oxford University Press, Oxford.

Isham, C. J. (1992), «Canonical Quantum Gravity and the Problem of Time» [gr-qc/9210011].

Isham, C. J., y Butterfield, J. (2000), «Some Possible Roles for Topos Theory in Quantum Theory and Quantum Gravity» [gr-qc/9910005].

Israel, W. (1967), «Event horizons in static vacuum space-times», *Phys. Rev.* **164**, 1776-1779.

Jackson, J. D. (1998), *Classical Electrodynamics*, John Wiley & Sons, Nueva York y Chichester.

Jennewein, T., Weihs, G., Pan, J., y Zeilinger, A. (2002), «Experimental Non-locality Proof of Quantum Teleportation and Entanglement Swapping», *Phys. Rev. Lett.* **88**, 017903.

Jensen, G. (1973), «Einstein Metrics on Principal Fibre Bundles», *J. Diff. Geom.* **8**, 599-614.

Johnson, C. (2003), *D-Branes*, Cambridge University Press, Cambridge.

Jones, H. F. (2002), *Groups, Representations, and Physics*, Institute of Physics Publishing, Bristol.

Jozsa, R. (1981), *Models in Categories and Twistor Theory*, Oxford, tesis doctoral no publicada.

Jozsa, R. O. (1998), «Entanglement and quantum computation», en *The Geometric Universe* (eds. S. A, Huggett, L. J. Mason, K. P. Tod, S. T. Tsou y N. M. J. Woodhouse), pp. 369-379, Oxford University Press, Oxford.

Jozsa, R., y Linden, N. (2002), «On the role of entanglement in quantum computational speed-up» [quant-ph/0201143].

Kahn, D. W. (1995), *Topology: An Introduction to the Point-Set and Algebraic Areas*, Dover Publications, Nueva York.

Kaku, M. (1993), *Quantum Field Theory: A Modern Introduction*, Oxford University Press, Oxford.

Kamberov, G., *et al.* (2002), *Quaternions, Spinors and Surfaces (Contemporary Mathematics (American Mathematical Society), v. 299)*, American Mathematical Society.

Kamberov, G., *et al.* (2002), *Quaternions, Spinors, and Surfaces* (Contemporary Mathematics (American Mathematical Society), v. 299), American Mathematical Society, Providence, RIO.

Kane, G., ed. (1999), *Perspectives on Supersymmetry* (Advanced Series on Directions in High Energy Physics), World Scientific Pub. Co, Singapur.

Kane, G. (2001), *Supersymmetry: Unveiling the Ultimate Laws of Nature*, Perseus Publishing, Nueva York.

Kapusta, J. I. (2001), «Primordial Black Holes and Hot Matter» [astro-ph/0101515].

Károlyházy, F. (1966), «Gravitation and quantum mechanics of macroscopic bodies», *Nuovo Cim.* **A42**, 390.

Károlyházy, F. (1974), «Gravitation and Quantum Mechanics of Macroscopic Bodies», *Magyar Fizikai Folyóirat.* **22**, 23-24 [tesis en húngaro].

Károlyházy, F., Frenkel, A., y Lukács, B. (1986), «On the possible role of gravity on the reduction of the wave function», en *Quantum Concepts in Space and Time* (eds. R. Penrose y C. J. Isham), pp. 109-28, Oxford University Press, Oxford.

Kasper, J. E., y Feller, S. A. (2001), *The Complete Book of Holograms: How They Work and How to Make Them*, Dover Publications.

Kauffman, L. H. (2001), *Knots and Physics*, World Scientific Publishing, Singapur.

Kay, B. S. (1998a), «Entropy defined, entropy increase and decoherence understood, and some black hole puzzles solved» [hep-th/9802172].

Kay, B. S. (1998b), «Decoherence of Macroscopic Closed Systems within NewtonianQuantum Gravity», *Class. Quant. Grav.* **15**, L89-98 [hep-th/9810077].

Kay, B. S. (2000), «Application of linear hyperbolic PDE to linear quantum in curved space-times: especially black holes, time machines, and a new semilocal vacuum concept», en *Journel des équations aux Dérivées Partielles, Nantes 5-9 Juin 2000*, Groupement de Recherche 1151 du CNRS [gr-qc/0103056].

Kay, B. S., Radzikowski, M. J., y Wald, R. M. (1996), «Quantum Field Theory on Spacetimes with a Compactly Generated Cauchy Horizon», *Commun. Math. Phys.* **183** (1997), 533-556 [gr-qc/9603012].

Kay, B. S., y Wald, R. M. (1991), «Theorems on the uniqueness and thermal properties of stationary, nonsingular, quasifree states on space-times with a bifurcate Killing horizon», *Phys. Rept.* **207**, 49-136.

Kelley, J. L. (1965), *General Topology*, Van Nostrand, Princeton, New Jersey.

Kemmer, N. (1938), «Quantum theory of Einstein-Bose particles and nuclear interaction», *Proc. R. Soc.* **A166**, 127.

Kemmer, N. (1939), «The particle aspect of meson theory», *Proc. R. Soc.* **A173**, 91.

Kerr, R. P. (1963), «Gravitational field of a spinning mass as an example of algebraically special metrics», *Phys. Rev. Lett.* **11**, 237-238.

Ketov, S. V. (2000), *Quantum Non-Linear Sigma-Models: From Quantum Field Theory to Supersymmetry, Conformal Field Theories, Black Holes, and Strings*, Springer-Verlag, Berlín y Londres.

Kibble, T. W. B. (1961), «Lorentz invariance and the gravitational field», *J. Math. Phys.* **2**, 212-221.

Kibble, T. W. B. (1979), «Geometrization of quantum mechanics», *Commun. Math. Phys.* **65**, 189.

Kibble, T. W. B. (1981), «Is a semi-classical theory of gravity viable?», en *Quantum Gravity 2: A Second Oxford Symposium* (eds. C. J. Isham, R. Penrose y D. W. Sciama), pp. 63-80, Oxford University Press, Oxford.

Killing, W. (1983), *Einfuehrung in die Grundlagen der Geometrie*, Paderborn.

Klein, F. (1898), «Über den Stand der Herausgabe von Gauss», *Werken. Math. Ann.* **51**, 128-133.

Knott, C. G. (1900), «Professor Kleins view of quaternions: A criticism», *Proc. Roy. Soc. Edinb.* **23**, 24-34.

Kobayashi, S., y Nomizu, K. (1963), *Foundations of Differential Geometry*, Interscience Publishers, Nueva York y Londres.

Kochen, S., y Specker, E. P. (1967), «The Problem of Hidden Variables in Quantum Mechanics», *Journal of Mathematics and Mechanics* **17**, 59-88.

Kodaira, K. (1962), «A theorem of completeness of characteristic systems for analytic submanifolds of a complex manifold», *Ann. Math.* **75**, 146-162.

Kodaira, K., y Spencer, D. C. (1958), «On deformations of complex analytic structures I, II», *Ann. Math.* **67**, 328401, 403-466.

Kolb, E. W., y Turner, M. S. (1994), *The Early Universe*, Perseus Publishing, Nueva York.

Komar, A. B. (1964), «Undecidability of macroscopically distinguishable states in quantum field theory», *Phys. Rev.* **133B**, 542-544.

Kontsevich, M. (1994), «Homological algebra of mirror symmetry», *Proceedings of the International Congress of Mathematicians*, vols. 1, 2. (Zurich, 1994), Birkhaüser, Basilea.

Kontsevich, M. (1995), «Enumeration of rational curves via toric actions», en *The Moduli Space of Curves* (eds. R. Dijkgraaf, C. Faber y G. van der Geer), *Progress in Math.* **129**, 335-368 [hep-th/9405035].

Koruga, D., Hameroff, S., Withers, J., Loutfy, R., y Sundareshan, M. (1993), *Fullerene C60: History, physics, nanobiology, nanotechnology*, North-Holland, Amsterdam.

Kraus, K. (1983), «States, effects and operations: fundamental notions of quantum theory», *Lecture Notes in Physics*, vol. 190, Springler-Verlag, Berlín.

Krauss, L. M. (2001), *Quintessence: The Mystery of the Missing Mass*, Basic Books, Nueva York.

Kreimer, D. (2000), *Knots and Feynman Diagrams*, Cambridge University Press, Cambridge.

Kronheimer, E. H., y Penrose, R. (1967), «On the structure of causal spaces», *Proc. Camb. Phil Soc.* **63**, 481-501.

Kruskal, M. D. (1960), «Maximal Extension of Schwarzschild Metric», *Phys. Rev.* **119**, 1743-1745.

Kuchar, K. (1981), «Canonical methods of quantization», en *QuantumGravity2* (eds. D. W. Sciama, R. Penrose y C. J. Isham), Oxford University Press, Oxford.

Kuchar, K. V. (1992), «Time and interpretations of quantum gravity», en *Proceedings of the 4th Canadian Conference on General Relativity and Relativistic Astrophysics* (eds. G. Kunstatter, D. Vincent y J. Williams), World Scientific, Singapur.

Labastida, J. M. F., y Lozano, C. (1998), «Lectures in Topological Quantum Field Theory» [hep-th/9709192].

Landsman, N. P. (1998), *Mathematical Topics Between Classical and Quantum Mechanics*, Springer-Verlag, Berlín.

Lang, S. (1972), *Differentiable Manifolds*, Addison-Wesley, Reading, MA.

Lange, A. E., *et al.* (2001), «First Estimation of Cosmological Parameters from BOOMERANG», *Phys. Rev.* **D63**, 042001 [astro-ph/0005004].

Laplace, P. S. (1799), «Allgemeine geographische Ephemeriden herausgegeben von F. von Zach. iv Bd. 1 st», 1 Abhandl., Weimar.

Laporte, O., y Uhlenbeck, G. E. (1931), «Application of spinor analysis to the Maxwell and Dirac equations», *Phys. Rev.* **37**, 1380-1552.

Lasenby, J., Lasenby, A. N., y Doran, J. L. (2000), «A unified mathematical language for physics and engineering in the 21st century», *Phil. Trans. Roy. Soc. Lond.* **A358**, 21-39.

Lawrie, I. (1998), *A Unified Grand Tour of Theoretical Physics*, Institute of Physics Publishing, Bristol.

Lawson, H. B., y Michelson, M. L. (1990), *Spin Geometry*, Princeton University Press, Princeton.

Lawvere, W., y Schanuel, S. (1997), *Conceptual Mathematics: A First Introduction to Categories*, Cambridge University Press, Cambridge.

LeBrun, C. R. (1985), «Ambi-twistors and Einstein's equations», *Class. and Quantum Grav.* **2**, 555-263.

LeBrun, C. R. (1990), «Twistors, ambitwistors, and conformal gravity», en *Twistors in Mathematical Physics* (eds. T. N. Bailey y R. J. Baston), LMS Lecture Note Series 156, Cambridge University Press, Cambridge.

Lebrun, C., y Mason, L.J. (2002), «Zoll manifolds and complex surfaces», *J. Diff. Geom.* **61** (3), 453-535.

Lefshetz, J. (1949), *Introduction to Topology*, Princeton University Press, Princeton, New Jersey.

Leggett, A. J. (2002), «Testing the limits of quantum mechanics: motivation, state of play, prospects», *J. Phys. CM* 14, R415-451.

Lasenby, J., Lasenby, A. N., y Doran, J. L. (2000), «A unified mathematical language for physics and engineering in the 21st century», *Phil. Trans. Roy. Soc. Lond.*, **A358**, 21-39.

Levitt, M. H. (2001), *Spin Dynamics: Basics of Nuclear Magnetic Resonance*, John Wiley & Sons, Nueva York.

Lichnerowicz, A., ed. (1994), *Physics on Manifolds: Proceedings of the International Colloquium in Honour of Yvonne Choquet-Bruhat, Paris, June 3-5, 1992*, Kluwer Academic Publishers, Boston y Dordrecht, Países Bajos.

Liddle, A. R. (1999), *An Introduction to Modern Cosmology*, John Wiley & Sons, Nueva York.

Liddle, A. R., y Lyth, D. H. (2000), *Cosmological Inflation and Large-Scale Structure*, Cambridge University Press, Cambridge.

Lifshitz, E. M., y Khalatnikov, I. M. (1963), «Investigations in relativistic cosmology», *Adv. Phys.* **12**, 185-249.

Linda, A. (1993), «Comments on Inflationary Cosmology» [astro-ph/9309043].

Linde, A. (1995), «Inflation with Variable Omega», *Phys. Lett.* **B351** [hep-th/9503097].

Littlewood, J. E. (1949), *Littlewood's Miscellany*, reimpresión en 1986, Cambridge University Press, Cambridge.

Livio, M. (2000), *The Accelerating Universe*, John Wiley & Sons, Nueva York.

Llewellyn Smith, C. H. (1973), «High energy behaviour and gauge symmetry», *Phys. Lett.* **B46** (2), 233-236 [disponible online].

Lockwood, M. (1989), *Mind, Brain and the Quantum; the Compound I*, Basil Blackwell, Oxford.

Lorentz, H. A., Einstein, A., Minkowski, H., y Weyl, H. (1952), *The Principle of Relativity: A Collection of Original Memoirs on the Special and General Theory of Relativity*, Dover, Nueva York.

Lounesto, P. (2001), *Clifford Algebras and Spinors*, Cambridge University Press, Cambridge.

Lüders, G. (1951), «Über die Zustandsaänderung durch den Messprozess», *Ann. Physik* **8**, 322-328.

Ludvigsen, M. (1999), *General Relativity: A Geometric Approach*, Cambridge University Press, Cambridge.

Ludvigsen, M., y Vickers, J. A. G. (1982), «A simple proof of the positivity of the Bondi mass», *J. Phys.* A15, L67-70.

Luminet, J.-P., *et al.* (2003), «Dodecahedral space topology as an explanation for weak wide-angle temperature correlations in the cosmic microwave back-ground», *Nature* **425**, 593-595.

Lyttleton, R. A., y Bondi, H. (1959), *Proc. Roy. Soc. (London)* **A252**, 313.

MacDuffee, C. C. (1933), *The Theory of Matrices*, Springer-Verlag, Berlín (reimpresión en Chelsea).

MacLane, S. (1988), *Categories for the Working Mathematician*, Springer-Verlag, Berlín.

McLennan, J. A., Jr. (1956), «Conformal invariance and conservation laws for relativistic wave equations for zero rest mass», *Nuovo. Cim.*, **3**, 1360-1379.

MACRO Collaboration (2002), «Search for massive rare particles with MACRO», *Nucl. Phys. Proc. Suppl.* **110**, 1868 [hep-ex/0009002].

Magueijo, J. (2003), *Faster Than the Speed of Light: The Story of a Scientific Speculation*, Perseus Publishing, Nueva York.

Magueijo, J., y Smolin, L. (2002), «Lorentz invariance with an invariant energy scale» [gr-qc/0112090].

Mahler (1981), *P-Adic Numbers and Functions*, Cambridge University Press, Cambridge.

Majorana, E. (1932), «Teoria relativistica di particelle con momento intrinsico arbitrario», *Nuovo Cimento*, **9**, 335-344.

Majorana, E. (1937), «Teoria asimmetrica dell elettrone del positrone», *Nuovo Cimento*, **14**, 171-184.

Maldacena, J. (1997), «The Large N Limit of Superconformal Field Theories and Supergravity» [hep-th/9711200].

Manogue, C. A., y Dray, T. (1999), «Dimensional Reduction», *Mod. Phys. Lett.***A14**, 93-97 [hep-th/9807044].

Manogue, C. A., y Schray, J. (1993), «Finite Lorentz transformations, auto-morphisms, and division algebras», *J. Math. Phys.* **34**, 3746-3767.

Markopoulou, F. (1997), «Dual formulation of spin network evolution» [gr-qc/970401].

Markopoulou, F. (1998), «The internal description of a causal set: What the universe looks like from the inside», *Commun. Math. Phys.* **211**, 559-583 [gr-qc/9811053].

Markopoulou, F., y Smolin, L. (1997), «Causal evolution of spin networks», *Nucl. Phys.* **B508**, 409-430 [gr-qc/9702025].

Marsden, J. E., y Tromba, A. J. (1996), *Vector Calculus*, W. H. Freeman & Co., Nueva York, nueva edición 2004 (hay traducción castellana: *Cálculo vectorial*, Addison Wesley Iberoamericana España, Madrid, 2004).

Marshall, W., Simon, C., Penrose, R., y Bouwmeester, D. (2003), «Towards Quantum Superpositions of a Mirror», *Phys. Rev. Lett.* **91**, 13.

Mason, L. J., y Woodhouse, N. M. J. (1996), *Integrability, Self-Duality, and Twistor Theory*, Oxford University Press, Oxford.

Mattuck, R. D. (1976), *A Guide to Feynman Diagrams in the Many-Body Problem*, Dover, Nueva York.

McLennan, J. A., Jr. (1956), «Conformal invariance and conservation laws for relativistic wave equations for zero rest mass», *Nuovo. Cim.* **3**, 1360-1379.

Merkulov, S. A., y Schwachhöfer, L. J. (1998), «Twistor solution of the holonomy problem», en *The Geometric Universe: Science, Geometry, and the Work of Roger Penrose* (eds. S. A. Huggett, L. J. Mason, K. P. Tod, S. T. Tsou y N. M. J. Woodhouse), Oxford University Press, Oxford.

Michell, J. (1784), «On the means of discovering the distance, magnitude, etc., of the fixed stars, in consequence of the diminution of their light, in case

such a diminution should be found to take place in any of them, and such other data should be procured from observations, as would be further necessary for that purpose», *Phil. Trans. Roy. Soc. Lond.* **74**, 35-57.

Mielnik, B. (1974), «Generalized Quantum Mechanics. *Commun*», *Math. Phys.* **37**, 221.

Milgrom, M. (1994), «Dynamics with a non-standard inertia-acceleration relation: an alternative to dark matter», *Annals Phys.* **229** [astro-ph/9303012].

Miller, A. (2003), «Erotica, Aesthetics, and Schröedinger's Wave Equation», en *It Must Be Beautiful* (ed. G. Farmelo), Granta, Londres.

Minassian, E. (2002), «Spacetime singularities in (2+1)-dimensional quantum gravity», *Class. Quant. Grav.* **19**, 5877-5900.

Minkowski, H. (1952), en Lorentz *et al.* (1952).

Misner, C. W. (1969), «Mixmaster Universe», *Phys. Rev. Lett.* **22**, 1071-1074.

Misner, C. W., Thorne, K. S., y Wheeler, J. A. (1973), *Gravitation*, Freeman, San Francisco.

Mohapatra, R. N. (2002), *Unification and Supersymmetry*, Springer-Verlag, Berlín y Londres.

Montgomery, D., y Zippin, L. (1955), *Topological Transformation Groups*, Interscience, Nueva York y Londres.

Moore, A. W. (1990), *The Infinite*, Routledge, Londres y Nueva York.

Moroz, I. M., Penrose, R., y Tod, K. P. (1998), «Spherically-symmetric solutions of the Schrödinger-Newton equations», *Class. Quant. Grav.* **15**, 2733-2742.

Mott, N. F. (1929), «The wave mechanics of a-ray tracks», en *Quantum Theory and Measurement* (eds. J. A. Wheeler y W. H. Zurek), Princeton University Press, Princeton, New Jersey, 1983; originalmente en *Proc. Roy. Soc. Lond.* **A126**, 79-84.

Moussouris, J. P. (1983), «Quantum models of space-time based on recoupling theory», Oxford, tesis doctoral no publicada.

Mukohyama, S., y Randall, L. (2003), «A Dynamical Approach to the Cosmological Constant», *Phys. Rev. Lett.* **92** (2004) 211302 [hep-th/0306108].

Munkres, J. R. (1954), *Elementary Differential Topology. Annals of Mathematics Studies, 54*, Princeton University Press, Princeton, New Jersey.

Myrheim, J. (1978), «Statistical Geometry», CERN preprint, TH-2538, no publicado.

Nahin, P. J. (1998), *An Imaginary Tale: The Story of Root sqrt-1 \emph*, Princeton University Press, Princeton.

Nair, V. (1988), «A Current Algebra For Some Gauge Theory Amplitudes», *Phys. Lett.* **B**214, 215.

Nambu, Y. (1970), *Proceedings of the International Conference on Symmetries and Quark Models*, Wayne State University, p. 269, Gordon and Breach Publishers.

Narayan, R., *et al.* (2003), «Evidence for the Black Hole Event Horizon», *Astronomy & Geophysics* **44(6)**, 6.22-6.26.

Needham, T. R. (1997), *Visual Complex Analysis*, Clarendon Press, Oxford University Press, Oxford.

Negrepontis, S. (2000), «The Anthyphairetic Nature of Platos Dialectics», en *Interdisciplinary Approach to Mathematics and their Teaching*, vol. 5, pp. 15-77, University-Gutenberg, Atenas (en griego).

Nester, J. M. (1981), «A new gravitational energy expression, with a simple positivity proof», *Phys. Lett.* **83A**, 241-242.

Newlander, A., y Nirenberg, L. (1957), «Complex Analytic Coordinates in Almost Complex Manifolds», *Ann. of Math.* **65**, 391-404.

Newman, E. T. (2002), «On a Classical Geometrical Origin of Magnetic Moments, Spin-Angular Momentum and the Dirac Gyromagnetic Ratio», *Phys. Rev.* **D65** 104005 [gr-qc/0201055].

Newman, E. T, y Penrose, R. (1962), «An approach to gravitational radiation by a method of spin coefficients», *J. Math. Phys.* **3**, 896902; errata (1963), 4, 998.

Newman, E. T., y Penrose, R. (1966), «Note on the Bondi-Metzner-Sachs group», *J. Math. Phys.* **7**, 863-870.

Newman, E. T., y Unti, T. W. J. (1962), «Behavior of asymptotically flat empty space», *J. Math. Phys.* **3**, 891-901.

Newman, E. T., Couch, E., Chinnapared, K., Exton, A., Prakash, A., y Torrence, R. (1965), «Metric of a rotating charged mass», *J. Math. Phys.* **6**, 918-919.

Newman, R. P. A. C. (1993), «On the Structure of Conformal Singularities in Classical General Relativity», *Proc. R. Soc. Lond.* **A443**, 473.

Newton, I. (1687), *The Principia: Mathematical Principles of Natural Philosophy*, reimpreso por University of California Press, 1999 (hay edición castellana de los *Principia*: *Principios matemáticos de la filosofía natural*, edición de Eloy Rada, Alianza, Madrid, 1994).

Newton, I. (1730), *Optiks*, Dover, 1952 (hay edición castellana: *Óptica*, edición de Carlos Solís, Alfaguara, Madrid, 1977).

Ng, Y. J. (2004), «Quantum Foam» [gr-qc/0401015].

Nicolai, H. (2003), «Remarks at AEI\ Symposium "String meet Loops"», 29-31 de octubre de 2003. www.aci-postdam.mpg.de/events/stringloop.html.

Nielsen, H. B. (1970). Enviado a *Proc. of the XV Int. Conf. on High Energy Physics*, Kiev (inédito).

Nielsen, M. A., y Chuang, I. L. (2000), *Quantum Computation and Quantum Information*, Cambridge University Press, Cambridge.

Nomizu, K. (1956), *Lie Groups and Differential Geometry*, The Mathematical Society of Japan, Tokio.

Novikov, I. D. (2001), *The River of Time*, Cambridge University Press, Cambridge.

O'Donnell, P. (2003), *Introduction to 2-Spinors in General Relativity*, World Scientific, Singapur.

O'Neill, B. (1983), *Semi-Riemannian Geometry: With Applications to Relativity*, Academic Press, Nueva York.

Oppenheimer, J. R. (1930), «On the theory of electrons and protons», *Phys. Rev.* **35**, 562-563.

Ozsvath, I., y Schucking, E. (1962), *Nature* **193**, 1168.

Ozsvath, I., y Schucking, E. (1969), *Ann. Phys.* **55**.

Page, D. (1995), «Sensible Quantum Mechanics: Are Only Perceptions Probabilistic?» [quant-ph/9506010].

Page, D. A. (1987), «Geometrical description of Berrys phase», *Phys. Rev.* **A36**, 3479-3481.

Page, D. N. (1976), «Dirac equation around a charged, rotating black hole», *Phys. Rev.* **D 14**, 1509-1510.

Pais, A. (1982), *Subtle is the Lord... The Science and the Life of Albert Einstein*, Clarendon Press, Oxford (hay traducción castellana: *El Señor es sutil; la ciencia y la vida de Albert Einstein*, Ariel, Barcelona, 1984).

Pais, A. (1986), *Inward Bound: Of Matter and Forces in the Physical World*, Clarendon Press, Oxford.

Parker, T., y Taubes, C. H. (1982), «On Witten's proof of the positive energy theorem», *Comm. Math. Phys.* **84**, 223-238.

Pars, L. A. (1968), *A Treatise on Analytical Dynamics*, reimpreso en 1981, Ox Bow Press.

Pearle, P. (1985), «Models for reduction», en *Quantum Concepts in Space and Time* (eds. C. J. Isham y R. Penrose), pp. 84-108, Oxford University Press, Oxford.

Pearle, P., y Squires, E. J. (1995), «Gravity, energy conservation and parameter values in collapse models», Durham University preprint, DTP/95/13.

Peitgen, H.-O., y Reichter, P. H. (1986), *The Beauty of Fractals: Images of Complex Dynamical Systems*, Springer-Verlag, Berlín y Heidelberg.

Peitgen, H.-O., y Saupe, D. (1988), *The Science of Fractal Images*, Springer-Verlag, Berlín.

Penrose, L. S., y Penrose, S. (1958), «Impossible Objects: A Special Type of Visual Illusion», *Brit. J. Psych.* **49**, 31-33.

Penrose, R. (1959), «The apparent shape of relativistically moving sphere», *Proc. Camb. Phil. Soc.* **55**, 137-139.

Penrose, R. (1960), «A spinor approach to general relativity», *Ann. Phys.* (Nueva York) 10, 171-201.

Penrose, R. (1962), «The Light Cone at Infinity», en *Proceedings of the 1962 Conference on Relativistic Theories of Gravitation*, Warsaw, Polish Academy of Sciences, Varsovia (publicado en 1965).

Penrose, R. (1963), «Asymptotic properties of fields and space-times», *Phys. Rev. Lett.* **10**, 66-68.

Penrose, R. (1964), «Conformal approach to infinity», en *Relativity, Groups and Topology: The 1963 Les Houches Lectures* (eds. B. S. DeWitt y C. M. DeWitt), Gordon and Breach, Nueva York.

Penrose, R. (1965a), «Zero rest-mass fields including gravitation: asymptotic behaviour», *Proc. R. Soc. Lond.* **A284**, 159-203.

Penrose, R. (1965b), «Gravitational collapse and space-time singularities», *Phys. Rev. Lett.* **14**, 57-59.

Penrose, R. (1966), *An analysis of the structure of space-time*, Adams Prize Essay, Cambridge University, Cambridge.

Penrose, R. (1967), «Twistor algebra», *J. Math. Phys.* **8**, 345-366.

Penrose, R. (1968a), «Structure of space-time», en *Battelle Rencontres, 1967* (eds. C. M. DeWitt y J. A. Wheeler), *Lectures in Mathematics and Physics*, Benjamin, Nueva York.

Penrose, R. (1968b), «Twistor quantization and curved space-time», *Int. J. Theor. Phys.* **1**, 61-99.

Penrose, R. (1969a), «Gravitational collapse: the role of general relativity», *Revista del Nuovo Cimento*, número especial 1, 252-276.

Penrose, R. (1969b), «Solutions of the zero rest-mass equations», *J. Math. Phys.* **10**, 38-39.

Penrose, R. (1971a), «Angular momentum: an approach to combinatorial space-time», en *Quantum Theory and Beyond* (ed. T. Bastin), Cambridge University Press, Cambridge.

Penrose, R. (1971b), «Applications of negative dimensional tensors», en *Combinatorial Mathematics and its Applications* (ed. D. J. A. Welsh), Academic Press, Londres.

Penrose, R. (1975), «Twistor theory: its aims and achievements», en *Quantum Gravity, an Oxford Symposium* (eds. C. J. Isham, R. Penrose, y D. W. Sciama), Oxford University Press, Oxford.

Penrose, R. (1976a), «The non-linear graviton», *Gen. Rel. Grav.* **7**, 171-176.

Penrose, R. (1976b), «Non-linear gravitons and curved twistor theory», *Gen. Rel. Grav.* **7**, 31-52.

Penrose, R. (1978), «Gravitational collapse: a Review», en *Physics and Astrophysics of Neutron Stars and Black Holes*, LXV Corso. Soc. Italiana di Fisica, Bolonia, Italia, pp. 566-582.

Penrose, R. (1979), «Singularities and time-asymmetry», en *General Relativiy: An Einstein Centenary* (eds. S. W. Hawking y W. Israel), Cambridge University Press, Cambridge.

Penrose, R. (1979b), «On the twistor description of massless fields», en *Complex Manifold Techniques in Theoretical Physics* (eds. D. E. Lerner y P. D. Sommers), Pitman, San Francisco. Véanse diversos artículos en L. P. Hughston y R. S. Ward, eds. (1979), *Advances in Twistor Theory*, Pitman, Advanced Publishing Program, San Francisco.

Penrose, R. (1980), «On Schwarzschild causalitya problem for Lorentz-covariant general relativity», en *Essays in General Relativity* (A. Taub Festschrift) (ed. F. J. Tipler), p. 112, Academic Press, Nueva York.

Penrose, R. (1982), «Quasi-local mass and angular momentum in general relativity», *Proc. Roy. Soc. Lond.* **A381**, 53-63.

Penrose, R. (1986), «Gravity and state-vector reduction», en *Quantum Concepts in Space and time* (eds. R. Penrose y C. J. Isham), pp. 129-146, Oxford University Press, Oxford.

Penrose, R. (1987a), «Quantum physics and conscious thought», en *Quantum Implications: Essays in Honour of David Bohm* (eds. B. J. Hiley y F. D. Peat), Routledge and Kegan Paul, Londres y Nueva York.

Penrose, R. (1987b), «On the origins of twistor theory», en *Gravitation and Geometry: a volume in honour of I. Robinson.* (eds. W. Rindler y A. Trautman), Bibliopolis, Nápoles.

Penrose, R. (1987d), «Newton, quantum theory, and reality», en *300 Years of Gravity* (eds. S. W. Hawking y W. Israel), pp. 17-49, Cambridge University Press, Cambridge.

Penrose, R. (1988), «Holomorphic linking», *Twistor Newsletter* **27**, 1-4.

Penrose, R. (1988a), «Topological QFT and Twistors: Holomorphic Linking; Holomorphic Linking: Postscript», *Twistor Newsletter* **27**, 14.

Penrose, R. (1988b), «Fundamental asymmetry in physical laws», Proceedings

of Symposia in *Pure Mathematics* **48**, American Mathematical Society, pp. 317-328.

Penrose, R. (1989), *The Emperor's New Mind: Concerning Computers, Minds, and the Laws of Physics*, Oxford University Press, Oxford (hay traducción castellana: *La nueva mente del emperador*, Mondadori, Barcelona, 1999).

Penrose, R. (1991), «On the cohomology of impossible figures [La cohomologie des figures impossibles]», *Structural Topology [Topologie structurale]* **17**, 11-16.

Penrose, R. (1992), «H-space and Twistors», en *Recent Advances in General Relativity*, Einstein Studies, vol. 4 (eds. A. I. Janis y J. R. Porter), p. 625, Birkhäuser, Boston.

Penrose, R. (1994), *Shadows of the Mind: An Approach to the Missing Science of Consciousness*, Oxford University Press, Oxford (hay traducción castellana: *Sombras de la mente: hacia una comprensión científica de la consciencia*, Crítica, Barcelona, 1996).

Penrose, R. (1996), «On gravity's role in quantum state reduction», *Gen. Rel. Grav.* **28**, 581-600.

Penrose, R. (1997a), *The Large, the Small and the Human Mind*, Cambridge University Press, Cambridge, Canto edition (2000) (hay traducción castellana: *Lo grande, lo pequeño y la mente humana*, Cambridge University Press, Madrid, 1999).

Penrose, R. (1997b), «On understanding understanding», *Internat. Stud. Philos. Sci.* **11**, 7-20.

Penrose, R. (1998a), «Quantum computation, entanglement and state-reduction», *Phil. Trans. Roy. Soc. Lond.* **A356**, 1927-1939.

Penrose, R. (1998b), «The question of cosmic censorship», en *Black Holes and Relativistic Stars* (ed. R. M. Wald), University of Chicago Press, Chicago, Illinois. Reimpreso en *J. Astrophys. Astr.* 20, 233-248 (1999).

Penrose, R. (2000), «Wavefunction collapse as a real gravitational effect», en *Mathematical Physics 2000* (eds. A. Fokas, T. W. B. Kibble, A. Grigouriou y B. Zegarlinski), pp. 26682, Imperial College Press, Londres.

Penrose, R. (2000b), «On Bell non-locality whitout probabilities: some curious geometry», en *Quantum Reflections* (eds. J. Ellis y D. Amati), Cambridge University Press, Cambridge

Penrose, R. (2001), «Towards a twistor description of general space-times; introductory comments», en *Further Advances in Twistor Theory, Vol. III: Curved Twistor Spaces* (eds. L. J. Mason, L. P. Hughston, P. Z. Kobak y K. Pulverer). Chapman & Hall/CRC Research Notes in Mathematics 424, Londres. 239-255.

Penrose, R. (2002), «John Bell, State Reduction, and Quanglement», en *Quantum [Un]speakables: From Bell to Quantum Information* (eds. R. A. Bertlmann y A. Zeilinger), Springer-Verlag, Berlín.

Penrose, R., y MacCallum, M. A. H. (1972), «Twistor theory: an approach to the quantization of fields and space-time», *Phys. Repts.* **6C**, 241-315.

Penrose, R. (2003), «On the instability of extra space dimensions», *The Future of Theoretical Physics and Cosmology, Celebrating Stephen Hawkings 60th Birthday* (eds. G. W. Gibbons, E. P. S. Shellard y S. J. Rankin), Cambridge University Press, Cambridge.

Penrose, R., y Rindler, W. (1984), *Spinors and Space-Time*, vol. I: *Two-Spinor Calculus and Relativistic Fields*, Cambridge University Press, Cambridge.

Penrose, R., y Rindler, W. (1986), *Spinors and Space-Time*, vol. II: *Spinor and Twistor Methods in Space-Time Geometry*, Cambridge University Press, Cambridge.

Penrose, R., Robinson, I., y Tafel, J. (1997), «Andrzej Mariusz Trautman», *Class. Quan. Grav.* 14, A1-A8.

Penrose, R., Sparling, G. A. J., y Tsou, S. T. (1978), «Extended Regge Trajectories», *J. Phys. A. Math. Gen.* **11**, L231-L235.

Penzias, A. A., y Wilson, R. W. (1965), «A Measurement of Excess Antenna Temperature at 4080 Mc/s», *Astrophys. J.* **142**, 419.

Percival, I. C. (1994), «Primary state diffusion», *Proc. R. Soc. Lond.* **A447**, 189-209.

Percival, I. C. (1995), «Quantum space-time fluctuations and primary state diffusion» [quant-ph/9508021].

Peres, A. (1991), «Two Simple Proofs of thr Kochen-Specker Theorem», *Journal of Physics A: Mathematical and General* **24**, L175- L178.

Peres, A. (1995), «Generalized Kochen-Specker Theorem» [quant-ph/9510018].

Peres, A. (2000), «Delayed choice for entanglement swapping», *J. Mod. Opt.* **47**, 531 [quant-ph/9904042].

Perez, A. (2001), «Finiteness of a spinfoam model for Euclidean quantum general relativity», *Nucl. Phys.* **B599**, 427-434.

Perez, A. (2003), «Spin foam models for quantum gravity», *Class. Quant. Grav.* **20**, R43-R104.

Perlmutter, S., *et. al.* (1998), «Cosmology from Type Ia Supernovae», *Bull. Am. Astron. Soc.* 29 [astro-ph/9812473].

Peskin, M. E., y Schröder, P. V. (1995), *Introduction to Quantum Field Theory*, Westview Press, Reading, MA & Wokingham.

Petiau, G. (1936), «Contribution à la théorie des equations d'ondes corpusculaires», *Adad. Roy. Belgique* (Cl. Sci. Mem, Collect. 16, n.° 2).

Pirani, F. A. E., y Schild, A. (1950), «On the Quantization of Einsteins Gravitational Field Equations», *Phys. Rev.* **79**, 986-991.

Pitkaenen, M. (1994), «p-Adic description of Higgs mechanism I: p-Adic square root and p-adic light cone» [hep-th/9410058].

Polchinski, J. (1998), *String Theory*, Cambridge University Press, Cambridge.

Polkinghorne, J. (2002), *Quantum Theory. A Very Short Introduction*, Oxford University Press, Oxford.

Popper, K. (1934), *The Logic of Scientific Discovery*, Routledge; nueva edición en marzo de 2002 (hay traducción castellana del original: *La lógica de la investigación científica*, Tecnos, Madrid, 1985).

Pound, R.V., y Rebka, G. A. (1960), *Phys. Rev. Lett.* **4**, 337.

Preskill, J. (1992), «Do black holes destroy information?» [hep-th/9209058].

Priestley, H. A. (2003), *Introduction to Complex Analysis*, Oxford University Press, Oxford.

Rae, A. I. M. (1994), *Quantum Mechanics*, Institute of Physics Publishing, 4.ª ed. 2002.

Raine, D. J. (1975), «Mach's principle in General Relativity», *Monthly Notices RAS* **171**, 507-528.

Randall, L., y Sundrum, R. (1999a), «A Large Mass Hierarchy from a Small Extra Dimension», *Phys. Rev. Lett.* **83**, 3370-3 [hep-ph/9905221].

Randall, L., y Sundrum, R. (1999b), «An Alternative to Compactification», *Phys. Rev. Lett.* **83**, 4690-3 [hep-th/9906064].

Rarita, W. and Schwinger, J. (1941), «On the theory of particles with half-integer spin», *Phys. Rev.* **60**, 61.

Redhead, M. L. G. (1987), *Incompleteness, Nonlocality and Realism*, Clarendon Press, Oxford.

Reed, M., y Simon, B. (1972), *Methods of Mathematical Physics*, vol. 1: *Functional Analysis*, Academic Press, Nueva York y Londres.

Reege, T. (1961), «General Relativity without Coordinates», *Nuovo Cimento A* **19**, 558-571.

Reeves, J. N., *et al.* (2002), «The signature of supernova ejecta in the X-ray afterglow of the gamma-ray burst 011211», *Nature* **416**, 512-515.

Reisenberger, M. P. (1997), «A lattice worldsheet sum for 4-d Euclidean general relativity» [gr-qc/9711052].

Reisenberger, M. P. (1999), «On relativistic spin network vertices», *J. Math. Phys.*, **40**, 2046-2054.

Reisenberger, M. P., y Rovelli, C. (2001), «Spacetime as a Feynman diagram: the connection formulation», *Class. Quant. Grav.* **18**, 121-140.

Reisenberger, M. P., y Rovelli, C. (2002), «Spacetime states and covariant quantum theory», *Phys. Rev.* **D65**, 125016.

Reula, O., y Tod, K. P. (1984), «Positivity of the Bondi energy», *J. Math. Phys.* **25**, 1004-1008.

Riemann, G. B. F. (1854), «Über die Hypothesen der Geometrie zu Grunde liegen (Habilitationsschrift, Göttingen)»; véase *Collected Works of Bernhardt Riemann* (ed. Heinrich Weber), 2.ª ed., Dover, Nueva York, 1953, pp. 272-287.

Rindler, W. (1977), *Essential Relativity*, Springer-Verlag, Nueva York.

Rindler, W. (1982), *Introduction to Special Relativity*, Clarendon Press, Oxford.

Rindler, W. (2001), *Relativity: Special, General, and Cosmological*, Oxford University Press, Oxford.

Rizzi, A. (1998), «Angular momentum in general relativity: A new definition», *Phys. Rev. Lett.* **81(6)**, 1150.

Robinson, D. C. (1975), «Uniqueness of the Kerr Black Hole», *Phys. Rev. Lett.* **34**, 905-906.

Rogers, A. (1980), «A global theory of supermanifolds», *J. Math. Phys.* **21**, 1352-1365.

Rolfsen, D. (2004), *Knots and Links*, American Mathematical Society, Providence, RI.

Roseveare, N. T. (1982), *Mercury's Perihelion from Le Verrier to Einstein*, Clarendon Press, Oxford.

Rovelli, C. (1991), «Quantum mechanics without time: A model», *Phys. Rev.* **D42**, 2638.

Rovelli, C. (1998), «Strings, loops and others: a critical survey of the present approach to quantum gravity», en *Gravity and Relativity: At the turn of the Millennium* (15 th International Conference on General Relativity and Gravitation, eds. N. Dadhich y J. Narlikar, Inter-University Centre for Astronomy and Astrophysics, Pune, India), 281-331.

Rovelli, C. (2003), «Quantum Gravity»; http://www.cpt.univ-mrs.fr/~rovelli/book.pdf.

Rovelli, C. y Smolin, L. (1990), «Loop representation for quantum general relativity», *Nucl. Phys.* **B331**, 80-152.

Runde, V. (2002), «The Banach-Tarski paradox or What mathematics and religion have in common», Pi in the Sky 2 (2000), 13-15 [math. GM/0202309].

Russell, B. (1903), *Principles of Mathematics*. La edición más reciente es la de W. W. Norton & Company, 1996 (hay traducción castellana: *Los principios de la matemática*, Espasa Calpe, Madrid, 1983, agotado).

Russell, B. (1927), *The Analysis of Matter*, Allen and Unwin; reimpreso 1954, Dover, Nueva York (hay traducción castellana: *Análisis de la materia*, Taurus, Madrid, 1976, agotado).

Ryder, L. H. (1996), *Quantum Field Theory*, Cambridge University Press, Cambridge.

Sabbagh, K. (2003), *The Riemann Hypothesis: The Greatest Unsolved Problem in Mathematics*, Farrar, Straus and Giraux.

Saccheri, G. (1733), «Euclides ab Omni NaevoVindicatus», traducción en Halsted, G.B. (1920), *Euclid Freed from Every Flaw*, Open Court, La Salle, Illinois.

Sachs, R. (1962), «Asymptotic symmetries in gravitational theory», *Phys. Rev.* **128**, 2851-2864.

Sachs, R., y Bergmann, P. G. (1958), «Structure of Particles in Linearized Gravitational Theory», *Phys. Rev.* **112**, 674-680.

Sachs, R. K. (1961), «Gravitational waves in general relativity,VI: the outgoing radiation condition», *Proc. Roy. Soc. Lond.* **A264**, 309-338.

Sachs, R. K. (1962a), «Gravitational waves in general relativity, VIII: waves in asymptoticaly flat space-time», *Proc. Roy. Soc. Lond.* **A270**, 103-126.

Sachs, R. K. (1962b), «Asymptotic symmetries in gravitational theory», *Phys. Rev.* **128**, 2851-2864.

Sakellariadou, M. (2002), «The role of topological defects in cosmology», Invited lectures in NATO ASI / COSLAB (ESF) School Patterns of Symmetry Breaking, septiembre de 2002, Cracovia [hep-ph/0212365]

Salam, A. (1980), «Gauge Unification of Fundamental Forces», *Rev. Mod. Phys.* **52(3)**, 515-523.

Salam, A., y Ward, J. C. (1959), «Weak and electromagnetic interaction», *Nuovo Cimento*, **11**, 568.

Sarkar, S. (2002), «Possible astrophysical probes of quantum gravity», *Mod. Phys. Lett.* **A17**, 1025-1036 [gr-qc/0204092].

Sato, M. (1958), «On the generalization of the concept of a function», *Proc. Japan Acad.* **34**, 126-130.

Sato, M. (1959), «Theory of hyperfunctions I», J. Fac. Sci. Univ. Tokyo, Sect. I, 8, 139-193.

Sato, M. (1960), «Theory of hyperfunctions II», J. Fac. Sci. Univ. Tokyo, Sect. I, 8, 387-437.

Schild, A. (1949), «Discrete space-time and integral Lorentz transformations», *Can. J. Math.* **1**, 29-47.

Schoen, R., y Yau, S. T. (1979), «On the proof of the positive mass conjecture in the general relativity», *Comm. Math. Phys.* **65**, 45-76.

Schoen, R., y Yau, S. T. (1982), «Proof that Bondi mass is positive», *Phys. Rev. Lett.* **48**, 369-371.

Schrödinger, E. (1930), «Sitzungber», Preuss. Akad. Wiss. Phys.-Math. Kl. 24, 418.

Schrödinger, E. (1935), «Probability relations between separated systems», *Proc. Camb. Phil. Soc.* **31**, 555-563.

Schrödinger, E. (1950), *Space-Time Structure*, Cambridge University Press, Cambridge (hay traducción castellana: *La estructura del espacio-tiempo*, Alianza, Madrid, 1993).

Schrödinger, E. (1956), *Expanding Universes*, Cambridge University Press, Cambridge.

Schrödinger, E. (1967), *«What is Life?» and «Mind and Matter»*, Cambridge University Press, Cambridge (hay traducción castellana: *¿Qué es la vida?*, Tusquets, 1984; *Mente y materia*, Tusquets, 1985).

Schouters, J. A. (1954), *Ricci-Calculus*, Springer, Berlín.

Schutz, B. (2003), *Gravity From the Ground Up: An Introductory Guide to Gravity and General Relativity*, Cambridge University Press, Cambridge.

Schutz, J. W. (1997), *Independent Axioms for Minkowski Space-Time*, Addison Wesley Longman Ltd., Harlow, Essex.

Schwartz, L. (1967), *Théorie des distributions*, Hermann, París.

Schwarz, J. H. (2001), «String Theory, *Curr. Sci.* **81(12)**, 1547-1553.

Schwarzschild, K. (1916), «Über das Gravitationsfeld eines Massenpunktes nach der Einsteinschen Theorie», Sitzber. Deut. Akad. Wiss. Berlin Math.-Phys. Tech. Kl. 189-196.

Schwinger, J. (1951), *Proc. Nat. Acad. Sci.* **37**, 452.

Schwinger, J., ed. (1958), *Quantum Electrodynamics*, Dover.

Sciama, D. W. (1959), *The Unity of the Universe*, Doubleday & Company, Inc., Nueva York.

Sciama, D. W. (1962), «On the analogy between charge and spin in general relativity», en *Recent Developments in General Relativity*, Pergamon & PWN, Oxford.

Sciama, D. W. (1972), *The Physical Foundations of General Relativity*, Heinemann, Londres.

Sciama, D. W. (1998), «Decaying neutrinos and the geometry of the universe», en *The Geometric Universe: Science, Geometry, and the Work of Roger Penrose*

(eds. S. A. Huggett, L. J. Mason, K. P. Tod, S. T. Tsou y N. M. J. Woodhouse), Oxford University Press, Oxford.

Seiberg, N., y Witten, E. (1994), «Electric-magnetic duality, monopole condensation, and confinement in N = 2 supersymmetric Yang-Mills theory, *Nucl. Phys.* B426 [hep-th/9407087].

Sen, A. (1982), «Gravity as a spin system», *Phys. Lett.* B119, 89-91.

Shankar, R. (1994), *Principles of Quantum Mechanics*, 2.ª ed., Plenum Press, Nueva York y Londres.

Shapiro, I. I., *et al.* (1971), *Phys. Rev. Lett.* 13, 789.

Shaw, W. T., y Hughston, L. P. (1990), «Twistors and strings», en *Twistors in Mathematics and Physics* (eds. T. N. Bailey y R. J. Baston), London Mathematical Society Lecture Notes Series, 156, Cambridge University Press, Cambridge.

Shawhan, P. (2001), «The Search for Gravitational Waves with LIGO: Status and Plans», *Intl. J. Mod. Phys.* A 16, supp. 01C, 1028-1030.

Shih, Y. H., *et al.* (1995), «Optical Imaging by Means of Two-Photon Entanglement», *Phys. Rev.* A, Rapid Comm. 52, R3429.

Shimony, A. (1998), «Implications of transience for spacetime structures», en *The Geometric Universe: Science, Geometry, and the Work of Roger Penrose* (eds. S. A. Huggett, L. J. Mason, K. P. Tod, S. T. Tsou y N. M. J. Woodhouse), Oxford University Press, Oxford.

Shrock, R. (2003), *Neutrinos and Implications for Physics Beyond the Standard Model*, World Scientific Pub. Co., Singapur.

Silk, J., y Rees, M. (1998), «Quasars and galaxy formation», *Astronomy and Astrophysics*, v. 331, p. L1-L4.

Simon, B. (1983), «Holonomy, the quantum adiabatic theorem, and Berry's phase», *Phys. Rev. Lett.* **51**, 2160-2170.

Singh, S. (1997), *Fermat's Last Theorem*, Fourth Estate, Londres.

Slipher, V. A. (1917), «Nebulae», *Proc. Am. Phil. Soc.* 56, 403.

Smolin, L. (1998), «The physics of spin networks», en *The Geometric Universe: Science, Geometry, and the Work of Roger Penrose* (eds. S. A. Huggett, L. J. Mason, K. P. Tod, S. T. Tsou y N. M. J. Woodhouse), Oxford University Press, Oxford.

Smolin, L. (1997), *The Life of the Cosmos*, Oxford University Press, Oxford.

Smolin, L. (2001), «The exceptional Jordan algebra and the matrix string» [hep-th/0104050].

Smolin, L. (2002), *Three Roads To Quantum Gravity*, Basic Books, Nueva York.

Smolin, L. (2003), «How far are we from the quantum theory of gravity?» [hep-th/0303185].

Smolin, L. (1991), «Space and time in the quantum universe», en *Conceptual Problems in Quantum Gravity* (eds. A. Ashtekar y J. Stachel), Birkhauser, Boston.

Smoot, G. F., *et al.* (1991), «Preliminary results from the COBE differential micro-wave radiometers: large-angular-scale isotropy of the Cosmic Microwave Background», *Astrophys. J.* 371, L1.

Snyder, H. (1947), «Quantized space-time», *Physical Review*, **71**, 38-41.

Sorabji, R. J. (1984), *Time, Creation and the Continuum*, Cornell University Press.

Sorabji, R. J. (1988), *Matter, Space and Motion*, Duckworth Publishing.

Sorkin, R. D. (1991), «Spacetime and Causal Sets», en *Relativity and Gravitation: Classical and Quantum* (eds. J. C. D'Olivo *et al.*), World Scientific, Singapur.

Sorkin, R. D. (1994), «Quantum Measure Theory and its Interpretation», en *Proceedings of 4th Drexel Symposium on Quantum Nonintegrability*, 811 Sep., Filadelfia, PA [gr-qc/9507057].

Spergel, D. N. (2003), «First Year Wilkinson Microwave Anisotropy Probe Observations: Determination of Cosmological Parameters», *Astrophys. J. Suppl.* **148**, 175.

Stachel, J. (1995), «History of relativity», en *History of 20th Century Physics* (eds. L. Brown, A. Pais y B. Pippard), cap. 4, American Institute of Physics (AIP) and British Institute of Physics (BIP).

Stairs, A. (1983), «Quantum logic, ralism and value-definiteness», *Phil. Sci.* **50(4)**, 578-602.

Stapp, H. P. (1971), «S-matrix Interpretation of Quantum Mechanics», *Phys. Rev.* D3, 1303-1320.

Stapp, H. P. (1979), «Whiteheadian Approach to Quantum Theory and the Generalized Bell's Theorem», *Found. Phys.* 9, 1-25.

Steenrod, N. E. (1951), *The Topology of Fibre Bundles*, Princeton University Press, Princeton.

Steinhardt, P. J., y Turok, N. (2002), «A Cyclic Model of the Universe», *Science* 296(5572) 143639 [hep-th/0111030].

Stoney, G. J. (1881), «On the Physical Units of Nature», *Philosophical Magazine*, vol. 11, 381.

Strauss, W. (1992), *Partial Differential Equations: An Introduction*, John Wiley and Sons.

Strominger, A., y Vafa, C. (1996), «Microscopic Origin of the Bekenstein-Hawking Entropy», *Phys. Lett.* B379, 99-104.

Strominger, A., Yau, S-T., y Zaslow, E. (1996), «Mirror symmetry is T-duality», *Nucl. Phys.* B479, 12, 243-259.

Struik, D. J. (1954), *A Concise History of Mathematics*, Dover, Nueva York.

Sudarshan, G., y Dhar, J. (1968), «Quantum Field Theory of Interacting Tachyons», *Phys. Rev.* 174, 1808.

Sudbery, A. (1987), «Division algebras (pseudo) orthogonal groups and spinors», *J. Phys.* A17, 939-955.

Susskind, L. (1970), «Structure of Hadrons Implied by Duality», *Nuovo Cimento*, A69, 457.

Susskind, L. (2003), «Twenty Years of debate with Stephen», en *The Future of Theoretical Physics and Cosmology* (eds. G.W. Gibbons, P. Shellard y S. Rankin), Cambridge University Press, Cambridge.

Susskind, L., Thorlacius, L., y Uglum, J. (1993), «The stretched horizon and black hole complementarity», *Phys. Rev.* 48, 3743 [hep-th/9306069].

Sutherland, W. A. (1975), *Introduction to Topology*, Oxfrord University Press, Oxford.

Swain, J. (2004), «The Majorana representations of spins and the relation between SU(1) and S Diff (S 2) [hep-th/0405004].

Synge, J. L. (1950), «The gravitational field of a particle», *Proc. Irish Acad.* **A53**, 83-114.

Synge, J. L. (1956), *Relativity: The Special Theory*, North-Holland Publ. Co., Amsterdam.

Synge, J. L. (1960), *Relativity: The General Theory*, North-Holland Publ. Co., Amsterdam.

Szekeres, G. (1960), «On the Singularities of a Riemannian Manifold», Publ. Mat. Debrecen, 7, 285-301.

't Hooft, G. (1978), «On the phase transition towards permanent quark confinement», *Nucl. Phys.* B138, 1.

't Hooft, G. (1978b), «Quantum gravity: a fundamental problem and some radical ideas», en *Recent Developments in Gravitation* (eds. M. Levy y S. Deser), Plenum, Nueva York.

Tait, P. G. (1900), «On the claim recently made for Gauss to the invention (not the discovery) of quaternions», *Proc. Roy. Soc. Edinb.* 23, 1723.

Taylor, E. F., y Wheeler, J. A. (1963), *Spacetime Physics*, W. H. Freeman, San Francisco.

Terrell, J. (1959), «Invisibility of the Lorentz contraction», *Phys. Rev.* 116, 1041-1045.

Thiele, R. (1982), *Leonhard Euler*, Leipzig (en alemán).

Thiemann, T. (1996), «Anomaly-free formulation of non-perturbative, four-dimensional Lorentzian quantum gravity», *Phys. Lett.* B380, 257-264.

Thiemann, T. (1998a), «Quantum spin dynamics (QSD)», *Class. Quant. Grav.* 15, 839-873.

Thiemann, T. (1998b), «QSD III: Quantum constraint algebra and physical scalar product in quantum general relativity», *Class. Quant. Grav.* 15, 1207-1247.

Thiemann, T. (1998c), «QSD V: Quantum gravity as the natural regulator of matter quantum field theories», *Class Quant. Grav.* 15, 1281-1314.

Thiemann, T. (2001), «QSD VII: Symplectic Structures and Continuum Lattice Formulations of Gauge Field Theories», *Class. Quant. Grav.* 18, 3293-3338.

Thirring, W. E. (1983), *A Course in Mathematical Physics: Quantum Mechanics of Large Systems*, Springer-Verlag, Berlín y Londres.

Thomas, I. (1939), *Selections Illustrating the History of Greek Mathematics*, vol. I: *From Thales to Euclid*, The Loeb Classical Library, Heinemann, Londres.

Thorne, K. (1986), *Black Holes: The Membrane Paradigm*, Yale University Press, New Haven.

Thorne, K. (1995a), *Black Holes and Time Warps*, W. W. Norton & Company (hay traducción castellana: *Agujeros negros y tiempo curvo*, Crítica, Barcelona, 1995).

Thorne, K. (1995b), «Gravitational Waves» [gr-qc/9506086].

Tipler, F. J. (1997), *The Physics of Immortality*, Anchor (hay traducción *castellana: La física de la inmortalidad*, Alianza, Madrid, 2005).

Tipler, F. J., Clarke, C. J. S., y Ellis, G. F. R. (1980), «Singularities and horizonsa review article», en *General Relativity and Gravitation*, vol. II (ed. A. Held), pp. 97206, Plenum Press, Nueva York.

Tittel, W., Brendel, J., Zbinden, H., y Gisin, N. (1998), «Violation of Bell Inequalities by Photons More Than 10 km Apart», *Phys. Rev. Lett.* **81**, 3563.

Tod, K. P., y Anguige, K. (1999a), «Isotropic cosmological singularities 1: Polytropic perfect fluid spacetimes», *Annals Phys.* **276**, 25793 [gr-qc/9903008].

Tod, K. P., y Anguige, K. (1999b), «Isotropic cosmological singularities 2: The Einstein-Vlasov system», *Annals Phys.* **276**, 294320 [gr-qc/9903009].

Tolman, R. C. (1934), *Relativity, Thermodinamics and Cosmology*, Clarendon Press, Oxford.

Tonomura, A., Matsuda, T., Suzuki, R., Fukuhara, A., Osakabe, N., Umezaki, H., Endo, J., Shinagawa, K., Sugita, Y., y Fujiwara, F. (1982), «Observation

of AharonovBohm effect with magnetic field completely shielded from the electronic wave», *Phys. Rev. Lett.* **48**, 1443.

Tonomura, A., Osakabe, N., Matsuda, T., Kawasaki, T., Endo, J., Yano, S., y Yamada (1986), «Evidence for AharonovBohm effect with magnetic field completely shielded from electron wave», *Phys. Rev. Lett.* **56**, 792-795.

Trautman, A. (1956), en Trautman, A., Pirani, F.A. E., y Bondi, H. (1965), «Lectures on General Relativity. *Brandeis 1964 Summer Institute on Theoretical Physics*, vol. I (Prentice Hall, Englewood Cliffs, NJ).

Trautman, A. (1958), «Radiation and boundary conditions in the theory of gravitation. *Bull. Acad. Polon. Sci. Ser. Sci. -Math., Astr. Phys.* 6, 407-412.

Trautman, A. (1962), «Conservation laws in general relativity», en *Gravitation: An Introduction to Current Research* (ed. L. Witten), Wiley, Nueva York.

Trautman, A. (1970), «Fibre bundles associated with space-time», *Rep. Math. Phys. (Torún)* 1, 29-34.

Trautman, A. (1972, 1973), «On the Einstein-Cartan equations I-IV», *Bull Acad. Pol. Sci.*, Ser. Sci. Math. Astron. Phys.**20**, 185-190; 503-506, 895-896, 21, 345-346.

Trautman, A. (1997), «Clifford and the Square Root' Ideas», en *Contemporary Mathematics* 203.

Turing, A. M. (1937), «On computable numbers, with an application to the Entscheidungsproblem», *Proc. Lond. Math. Soc.* 42(2), 230-265; una corrección (1937), 43, 544-546.

Unruh, W. G. (1976), «Notes on black hole evaporation», *Phys. Rev.* **D14**, 870.

Vafa, C. (1996), «Evidence for F-theory», *Nucl. Phys.* B469, 403.

Valentini, A. (2002), «Signal-Locality and Subquantum Information in Deterministic Hidden-Variables Theories», en *Non-Locality and Modality* (eds. T. Placek y J. Butterfield), Kluwer [quant-ph/0112151].

Van der Waerden, B. L. (1929), «Spinoranalyse», Nachr. Akad. Wiss. Götting. Math.-Physik, Kl. 100-109.

Van der Waerden, B. L. (1985), *A History of Algebra: From al-Khwrizmi to Emmy Noether*, pp. 166-174, Springer-Verlag, Berlín.

Van Heijenoort, J., ed. (1967), *From Frege to Gödel: A Source Book in Mathematical Logic*, 18791931, Harvard University Press, Cambridge, MA.

Van Kerkwijk, M. H. (2000), «Neutron Star Mass Determinations» [astro-ph/0001077].

Varadarajan, M. (2000), «Fock representations from U(1) holonomy algebras», *Phys. Rev.* **D61**, 104001.

Varadarajan, M. (2001), «M. Photons from quantized electric flux representations», *Phys. Rev.* **D64**, 104003.

Veneziano, G. (1968), *Nuovo Cimento*, 57A, 190.

Vilenkin, A. (2000), *Cosmic Strings and Other Topological Defects*, Cambridge University Press, Cambridge.

Vladimirov, V. S., y Volovich, I.V. (1994), *P-Adic Analysis and Mathematical Physics*, World Scientific Publishing Company, Inc:

Von Neumann, J. (1955), *Mathematical Foundations of Quantum Mechanics*, Princeton University Press, Princeton, New Jersey (hay traducción castellana del original: *Fundamentos matemáticos de la mecánica cuántica*, Consejo Superior de Investigaciones Científicas, Madrid, 1991).

Wald, R. M. (1984), *General Relativity*, University of Chicago Press, Chicago.

Wald, R. M. (1994), *Quantum Field Theory in Curved Spacetime and Black Hole Thermodynamics*, University of Chicago Press, Chicago.

Ward, R. S. (1977), «On self-dual gauge fields», *Phys. Lett.* **61A**, 81-82.

Ward, R.S., y Wells, R.O., Jr. (1989), *Twistor Geometry and Field Theory*, Cambridge University Press, Cambridge.

Weinberg, S. (1967), «A model of leptons», *Phys. Rev. Lett.* **19**, 1264-1266.

Weinberg, S. (1972), *Gravitation and Cosmology. Principles and Applications of the General Theory of Relativity*, Wiley, Nueva York.

Weinberg, S. (1989), «Precision Tests of Quantum Mechanics», *Phys. Rev. Lett.* **62**, 485-488.

Weinberg, S. (1992), *Dreams of a Final Theory: The Scientists Search for the Ultimate Laws of Nature,* Pantheon Books, Nueva York (hay traducción castellana: *El sueño de una teoría final: la búsqueda de las leyes fundamentales de la naturaleza*, Crítica, Barcelona, 1994).

Wells, R. O. (1991), *Differential analysis on complex manifolds*, Prentice Hall, Englewood Cliffs.

Werbos, P. (1989), «Bell's theorem: the forgotten loophole and how to exploit», en *Bell's Theorem, Quantum Theory, and Conceptions of the Universe* (ed. M. Kafatos), Kluwer, Dordrecht, Países Bajos.

Werbos, P. J., y Dolmatova, L. (2000), «The Backwards-Time Interpretation of Quantum Mechanics: Revisited With Experiment» [http://arxiv.org/ftp/quant-ph/papers/0008/0008036.pdf].

Wess, J., y Bagger, J. (1992), *Supersymmetry and Supergravity*, Princeton University Press, Princeton.

Wess, J., y Zumino, B. (1974), «Supergauge transformations in four dimensions», *Nucl. Phys.* **70**, 39-50.

Weyl, H. (1918), en Lorentz *et al.* (1952).

Weyl, H. (1928), *Gruppentheorie und Quantenmechanik*, Hirzel, Leipzig; traducción inglesa de la segunda edición, *The Theory of Groups and Quantum Mechanics*, Dover, Nueva York.

Weyl, H. (1929a), *Z. Phys.* 56, 330.

Weyl, H. (1929b), «Elektron und Gravitation I», *Z. Phys.* **56**, 330-352.

Wheeler, J. A. (1960), *Neutrinos, Gravitation and Geometry: contribution to Rendiconti della Scuola Internazionale di Fisica' Enrico Fermi-XI, Corso, July 1959*, Zanichelli, Bolonia (reimpresión en 1982.)

Wheeler, J. A. (1965), «Geometrodynamics and the issue of the final state», en *Relativity, Groups and Topology* (eds. B. S. y C. M. DeWitt), Gordon and Breach, Nueva York.

Wheeler, J. A. (1973), «From Relativity to Mutability», en *The Physicist's Conception of Nature* (ed. J. Mehra), D. Reidel, Boston, pp. 202-247.

Wheeler, J. A. (1975), «Assessment of Everett's Relative State' Formulation of Quantum Theory», *Rev. Mod. Phys.* **29**, 463-465.

Wheeler, J. A. (1983), «Law without law», en *Quantum Theory and Measurement* (eds. J. A. Wheeler y W. H. Zurek), Princeton Univ. Press, Princeton, 182-213.

Whittaker, E. H. (1903), «On the partial differential equations of mathematical physics», *Math. Ann.* **57**, 333-355.

Wick, G. C. (1956), «Spectrum of the Bethe-Salpeter equation», *Phys Rev.* 101, 1830.

Wigner, E. P. (1960), «The Unreasonable Effectiveness of Mathematics», *Commun. Pure Appl. Math.* **13**, 114.

Wilder, R. L. (1965), *Introduction to the foundations of mathematics*, John Wiley & Sons, Nueva York.

Wiles, A. (1995), «Modular elliptic curves and Fermat's Last Theorem», *Ann. Maths* 142, 443-551.

Williams, R.K. (1995), «Extracting X-rays, g-rays, and Relativistic e^-e^+.Pairs from Supermassive Kerr Black Holes Using the Penrose Mechanism», *Phys. Rev.* **D51**, 5387.

Williams, R.K. (2002), «Production of the High EnergyMomentum Spectra of Quasars 3C279 and 3C273 Using the Penrose Mechanism» [astro-ph/0306135]. Aceptado para publicación en *Astrophysical Journal*, 2004.

Williams, R.K. (2004), «Collimated Escaping Vortical Polar e^-e^+ Jets Intrinsically Produced by Rotating Black Holes and Penrose Processes [astro-ph/0404135].

Willmore, T. J. (1959), *An Introduction to Differential Geometry*, Clarendon Press, Oxford.

Wilson, K. (1975), *Phys. Reps.* 23, 331.

Wilson, K. (1976), «Quarks on a lattice, or the colored string, model», *Phys. Rep.* **23(3)**, 331-347.

Winicour, J. (1980), «Angular momentum in general relativity», en *General Relativity and Gravitation*, vol. 2 (ed. A. Held), pp. 71-96, Plenum Press, Nueva York.

Witten, E. (1978), «An interpretation of classical Yang-Mills theory», *Phys. Lett.* **77B**, 394-398.

Witten, E. (1981), «A new proof of the positive energy theorem», *Comm. Math. Phys.* **80**, 381-402.

Witten, E. (1982), «Supersymmetry and Morse theory», *J. Diff. Geom.* **17**, 661-692.

Witten, E. (1988), «Topological quantum field theory», *Commun. Math. Phys.* **118**, 411.

Witten, E. (1995), «String theory in various dimensions», *Nucl. Phys.* **B443**, 85.

Witten, E. (1996), «Reflections on the Fate of Spacetime», *Phys. Today*, abril de 1996.

Witten, E. (1998), «Anti de Sitter Space and Holography [hep-th/9802150].

Witten, E. (2003), «Perturbative Gauge Theory as a String Theory in Twistor Space» [hep-th/0312171].

Witten, L. (1959), «Invariants of general relativity and the classification of spaces», *Phys. Rev.* **113**, 357-362.

Wolf, J. (1974), *Spaces of Constant Curvature*, Publish or Perish Press, Boston, MA.

Woodhouse, N. M. J. (1991), *Geometric Quantization*, 2.ª ed., Clarendon Press, Oxford.

Woodin, W. H. (2001), «The Continuum Hypothesis. Parts I & II. Notices of the AMS», disponible online en: http://www.ams.org/notices/200106/fea-woodin. pdf.

Wooters, W. K., y Zurek, W. H. (1982), «A single quantum cannot be cloned», *Nature* **299**, 802-803.

Wykes, A. (1969), *Doctor Cardano: Physician Extraordinary*, Frederick Muller, Londres.

Yang, C. N., y Mills, R. L. (1954), «Conservation of Isotopic Spin and Isotopic Gauge Invariance», *Phys. Rev.* **96**, 191-195.

Yui, N., y Lewis, J. D. (2003), *Calabi-Yau Varieties and Mirror Symmetry*, Fields Institute Communications, V. 38, American Mathematical Society, Providence, RI.

Zee, A. (2003), *Quantum Field Theory in a Nutshell*, Princeton University Press, Princeton.

Zeilinger, A., Gaehler, R., Shull, C. G., y Mampe, W. (1988), «Single and double slit diffraction of neutrons», *Rev. Mod. Phys.* **60**, 1067.

Zel'dovich Ya, B. (1966), «Number of quanta as an invariant of the classical electromagnetic field», *Soviet Phys.-Doklady* **10**, 771-772.

Zimba, J., y Penrose, R. (1993), «On Bell non-locality without probabilities: more curious geometry», *Stud. Hist. Phil. Sci.* **24**, 697-720.

Zinn-Justin, J. (1996), *Quantum Field Theory and Critical Phenomena*, Oxford University Press, Oxford.

Índice alfabético

Abel, Niels Henrik, 357
abelianos, grupos, 357
aberración estelar, 584
absoluta, función, 175
acumulación, punto de, 338
Adams, W. S., 632
afín, espacio, 411
afín, parámetro, 426, *427*, 443
agujeros blancos, 1112-1115, 1127-1132, *1128*
agujeros de gusano, 1115, *1116*
agujeros negros, 951-957, *954*; agujero blanco como inverso temporal de, 1127-1132, *1128*; cuerdas y entropía de, 1225-1230; entropía de, 960-963; ergosfera, 1121, *1121*; evaporación de, *1124*; explosiones de Hawking, 1122-1127; horizonte de sucesos, *954*, 955, 957-959; localización del horizonte, *1125*; radiación de Hawking, 1119-1122, *1120*; singularidades, 957-959; sistemas de estrellas dobles, 960-963; temperatura a partir de periodicidad compleja, 1108-1115, *1110*, *1112*; temperatura de Hawking, 1103-1108, *1108*; vectores de Killing, 1106, *1107*, 1116-1117, *1117*
Aharonov-Bohm, efecto, 615-616, *617*
Ahmavaara, Y., 491
Alexander, polinomio de, 1264
α, desintegración, 851
álgebra, teorema fundamental del, 136
álgebra compleja, geometría de, 151-155, *152*
álgebra graduada, 314
álgebras de dimensión infinita, 887-890
Amati, Danielle, 1188
Amphos, 43-45
amplitudes de probabilidad, 729
Anderson, Carl, 840
ángulos rectos, 74, *74*
anillo con identidad, 294
anillo de convergencia, 242-243, *243*
aniquilación, operadores de, 885, *886*
anomalías, 1097
anti-De Sitter, espacio, 1004-1007, *1005*, 1231
antiholomorfas, funciones, 1215
antilinealidad, 397, 400
antineutrón, 845
antipartículas, 125-126, 819-820, 844-845; campos cuánticos, 821, *823*; positrones, 836-840, *839*, 856; teoría cuántica de campos (QFT), 890-892; viaje hacia atrás en el tiempo, 858
antipodal, punto, 306
antiprotones, 126, 844-845
antisimetría, 403
antrópico, principio, 1016-1021, *1017*, 1379
antrópico débil, pincipio, 1017, *1017*
antrópico fuerte, principio, 1017, *1017*
año luz, 553
aplicaciones conformes, 218-222, *219*
Argand, Jean Robert, 143
Aristóteles, 525-527, 540
armónicos esféricos, 755-761, *758*
armónicos superiores, *237*, 238, *239*
Ashtekar, Abhay, 1176, 1253, 1360
Ashtekar, variables de, 1252, 1288; forma de las, 1257-1260; ingrediente quiral de las, 1253-1257
Ashtekar-Rovelli-Smolin-Jacobson, teoría de variables de lazo de, 499
asíntota, 96-97, *96*
autoadjunto, propiedad, 727
autofunciones (o autoestados), 684, *688*; de interacciones fuertes, 873
autovalores y autovectores, 374-378, *376*
axiomas, 52, 75; de elección, 57-58, 502-503, *503*

Banach-Tarski, teorema de, 503
Barbero-Immirzi, parámetro de, 1259, 1274-1275
bariones, 866; número en el universo, 977
Barrett, John, 1279
Becquerel, Henri, 844
Bekenstein, Jakob, 1104
Bekenstein-Hawking, entropía de, 961, 977, 1103-1104, 1225, 1228, 1230, 1234, 1274
Bell, John S., 784, 1077; desigualdad de, 785, 1156, 1382
Beltrami, Eugenio, 88, 1384; disco de Poincaré de, 219; geometría de, *89*, 95-97, 579
β, desintegración, 851-855, *852*, *853*
Bethe, Hans Albrecht, 882
Bianchi, Luigi: identidad de, *424*, 424-425, 484, 626; simetría de, *424*, 424
big bang, 864, 944-950, *946*; extraordinariamente especial, 975-982; principio antrópico, 1022-1025; punto de origen, 969; radiación de fondo de microondas, 947, *976*, 1370
big crunch, 966, 977, *978*
blandas, estructuras, 447
Bloch, Ernst, esfera de, 1064-1069
Bohm, David, 784; experimentos EPR tipo, 787-792, 806
Bohm/Hiley, ontología de la «onda-piloto», 1054, 1087
Bohr, Niels, 697, 718, 1051, 1079, 1393
Boltzmann, Ludwig, 600, 1384; constante de, *679*, 929, 961; ecuación de, 936; entropía de, 929-930, *930*; estadística de, 932-933; fórmula de, 1225
Bolyai, János, 95, 1384
Bombelli, Raphael: *L'Algebra*, 131, 134, 137-138, 157
Bondi, Hermann, 572, 634, 1006
Born, Max, 725
Bose-Einstein, condensación de, 801, 1049
bosones, 169, 798-801, 845, 851; estados cuánticos, 801-804; grupo de simetría electrodébil, 860-865
Bouwmeester, Dik, 1148, 1355
BPS, estado, 1227, 1229, 1236
bra, vector, 721
Brahmagupta, matemático hindú, 121, 157
Broglie, Louis-Victor, príncipe de, 677-678; frecuencia de, 848
Brooks, R., 60

C^∞-suave, funciones, 182-186
C^ω-suave, funciones, 186
Cabibbo, ángulo de, 874, 1388
cadena, regla de la, 278-279
Calabi, Eugenio, 1218
Calabi-Yau, espacios de, 1208, *1212*, 1218-1222, 1241, 1263, 1327, 1394
cálculo espinorial, 834
cálculo exterior, 333; teoría fundamental del, 337
cálculo infinitesimal, 118-119, 173-176; teorema fundamental del, 173, 189, *190*
campo de covectores, 325, *327*, *328*
campo vectorial, 279-280, *280*, 654-655; constancia, 414; derivada covariante, 418-422, *419*
campos escalares, 274, *280*
campos espinoriales, 1219
campos finitos, 491-494
campos sin masa, ecuación de, 1318
Candelas, Philip, 1222, 1239, 1392
Cantor, Georg, 113, 115, 122, 499-500, 504-512; teoría del infinito, 1390
caóticos, sistemas, 926-927, *926*
Cardano, Girolamo: *Ars magna*, 131, 134, 136-137, 157
cardinalidad, 500-502
carga, conjugación de, 857-860
carga desnuda, 910
carga eléctrica, 124-125, 537, 611; conservación de la, 607-609, *608*
carga-corriente, vector de, 602
Carnot, Sadi, 600, 928
Cartan, Élie, 333-335, 363, 601, 1384; espaciotiempo newtoniano de, 539-546, 557, 1138
Carter, Brandon, 1016-1017, 1113
cartesianas, coordenadas, conversión a polares, 154, *155*, *156*
Casimir, operadores de, 763-764
categorías, teoría de, 1286
Cauchy, Augustin, 141, 198; fórmula de, 203-204
Cauchy-Riemann, ecuaciones de, 197, 199, 203, 218, 286-290, 349, 1316
Cayley, Arthur, 298, 370
censura cósmica, 959, 1029, *1030*
cero, introducción del, 121
Chadwick, James, descubrimiento del neutrón, 844
Chan Hong Mo, 875-876
Chandrasekhar, Subrahmanyan, 953, 960
Chen Ning Yang, 309, 852
Chew, Geoffrey, 1362
Chien-Shiun Wu, 309, 852
Christoffel, Elwin, 443

cinética, energía, 587-588
clase de equivalencia, 260
clases, 511-512
Clausius, Rudolf Julius Emanuel, 600; entropía según, 929
Clifford, William Kingdom, 297, 1359; álgebras de, 299, 306-310, 370, 831, 833, 834-835, 837, 850, 887, 1384; estructura de paralelos de, 563; fibrado de, 462-467, 462, *463*, 465-467, *466*, 476, 485, 757; paralelos de, 464, *465*, 1313, *1314*
Cohen, Paul, 509
cohomología, 1321-1327, *1326*; objetos imposibles, 1327, *1327*
cohomología de haces holomorfa, 1324
Coma, cúmulo, 530
combinatorio, enfoque, 120
compacto, término, 337-338, *337*, *339*
complejas, dimensiones, 269-271
complejo conjugado, 269-270, *270*
complexificación, 388, 394, 566
componentes, 284, *285*
Compton, Arthur, 680; proceso de dispersión de, *859*
computación, 513-516
computadores cuánticos, 782
conexión, 418-422; riemanniana, 443
congruencia, 76
conjugación hermítica, 399-400, *401*
conjunto nulo o vacío, 122
conjuntos, 122-123, 511-512
conmutación, relación de, 669-670
conmutadores del grupo, 380-381
Connes, Alain, 1286
conservación de la carga eléctrica, 607-609, *608*
constante cosmológica, 627-629, 1007, 1035
constante de acoplamiento de cuerda, 1228
constante de estructura fina, 913
constantes de estructura, 381-382, *382*
continuo, hipótesis del, 509
contorno, integración de, 199-203; contornos cerrados, 201-203, *202*; prolongación analítica, 206-211
convenio de suma, 342-343
Cooper, pares de, 801
coordenadas polares, 153-154, *155*, *156*; esféricas, 756, *757*
Copenhague, interpretación de, 1051, 1079, 1086, 1154, 1157, 1379; gato de Schrödinger, 1078-1081, *1078*
cosmología, 963-970; *véase también* constante cosmológica; parámetros cosmológicos
Cotes, Roger, 163
covariancia general, principio de, 623, *643*, *660*
cromodinámica cuántica (QCD), 866
cuadrados, 72-75, *73*, *74*, 77-78; geometría hiperbólica, 90-91, *91*, *92*; simetrías de, *357*, *362*
cuanlanzamiento, 556, 777, 810-816, *811*, *813*
cuántica de campos (QFT), teoría, 881-884; álgebras de dimensión infinita, 887-890; antipartículas en, 890-892; construyendo diagramas de Feynman, 903-907, *905*, *906*; diagramas de Feynman a partir de lagrangianos, 913-915; diagramas de Feynman y la elección del vacío, 915-918; integrales de camino divergentes, 900-903; interacciones, 894-900, *897*, *898*; operadores de creación y aniquilación, 883-887, *886*; renormalización, 907-913; vacíos alternativos, 893-894
cuantización, 671; geométrica, 673
cuantos, 119
cuásares, 959; emisión de energía, 1122, *1123*
cuaterniones, álgebra de los, 293-296; anillo de división, 295; conjugados, 295; forma cuadrática, 297; geometría de, 299-303, *300*; inversos, 295; papel físico para los, 296-299
cuerdas, teoría de, 654; como gravedad cuántica, 1193-1197; dimensiones espaciotemporales extra, 1191-1193; dinámica de, 1197-1201; entropía de agujero negro y, 1225-1230; estatus físico de la, 1238-1242; hacia una teoría del universo, 1187-1191; historia de cuerdas, *1216*; problema de los moduli, 1208; QFT de cuerdas finita, 1215-1217; segunda revolución de las supercuerdas, 1221-1223; supersimetría incorporada a la, 1192-1193; teoría hadrónica original, 1182-1187
curvatura, 429-432, *430*, *440*, *442*
curvatura de funciones, *175*, 177, *177*, 180-181, *180*; punto de inflexión, 181
curvatura gaussiana, 431

D'alembertiano (operador de onda), 830
Davisson, C. J., 676
Davisson-Germer, experimento, *677*
D-brana, perspectiva de la, 1234-1238, *1236*, *1237*
De Broglie, ontología de la «onda-piloto», 1054, 1087
De Broglie-Bohm, interpretación cuántica de, 1381
De la Ossa, Pedro, 1239

De Sitter, espacio de, 1003, 1005-1006
De Sitter, espaciotiempo de, *1004*, 1007
decoherencia por el entorno, 1075-1077; filosofía FAPP de la, 1075-1077
Dedekind, Richard, 113, 115, 120, 123, 133
definida positiva, 402
deformaciones homólogas, 200, *201*
deformaciones homotópicas, 200
degeneración, presión de, 952
degeneraciones, 376
demostración por contradicción, 93
densidad de un objeto, 332
derivada exterior, 335-341, *337*
derivadas parciales, 275
desacoplamiento del universo, 967-968, *968*
Desargues, Gérard, teorema de, 471-472, *473*, 495, *496*
desarrollo perturbativo, 913
desplazamiento hacia el rojo, 947, *954*
determinantes de una matriz, 372-374, *372*
dextrógiro, gravitón no lineal, 1336-1337
diagramas conformes, 971-975, *973*, *975*
diagramas de árbol, 907, *907*
Dicke, Robert, 1016
diferenciabilidad, 274-276
diferenciación, 173-174, 177-178; derivada covariante, 418-422, *429*; derivada de Lie, 433-440, *433*; derivada exterior, 335-341, *337*; derivada segunda y de orden superior, 178-184, *179*, *180*, 274-275; en una variedad, 411-413; funcional, 658-659; método de Heaviside, 668; reglas de, 186-189; relación de conmutación, 669
dimensiones internas, 452
dinámica newtoniana, 531-535, *534*; evolución fuera de la, 599-602
Diofanto: *Arithmetica*, 55
Dirac, Paul Adrien Maurice, 28, 118, 125, 308, 680, 721, 785, 1016, 1114, 1165, 1252, 1357, 1363, 1366, 1384; antipartículas de, 828, 831-840, 849, 881; cable de, 993, *994*; ecuación de onda para el electrón de, 833-836, 838, 843, 845-846, 849-850, 854-855, 891, 901, 1010, 1190, 1389, 1393; función delta de, 263-264, 702, 1261; mar de, 891-894; mecánica cuántica de, 769, 1093; ruta al positrón de, 836-840; teoría cuántica de campos (QFT), 882; *Los principios de la mecánica cuántica*, 725
Dirac-Weyl, ecuación del neutrino, 1318
Dirichlet, Peter G. Lejeune, 1236
Dirichlet: problema de, 1236; series de, 211
disco cerrado, 207
discos mágicos, 495, *496*
distancia en el espacio físico, 114, 117
distribución gaussiana, 935, *935*
distribuciones, 194, 264
Doppler, desplazamiento, 1039
Doppler, efecto, 947
dos rendijas, experimento de las, 681-682, *681*
Dowker, F., 1086
Duffing-Kemmer-Petiau, ecuación de, 1363-1364
Duggins, Andrew, 1382

e, base de los logaritmos naturales, 159
Eddington, sir Arthur, 632, 937-938, 1027, 1165
Eilenberg, Samuel, 1286
Einstein, Albert, 118, 120, 533, 535, 539, 546, 599-600, 1029, 1037, 1168, 1198-1199, 1283, 1357, 1369, 1372, 1384, 1403-1404; campo gravitatorio de, 599, 1393; cilindro o universo de, 1234; constante cosmológica de, 627, 1041; convenio de suma de, 342-343; ecuación de vacío de, 1200, 1336, 1344; ecuación $E=mc^2$, 617, 707, 821, 1141; ecuaciones de campo de, 425, 559, 622-628, 630, 958, 1006-1007, 1108, 1251; efecto fotoeléctrico, 678; espaciotiempo curvo de, 412; principio de covariancia general, 1136-1137; relatividad general de, 28, 52, 97, 101, 114, 214, 217, 272, 412, 416, 425, 432, 451, 535, 543, 546, 557-559, 574, 628, 632, 667, 883, 911-912, 1094, 1096, 1135, 1150, 1178, 1190, 1196, 1199, 1227, 1256, 1337, 1351-1353, 1371, 1373; y el big bang, 946
Einstein-Cartan-Sciama-Kibble, teorías de espín/torsión de, 423
Einstein-Maxwell, teoría combinada de, 1181-1182, 1227
Einstein-Podolski-Rosen (EPR), experimentos, 784-785, *785*, 787-792, 804, 807-808, 814
electrodébiles, interacciones, 850-857
electrodinámica cuántica (QED), 849
electromagnético, potencial, 610-611
electrones, 125, 866; ecuación de Dirac, 833-834; funciones de onda, 689; imagen zigzag del, 845-850, *848*; pares de Cooper, 801; polarización del vacío y la carga de los, 909-910
elementos discretos, geometría con, 1283-1288, *1285*
elipsoide de revolución, 542-543, *543*, *544*

elipsoides, 366, *367*
Elitzur, Avshalom, 734
Elitzur-Vaidman, prueba de la bomba de, *736*, 812, *813*
Ellingstrud, Geir, 1222
enanas blancas, estrellas, 953
energía, 119; dificultades con la formulación relativista, 826-828; equipartición de la, 935; formulación de Newton, 587; formulación relativista, 590-591; positividad en mecánica cuántica, 823-825
energía oscura, 627-628
energía potencial, 587-588
energía térmica, 588
energía-momento, tensor, 617-622
enlaces y nudos, 1264-1267, *1265*
ensilladura, punto de, 643, *644*
entrelazamiento cuántico: desigualdades de Bell, 783-787; dos misterios del, 795-798, *796*; *véase también* cuanlanzamiento
entropía, 928-931, *930*; asimetría temporal, 938, *939*; carácter robusto del concepto de, 931-936; de agujeros negros, 960-963; sistemas aislados, 940-944
equilibrio térmico, 934, *934*
equivalencia, principio de, 535-539
ergosfera, 1121, *1121*
escalares, cantidades, 124
escalares, productos, 285
Escher, M. C., 969, *82*, *83*, *88*, *99*
esfera celeste, como una esfera de Riemann, 583-586
esferas: simetría rotacional, *359*, *362*; transformaciones lineales hacia un elipsoide, 366, *367*
espacio base, 455
espacio de configuración, 317, *318*, 640, *641*
espacio de momentos, 703-707, *706*
espacio dual, 383-385, *386*
espacio orientado, 338-340, *339*
espacio tangente, 326
espacio vectorial, 294; estructura del, 411
espacios de fases, 321-322, *321*, *640*, 929, *930*
espaciotiempo aristotélico, 525-528, *526*, *528*
espaciotiempo, 118, 296; abandono del tiempo absoluto, 552-556, *553*; ausencia aparente de dimensiones especiales extra, 1201-1207; conos de luz, 548-551, *549*, *550*, *552*, *553*; de dimensiones más altas, 1177-1182, *1179*, 1191-1193; de la física aristotélica, 525-528, *526*, *528*; de la relatividad general de Einstein, 557-559, *557*, *558*; dinámica newtoniana, 531-535, *534*; espacio minkowskiano, 554-555, *555*, *565*, *567*; estabilidad cuántica, 1207-1211; inestabilidad clásica de las dimensiones extra, 1211-1215; newtoniano, 539-546, *541*, *543*, *544*; principio de la equivalencia, 535-539; principio holográfico, 1231-1234; relatividad galileana, 528-531, *530*; teoría M, 1224-1225; velocidad fija de la luz, 546-548, 552
espaciotiempo newtoniano, 539-546, *540*, *543*, *544*; energía y momento angular, 587-590
espín, 712, 735-736, 739-745; bosones y fermiones, 798-799; espín más alto, 752-755, *752*; grados de libertad de, 745-747; mecanocuántico, 120; total, 1268-1269; *véase también* espumas de espín; redes de espín
espinores y objetos espinoriales, 301-303, *302*, 309, 452, 468, 739-745; fermiones, 798-799; twistores como espinores de dimensión superior, 1301-1304
espiral logarítmica, 164, *165*
espirales equiangulares, 164, *165*
espumas de espín, 1278
estacionarios, puntos, 642-643, *643*
estados de posición, 701-703, *706*
estrellas de neutrones, 631, 953; ondas gravitatorias emitidas por el sistema PSR 1913+16, 631-632, *631*, 633
estructura hermítica, 411
euclideanización, 1031, 1033
Euclides, geometría de, 52, 100, 109, 117, 118, 411, 442, 576, 1383; postulados de, 75-78, 93-94, 105, 414, 577
Eudoxo, 113, 157
Euler, Leonhard, 140, 194, 206, 265, 639, 908, 1317, 1384; fórmula de, 162-163, 174, 182-186, 198; operador homogeneidad de, 1317
Euler-Lagrange, ecuaciones de, 642, *643*, 644, 658-659, 661
Euler-Riemann, función zeta de, 211
Everett III, Hugh, 1052
evolución unitaria, *véase* Schrödinger, evolución de
excitones, 1209-1210
existencia, naturaleza de la, 54-55; tres mundos y tres profundos misterios, 61-66, *61*, *65*
exponenciación, 157
extirpación, teorema de, 261, *262*

factoriales, 159
Fano, plano de, 497, *497*
FAPP, filosofía, de la decoherencia por el entorno, 1075-1078, 1085, 1127, 1144, 1154-1155

Faraday, Michael, 599-600
Fermat, Pierre de, 55-57, 173, 1359, 1384; último teorema de, 55, 518, 1241, 1390
Fermi, Enrico, 851-852
Fermi-Dirac, estadística de, 932
fermiones, 169, 798-801; estados cuánticos de, 801-804
ferromagnetismo, 988-991, *990*, *993*
Feynman, Richard P., 833, 1285, 1357; diagramas de, 822, 847, 849-850, 858, 860-863, 872, 882, 902-907, 909, 913-918, 1178, 1183-1185, *1183*, 1209, 1339; integrales de camino de, 896, 1343, 1389; propagador de, 902
fibrado cotangente, 470-471, *470*, 640
fibrado, 451-454, *455*; cotangentes, 470, *470*; curvatura, 481-486, *482*, *485*; de Clifford, 462-467, *463*, *466*; espacios proyectivos, 471-476, *472*, *473*; espaciotiempo galileano, 530-532; idea matemática de, 454-459, *456*, *457*; no trivialidad en las conexiones, 476-480, *477*, *480*; secciones transversales de, 459-462, *460*; tangentes, 469, *470*; vectoriales complejos, 467-471, *468*
Finkelstein, David, 1285
física cuántica: armónicos esféricos, 755-761, *758*; asimetría temporal en la reducción del estado cuántico, 1098-1103, *1099*; base experimental de la, 676-682; bosones y fermiones, 798-801; conflicto con los principios de Einstein, 1136-1141, *1139*; decoherencia por el entorno, 1053, 1075-1077; descripción en el espacio de momentos, 703-707, *706*; desigualdades de Bell, 783-787; distribución de probabilidad en una función de onda, 698-701; dos misterios del entrelazamiento cuántico, 795-798, *796*; dualidad onda-partícula, 682-685, 690; ecuación de Schrödinger, 674-676, 823-824; ejemplo EPR de Hardy, 792-794, *794*; enormidad del espacio de estados de muchas partículas, 780-783; esfera de Riemann de los sistemas de dos estados, 745-752, *747*, *749*; espacio de Hilbert, 720, 730, *730*; espín más alto, 752-755, *754*; espín y espinores, 739-745; estados cuánticos de bosones y fermiones, 801-804; estados de posición, 701-703; evolución unitaria, 722-726; experimentos EPR tipo Bohm, 787-792; funciones de onda, 674; helicidad, 737-739, 765, 1253; imagen de Heisenberg, 723-726; imagen de Majorana, 752-755; linealidad de U y sus problemas para R, 715-178; matrices densidad para espín 1/2, 1064-1069; matriz densidad en situaciones EPR, 1069-1074, *1074*; matriz densidad, 1061-1064; medidas nulas, 733-739, *735*; medidas sí/no, 730-733, 738-739; momento angular cuántico relativista, 761-765, *763*; naturaleza holística de una función de onda, 691-696; notación de Dirac, 720-721; objeto cuántico aislado general, 765-771; observables, 726-730; ontologías convencionales, 1049-1053; ontologías no convencionales, 1054-1061; positividad de la energía, 823-825; procedimientos **U** y **R**, 711-715, *714*, 1050, 1059; proyectores, 730-733, *733*; realidad cuántica, 685-690; saltos cuánticos, 696-698; sistema de muchas partículas, 777-780; teleportación, 804-809, *809*; tensión con la relatividad, 819-821, 826-828; teoría cuántica twistorial, 1314-1318
física de partículas, 168-170; conjugación de carga, 857-860; grupo de simetría electrodébil, 860-865; imagen zigzag del electrón, 845-850, *847*, *848*; interacciones electrodébiles, 850-857; más allá del modelo estándar, 874-876; orígenes del modelo estándar, 843-845; partículas fuertemente interactuantes, 866-869; quarks coloreados, 870-874
FitzGerald-Lorentz, efecto de achatamiento de, *585*, 586, *587*
FLRW, modelo cosmológico, 1006-1008, 1015, 1028, 1030, 1151, 1153-1156
Fock, V. A., espacio de, 895, 1233
Fontana, Niccolò, 137
formas hermíticas, 397-399
fotoeléctrico, efecto, 678
fotones, 549, 555, 583, 594, *594*, 606; efecto fotoeléctrico, 678; espín del, 737; experimento de la doble rendija, 681-682, *681*; grupo de simetría electrodébil, 860-861; helicidad, de, 737-739; polarización, *737*, 750, *751*, *752*; subconversión paramétrica, 811, *811*
Fourier, Joseph, 255 análisis de, 236, 653, 761; componentes de, 648; descomposición de, *239*, 240-241; series de, 199, 235-240, *252*, 258, 289, 1205; transformada de, 248-254, 705, 729, 761
Fowler, William, 1018
fracciones: continuas, 110; desarrollo decimal periódico de, 109
frecuencia angular, 238
frecuencia normal de vibración, 649, 652

Frege, Gottlob, 501
Friedmann, Alexandr, 946, 964, 967-968
Friedmann, modelo de, 1020
Friedmann-Lemaître-Robertson-Walker (FLRW), modelos de, 964-965, *966*, 977, 980-981, 999
fuerza débil, 844
fuerza fuerte, 843-844; libertad asintótica de la, 912
fuerzas fundamentales, 47-50
fuerzas, ley de, 533, *534*
fullerenos, 1354-1355
función analítica, 185-186
función exponencial, 159-162, *161*; potencias complejas, 164-167
funciones, 173-176, *174*; C^∞-suave, 182-186; C^ω-suave, 186; de transición, 277, 323, *323*; hiperfunciones, 258-265, *260*; noción euleriana de, 184-186; pendientes de, *175*, 176-178, *177*; periódicas, 236, *237*, *239*
funciones complejas: integración de contorno, 199-203; series de potencias a partir de la suavidad compleja, 203-211; suavidad compleja, 197-199, 324
funciones continuas y discontinuas, *175*, 176-178, *177*
funciones de onda normalizadas, 698
funciones suaves, 176-178, *177*; C^∞-suave, 182-186; C^ω-suave, 186; derivadas de orden superior, 178-183, *179*, *180*

Galileo Galilei, 118, 500, 525, 528-529, 531, 535, 538, 540, 546, 649, 923, 1371; principio de relatividad de, 531, 588
Galileo-Newton, dinámica de, 531
gauge, conexiones, 451-452, 611
gauge, *véase* invariancia; monopolos; transformación
gauge no abeliana, teoría, 872, *873*
Gauss, Carl Friedrich, 95, 143, 334, 601, 639, 1384; ley de, 608-609, *609*
Gell-Mann, Murray, 867-868
género de una superficie, 226-227, *226*, *227*
geodésicas, 425-428, 443; construcción de paralelogramos, 427-429, *428*, 430; desviación, 432, *432*
geometría elíptica, 98, *99*
geometría hiperbólica: aspectos históricos, 92-97; cuadrados en, 90-91, *91*; distancia hiperbólica, 84, 86-87; en el espacio de Minkowski, 576-582, *576*, *577*, *578*, *579*; existencia de la, 90; imagen conforme, 81-86, *82*, *83*, *84*; pseudoesferas, 96-97, *96*; pseudorradio de la, 85; relación con el espacio físico, 97-101, *99*; representación hemisférica, 89-90, *89*, *91*; representación proyectiva, 86-87, *87*, *88*
geometría infinita, 494-499
geometría no conmutativa, 1287-1288
geometría plana uniforme, *99*
geometría tridimensional, 49
geometrías riemannianas, 100-101
Germer, L. H., 676
Geroch, Robert, 1363
Gibbons-Perry, argumento de, 1108, 1111, *1112*, 1115
giromagnética, razón, 1114
gluones, 125, 845, 872, *873*
Gödel, Kurt, 509; teorema de la incompletitud de, 507, 510, 517-518, 1382, 1390
Gold, Thomas, 1006
Goldbach, conjetura de, 518
googly, problema, 1337
Goudsmit, Samuel, 836
gran unificación, teoría de (GUT), 874, 992, 1171
granulado grueso, 929-930, *930*
Grassmann, Hermann, 299; álgebras de, 310-314, 370, 1170, 1174; producto de, 329-332
gravedad, 48, 950, 981; ausencia de, 977; autoenergía, 1139, *1140*; densidad de energía, 617; efecto de marea, 542-546, *543*, *544*; Einstein sobre la, 535, 623; energía del campo gravitatorio, 629-635; experimento de Galileo en la torre de Pisa, 535-539, *536*; ley de Newton, 534; ondas gravitatorias, 630-632, *631*; velocidad de escape, 952
gravedad cuántica, 1251-1255; estatus de variables de lazo, 1275-1279; quiralidad de gravitones, 1253-1257; teoría de cuerdas como, 1193-1197; variables originales de Ashtekar, 1253-1260, *1258*
Graves, John, 298
gravitación cuántica, 120, 959, 1031
gravitones: helicidad, 1253-1257; no lineales, 1330-1336, *1331*, *1332*
Green, Michael, 334, 601, 1187-1188, 1192
grupo cíclico, 166
grupo lineal general, 371
grupo producto, 364, 456
grupos clásicos, 363, 407
grupos excepcionales, 363-364
grupos simples, 363-364
grupos unitarios, 396-402
Gurzadyan, Vahe, 1040, 1370
Guth, Alan, 1002, 1011; *El universo inflacionario*, 1010

haces twistorial, cohomología de, 1321-1322, *1322*
hadrones, 169, 866
Hall, Aspeth, 1094
halográfico, principio, 1231-1234, *1231*
Hamerof, Stuart, 1381
Hamilton, William Rowan, 293-297, 303-304, 639, 831-832, 1358; ecuaciones de, 446; ecuaciones del Puente Brougham de, 303, 308; principio cuántico de, *899*, 899, 1358
Hamilton, principio de, o de acción estacionaria, 642, *643*, 645-646, 659-660, *660*
hamiltoniana, imagen, *641*, 641, 644-648, *647*; geometría simpléctica, 654-657, *657*; pequeñas oscilaciones, 648-654
hamiltonianos cuánticos, 671-673
Hamilton-Jacobi, ecuación de, 1358
Hardy, G. H., 93
Hardy, Lucien, 785, 1072, 1084; ejemplo EPR de, 792-794, 796, 813
Hariot, Thomas, 94, *95*
Hartle, James, 1026
Hartle-Hawking, «ausencia de frontera» de, 1030-1034, 1042, 1095, 1369
Hausdorff, espacio de, *323*, 324
Hawking, Stephen, 1026, 1034, 1053, 1216, 1360-1361; agujeros negros, 1077; caja de, *1130*, *1131*; radiación de, 1107, 1129, 1229; temperatura de agujero negro de, 1103-1109, *1108*, *1112*, 1113, 1119-1120, *1120*, 1226-1227
Hawking, explosiones de, 1122-1127
Heaviside, Oliver, 668-669; función escalón de, 175, 263
Heisenberg, Werner Karl, 1242, 1384; ecuaciones de movimiento de, 724; principio de incertidumbre de, 705, *706*, 707, 895, 908, 1141-1142, 1151-1153, 1293, 1354
helicidad, 736-739, 765, 1253; twistores, 1311
hemisférica, representación, de la geometría hiperbólica, 89-90, *89*
Hermite, Charles, 397
Hertz, Heinrich, 548, 607, 678
Higgs, bosón de, 1172, 1288, 1296
Higgs, partícula (campo) de, 845, 864-865, *865*, 1008, 1164
Hilbert, David, 662
Hilbert, espacio de, 38, 520, 720-724, 727, 729-730, 732, 741, 748, 782-783, 786, 795, 887-888, 892-893, 1055-1056, 1063, 1067, 1074, 1131, 1288, 1384; lagrangiano gravitatorio de, 662
hiperfunciones, 258-265, *260*, 520; teorema de extirpación, 261, *261*
hipersuperficie, 1175, *1176*
Hodge, duales de, 603-605, *604*, *605*
Hodges, Andrew, 1339
holomorfa, sección transversal, 469
holomorfas, funciones, 198, 220-221, *220*, 1215
HOMFLY, polinomios, 1265
Hopf, Heinz, 462; fibración de, 462
horizonte, problema del, 1003
horizonte de sucesos, 954, *954*, 957-959
Hoyle, Fred, 1006, 1018
Hubbard, L. Ronald, 148
Hubble, Edwin, 946
Hughston, Lane, 1339
Hulse, Russell, 631, 633
Hurwitz, Adolf, 298
hyperkhaler, variedades, 1341

i, raíz cuadrada de −1, 127, 131-134, 153, *154*; a la potencia i, 165; potencias de, 356
identidad, elemento, 356
IGUS, sistema de recogida y utilización de información, 1085
imagen conforme de la geometría hiperbólica, 81-86, *82*, *83*, *85*
infinitesimales, 118
infinitesimales, elementos de grupos, 379
infinito, 500-504; corte diagonal de Cantor, 504-509, *505*, 515-516; tamaños de infinitos en física, 518-521
inflación, 1002-1009, *1005*, *1008*; validez de los motivos para la, 1010-1015, *1023*
inflexión, puntos de, *180*, 181
información cuántica, 810
Instituto Federal de Tecnología de Zurich (ETH), 554
integración, 173, 189-195, *190*, *191*; de contorno, 199-203
integral definida, 192
integral indefinida, 192
integrales de formas, 332-335
interacciones fuertes, 867, 912; autoestado propio, 873
interferencia cuántica, 616
interferencia, modelos de, 681-682, *681*
invariancia gauge, 612, 662-663
irreductibilidad, 387
Isham, Christopher, 1285
isotropía del universo, 76
isotrópico, universo, 964-965
iteración, 147

Jacobi, Moritz Hermann von, 382, 447, 639; ecuación de, 432
Jacobson, Ted, 1261
Jones, polinomio de, 1265
Jordan, Pascual, 725, 843, 882, 1242
Jozsa, Richard, 1285
Júpiter-Sol, sistema, 630

Kähler, variedades de, 1218
Kaluza, Theodor, 452, 1178-1179
Kaluza-Klein, espaciotiempo de, 453-454, 612, 1179-1182, *1179*, 1192, 1197, 1201, 1233, 1235, 1237, 1295, 1304
Kauffmann, álgebra de, 1265, *1265*
Kauffmann, polinomio de, 1265
Kepler, Johannes, 923
Kepler-Newton, órbitas elípticas de, 769, *770*, 1359
Kerr, John: métrica de, 960, 962, 1111, 1113-114, 1226
Kerr-Newman, métrica de, 1114-1115
ket, vectores, 721
Killing, vector de, *445*, 445, 620-621, 1106, *1107*, *1110*, *1112*, 1115, 1120-1121, 1136, 1138, 1181-1182, 1197, 1201, 1235; conflicto entre relatividad y física cuántica, 1136-1141, *1139*; de género tiempo, 1106, *1107*, 1134
Killing, Wilhelm, 363, 388; campo de, *1134*, 1233
Kirschner, Robert, 1037
Klein, Oskar, 96, 452, 1167, 1179
Klein-Gordon, ecuación de onda de, 830, 1206-1207
Kodaira, Kunihiko, 1333; curvas de, *1334*; teorema de, 1335
Kronecker, delta de, *346*, 369, 392-393, 400, 442, 670, 729

Lagrange, Joseph-Louis, 111-112, 514, 518, 639, 899
lagrangiano, formalismo, 639-644, *641*; espacio de configuración, 640, *640*; principio de Hamilton, 642, *643*, *660*; teorías modernas, 661-664; tratamiento de los campos, 658-660
lambdas, partículas, 844
Lambert, Johann Heinrich, 84, 94-95, 98, 396, 1384; plano no euclídeo de, 577-578
Laplace, ecuación de, 288; 2-dimensional, 289, 293
Laplace, Pierre-Simon, 601, 639, 951
Laurent, serie de, 240, 242-245, *252*, 253, 256-257; anillos de convergencia, 242-244, *243*
lazo, variables de, 1260-1264, *1262*, *1264*
Leggett, Tony, 798
Leibniz, Gottfried Wilhelm, 173, 333, 1384; ley de, 188, 420-421; cálculo infinitesimal, 1360; propiedad para la derivada de un producto, 722
Lenard, Philipp, 678
leptones, 866-867
Levi-Civita, Tulio, 443, 445; tensores asimétricos de, 370, 372, 603, 1255
levógiro, gravitón no lineal, 1336-1338
ley asociativa de la multiplicación, 356
ley conmutativa de la suma y de la multiplicación, 24, 357-358
leyes conmutativa y asociativa de la suma, 294
leyes distributivas de la multiplicación respecto a la suma, 294
Lie, grupos de, 363, 379, 383, 716, 1220
Lie, Sophus, 379; álgebra de, 380-382, *383*, 388, 405-406, 435, *435*, 436, 569, 716, 740-741, 743, 1169-1170; derivada de, 433, 438-439, 445; paréntesis de, 381, 388, 432-433-434, 438, 447
línea del universo de la partícula, 532
Liouville, Joseph, 639; teorema de, 655, *656*, 1130, *1131*
Lobachevski, Nicolái Ivánovich, 95, 1384
logaritmos, 85, 156-158; naturales, 159-162, *161*; reglas de cálculo con escala de, *157*
Lorentz, Hendrick Antoon, 392, 547-548, 1037, 1150, 1373; grupo de, 394, 406, 568-569, 612, 620, 1031, 1301-1302, 1344, 1388; movimientos rotacionales de, 761; transformaciones de, 834, 1314
Lorentz, invariancia, *1284*, 1285
Lüders, punto de, *739*
luz: conos de, 548-551, *549*, *550*, *552*, *553*; forma de ondas, 548; velocidad fija de la, 546-548, 553
luz, twistores como rayos de, 1288-1296, *1291*

M, teoría, 1224-1225
Mach, principio de, 1010, 1268
Mach-Zehnder, interferómetro de, *695*, 695, 734
MacLane, Saunders, 1286
Mahavira, 121
Majorana, Ettore: campo espinorial de, 891; imagen de, 745, 752-755, 757, 759, 793, 1072, 1210, 1289
Maldacena, Juan, 1231, 1234, 1240
Maldacena (ADS/CFT), conjetura de, 1231-1232, *1232*, 1234

Mandelbrot, conjunto de, 58-60, 63, *59*, 147-148, 1014
Manogue, Corinne, 1285
mapas, transformaciones de, 152-153, *153*, *154*
marea, efecto gravitatorio de, 542-545, *543*, *544*
marea, fuerzas de, 622
Marshall, William, 1148
Marte, 633
masa, 126, 536; conservación de la, 48, 590, *590*; formulación relativista, 591; *véase también* partículas sin masa
Matelski, J. P., 60
matemáticas y verdad, 50-53, 67; demostración matemática, 50; demostración por contradicción, 93; modelos matemáticos, 54; nociones, 58; números reales en el mundo físico, 116-120; objetividad, 55-56
materia oscura, 1035-1036
matrices, 365-372; autovalores y autovectores, 374-378, *376*; determinantes, 372-374, *372*; hermíticas, 400-401; ortogonales, 394; traspuesta conjugada de, 400; traspuesta de, 384; trazas y, 373-374
matriz de dispersión (matriz S), 906
matriz densidad, 1061-1064, *1063*; en situaciones EPR, 1069-1073, *1074*; para espín 1/2, 1064-1069
matriz hermítica, 400
Maxwell, James Clerk, 118, 547-549, 599-601, 1181, 1384; campo de, 1227; distribución de, 934-935; ecuaciones de, 601, 605-608, 610, 770, 843, 862, 889-890, 938, 1050, 1319; leyes de conservación y de flujo en la teoría de, 607; ondas de, 690; tensor de, 875; teoría electromagnética de, 233, 453, 486, 599, 602-607, 613, 1351, 1389; teoría física de campos de, 600, 610-611, 658, 781, 1296
medidas nulas, 733-739, *735*
medidas sí/no, 730-733, 738-739
membranas (*p*-branas, branas), 1163, 1225; D-branas, 1234-1238, *1236*, *1237*
mentes, existencia como concepto dentro de nuestras, 54
Mercurio, 633; perihelio de, 632, 1094, 1393
Merkulov, Sergei A., 1341
mesones, 866
mesones-π, 169
métrica, 310, 392, 441-446; lorentziana, 443
métrica cuantizada, 1293, *1294*
Michell, John, 951-953, 955
Mie, teoría de, 662-663
Millikan, Robert, 678
Minding, Ferdinand, 97
Minkowski, espacio de, 578, 580, 595, 607, 609, 611-612, 620, 648, 761, *898*, 952, 972, 1004, 1006, 1031, 1105, 1202, 1256, 1263, 1298-1299, *1299*, 1305, 1373; grupos de simetría del, 567-570; geometría hiperbólica en el, 576-582
Minkowski, espacio: compactificado, 1296-1301, *1299*, *1300*
Minkowski, Hermann, 554; geometría espaciotemporal de, 394, *552*, 554-556, 563-565, 574, 582, 601, 608, 832, 995, 1031, 1096, 1206, 1291, 1384
minkowskiana, geometría: 4-espacios euclídeo y minkowskiano, 563-566, *567*; energía y momento angular newtonianos, 587-590; energía y momento angular relativista, 590-595; esfera celeste como una esfera de Riemann, 582-587; geometría hiperbólica en, 576-582, *576*, *577*, *579*, *581*; grupos de simetría, 567-570, *568*; ortogonalidad lorentziana, 570-574, *571*, *573*; paradoja del reloj o de los gemelos, 572-574, *575*
modos normales de vibración, 649, 652
moduli, problema de los, 1208
Moebius, cinta de, 338, 457-458, *458*; fibrado de, 459, 461, 477
Moebius, transformación lineal o de, 222, 224, 394
momento: de las partículas clásicas, 321, *321*; de una partícula mecanocuántica, 250; formulación newtoniana, 589; formulación relativista, 591, *593*; función hamiltoniana, 54
momento magnético, 835
momentos angulares, 740, 1278; cuántica relativista, 761-765; Newton sobre, 589-590; twistores, 1310-1311
monopolos gauge, 996
monopolos magnéticos, 610; de Dirac, 993, *994*; gauge, 996
monstruo, grupo, 363
moralidad, 67
movimiento armónico simple, 649, *649*
movimiento inercial, 532-533, 539
muchos universos, teoría de, 718
multiplicativo finito, grupo, 166
mundo físico: distancias, 114, 117; números discretos en el, 123-127; números naturales, 120-123; números reales en el, 116-120; relación con la geometría euclídea, 97-98; relación con la geometría hiperbólica, 97-101

muones, 844, 866
musicales, tonos, *237*, 238, *239*

Nambu, Yoichiro, 1184, 1192
Ne'eman, Yuval, 867-868
Neumann, John von, 122, 731, 1064, 1384
neutrinos, 844, 854-855, 867
neutrones, 169, 844, 866
Newlander-Nirenberg, teorema de, 350
Newman, Ezra T., 1114, 1338
Newton, Isaac, 96, 118, 173, 525, 531, 546, 1384; cálculo infinitesimal, 1360; constante gravitatoria de, 534, 961, 1165, 1228; mecánica de, 639; movimientos inerciales de, 533; primera ley de, 532; segunda ley de, 533, 536-537; teoría gravitatoria de, 539, 545, 624, 1093-1094, 1180; tercera ley de, 533, *534*, 589, 663; *Principia*, 531, 599, 923
Newton-Cartan, esquema gravitatorio de, 1138
no trivialidad topológica, 318
Nöther, teorema de, 595, 620, 661-662, 670
nudos y enlaces, 1264-1267, *1265*
número, concepto de, 52, 58, 114
número entero, 124-126
números cardinales, 501-502
números complejos, 126-127, 131-134; argumento (fase) de, 154-155; construcción del conjunto de Mandelbrot, 147-148; convergencia de las series de potencias, 144-146, *146*; módulos de, 154, *155*; potencias complejas, 164-167; raíces cuadradas de, 135; resolviendo ecuaciones con, 135-138
números discretos, 123-127
números irracionales, 105-108, 115; cuadráticos, 111
números naturales, 121-123
números negativos, 124
números ordinales, 122
números racionales, 105, 111
números reales, 108-115; en el mundo físico, 116-120; necesidad de, 108

O(3), grupo espacial, 359-361
observables cuánticos, 726-730; proyectores, 730-733, *733*
octoniones, 298, 497
omega menos, partícula, 844
omnium, 1051
onda, función de, 674, 778; distribución de probabilidad en una, 698-701; electrones, 689; fase, 687-688, *689*; funciones twistoriales, 1318, *1325*; naturaleza holística de una, 691-696; normalizada, 698-699
onda de diente de sierra, 255, *256*
onda-partícula, dualidad, 682-685, 690
ondas, propagación de, 235-236
ondas cuadradas, 255-257, *256*
operador de ondas (D'Alembertiano), 830; raíz cuadrada de Clifford-Dirac de un, 831-833
operador hermítico, 400
operadores de creación, 883-886, *886*
Oppenheimer, J. Robert, 957
orden de factores, problema del, 673
ortogonal, 83; complemento, 570-571, *571*, 605
ortogonales, grupos, 389-396, *389*
ortogonales, matrices, 394
ortogonalidad, 552, 570-574, *571*

palindrómica, secuencia, 112
Pappos, teorema de, 471-472, *473*, 495, *496*
paradoja del reloj (o de los gemelos), 572-574, *575*
paralelogramo, ley del, 151, *152*, 411-412
parámetros cosmológicos, 1034-1043
paridad, 168
partículas sin masa, 583, 594, *594*, 854-855, 861; twistores como partículas sin masa con espín, 1309-1314
partículas virtuales, 126, *851*, 905, 908-909
Paschen-Back, efecto, 1242
Pauli, matrices de, 741, 836
Pauli, principio de exclusión de, 801-802, 837, 856, 869, 884-885, 952
Pauli, Wolfgang, 836, 843, 882
Pauli-Lubanski, vector de, 765
periódico, término, 236; funciones, *239*
Perjés, Zoltan, 1339
Perlmutter, Saul, 1037
Philiponos, Ioannes, 535
pi (π), 80, *80*
piones, 844; neutros, *594*
Pitágoras de Samos, 45, 50-53, 157
Pitágoras, teorema de, 51, *71*, 71-75, 78-81, *79*, *80*, 100, 106, 108, 111, 442; aplicabilidad en la geometría hiperbólica, 100-101; demostración por áreas semejantes, 78-81, *79*, *80*; fallo del, 442; raíz cuadrada de 2, 105-109, *106*; teselación, prueba de la, 71-75, *72*, *73*
plana, polarización, 750, *751*
Planck, constante de, 615, 671, 677, *679*, 679-680, 705, 961, 1165-1166; escala de, 117; unidades de, 742, 931, 961, 1114
Planck, distancia de, 1179, 1204, 1356
Planck, escala de, *1077*, 1152, 1167, 1201-1203, 1208, 1213, 1218, 1228

Planck, Max, 599, 678
Planck, unidades de, 1016, 1165-1166
planitud, problema de la, 1003
plano, elemento, 310-313, *313*
plano complejo, 143-144, *144*, *146*; conjunto de Mandelbrot, 147-148; mapas de operaciones aritméticas, *152*, *154*
Platón, 53-54, 58, 60, 63, 66, 68, 501
platónico de las formas matemáticas, mundo, 53, 114; lo bello y lo bueno, 66; realidad del, 53-60, 1376-1377, *1377*
Poincaré, Henri, 547-548, 569, 595, 761-762, 1031, 1373; disco de, 83, 219; grupo de, 1295-1296, 1301; lema de, 336
Poisson, Siméon Denis, 446, 639; paréntesis de, 654, 1097
polares esféricas, coordenadas, 756, *757*
polarización circular, 737, 750, *751*
Polchinski, Joseph, 1194
Popper, Karl, 1365, 1367-1368
positrones, 836-840, *839*, 856
postulado de las paralelas de Euclides, 75, 77, 77
potencias, series de: convergencia de, 138-143, *140*, *143*; convergencia usando números complejos, 144-146, *146*; divergencia de, 138-141, *140*, *143*, 143
Powell, C. F., 866
producto cuña, 311
producto de grupo, 380
producto escalar hermítico, 400-401
prolongación analítica, 208-211
protones, 125-126, 169, 867
proyección canónica, 456
proyección estereográfica, 89
proyectiva de la geometría hiperbólica, representación, 86, *88*, *89*
proyectivos, espacios, 471-476, *472*, *473*
proyectores, 730-733, *733*
pseudoesferas, 96-97, *96*
pseudométrica, 392
pseudoortogonales, grupos, 394
pseudoortonormal, base, 401
pseudorradio de la geometría, 85
púlsar, 631
punto de fuga, 471, *472*, 474
puntos cuánticos, 782

q-deformadas, estructuras, 1266
QED, *véase* electrodinámica cuántica
QFT, *véase* cuántica de campos, teoría
QFT topológica, 1266-1267, *1267*
quarkedad, 169
quarks, 125, 169, 845, 866; coloreados, 870-874; confinamiento, 912; sabores, 867
quiralidad del espacio, 309, 339

radiación de cuerpo negro, 678-679, *679*
radianes, 80
raíces n-ésimas de la unidad, 166, *167*, 168
raíz cuadrada de 2, 105-109, *106*
rama, puntos, 215, *216*
rapidez, 581, *581*
Rayleigh-Jeans, curva de, *679*, 679
razón giromagnética, 1114
realidad, búsqueda de la: belleza y milagros, 1389-1395; camino matemático a la realidad, 1383-1388; física fundamental matemáticamente dirigida, 1356-1360; grandes teorías de la física del siglo XX, 1351-1356; mundo matemático de Platón, 1377, *1377*; papel de las modas en la teoría física, 1361-1365; papel de la mentalidad en la teoría física, 1378-1383; preguntas profundas respondidas y más profundas planteadas, 1395-1398; próxima revolución en física, 1371-1375; refutaciones experimentales de una teoría errónea, 1365-1371; ¿qué es la realidad?, 1375-1378
Reasenberg, 633
Rebka, Glen, 632
recursivos, conjuntos, 515
redes de espín, 1267-1275, *1270*, *1271*, *1273*, *1274*, 1289, *1289*
reducción de estado (proceso R), 713-714, *714*, 1050-1051; asimetría temporal, 1098-1103, *1099*; linealidad de U, 715-718
reductio ad absurdum (demostración por contradicción), 93
reflexión especular, 852
Regge, Tullio, 1285
Regge: cálculo de, *1262*, *1286*; recurrencias de, 867
Regge, trayectoria de, 1185, *1186*, 1362
región abierta, 206, *207*
Reichter, P. H., 148
Reissner-Nordstrøm, solución de, 1227, 1230
renormalización, 907-913
representación, teoría de la, 378-383; espacios de representación tensoriales, 383-389
representación geométrica del álgebra, 151-155, *152*
Ricci: escalar de, 626; curvatura de, 1028, *1028*; tensor de, 623-624, 629, 1006-1007, 1026, 1180, 1196, 1214, 1266
Ricci-planitud, métrica lorentziana, 1182, 1196-1197, 1200-1203, 1205, 1212-1214, 1218-1219, *1334*, 1335
Riemann, Bernhard, 214, 441, 443, 1384
Riemann, esfera de, 223, *224*, 225-226, 231, 243-246, *246*, 252, 254, 258, 262, 269, 272, 289, 394, 463-464, 467, *468*, 469, 474-475, 569, 727, 753, 808, 811, 824, 836, 1289, 1292, 1294-1295, 1303, 1307-1308, *1308*, 1327-1329, *1328*, *1330*, 1344, 1387, 1394-1395; de los sistemas de dos estados, 745-752; esfera celeste como una, 582-586
Riemann, hipótesis de, 211, 1391
Riemann, superficies de, 213-233, *216*, 318, 348, 412, 1186, 1208, 1215-1216, *1216*, 1220, 1222, *1330*, 1394
Riemann, tensor de curvatura de, 629
Riemann, teorema de la aplicación de, 230-233
Riemann-Christoffel, tensor de, 444, 623-624
RO gravitatorio, esquema, 1143-1144, 1147-1150, 1156, 1354, 1381-1383, 1386
Robinson, congruencia de, 1313, *1313*, *1314*
rotaciones en la geometría de cuaterniones, 300-303, *300*; composición de, 304-306, *305*; espinor y objetos espinoriales, 301-303, *302*, 309; paridad de la, 301
Rovelli, Carlo, 1261, 1361
Russell, Bertrand, 510; paradoja de, 510-511, 517

Saccheri, Girolamo, 92-94
Sachs, Rayner, 634
Salam, Abdus, 882, 1364; teoría electrodébil de, 851
Sato, Katsuoko, 1002
Sato, Mikio, 259
Scherk, Joël, 1192
Schmidt, Brian, 1037; descomposición de, 1084
Schrödinger, Erwin, 118-119, 697-698, 701, 718, 729, 783, 1357; bulto de, 1132-1136; ecuación de, 674-676, 699, 701-702, 712-715, 717, 723, 768, 777-781, 795, 828-830, 834, 838, 925, 1050, 1134, 1139, 1142-1143, 1152, 1385; función de onda de, 725; gato de, 760, 1078-1085, *1078*, 1088, 1132, *1133*, 1144, 1380; imagen de, 724-726, 779; mecánica cuántica de, 769, 1093, 1342; *¿Qué es la vida?*, 1396
Schrödinger-Newton, estados de, 1141-1144
Schwachhöfer, L. J., 1341
Schwarz, John, 1192
Schwarzschild, Karl, 954, 957, 972; agujero de, 1230; espaciotiempo de, 1111, *1114*; geometría de, 1111; métrica de, 1113; solución de, 1108, 1274
Sen, Amitabha, 1253
Shapiro, Erwin, 633
sigma, partículas, 844
signatura, 395, 404
simetría, grupos de: axiomas del grupo, 357; espacio de Minkowski, 567-570, *568*; espacios de representación tensoriales, 383-389, *386*; grupos de transformaciones, 355-359; grupos ortogonales, 389-396; grupos simplécticos, 402-407; grupos unitarios, 396-402; orden del grupo, 358; reducibilidad, 385-387; subgrupos y grupos simples, 359-364; teoría de la representación, 378-383; transformaciones lineales y matrices, 364-372, *365*
simetría, ruptura de, 863; en el universo primitivo, 987-992, *990*, *993*
simetría electrodébil, grupo de, 860-865
simetrías no reflexivas, 359, *359*
simpléctica, estructura, 446
simplécticas, variedades, 446-448
simplécticos, grupos, 402-407; dinámica hamiltoniana como, 654-657, *657*;
simultaneidad, 527, 533
singularidades de los agujeros negros, 957-959
singularidades desnudas, 1029
Sirio, estrella, 632
Smolin, Lee, 1261
Snyder, Hartland, 957
Snyder-Schild, espaciotiempo, *1284*
SO(3), grupo ortogonal, 359-361
Sol, 947-948; campo gravitatorio del, 1027; energía devuelta a la Tierra, *949*
Sorkin, Raphael, 1284
Sparling, George, 1339
Stapp, Henry, 785, 796, 813
Starobinski, Alexei, 1002
Stein, conjuntos de, 1324
Steinhardt, P. J., 1369
Stern-Gerlach, medida de, 754, *755*
Stevin, Simon, 535
Stokes, George Gabriel, 334
Stokes, vector de, 751, *752*
Stoney, George Johnstone, 1166
Strominger, Andrew, 1225-1226
Strømme, Stein Arilde, 1222
suavidad, problema de la, 1003
subgrupos, 360-363; no normales, 361; normales, 360-361
submicroscópicos, ingredientes, 925-928
superconductividad, 991

supercuerdas, 1192-1193; segunda revolución de las, 1221-1223
supergravedad, 423, 1168
Superkamiokande, detector japonés, 855
superlumínico, viaje espacial, 1115-1118, *1116*, *1118*
supernovas, 953
superposiciones, 896, 1153, 1277
supersimetría, 299, 1163, 1168-1173, *1173*; álgebra y geometría de la, 1173-1177, *1176*; teoría de cuerdas incorporada a la, 1192-1193
supervariedad, 1174, *1176*
Susskind, Leonard, 1186

Tales, filósofo, 50
tangente, espacio, 326, *326*, 412
tangente a la curva, línea, 178
tangentes, fibrados, 469-470, *470*, 640
Taniyama-Shimura, conjetura de, 1241
tauones, 866-867
Taylor, Joseph, 631, 633; fórmula de, 204; serie de, 198, 205; teorema de, 436
Teeteto, 112
teleportación cuántica, 804-809, *809*
tensores, 344-348, *346*; álgebra tensorial, 341; curvatura, 422-424, 430-431, *430*; espacios de representación, 383-389, *386*; notación de índices abstractos, 345, 368; notación tensorial diagramática, 345, *346*, *347*, 368, *369*; productos tensoriales, 385; tensor de Riemann-Christoffel, 444; torsión, 423
teorema, definición de, 52
termodinámica, 600
termodinámica, primera ley de la, 928
termodinámica, segunda ley de la, 927; derivación de la, 936-940, *936*
teselación, *72*
Thiemann, Thomas, 1273
't Hooft, Gerard, 875, 882, 911
Thomas, Richard, 1239, 1395
tiempo: abandono del tiempo absoluto, 552-556, *553*; asimetría de la entropía, 936-938, *939*; asimetría temporal cosmológica, 1095-1098; asimetría temporal en la reducción del estado cuántico, 1098-1103, *1099*; física aristotélica, 526-527; paradoja del reloj (o de los gemelos), 572-574, *575*; simetría en la evolución dinámica, 923-925; teoría gauge de Weyl, 612, *614*; violación de causalidad, 557
Tierra, 1069-1071, 1397-1398; campo gravitatorio de la, 537, 632; rotación de la, 529-530; velocidad orbital de la, 584; vida en la, 1024
topología, 215-217, 226, 318
topológica, teoría cuántica de campos, 1266-1267, *1267*
topos, teoría de, 1285
toro, 215-216, *216*, 319, *319*
torsión topológica, 320
tractriz, 96, *96*
transformación gauge, 612
transformación lineal ortogonal, 365-366, 393-396, *393*
transformaciones lineales, 364-371, *365*, 378; ortogonales, 366, 393-395, *393*
transformaciones pasivas, 390, *390*
transición de fase, 991
transición, funciones de, 277, 322-323, *323*
transporte paralelo, 414-418, *415*, *417*
transversal, 76, 77
traslación, operación geométrica, 72, 152, 304, *305*
traza de una matriz, 373-374, *374*
trébol, nudo, *1264*
triángulo semejante, ley del, 151, *152*
triángulos: esféricos, 94-95, *95*; semejantes, 79, *79*; suma de los ángulos, 79-80, *80*, 83-84
Tsou Sheung-Tsun, 875-876
Tsou, Florence (Tsou Seung Tsun), 1339
Tsung Dao Lee, 309, 852
tubo avanzado, 1320, 1329
Turing, Alan, 507, 513; máquina de, 513-516, 518
Turing: máquina de, 1383; noción de computabilidad de, 1382
twistores, teoría de: campos sin masa, 1318-1321; cohomología de haces twistorial, 1321-1327, *1322*, *1323*; como rayos de luz, 1288-1296, *1291*; de la física de partículas, 1339-1340; diagramas twistoriales, 1339; espacio de Minkowski compactificado, 1296-1301, *1299*, *1300*; espacio twistorial, 1293, *1293*, 1301, *1306*; espaciotiempo complejo, 1311-1312, *1312*; funciones twistoriales, 1318, *1325*; futuro de la, 1340-1345; geometría twistorial básica y coordenadas, 1304-1309, *1306*, *1308*; gravitón no lineal, 1330-1336, *1331*, *1332*; grupo conforme, 1296-1298, *1297*; helicidad, 1311; separación en frecuencia positiva/negativa, 1327-1330, *1328*, *1330*; teoría cuántica, 1314-1318; twistores como espinores de dimensión superior, 1301-1304; twistores como partículas sin masa con espín, 1309-1314

Uhlenbeck, George, 836
Unhru, efecto, 1106-1107
universo cerrado, 969
universo de juguete, 781-782, *782*
universo primitivo, desarrollo del: cosmología inflacionaria, 1002-1009, *1004*; defectos topológicos cósmicos, 992-997, *994*, *995*; hipótesis de curvatura de Weyl, 1026-1030, *1030*; origen de las fluctuaciones en el, 1151-1157; parámetros cosmológicos, 1034-1043; principio antrópico aplicado al big bang, 1022-1025; principio antrópico, 1016-1021, *1017*, 1379; propuesta de ausencia de frontera de Hartle-Hawking, 1030-1034, *1033*; ruptura espontánea de simetría, 987-992, *990*, *993*; ruptura espontánea de simetría, problema para la, 997-1001; validez de la cosmología inflacionaria, 1010-1015, *1023*
universo: cerrado, 969; creación del, 980, *980*; diagramas conformes, 971-974, *974*, *975*; espacialmente infinito, 980-982; expansión del, 969-970; isótropo, 964-965
1-formas, 277-278, 284-286, 325; densidad de, 332-333

vacío, polarización del, 908-909, *909*
vacío térmico, 1106
vacíos, estados: diagramas de Feynman, 915-918; «mar» de Dirac, 892; teoría cuántica de campos (QFT), 892-894
Vafa, Cumrun, 1226
Vaidman, Lev, 734
Van der Waerden, B. L., 1363
variables canónicas conjugadas, 1315
variables no conmutativas, 667-671
variedad riemanniana, 441
variedades, 216, 271, *282*, *283*; cartas de coordenadas y, 322-324; complejas, 348-350; convenio de suma de *n*-variedad, 342-343; de dimensiones más altas, 317-322; diferenciación en, 411-413; elemento de volumen de *n*-variedad, 341-344; escalares, vectores y covectores, 324-329; múltiplemente conexos, *319*, 319, *320*; parciales, 271-272, *273*; riemannianas, 441; simplécticas, 446-448; simplemente conexos, 319, *319*
vector de estado, 712
vectores: bra, 721; ket, 721; representación de rotaciones, 304, *305*
velocidad de la luz, 546-548
Veneziano, Gabriele, 1184, 1186, 1192
Venus, 633
Vía Láctea, 530, 957, 963
viaje en el tiempo, 558
vibración, frecuencia de, 648-649
violación de causalidad, 557
Virgo, cúmulo, 530
Voigt, Woldemar, 1242
volúmenes, 118

W, bosones, 845, 851, 859-860; grupo de simetría electrodébil, 860-865, *861*
Ward, Richard, 882, 1338, 1340
Ward, teoría electrodébil de, 851
Warren, John, 143
Weinberg, Steven, 882; ángulo de, o de mezcla débil, 873-874, 1388; teoría electrodébil de, 851
Wessel, Caspar, 143
Weyl, Hermann, 612-614, 843; curvatura de, 959, 1026-1030, 1033, 1127-1129, *1128*, 1254, 1256, *1334*, 1335, 1343; distorsiones de, *1028*; ecuación de, 855; tensor de, 629, 1026-1027, 1255-1256, 1266
Weyl-plana, curvatura, 1335
Wheeler, John A., 858, 1020, 1152-1153, 1166
Wheeler/Smolin, propuesta de tipo, 1388
Whitehead, enlace de, *1264*
Wick, Gian Carlo, 1031; rotación de, 1031-1032, *1216*, 1216
Wiles, Andrew, 56, 518, 1241, 1390
Wilson, Kenneth, 882, 1261
Wilson, lazo de, 1261
Witten, Edward, 1187-1188, 1199, 1221, 1223-1224, 1240-1241, 1266, 1394
WMAP, sonda espacial, 1038, 1040, 1156, 1370
Woodhouse, Nick, 282

Yang-Mills, teorías de, 616-617, 871, 996-997, 1221, 1226-1227, 1229-1230, 1233, 1236, 1240, 1343, 1394
Yukawa, Hideki, 844, 866

Z, bosones, 845, 851, 859; grupo de simetría electrodébil, 861-865, *861*
Zeilinger, Anton, 1354-1355
zeta, función, 211
Zhoukowski (o Joukowski), transformación del alerón de, 231-232, *232*
Zoll, espacios de, 1341
Zumino, Bruno, 1364

INTRODUCCIÓN AL CÁLCULO Y AL ANÁLISIS MATEMÁTICO

INTRODUCCIÓN AL CÁLCULO Y AL ANÁLISIS MATEMÁTICO
Vol. I

RICHARD COURANT y FRITZ JOHN

DEL COURANT INSTITUTE OF MATHEMATICAL SCIENCES, UNIVERSIDAD DE NUEVA YORK

LIMUSA
NORIEGA EDITORES
MÉXICO • España • Venezuela • Colombia

Versión autorizada en español de la obra publicada en inglés con el título:
INTRODUCTION TO CALCULUS AND ANALYSIS. Volume 1
© John Wiley & Sons, Inc.

Colaboradores en la traducción:
SAÚL HAHN GOLDBERG
Ingeniero en comunicaciones y electrónica. Licenciado en ciencias fisicomatemáticas y profesor de matemáticas en la Escuela Superior de Física y Matemáticas del Instituto Politécnico Nacional, México.

ROLANDO V. JIMÉNEZ DOMÍNGUEZ
Ingeniero en comunicaciones y electrónica. Licenciado en fisicomatemáticas y profesor de física en la Escuela Superior de Física y Matemáticas del Instituto Politécnico Nacional, México.

JOSÉ S. FLORIO
Profesor y doctor en ciencias matemáticas. Jefe de la división de ciencias y director del Instituto Tecnológico, Universidad de las Américas.

La presentación y disposición en conjunto de

INTRODUCCIÓN AL CÁLCULO Y AL ANÁLISIS MATEMÁTICO. Volumen 1

son propiedad del editor. Ninguna parte de esta obra puede ser reproducida o transmitida, mediante ningun sistema o método, electrónico o mecánico (incluyendo el fotocopiado, la grabación o cualquier sistema de recuperación y almacenamiento de información), sin consentimiento por escrito del editor.

Derechos reservados:

© 1999, EDITORIAL LIMUSA, S.A. de C.V.
GRUPO NORIEGA EDITORES
Balderas 95, México, D.F.
C.P. 06040
☎ (5)521-21-05
01(800) 7-06-91-00
(5)512-29-03
limusa@noriega.com.mx
www.noriega.com.mx

CANIEM Núm. 121

Decimosexta reimpresión

Impreso en México
ISBN 968-18-0639-5

Prólogo

En la segunda mitad del siglo XVIII el nuevo análisis matemático surgió como la fuerza dominante en la Matemática. Este análisis se caracteriza por la extraordinariamente eficaz operación que se funda en procesos infinitos, o sea, límites. Dos de estos procesos, la derivación y la integración, se convirtieron en la esencia de la disciplina, llamada Cálculo Diferencial e Integral o, a menudo, simplemente "Cálculo", que es la base de todo el análisis matemático.

Se apreció, de inmediato, la importancia de los nuevos descubrimientos y métodos, los cuales causaron una profunda conmoción intelectual. Sin embargo, el dominio de este poderoso arte matemático constituyó, al principio, un objetivo formidable y difícil de lograr, ya que los textos sobre la materia eran escasos, poco sistemáticos y, con frecuencia carentes de claridad. Por tanto, fue venturoso para la Matemática y para la Ciencia, en general, que las figuras principales del nuevo movimiento reconocieran pronto la vital necesidad de escribir tratados con el fin de que la materia fuera accesible a un público mucho más numeroso que el pequeño círculo intelectual de esa época. Uno de los matemáticos más destacados de los tiempos modernos, Leonhard Euler, estableció los fundamentos de una larga tradición en sus libros introductorios; y aunque se ha progresado mucho en la aclaración y simplificación del material del Cálculo, estos libros del siglo XVIII, en la actualidad aún son fuente de inspiración.

Después de Euler, los autores, uno tras otro, se solidarizaron con la separación entre cálculo diferencial y cálculo integral y, al hacerlo, oscurecieron un punto clave: la reciprocidad entre derivación e integración. Fue en 1927, cuando apareció, publicada por "Springer-Verlag", la primera edición alemana de la obra de R. Courant "Vorlesungen über Differential and Integralrechnung", que se eliminó tal separación y el Cálculo se presentó como una disciplina unificada.

La presente obra se originó en aquel primer libro alemán y sus ediciones subsiguientes. A partir de 1939, con la colaboración de James y Virginia

McShane, la casa "Blackie & Sons" de Glasgow, preparó y publicó una versión inglesa muy ampliada y modificada, de la cual la Interscience-Wiley distribuyó numerosas reimpresiones en los Estados Unidos de Norteamérica.

Con el transcurso de los años, se hizo evidente que los programas de la enseñanza superior en los Estados Unidos, requerían una nueva revisión de la obra. Por otra parte, parecía poco prudente alterar las versiones originales que todavía son útiles.

Así, en vez de modificar la obra existente, se optó por complementarla con un libro esencialmente nuevo, relacionado en muchos sentidos con los originales europeos pero, en términos más específicos, orientado para satisfacer las necesidades de la actual y futuras generaciones de estudiantes en los Estados Unidos. Tal plan resultó factible cuando Fritz John, que ya había ayudado mucho en la preparación de la primera edición inglesa, convino en escribir de nuevo el libro con R. Courant.

Aunque este libro difiere notablemente del original, tanto en forma como en contenido, tiene el mismo propósito: guiar al estudiante directamente a la esencia del tema y capacitarlo para aplicar sus conocimientos. La obra evita el estilo dogmático que oculta la motivación de los conceptos y las raíces que el Cálculo tiene en la realidad intuitiva. Un importante objetivo que se persigue en este libro es mostrar la relación recíproca entre el análisis matemático y sus diversas aplicaciones y destacar el papel de la intuición. Esperamos que cierto énfasis en la precisión no interfiera con este objetivo.

La Matemática presentada como un sistema de verdades, acabado y ordenado, sin referencia al origen y propósito de sus conceptos y teorías, tiene su encanto y satisface una necesidad filosófica. Pero esa actitud introvertida en el campo de la ciencia, no es adecuada para los estudiantes que buscan independencia intelectual, más bien que indoctrinación. Y menospreciar las aplicaciones e intuición conduce al aislamiento y atrofia en la Matemática. Entonces, resulta sumamente importante que estudiantes y maestros estén a salvo del purismo presumido.

El libro está dirigido a estudiantes de varios niveles, a matemáticos, científicos en general e ingenieros. No pretende facilitar la materia suavizando las dificultades, sino que, más bien, trata de ayudar al lector sinceramente interesado, a base de aclarar las interrelaciones y propósitos de la obra en conjunto.

En lugar de obstaculizar el acceso al caudal de hechos y propiedades por medio de largas exposiciones de carácter fundamental, éstas se han pospuesto, a veces, relegándolas a apéndices de los diversos capítulos.

Se dan numerosos ejemplos y problemas al final de los capítulos. Algunos son fascinantes, otros llegan a ser verdaderamente difíciles; y la mayoría de ellos complementan el material teórico del texto. Se preparará un manual complementario de esta obra con problemas y ejercicios de carác-

ter más rutinario, y, además, se agregarán respuestas y sugerencias para resolverlos.

Muchos colegas y amigos han ayudado. Albert A. Blank no sólo aportó críticas perspicaces y constructivas, sino que también desempeñó un papel importante en la labor de ordenar, aumentar y seleccionar los problemas y ejercicios y, además, asumió la responsabilidad principal de preparar el manual mencionado.* Alan Solomon ayudó, muy desinteresada y eficazmente, en todas las etapas de la preparación del libro. También se les agradece a Charlotte John, Anneli Lax, R. Richtmyer, y a otros amigos, entre los que figuran James y Virginia McShane.

Este volumen trata principalmente de funciones de una sola variable, mientras que en el Volumen II se estudiarán las teorías del Cálculo más diversificadas para funciones de varias variables. Al lector se le debe hacer una observación final, ya que podría resultar un fracaso tratar de dominar la materia estudiando un libro de este tipo página por página, sin desviarse del camino trazado. Se puede alcanzar gradualmente una mejor comprensión del tema y una visión más amplia, sólo tomando atajos en la primera lectura y, luego, volviendo repetidas veces a los mismos temas, preguntas y dificultades.

Se ha tratado de ayudar al lector señalando con asteriscos ciertos tópicos que podrían causarle trastornos al estudiarlos por primera vez, así como algunos de los problemas más difíciles.

Esperamos que la obra, con este nuevo enfoque, será útil a la joven generación de científicos. Estamos conscientes de sus muchas imperfecciones, y sinceramente solicitamos comentarios críticos que podrían ser beneficiosos para mejoras posteriores.

RICHARD COURANT
FRITZ JOHN

* Manual de Problemas de Cálculo y Análisis Matemático, por Albert A. Blank, publicado en español por la Editorial Limusa

Contenido

CAPITULO 1 Introducción 25

1.1 El continuo de números
a. El sistema de números naturales y su extensión. Numeración y medición, **26** **b.** Números reales e intervalos encajados, **31** **c.** Fracciones decimales. Bases distintas a la de diez, **33** **d.** Definición de vecindad o entorno, **36** **e.** Desigualdades, **36**

1.2 El concepto de función 41
a. Función-Gráfica, **42** **b.** Definición del Concepto de funciones de una variable continua. Dominio y rango de una función, **45** **c.** Representación gráfica. Funciones monótonas, **48** **d.** Continuidad, **55** **e.** El teorema del valor intermedio. Funciones inversas, **68**

1.3 Las funciones elementales 71
a. Funciones racionales, **71** **b.** Funciones algebraicas, **72** **c.** Funciones trigonométricas, **73** **d.** La función exponencial y el logaritmo, **74** **e.** Funciones compuestas, productos simbólicos, funciones inversas, **76**

1.4 Sucesiones 79

1.5 Inducción matemática 80

1.6 El límite de una sucesión 84
a. $a_n = \frac{1}{n}$, **84** **b.** $a_{2m} = \frac{1}{m}$; $a_{2m-1} = \frac{1}{2m}$, **85**
c. $a_n = \frac{n}{n+1}$, **86** **d.** $a_n = \sqrt[n]{p}$, **87**

e. $a_n = \alpha^n$, **88**
f. Ilustración geométrica de los límites α^n y $\sqrt[n]{p}$, **89**
g. La serie geométrica, **90** **h.** $a_n = \sqrt[n]{n}$, **92**
i. $a_n = \sqrt{n+1} - \sqrt{n}$, **92**
j. $a_n = \alpha^n n$, para $\alpha > 1$, **93**

1.7 Discusión del concepto de límite 93
a. Definición de convergencia y divergencia, **93** **b.** Operaciones racionales con límites, **95** **c.** Pruebas intrínsecas de convergencia. Sucesiones monótonas, **96** **d.** Series infinitas y el símbolo de sumatoria, **98** **e.** El número e, **100** **f.** El número π como límite, **103**

1.8 El concepto de límite para funciones de una variable continua 105
a. Algunas observaciones sobre las funciones elementales, **109**

SUPLEMENTO 110

S.1. Los límites y el concepto de número 112
a. Los números racionales, **112** **b.** Números reales determinados por encaje de intervalos racionales, **113** ***c.** Orden, límites y operaciones aritméticas para números reales, **115** **d.** Plenitud de continuo de números. Compacidad de intervalos cerrados. Criterios de convergencia, **117** **e.** Cotas. Extremo superior y extremo inferior, **120** **f.** Numerabilidad de los números racionales, **121**

S.2. Teoremas sobre funciones continuas 123

S.3 Coordenadas polares 125

S.4 Observaciones sobre los números complejos 126

PROBLEMAS, 129

CAPITULO 2 Las ideas fundamentales del cálculo integral y diferencial 141

2.1 La integral 142
a. Introducción, **142** **b.** La integral como un área, **143** **c.** Definición analítica de la integral. Notaciones, **145**

2.2 **Ejemplos elementales de integración** 150
a. Integración de una función lineal, 150 **b. Integración** de x^2, 152 **c.** Integración de x^α para enteros $\alpha \neq -1$, 153 **d.** Integración de x^α para α racional distinto de -1, 156 **e.** Integración de sen x y cos x, 157

2.3 **Reglas fundamentales de integración** 158
a. Aditividad, 158 **b.** Integral de una suma y de un producto con una constante, 159 **c.** Estimación de integrales, 160 **d.** El teorema del valor medio para integrales, 161

2.4 **La integral como función del límite superior (integral indefinida)** 165

2.5 **El logaritmo definido mediante una integral** 167
a. Definición de la función logaritmo, 167 **b.** El teorema de adición para logaritmos, 169

2.6 **Función exponencial y potencias** 171
a. El logaritmo del número e, 171 **b.** La función inversa del logaritmo. La función exponencial, 172 **c.** La función exponencial como límite de potencias, 173 **d.** Definición de potencias arbitrarias de números positivos, 174. **e.** Logaritmos de base arbitraria, 175

2.7 **La integral de una potencia arbitraria de x** 176

2.8 **La derivada** 177
a. La derivada y la tangente, 178 **b.** La derivada como una velocidad, 183 **c.** Ejemplos de diferenciación, 184 **d.** Algunas reglas fundamentales para la derivación, 187 **e.** Derivabilidad y continuidad de funciones, 187 **f.** Derivadas superiores y su significado, 190, **g.** Derivada y cociente de incrementos. Notación de Leibnitz, 192. **h.** El teorema del valor medio del cálculo diferencial, 194 **i.** Demostración del teorema, 196 **j.** La aproximación de funciones mediante funciones lineales. Definición de diferenciales, 200 **k.** Observaciones sobre aplicaciones a las ciencias naturales, 205

2.9 **La integral, la función primitiva y los teoremas fundamentales del cálculo** 206
a. La derivada de la integral, 206 **b.** La función primitiva y su relación con la integral, 208 **c.** El uso de la función primitiva para la evaluación de integrales definidas, 211 **d.** Ejemplos, 212

Suplemento. La existencia de la integral definida de una función continua 213

PROBLEMAS 217

CAPITULO 3 Las técnicas del cálculo 223

PARTE A Derivación e integración de las funciones elementales 223

3.1 **Las reglas más simples para derivar y sus aplicaciones** 223
a. Reglas de la derivación, 223 **b.** Derivación de las funciones racionales, 226. **c.** Derivación de las funciones trigonométricas, 227

3.2 **La derivada de la función inversa** 228
a. Fórmula general, 228 **b.** La inversa de la n-ésima potencia; la n-ésima raíz, 228 **b.** Las funciones trigonométricas inversas. Multivalencia, 232 **d.** Las fórmulas integrales correspondientes, 237 **e.** Derivada e integral de la función exponencial, 238

3.3 **Derivación de funciones compuestas** 239
a. Definiciones, 239 **b.** La regla de la cadena, 239 **c.** El teorema generalizado del valor medio del cálculo diferencial, 243

3.4 **Algunas aplicaciones de la función exponencial** 244
a. Definición de la función exponencial por medio de una ecuación diferencial, 244 **b.** Interés compuesto continuamente. Desintegración radiactiva, 245 **c.** Enfriamiento o calentamiento de un cuerpo por el medio ambiente, 246 **d.** Variación de la presión atmosférica con la altura sobre la superficie de la Tierra, 247 **e.** Progreso de una reacción química, 248 **f.** Apertura y cierre de un circuito eléctrico, 249

3.5 **Las funciones hiperbólicas** 250
a. Definición analítica, 250 **b.** Teoremas de adición y fórmulas para derivar, 252 **c.** Las funciones hiperbólicas inversas, 253 **d.** Otras analogías, 255

3.6 Máximos y mínimos 257
a. Convexidad y concavidad de curvas, 257 **b.** Máximos y mínimos—Extremos relativos. Puntos estacionarios, 259

***3.7 El orden de magnitud de las funciones** 269
a. El concepto de orden de magnitud. Los casos más simples, 269 **b.** Los órdenes de magnitud de la función exponencial y del logaritmo, 270 **c.** Comentarios generales, 272 **d.** El orden de magnitud de una función en la vecindad de un punto arbitrario, 273 **e.** El orden de magnitud (pequeñez) de una función que tiende a cero, 273 **f.** Las notaciones "O" y "o" para órdenes de magnitud, 274

APENDICE 276

A.1 Algunas funciones especiales 276
a. La función $y = e^{-1/x^2}$, 277 **b.** La función $y = e^{-1/x}$, 278 **c.** La función $y = \tanh 1/x$, 278 **d.** La función $y = x \tanh 1/x$, 279 **e.** La función $y = x \operatorname{sen} 1/x$, $y(0) = 0$, 280

A.2 Comentarios sobre la derivabilidad de funciones 280

PARTE B Técnicas de integración 282

3.8 Tabla de integrales elementales 285

3.9 El método de substitución 285
a. La fórmula de substitución. Integral de una función compuesta, 285 **b.** Una deducción alternativa de la fórmula de substitución, 289 **c.** Ejemplos. Fórmulas de integración, 291

3.10 Otros ejemplos del método de substitución 292

3.11 Integración por partes 295
a. Fórmula general, 295 **b.** Otros ejemplos de integración por partes, 297 **c.** Fórmula integral para $f(b) + f(a)$, 298 **d.** Fórmulas recursivas, 298 **e.** El producto infinito de Wallis para π, 300

3.12 Integración de funciones racionales 302
a. Los tipos fundamentales, 303 **b.** Integración de los tipos fundamentales, 304 **c.** Fracciones parciales, 306

d. Ejemplos de resolución en fracciones parciales. El método de los coeficientes indeterminados, 308

3.13 Integración de algunas otras clases de funciones 310
a. Comentarios preliminares sobre la representación racional de la circunferencia y de la hipérbola, 310 **b.** Integración de $R(\cos x, \operatorname{sen} x)$, 312 **c.** Integración de $R(\cosh x, \operatorname{senh} x)$, 313 ***d.** Integración de $R(x, \sqrt{1-x^2})$, 314 ***e.** Integración de $R(x, \sqrt{x^2-1})$, 314 ***f.** Integración de $R(x, \sqrt{x^2+1})$, 314 ***g.** Integración de $R(x, \sqrt{ax^2+2bx+c})$, 315 ***h.** Otros ejemplos de reducción a integrales de funciones racionales, 315 **i.** Comentarios sobre los ejemplos, 316

PARTE C Otros pasos en la teoría del cálculo integral 317

3.14 Integrales de funciones elementales 317
a. Definición de funciones por medio de integrales. Integrales y funciones elípticas, 317 **b.** Sobre la derivación y la integración, 320

3.15 Extensión del concepto de integral 320
a. Introducción. Definición de integrales "impropias", 320 **b.** Funciones con discontinuidades infinitas, 323 **c.** Interpretación de la integral como un área, 323 **d.** Criterios para determinar la convergencia, 324 **e.** Intervalo de integración infinito, 325 **f.** La función gamma, 327 **g.** La integral de Dirichlet, 328 **h.** Substitución. Integrales de Fresnel, 329

3.16 Las ecuaciones diferenciales de las funciones trigonométricas 331
a. Comentarios introductorios sobre las ecuaciones diferenciales, 331 **b.** Sen x y cos x definidos mediante una ecuación diferencial y condiciones iniciales, 331

PROBLEMAS 333

CAPITULO 4 Aplicaciones en Física y Geometría 343

4.1 Teoría de curvas planas 343
a. Representación paramétrica, 343 **b.** Cambios de parámetros, 345 **c.** Movimiento a lo largo de una curva. El

tiempo como un parámetro. El ejemplo de la cicloide, 347 **d.** Clasificación de curvas. Orientación, 351 **e.** Derivadas, tangentes y normales en representación paramétrica, 361 **f.** La longitud de una curva, 366 **g.** La longitud de arco como parámetro, 370 **h.** Curvatura, 372 **i.** Cambio de ejes de coordenadas. Invariancia, 378 ***j.** Movimiento uniforme en la teoría especial de relatividad, 381 **k.** Integrales que expresan áreas dentro de curvas cerradas, 382 **l.** Centro de masa y momento de una curva, 390 **m.** Area y volumen de una superficie de revolución, 391 **n.** Momento de inercia, 392

4.2 **Ejemplos** 393
a. La cicloide común, 394 **b.** La catenaria, 395 **c.** La elipse y la lemniscata, 396

4.3 **Vectores en dos dimensiones** 397
a. Definición de vectores mediante traslaciones. Notaciones, 398 **b.** Adición y multiplicación de vectores, 402 **c.** Vectores variables, sus derivadas e integrales, 410 **d.** Aplicación a curvas planas. Dirección "rapidez" y aceleración, 411

4.4 **Movimiento de una partícula bajo la acción de fuerzas especificadas** 415
a. Ley de movimiento de Newton, 415 **b.** Movimiento de cuerpos en caída, 417 **c.** Movimiento de una partícula restringido a una curva dada, 418

4.5 **Caída libre de un cuerpo venciendo la resistencia del aire** 420

4.6 **El tipo más simple de vibración elástica** 423

4.7 **Movimiento sobre una curva dada** 424
a. La ecuación diferencial y su solución, 424 **b.** Partícula que se desliza hacia abajo sobre una curva, 426 **c.** Discusión del movimiento, 428 **d.** El péndulo ordinario, 429 **e.** El péndulo cicloidal, 430

*4.8 **Movimiento en un campo gravitacional** 431
a. Ley universal de la gravitación de Newton, 431 **b.** Movimiento circular en torno al centro de atracción, 433 **c.** Movimiento radial-Velocidad de escape, 435

4.9 **Trabajo y energía** 436

a. Trabajo realizado por fuerzas durante un movimiento, 436 **b.** Trabajo y energía cinética. Conservación de la energía, 438 **c.** La atracción mutua de dos masas, 440 **d.** El estiramiento de un resorte, 441 ***e.** La carga de un condensador, 442

APENDICE 442

A.1 Propiedades de la evoluta 442

A.2 Areas limitadas por curvas cerradas. Indices 448

PROBLEMAS 453

CAPITULO 5 **Desarrollo de Taylor** 459

5.1 **Introducción: Series de potencias** 459

5.2 **Desarrollo del logaritmo y de la tangente inversa** 461
a. El logaritmo, 461 **b.** La tangente inversa, 463

5.3 **Teorema de Taylor** 464
a. Representación de Taylor para polinomios, 464
b. Fórmula de Taylor para funciones no polinomiales, 465

5.4 **Expresiones y estimaciones para el residuo** 466
a. Expresiones de Cauchy y de Lagrange, 466 **b.** Una derivación alternativa de la fórmula de Taylor, 469

5.5 **Desarrollos de funciones elementales** 472
a. La función exponencial, 472 **b.** Desarrollos de sen x, cos x, senh x, cosh x, 473 **c.** La serie binomial, 475

5.6 **Aplicaciones geométricas** 476
a. Contactos entre curvas, 477 **b.** Sobre la teoría de máximos y mínimos relativos, 480

APENDICE I 481

A.I.1 Ejemplo de una función que no se puede desarrollar en una serie de Taylor 481

A.I.2 **Ceros e infinitos de funciones** 482
a. Ceros de orden n, 482 **b.** Infinito de orden ν, 482

A.I.3 **Expresiones indeterminadas** 483

*A.I.4 **La convergencia de la serie de Taylor para una función con derivadas no negativas de todos los órdenes** 486

APENDICE II INTERPOLACION 489

*A.II.1 **El problema de la interpolación. Unicidad** 489

A.II.2 **Construcción de la solución. Fórmula de interpolación de Newton** 490

A.II.3 **La estimación del residuo** 493

A.II.4 **La fórmula de interpolación de Lagrange** 495

PROBLEMAS 496

CAPITULO 6 **Métodos numéricos** 501

6.1 **Cálculo de integrales** 502
a. Aproximación mediante rectángulos, 502 **b.** Aproximaciones refinadas-Regla de Simpson, 503

6.2 **Otros ejemplos de métodos numéricos** 509
a. El "cálculo de errores", 509 ***b.** Cálculo de π, 512 ***c.** Cálculo de logaritmos, 512

6.3 **Solución numérica de ecuaciones** 514
a. Método de Newton, 514 ***b.** La regla de la posición falsa, 517 **c.** El método iterativo, 518 **d.** Iteraciones y procedimiento de Newton, 521

APENDICE 523

*A.1 **Fórmula de Stirling** 523

PROBLEMAS 526

CAPITULO 7 **Sumas y productos infinitos** 529

7.1 **Los conceptos de convergencia y divergencia** 530
a. Conceptos básicos, 530 **b.** Convergencia absoluta y convergencia condicional, 532 ***c.** Reordenamiento de términos, 536 **d.** Operaciones con series infinitas, 538

7.2 **Criterios de convergencia absoluta y de divergencia** 539
a. El criterio de comparación. Mayorantes, 539 **b.** Investigación de la convergencia mediante la comparación con la serie geométrica, 540 **c.** Comparación con una integral, 543

7.3 **Sucesiones de funciones** 545
a. Procesos de límite con funciones y curvas, 545

7.4 **Convergencia uniforme y convergencia no uniforme** 547
a. Comentarios generales y definiciones, 547 **b.** Un criterio de convergencia uniforme, 552 **c.** Continuidad de la suma de una serie uniformemente convergente de funciones continuas, 554 **d.** Integración de series uniformemente convergentes, 555 **e.** Derivación de series infinitas, 557

7.5 **Series de potencias** 558
a. Propiedades de convergencia de las series de potencias-Intervalo de convergencia, 558 **b.** Integración y derivación de series de potencias, 560 **c.** Operaciones con series de potencias, 561 **d.** Unicidad del desarrollo, 562 ***e.** Funciones analíticas, 563

7.6 **Desarrollos en series de potencias de funciones dadas. El método de los coeficientes indeterminados. Ejemplos** 564
a. La función exponencial, 564 **b.** La serie del binomio, 565 **c.** La serie para $\operatorname{arc\,sen} x$, 566 **d.** La serie para $\operatorname{ar\,senh} x = \log[x + \sqrt{(1 + x^2)}]$, 567 **e.** Ejemplo de multiplicación de series, 567 **f.** Ejemplo de integración término a término (integral elíptica), 567

7.7 **Series de potencias con términos complejos** 568
a. Introducción de términos complejos en una serie de potencias. Representaciones complejas de funciones trigonométricas, 568 ***b.** Un vistazo a la teoría general de funciones de una variable compleja, 571

APENDICE 572

A.1 **Multiplicación y división de series** 572
a. Multiplicación de series absolutamente convergentes, 572 ***b.** Multiplicación y división de series de potencias, 573

A.2 **Series infinitas e integrales impropias** 574

*A.3 **Productos infinitos** 576

*A.4 **Series en que aparecen números de Bernoulli** 579

PROBLEMAS 581

CAPITULO 8 **Series trigonométricas** 589

8.1 **Funciones periódicas** 590
a. Comentarios generales. Extensión periódica de una función, 590 **b.** Integrales calculadas sobre un período, 591 **c.** Vibraciones armónicas, 592

8.2 **Superposición de vibraciones armónicas** 594
a. Armónicas. Polinomios trigonométricos, 594 **b.** Batidos, 595

8.3 **Notación compleja** 600
a. Comentarios generales, 600 ***b.** Aplicación al caso de corrientes alternas, 601 **c.** Notación compleja para polinomios trigonométricos, 602 **d.** Una fórmula trigonométrica, 603

8.4 **Series de Fourier** 604
a. Coeficientes de Fourier, 604 **b.** Lema básico, 605 **c.** Demostración de que $\int_0^\infty \frac{\operatorname{sen} z}{z}\,dz = \frac{\pi}{2}$, 606 **d.** Desarrollo de Fourier para la función $\phi(x) = x$, 609 **e.** El teorema principal sobre desarrollos de Fourier, 611

8.5 **Ejemplos de series de Fourier** 615
a. Comentarios preliminares, 615 **b.** Desarrollo de la función $\phi(x) = x^2$, 616 **c.** Desarrollo de $x \cos x$, 616 **d.** La función $f(x) = |x|$, 617 **e.** Una función constante por trozos, 618 **f.** La función $|\operatorname{sen} x|$, 619 **g.** Desarrollo de $\cos \mu x$. Descomposición de la cotangente en fracciones parciales. El producto infinito para el seno, 619 **h.** Ejemplos adicionales, 621

8.6 **Discusión adicional sobre la convergencia** 621
a. Resultados, 621 **b.** Desigualdad de Bessel, 622 ***c.** Demostración de los corolarios (a), (b) y (c), 622 **d.** Orden de magnitud de los coeficientes de Fourier. Derivación de series de Fourier, 624

*8.7 **Aproximación mediante polinomios trigonométricos y racionales** 625
a. Comentario general sobre las representaciones de funciones, 625 **b.** Teorema de aproximación de Weierstrass, 626 ***c.** Aproximación trigonométrica de Fejer para los polinomios de Fourier, usando medias aritméticas, 627 ***d.** Aproximación en media y relación de Parseval, 629

APENDICE I 632

*A.I.1 **Alargamiento del intervalo del período. Teorema de la integral de Fourier** 632

*A.I.2 **Fenómeno de Gibbs en puntos de discontinuidad** 633

*A.I.3 **Integración de series de Fourier** 635

APENDICE II 636

*A.II.1 **Polinomios de Bernoulli y sus aplicaciones** 636
a. Definición y desarrollo de Fourier, 636 ***b.** La función generadora. La serie de Taylor para la cotangente trigonométrica e hiperbólica, 639 **c.** La fórmula de la suma de Euler-Maclaurin, 642 **d.** Aplicaciones. Desarrollos asintóticos, 644 **e.** Sumas de potencias. Fórmula de recurrencia para los números de Bernoulli, 645 **f.** Constante de Euler y serie de Stirling, 646

PROBLEMAS 649

CAPITULO 9 **Ecuaciones diferenciales para los tipos más simples de vibraciones** 651

9.1 **Problemas de vibración en Mecánica y en Física** 652
a. Las vibraciones mecánicas más simples, 652 **b.** Oscilaciones eléctricas, 653

9.2 **Solución de la ecuación homogénea. Oscilaciones libres** 654
a. La solución formal, 654 **b.** Interpretación de la solución, 656 **c.** Satisfacción de condiciones iniciales dadas. Unicidad de la solución, 657

9.3 **La ecuación no homogénea. Oscilaciones forzadas** 658
a. Comentarios generales. Superposición, 658 **b.** Solución de la ecuación no homogénea, 660 **c.** La curva de resonancia, 661 **d.** Una discusión adicional sobre las oscilaciones, 664 **e.** Comentarios sobre la construcción de instrumentos de registro, 665

LISTA DE FECHAS BIOGRAFICAS 667

INDICE 669

INTRODUCCIÓN AL CÁLCULO Y AL ANÁLISIS MATEMÁTICO

1

Introducción

Desde la antigüedad las nociones intuitivas de cambio continuo, crecimiento y movimiento, han atraído la atención de las mentes científicas. Empero, el camino hacia el entendimiento de variación continua no fue logrado sino hasta el siglo XVII, cuando la ciencia moderna surgió y se desarrolló rápidamente en cercana conjunción con el cálculo integral y diferencial, llamado brevemente Cálculo, y el análisis matemático.

Las nociones básicas del Cálculo son las de derivada e integral: la derivada es una medida de la "rapidez" de una variación; la integral, una medida del efecto total de un proceso de cambio continuo. Una comprensión precisa de estos conceptos y de su abrumadora fecundidad descansa sobre los conceptos de límite y función; los cuales a su vez dependen de un entendimiento del continuo de números. Sólo gradualmente, penetrando más y más en la esencia del Cálculo, puede uno apreciar su potencia y belleza. En este capítulo introductorio se explicarán los conceptos básicos de número, función y límite, primero de manera simple e intuitiva, y después mediante argumentos cuidadosos.

1.1 El continuo de números

Los enteros positivos o *números naturales* 1, 2, 3, ... son símbolos abstractos para indicar "cuántos" objetos hay en una *colección* o *conjunto* de elementos discretos.

Estos símbolos están despojados de toda referencia a las cualidades concretas de los objetos contados, ya sea que se trate de personas, átomos, casas o cualesquiera objetos.

Los números naturales son el instrumento adecuado para contar elementos de una colección o "conjunto". Sin embargo, no bastan para otro objetivo igualmente importante: *medir* cantidades tales como la longitud de una curva y el volumen o peso de un cuerpo. La pregunta "¿cuánto?"

no puede ser contestada inmediatamente en términos de los números naturales. La profunda necesidad de expresar medidas de cantidades en términos de lo que nosotros quisiéramos llamar números nos obliga a extender el concepto de número de manera que podamos describir una graduación continua de las medidas. Esta extensión es llamada el *continuo de números* o el sistema de "números reales" (un nombre no descriptivo pero aceptado generalmente). La extensión del concepto de número al de continuo es de tal manera convincente que fue utilizada por todos los grandes matemáticos y científicos de épocas anteriores sin hacer preguntas escudriñadoras. No fue sino hasta el siglo XIX que los matemáticos se vieron impulsados a buscar un fundamento lógico más firme para el sistema de los números reales. La subsiguiente formulación precisa de estos conceptos condujo, a su vez, a un progreso mayor en matemáticas. Principiaremos con un tratamiento intuitivo sin complicaciones, y posteriormente se dará un análisis más profundo del sistema de los números reales.[1]

a. El sistema de números naturales y su extensión. Numeración y medición

Los números naturales y los racionales. La sucesión de números "naturales" 1, 2, 3, ... se considera como dada. No se requiere discutir cómo estos entes abstractos, los números, pueden ser catalogados desde un punto de vista filosófico. Para el matemático, y para cualquiera que trabaje con números, es importante conocer solamente las reglas o leyes mediante las cuales éstos pueden combinarse para obtener otros números naturales. Estas leyes forman la base de las reglas conocidas para sumar y multiplicar números en el sistema decimal; incluyen las *leyes conmutativas* $a + b = b + a$ y $ab = ba$, las *leyes asociativas* $a + (b + c) = (a + b) + c$ y $a(bc) = (ab)c$, la *ley distributiva* $a(b + c) = ab + ac$, la *ley de la cancelación*, esto es que $a + c = b + c$ implica $a = b$, etc.

Las operaciones inversas, substracción y división, no son siempre posibles dentro del conjunto de los números naturales; no podemos restar 2 de 1 o dividir 1 entre 2 y permanecer dentro de ese conjunto. Para hacer posibles estas operaciones sin restricción, nos vemos obligados a extender el concepto de número inventando el número 0, los enteros "negativos" y las fracciones. La totalidad de todos estos números se llama la clase o conjunto de *números racionales;* todos ellos se obtienen de la unidad utilizando las "operaciones racionales" de cálculo, a saber: la adición, substracción, multiplicación y división.[2]

---

[1] Una exposición más completa se da en *What Is Mathematics?* por Courant y Robbins, Oxford University Press, 1962.

[2] Aquí, la palabra "racional" no significa razonable o lógico sino que se deduce de la palabra "razón", significando la proporción relativa de dos magnitudes.

Un número racional puede expresarse siempre en la forma p/q, donde p y q son enteros y $q \neq 0$. Podemos hacer que esta representación sea única exigiendo que q sea positivo y que p y q no tengan un factor común mayor que 1.

Dentro del dominio de los números racionales todas las *operaciones racionales,* adición, multiplicación, substracción y división (excepto la división por cero), pueden efectuarse y producir nuevamente números racionales. Como se sabe de la aritmética elemental, las operaciones con números racionales obedecen las mismas leyes que las operaciones con números naturales; por ello los números racionales extienden el sistema de enteros positivos en una forma completamente natural.

Representación gráfica de los números racionales. Los números racionales son usualmente representados en forma gráfica mediante puntos en una línea recta L, el *eje numérico* (o *recta numérica).* Tomando un

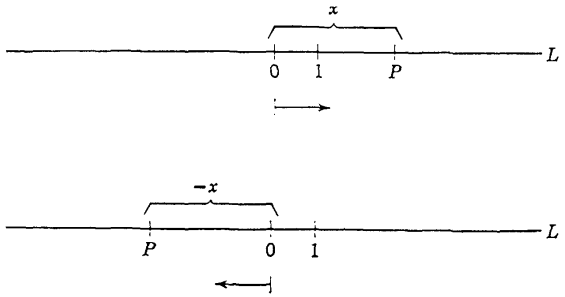


Figura 1.1 El eje numérico.

punto arbitrario de L como el origen o punto 0 y otro punto arbitrario como el punto 1, utilizamos la distancia entre estos dos puntos como escala o unidad de medida y definimos la dirección de 0 a 1 como "positiva". La línea recta con una dirección así impuesta se llama recta dirigida. Es costumbre dibujar a L de manera que el punto 1 se encuentre a la derecha del punto 0 (Fig. 1.1). La localización de cualquier punto P en L está completamente determinada por dos partes de información: la distancia a P desde el origen 0 y la dirección de 0 a P (a la derecha o izquierda del 0). El punto P en L que representa un número racional positivo se encuentra a una distancia de x unidades a la derecha del 0. Un número racional negativo x es representado por el punto $-x$ unidades a la izquierda del 0. En cualquier caso la distancia del 0 al punto que representa a x se llama el valor absoluto de x, denotado por $|x|$, y se tiene

$$|x| = \begin{cases} x, & \text{si } x \text{ es positivo o cero,} \\ -x, & \text{si } x \text{ es negativo} \end{cases}$$

Se observa que x nunca es negativo y es igual a cero sólo cuando $x = 0$.

Recordemos de la geometría elemental que con regla y compás es posible subdividir la unidad de longitud en un número cualquiera de partes iguales. Se sigue que cualquier longitud racional puede ser construida, y, por lo tanto, que el punto que representa un número racional x puede ser encontrado por métodos puramente geométricos.

De esta forma se obtiene una representación geométrica de números racionales mediante puntos en L, los llamados *puntos racionales.* Conforme a nuestra notación para los puntos 0 y 1, nos tomamos la libertad de denotar tanto los números racionales como los puntos correspondientes en L por el mismo símbolo x.

La relación $x < y$ para dos números racionales significa geométricamente que el punto x se encuentra a la izquierda del punto y. En ese caso la distancia entre los puntos es $y - x$ unidades. Si $x > y$, la distancia es $x - y$ unidades. En cualquier caso la distancia entre dos puntos racionales x, y de L es $|y - x|$ unidades y es también un número racional.

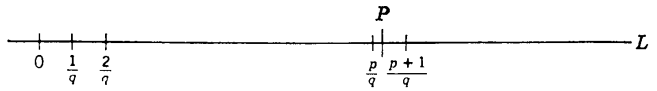


Figura 1.2

Un segmento en L con puntos extremos a, b con $a < b$ se denominará *intervalo.* El segmento particular con puntos extremos 0, 1 se llama *intervalo unitario.* Si los puntos extremos están incluidos en el intervalo, se dirá que el intervalo es *cerrado;* si los puntos extremos se excluyen, el intervalo se llama *abierto.* El intervalo abierto, denotado por (a, b), consiste en aquellos puntos x para los cuales $a < x < b$, esto es, en aquellos puntos que se encuentran "entre" a y b. El intervalo cerrado, denotado por $[a, b]$, consiste en los puntos x para los cuales $a < x < b$.[1] En cualquier caso la *longitud del intervalo* es $b - a$.

Los puntos que corresponden a los enteros 0, ± 1, ± 2, ... subdividen el eje numérico en intervalos de longitud unitaria. Todo punto en L es un punto extremo o bien un punto interior de uno de los intervalos de la subdivisión. Si se sigue subdividiendo todo intervalo en q partes iguales, se obtiene una subdivisión de L en intervalos de longitud $1/q$ por puntos racionales de la forma p/q. Todo punto P de L es entonces un punto racional de la forma p/q, o bien se encuentra entre dos puntos racionales sucesivos p/q y $(p + 1)/q$ (ver Fig. 1.2). Puesto que puntos sucesivos de la subdivisión distan en $1/q$ unidades, se sigue que podemos encontrar un punto

[1] La relación $a \leq x$ (léase "a es menor que o igual a x") es interpretada como "$a < x$, o bien $a = x$". Los dobles signos $\geq$ y $\pm$ se interpretan de manera análoga.

racional p/q cuya distancia desde P no excede $1/q$ unidades. El número $1/q$ puede hacerse tan pequeño como se desee escogiendo q como un entero positivo suficientemente grande. Por ejemplo, escogiendo $q = 10^n$ (donde n es cualquier número natural), podemos encontrar una "fracción decimal" $x = p/10^n$ cuya distancia desde P es menor que $1/10^n$. Aunque no se afirma que todo punto de L es un punto racional, se ve por lo menos que pueden encontrarse puntos racionales arbitrariamente próximos a cualquier punto P de L.

Densidad

La proximidad arbitraria de puntos racionales a un punto P dado de L se expresa diciendo: *Los puntos racionales son densos en el eje numérico.* Es claro que aun conjuntos menores de números racionales son densos, por ejemplo, los puntos $x = p/10^n$, para todos los números naturales n y enteros p.

La densidad implica que entre dos puntos racionales cualesquiera a y b existe una infinidad de otros puntos racionales. En particular, el punto medio entre a y b, $c = \frac{1}{2}(a + b)$, correspondiente a la media aritmética de los números a y b, es nuevamente racional. Tomando los puntos medios de a y c, de b y c, y continuando de esta manera, se puede obtener cualquier número de puntos racionales entre a y b.

Un punto arbitrario P en L puede ser localizado con cualquier grado de precisión utilizando puntos racionales. Puede parecer entonces a primera vista que la labor de localizar P por un número ha sido lograda introduciendo los números racionales. Después de todo, en la realidad física las cantidades nunca se dan o conocen con absoluta precisión sino que siempre con sólo un grado de incertidumbre y por ello pueden muy bien ser consideradas como medidas por números racionales.

Cantidades inconmensurables. Los números racionales, aunque densos, no bastan como base teórica de medición por medio de números. Dos cantidades cuya razón es un número racional se llaman *conmensurables,* debido a que pueden ser expresadas como múltiplos enteros de una unidad común. Desde época tan remota como la de los siglos v o vi A.C. los matemáticos y filósofos griegos hicieron el sorprendente y profundo descubrimiento de que existen cantidades que no son conmensurables con una unidad dada. En particular, existen segmentos de recta que no son múltiplos racionales de un segmento unitario dado.

Es fácil dar un ejemplo de una longitud inconmensurable con la longitud unidad: la diagonal l de un cuadrado con lados de longitud unidad. Pues, por el teorema de Pitágoras, el cuadrado de esta longitud l debe ser

igual a 2. Por ello, si l fuese un número racional y consecuentemente igual a p/q, donde p y q son enteros positivos, tendríamos $p^2 = 2q^2$. Podemos suponer que p y q no tienen factores comunes, pues tales factores comunes podrían ser cancelados en primer lugar. De acuerdo con la ecuación anterior, p^2 es un número par; de aquí que p mismo debe ser par, digamos $p = 2p'$. Sustituyendo p por $2p'$ se obtiene $4p'^2 = 2q^2$, o bien $q^2 = 2p'^2$; consecuentemente, q^2 sería par y por lo tanto q también par. Esto es lo mismo que decir, que p y q tienen ambos el factor 2. Sin embargo, esto contradice nuestra hipótesis de que p y q no tienen un factor común. Puesto que la suposición de que la diagonal puede ser representada por una fracción p/q da lugar a una contradicción, la suposición es falsa.

Este razonamiento, ejemplo característico de *prueba indirecta,* muestra que el símbolo $\sqrt{2}$ no puede corresponder a ningún número racional. Otro ejemplo es π, la razón de la circunferencia de un círculo a su diámetro. La demostración de que π no es racional es mucho más complicada y se obtuvo solamente en tiempos modernos (Lambert, 1761). Es fácil encontrar muchas cantidades inconmensurables (véase problema 1, p. 129); de hecho, las cantidades inconmensurables son, en cierto sentido, mucho más comunes de las conmensurables (véase p. 122).

Números irracionales

Debido a que el sistema de números racionales no es suficiente para la geometría, es necesario inventar nuevos números como medidas de cantidades inconmensurables: estos nuevos números se llaman "irracionales". Los antiguos griegos no pusieron de relieve el concepto de número abstracto; sin embargo, consideraron entes geométricos, tales como segmentos de recta, como los elementos básicos. De manera puramente geométrica ellos desarrollaron un sistema lógico para tratar y operar con cantidades inconmensurables, así como también con las conmensurables (racionales). Este logro tan importante, iniciado por los pitagóricos, fue desarrollado grandemente por Eudoxio y está expresado extensamente en el famoso *Elementos* de Euclides. En los tiempos modernos las matemáticas fueron creadas de nuevo y vastamente desarrolladas sobre el fundamento de conceptos numéricos en lugar de conceptos geométricos. Con la introducción de la geometría analítica, se desarrolló una reversión de orientación e importancia en la relación antigua entre números y cantidades geométricas, y así la teoría clásica de los inconmensurables fue poco menos que olvidada u omitida. Se supuso como un hecho el que a todo punto del eje numérico le corresponde un número racional o irracional y que esta totalidad de números "reales" obedece las mismas leyes aritméticas que los números racionales. Sólo posteriormente, en el siglo xix, se sintió la necesidad de justificar

tal hipótesis y finalmente fue satisfecha completamente en un notable folleto de Dedekind cuya lectura es fascinante aún hoy en día.[1]

En efecto, Dedekind mostró que el tratamiento intuitivo e "ingenuo" seguido por todos los grandes matemáticos desde Fermat y Newton hasta Gauss y Riemann estaba sobre la pista correcta: que el sistema de números reales (como símbolos para las longitudes de segmentos, o definidos de alguna otra manera) es un instrumento coherente y completo para la medición científica, y que en este sistema las reglas de cálculo del sistema de números racionales siguen siendo válidas.

Sin perjuicio, podríamos dejarlo así y dirigirnos directamente a la esencia del cálculo. Sin embargo, para un entendimiento más profundo del concepto de número real, el cual es necesario para nuestro trabajo posterior, la siguiente explicación, así como el Suplemento a este capítulo, deberían ser estudiados.

b. Números reales e intervalos encajados

Por el momento imaginemos los puntos en una recta L como los elementos básicos del continuo. Postulamos el que a cada punto en L le corresponde un "número real" x, su *coordenada,* y que para estos números x, y las relaciones recién descritas para los números racionales conservan su significado. En particular, la relación $x < y$ indica orden en L y la expresión $|y - x|$ significa la distancia entre el punto x y el punto y. El problema básico es el de relacionar estos números (o mediciones en el continuo de puntos dado geométricamente) con los números racionales considerados originalmente y, por tanto, finalmente con los enteros. Además, debemos explicar cómo operar con los elementos de este "continuo de números" en la misma forma que con los números racionales. Con el tiempo, formularemos el concepto del continuo independientemente de los conceptos geométricos intuitivos, pero por ahora postergamos algo de la discusión más abstracta hasta el Suplemento.

¿Cómo podemos describir un número real irracional? Para algunos números, tales como $\sqrt{2}$ o π, podemos dar una caracterización geométrica simple, pero esto no es siempre factible. Un método suficientemente flexible para obtener cualquier punto real x consiste en describir el valor x por una sucesión de aproximaciones racionales de precisión cada vez mayor. Específicamente, aproximaremos x simultáneamente desde la derecha y desde la izquierda incrementando sucesivamente la precisión y de tal manera que el margen de error se aproxime a cero. En otras palabras, utilizamos una

[1] R. Dedekind, "Nature and Meaning of Number", en *Essays on Number,* Londres y Chicago, 1901. (El primero de estos ensayos, "Continuity and Irrational Numbers", proporciona una información detallada de la definición y leyes de operación con números reales.) Reimpreso bajo el título *Essays on the Theory of Numbers,* Dover, Nueva York, 1964. El original de estas traducciones apareció en 1887 bajo el título "Was sind und was sollen die Zahlen?"

"sucesión" de intervalos racionales que contienen a x, con cada intervalo de la sucesión conteniendo al siguiente de tal manera que la longitud del intervalo, y con él el error de aproximación, puede hacerse, tomando intervalos suficientemente alejados en la sucesión, menor que cualquier número positivo especificado.

Para comenzar, confínese x a un intervalo cerrado $I_1 = [a_1, b_1]$, esto es,

$$a_1 \leq x \leq b_1,$$

donde a_1 y b_1 son racionales (ver Fig. 1.3). Dentro de I_1 consideramos un "subintervalo" $I_2 = [a_2, b_2]$ que contiene a x, esto es,

$$a_1 \leq a_2 \leq x \leq b_2 \leq b_1,$$

donde a_2 y b_2 son racionales. Por ejemplo, podemos escoger para I_2 una de las mitades de I_1, pues x debe encontrarse en uno o ambos de los semiintervalos. Dentro de I_2 consideramos un subintervalo $I_3 = [a_3, b_3]$ que también contiene a x:

$$a_1 \leq a_2 \leq a_3 \leq x \leq b_3 \leq b_2 \leq b_1,$$

donde a_3 y b_3 son racionales, etc. Se requiere que la longitud del intervalo I_n *tienda a cero* con n creciente; esto es, que la longitud de I_n sea menor que cualquier número positivo preasignado, para toda n suficientemente grande. Un conjunto de intervalos cerrados $I_1, I_2, I_3, \ldots$, cada uno conteniendo al siguiente y tal que las longitudes tiendan a cero se denominará "sucesión anidada de intervalos" o "encaje" de intervalos. El punto x está determinado de manera única por la sucesión o encaje; esto es, ningún otro

Figura 1.3 Una sucesión anidada de intervalos.

punto y puede pertenecer a todo I_n, puesto que la distancia entre x e y excede la longitud de I_n una vez que n es suficientemente grande. Puesto que aquí escogemos siempre puntos racionales para los puntos extremos de los I_n y puesto que todo intervalo con puntos extremos racionales es descrito por dos números racionales, vemos que todo punto x de L, esto es, todo número real, puede ser precisamente determinado con la ayuda de una infinidad de números racionales. La aseveración inversa no es tan obvia; la aceptaremos como un *axioma* básico.

POSTULADO DE LOS ENCAJES DE INTERVALOS. *Si $I_1, I_2, I_3, \ldots$ forman una sucesión "encaje" de intervalos con puntos extremos racionales, existe un punto x contenido en todo I_n.*[1]

[1] Es importante destacar que para un encaje los intervalos I_n son *cerrados.* Si, por ejemplo, I_n denota el intervalo abierto $0 < x < 1/n$, entonces cada I_n contiene al siguiente y las longitudes de los intervalos tienden a cero, pero no existe x contenido en todo I_n.

Como veremos, éste es un *axioma de continuidad:* garantiza que no existen vacíos en el eje real. Usaremos el axioma para caracterizar el continuo real y justificar todas las operaciones con límites que son básicas para el cálculo y el análisis. (Como veremos posteriormente, hay también muchas otras maneras de formular este axioma.)

c. Fracciones decimales. Bases distintas a la de diez

Fracciones decimales infinitas. Una de las muchas maneras de definir números reales es la descripción familiar en términos de decimales infinitos. Es enteramente posible tomar los decimales infinitos como los objetos básicos, en lugar de los puntos del eje numérico; pero procederíamos tal vez en una forma geométrica más sugestiva definiendo la representación decimal infinita de números reales en términos de encajes de intervalos.

Consideremos el eje numérico subdividido en intervalos unitarios por los puntos correspondientes a enteros. Un punto x se encuentra entre dos puntos sucesivos de subdivisión, o bien es él mismo uno de los puntos de división. En cualquier caso existe por lo menos un entero c_0 tal que

$$c_0 \leq x \leq c_0 + 1,$$

de modo que x pertenece al intervalo cerrado $I_0 = [c_0, c_0 + 1]$. Dividimos I_0 en diez partes iguales por puntos $c_0 + \frac{1}{10}$, $c_0 + \frac{2}{10}$, ..., $c_0 + \frac{9}{10}$. El punto x debe entonces pertenecer por lo menos a uno de los subintervalos cerrados de I_0 (posiblemente a dos adyacentes, si x es uno de los puntos de subdivisión). En otras palabras, existe un dígito c_1 (esto es, uno de los enteros 0, 1, 2, ..., 9) tal que x pertenece al intervalo cerrado I_1 dado por

$$c_0 + \tfrac{1}{10}c_1 \leq x \leq c_0 + \tfrac{1}{10}c_1 + \tfrac{1}{10}.$$

Dividiendo a su vez I_1 en diez partes iguales encontramos un dígito c_2 tal que x se encuentra en el intervalo I_2 dado por

$$c_0 + \tfrac{1}{10}c_1 + \tfrac{1}{100}c_2 \leq x \leq c_0 + \tfrac{1}{10}c_1 + \tfrac{1}{100}c_2 + \tfrac{1}{100}.$$

Repetimos este proceso. Después de n pasos x está confinado en un intervalo I_n dado por

$$c_0 + \frac{1}{10}c_1 + \cdots + \frac{1}{10^n}c_n \leq x \leq c_0 + \frac{1}{10}c_1 + \cdots + \frac{1}{10^n}c_n + \frac{1}{10^n},$$

donde $c_1, c_2, \ldots$ son todos dígitos. El intervalo I_n tiene longitud $1/10^n$, la cual tiende a cero para n creciente. Es claro que los I_n forman un encaje de intervalos y, por lo tanto, que x está determinado de manera única por los I_n. Puesto que los I_n son conocidos, una vez que los números $c_0, c_1, c_2, \ldots$ están dados, encontramos que un número real arbitrario puede ser descrito completamente por una sucesión infinita de enteros $c_0, c_1, c_2, \ldots$, donde

todos excepto el primero son dígitos que tienen valores desde cero hasta nueve solamente. En la notación decimal ordinaria, la conexión entre x y $c_0, c_1, c_2, \ldots$ se indica escribiendo

$$x = c_0 + 0.c_1c_2c_3\cdots.$$

(Usualmente el entero c_0 mismo es también escrito en notación decimal si c_0 es positivo.) Inversamente, por el axioma de continuidad, toda aquella expresión que denota una fracción decimal infinita representa un número real.

Es posible que existan dos representaciones decimales diferentes del mismo número; por ejemplo,

$$1 = 0.99999\cdots = 1.00000\cdots.$$

En nuestra construcción el entero c_0 está determinado de manera única por x, a menos que el propio x sea un entero. En este caso podríamos elegir $c_0 = x$ o bien $c_0 = x - 1$. Una vez efectuada la elección, el dígito c_1 es único a menos que x sea uno de los nuevos puntos que subdividen I_0 en diez partes iguales. Continuando así, encontramos que c_0 y todos los c_k están determinados de manera única por x a menos que x resulte un punto de subdivisión en algún paso. Si esto sucediera por primera vez en el n-ésimo paso, entonces

$$x = c_0 + \frac{1}{10}c_1 + \cdots + \frac{1}{10^n}c_n,$$

donde $c_1, c_2, \ldots, c_n$ son dígitos y donde $c_n > 0$, puesto que de otra manera x hubiera sido un punto de subdivisión en algún paso anterior. Se sigue que I_{n+1} es el intervalo $[x, x + 1/10^{n+1}]$ o bien el intervalo $[x - 1/10^{n+1}, x]$. En el primer caso x será el punto extremo izquierdo de todos los intervalos posteriores $I_{n+2}, I_{n+3}, \ldots$, y en el segundo caso, el punto extremo derecho. Esto nos lleva entonces a la representación decimal

$$x = c_0 + 0.c_1c_2\cdots c_n000\cdots$$

o bien a la representación

$$x = c_0 + 0.c_1c_2\cdots(c_n - 1)99999\cdots.$$

De aquí que el único caso en el cual puede surgir una ambigüedad es para un número racional x que pueda ser escrito como una fracción que tiene una potencia de diez para su denominador. Podemos aún eliminar esta ambigüedad excluyendo representaciones decimales en las cuales todos los dígitos a partir de cierto punto son nueves.

En la representación decimal de números reales el papel especial desempeñado por el número diez es puramente incidental. La única razón evidente para el extenso uso del sistema decimal es la facilidad de contar por

decenas con nuestros dedos (dígitos). Todo entero p mayor que uno puede servir igualmente bien. Podrían utilizarse p subdivisiones iguales en cada paso. Un número real x sería entonces representado en la forma

$$x = c_0 + 0.c_1c_2c_3\cdots,$$

donde c_0 es un entero, y ahora $c_1, c_2, \ldots$ tienen uno de los valores 0, 1, 2, ..., $p-1$. Esta representación caracteriza nuevamente a x por un encaje de intervalos, a saber

$$c_0 + \frac{1}{p}c_1 + \cdots + \frac{1}{p^n}c_n \leq x \leq c_0 + \frac{1}{p}c_1 + \cdots + \frac{1}{p^n}c_n + \frac{1}{p^n}.$$

Si x es positivo o cero, el entero c_0 es también positivo o cero, y el propio c_0 tiene un desarrollo finito de la forma

$$c_0 = d_0 + pd_1 + p^2d_2 + \cdots + p^kd_k,$$

donde $d_0, d_1, \ldots, d_k$ toman uno de los valores 0, 1, ..., $p-1$. La representación completa de x "en la base p" toma la forma

$$x = d_kd_{k-1}\cdots d_1d_0.c_1c_2c_3\cdots.$$

Si x es negativo podemos utilizar este tipo de representación para $-x$.

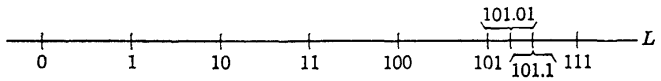


Figura 1.4 La fracción $2\frac{1}{4}$ en el sistema binario.

Bases distintas a la de 10 han sido utilizadas extensamente en la actualidad. Siguiendo la dirección de los antiguos babilonios, los astrónomos durante muchos siglos representaron sistemáticamente los números como fracciones "sexagesimales", con $p = 60$ como la base.

Representación binaria. El sistema "binario" con la base $p = 2$ tiene un especial interés teórico y es útil en el diseño lógico de máquinas calculadoras electrónicas. En el sistema binario los dígitos tienen solamente dos valores posibles, el cero y el uno. El número $2\frac{1}{4}$, por ejemplo, sería escrito 101.01, de acuerdo con la fórmula

$$\frac{21}{4} = 2^2\cdot 1 + 2^1\cdot 0 + 1\cdot 1 + \frac{1}{2}\cdot 0 + \frac{1}{2^2}\cdot 1 \qquad \text{(ver Fig. 1.4)}.$$

Cálculo con números reales. Aunque la definición de números reales y sus representaciones decimal infinita, binaria, etc., son naturales, puede no parecer obvio el que uno pueda operar con el continuo de números exactamente como con números racionales, efectuando las operaciones racionales

y reteniendo las leyes de la aritmética, tales como las leyes asociativa, conmutativa y distributiva. La demostración es simple aunque algo tediosa. En lugar de dificultar el camino hacia la esencia viva del análisis considerando la cuestión aquí, aceptaremos temporalmente la posibilidad de cálculo aritmético ordinario con los números reales. Una comprensión más profunda de la estructura lógica subyacente al concepto de número será alcanzada cuando se descubra la idea de límite y sus implicaciones. (Ver el Suplemento a este capítulo, p. 112.)

d. Definición de vecindad o entorno

No sólo las operaciones racionales, sino también relaciones de orden o desigualdades para números reales obedecen las mismas reglas que los números racionales.

Pares de números reales a y b con $a < b$ dan nuevamente lugar a intervalos cerrados $[a, b]$ (dados por $a \leq x \leq b$) e intervalos abiertos (a, b) (dados por $a < x < b$). Frecuentemente tendremos necesidad de asociar con un punto x_0 los diferentes intervalos abiertos que contienen a ese punto o bien, específicamente, tenerlo como centro; a esos intervalos los llamaremos *vecindades* o *entornos* del punto. Más precisamente, para todo positivo ε, la *ε-vecindad* del punto x_0 consiste de los valores x para los cuales $x_0 - \varepsilon < x < x_0 + \varepsilon$; esto es, es el intervalo $(x_0 - \varepsilon, x_0 + \varepsilon)$. Todo intervalo abierto (a, b) que contiene a un punto x_0, también contiene a una vecindad completa de x_0.

Habiendo definido intervalos con puntos extremos reales, podemos formar ahora *encajes* de intervalos utilizando la misma definición como en el caso de puntos extremos racionales. Es de lo más importante para la coherencia lógica del cálculo que para todo encaje de intervalos con puntos extremos reales exista un número real contenido en todos ellos. (Ver Suplemento, p. 118.)

e. Desigualdades

Reglas básicas

Las desigualdades juegan un papel mucho mayor en matemáticas superiores que en matemáticas elementales. A menudo el valor preciso de una cantidad x es difícil de determinar, mientras que puede ser fácil efectuar una estimación de x, esto es, mostrar que x es mayor que alguna cantidad conocida a y menor que alguna otra cantidad b. Para muchos propósitos sólo la información contenida en tal estimación de x es significativa. Recordaremos por ello brevemente algunas de las reglas elementales acerca de las desigualdades.

El hecho básico es que la suma y producto de dos números reales positivos son también números positivos,* esto es, si $a > 0$ y $b > 0$, entonces $a + b > 0$ y $ab > 0$. Más aún, contamos con el hecho de que la desigualdad $a > b$ es equivalente con $a - b > 0$. Consecuentemente, dos desigualdades $a > b$ y $c > d$ pueden ser *sumadas* para dar la desigualdad $a + c > b + d$, puesto que

$$(a + c) - (b + d) = (a - b) + (c - d)$$

es positivo por ser la suma de dos números positivos. (Restar las desigualdades para obtener $a - c > b - d$ no es legítimo. ¿Por qué?) Una desigualdad puede ser *multiplicada por un número positivo;* esto es, si $a > b$ y $c > 0$, entonces $ac > bc$. Para la demostración, observamos que

$$ac - bc = (a - b)c$$

es positivo puesto que es el producto de números positivos. Si c es *negativo,* podemos concluir de $a > b$ que $ac < bc$. Más generalmente, se sigue de $a > b > 0$ y $c > d > 0$ que $ac > bd$.

Es obvio geométricamente que la desigualdad es *transitiva;* esto es, si $a > b$ y $b > c$, entonces $a > c$. La transitividad[1] se sigue también inmediatamente a partir de la positividad de la suma

$$(a - b) + (b - c) = a - c.$$

Las reglas anteriores también se cumplen si se reemplaza el signo $>$ por $\geq$ en todas partes.

Sean a y b números *positivos* y obsérvese que

$$a^2 - b^2 = (a + b)(a - b).$$

Puesto que $a + b$ es positivo, concluimos que $a^2 > b^2$ es consecuencia de tener $a > b$. Por tanto, una desigualdad entre números positivos puede ser "elevada al cuadrado". Análogamente, $a^2 \geq b^2$ siempre que $a \geq b \geq 0$. De la ecuación

$$a - b = \frac{1}{a + b}(a^2 - b^2),$$

válida para todos a y b positivos, se sigue que la recíproca también es cierta; esto es, para a y b positivos, $a^2 > b^2$ implica $a > b$. Aplicando este resultado a los números $a = \sqrt{x}$, $b = \sqrt{y}$, para números positivos arbitrarios x, y,

* (N. del T.) El conjunto de los números positivos, con las operaciones $+$ y $\cdot$, constituye un *anillo,* que es un *subanillo* de los reales. Para estos conceptos véase cualquier libro de introducción al álgebra moderna.

[1] La transitividad justifica el uso de la fórmula compuesta "$a < b < c$" para expresar "$a < b$ y $b < c$, etc.". Evítense arreglos no transitivos como $x < y > z$; éstos dan lugar a confusiones y engaños.

encontramos[1] que $\sqrt{x} > \sqrt{y}$ cuando $x > y$. Más generalmente, $\sqrt{x} \geq \sqrt{y}$ siempre que $x \geq y \geq 0$. De aquí que es legítimo tomar raíz cuadrada de ambos miembros de una desigualdad entre números reales no negativos.

Supóngase que a y b son positivos y n es un entero positivo. En la factorización

$$a^n - b^n = (a - b)(a^{n-1} + a^{n-2}b + \cdots + b^{n-1})$$

el segundo factor es positivo. Por ello $a^n - b^n$ tiene el mismo signo que $a - b$; si $a^n > b^n$, entonces $a > b$ y si $a^n < b^n$, entonces $a < b$.

Muchas desigualdades que encontraremos aparecen en la forma de estimaciones para el valor absoluto de un número. Recordemos que $|x|$ se define como x para $x \geq 0$ y $-x$ para $x < 0$. Puede decirse también que $|x|$ es el mayor de los números x y $-x$ cuando x no es cero y es igual a ambos cuando x es cero. La desigualdad $|x| \leq a$ establece entonces que ni x ni

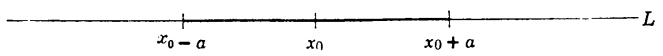


Figura 1.5 El intervalo $|x - x_0| \leq a$.

$-x$ excede a, esto es, que $x \leq a$ y $-x \leq a$. Puesto que $-x \leq a$ es equivalente a $x \geq -a$, vemos que la desigualdad $|x| \leq a$ significa que x se encuentra en el intervalo cerrado $-a \leq x \leq a$ con centro 0 y longitud $2a$. La desigualdad $|x - x_0| \leq a$ establece entonces que $-a \leq x - x_0 \leq a$, o bien que $x_0 - a \leq x \leq x_0 + a$; por lo tanto, que x se encuentra en el intervalo cerrado con centro x_0 y longitud $2a$ (ver Fig. 1.5). Análogamente, la ε-vecindad $(x_0 - \varepsilon, x_0 + \varepsilon)$ de un punto x_0, esto es, el intervalo abierto $x_0 - \varepsilon < x < x_0 + \varepsilon$, puede ser descrita por la desigualdad $|x - x_0| < \varepsilon$.

Desigualdad del triángulo

Una de las desigualdades más importantes que involucra valores absolutos es la llamada desigualdad del triángulo

$$|a + b| \leq |a| + |b|$$

para a, b, reales cualesquiera. El nombre "desigualdad del triángulo" es más apropiado para el aserto equivalente

$$|\alpha - \beta| \leq |\alpha - \gamma| + |\gamma - \beta|$$

para el cual hemos hecho $a = \alpha - \gamma$, $b = \gamma - \beta$. La interpretación geométrica de esta desigualdad es que la distancia directa desde α hasta β es menor

[1] Aquí y en lo sucesivo el símbolo $\sqrt{z}$ para $z \geq 0$ denota aquel número no negativo cuyo cuadrado es z. Con esta convención $|c| = \sqrt{c^2}$ para todo real c puesto que $|c| \geq 0$ y $|c|^2 = c^2$. De esto se obtiene la importante identidad $|xy| = |x| \cdot |y|$ puesto que

$$|xy|^2 = (xy)^2 = x^2y^2 = (|x| \cdot |y|)^2.$$

que o igual a la suma de las distancias vía un tercer punto γ; (esto corresponde también al hecho de que en cualquier triángulo la suma de dos de sus lados excede al tercero).

Una demostración formal de la desigualdad del triángulo es fácil de dar. Distinguimos los casos $a + b \geq 0$ y $a + b < 0$. En el primer caso la desigualdad establece que $a + b \leq |a| + |b|$; pero esto se obtiene trivialmente mediante la suma de las desigualdades $a \leq |a|$ y $b \leq |b|$. En el segundo caso la desigualdad del triángulo se reduce a $-(a + b) \leq |a| + |b|$, la cual se obtiene nuevamente mediante la suma de $-a \leq |a|$, $-b \leq |b|$.

Obtenemos inmediatamente una desigualdad análoga para tres cantidades:

$$|a + b + c| \leq |a| + |b| + |c|;$$

pues, aplicando la desigualdad del triángulo dos veces, resulta

$$|a + b + c| = |(a + b) + c| \leq |a + b| + |c| \leq |a| + |b| + |c|.$$

De la misma manera se obtiene la desigualdad más general

$$|a_1 + a_2 + \cdots + a_n| \leq |a_1| + |a_2| + \cdots + |a_n|$$

Ocasionalmente se requieren estimaciones para $|a + b|$ por debajo. Observamos que

$$|a| = |(a + b) + (-b)| \leq |a + b| + |-b| = |a + b| + |b|,$$

y por lo tanto, que la desigualdad

$$|a + b| \geq |a| - |b|$$

se satisface.

La desigualdad de Cauchy-Schwarz

Algunas de las desigualdades más importantes utilizan el hecho obvio de que el cuadrado de un número real nunca es negativo y que, consecuentemente, una suma de cuadrados tampoco puede ser negativa. Uno de los resultados más frecuentemente utilizados, obtenido de esta manera, es la desigualdad de Cauchy-Schwarz:

$$(a_1b_1 + a_2b_2 + \cdots + a_nb_n)^2 \leq (a_1^2 + a_2^2 + \cdots + a_n^2)(b_1^2 + b_2^2 + \cdots + b_n^2).$$

Haciendo

$$A = a_1^2 + a_2^2 + \cdots + a_n^2,$$
$$B = a_1b_1 + a_2b_2 + \cdots + a_nb_n,$$
$$C = b_1^2 + b_2^2 + \cdots + b_n^2,$$

la desigualdad se transforma en $AC \geq B^2$. Para probarla observamos que para todo t real es

$$0 \leq (a_1 + tb_1)^2 + (a_2 + tb_2)^2 + \cdots + (a_n + tb_n)^2$$

puesto que el miembro derecho es una suma de cuadrados. Desarrollando cada cuadrado y ordenando según las potencias de t, se encuentra que

$$0 \leq A + 2Bt + Ct^2$$

para todo t, donde A, B, C tienen el mismo significado anterior. Aquí es $C \geq 0$. Podemos suponer que $C > 0$, pues, ciertamente, $B^2 = AC = 0$ cuando $C = 0$. Substituyendo entonces para t el valor especial $t = -B/C$ [que corresponde al mínimo de la expresión cuadrática

$$A + 2Bt + Ct^2 = C\left[\left(t + \frac{B}{C}\right)^2 + \left(A - \frac{B^2}{C}\right)\right],$$

encontramos que

$$0 \leq A - \frac{2B^2}{C} + \frac{B^2}{C} = \frac{AC - B^2}{C},$$

y por lo tanto $AC - B^2 \geq 0$.

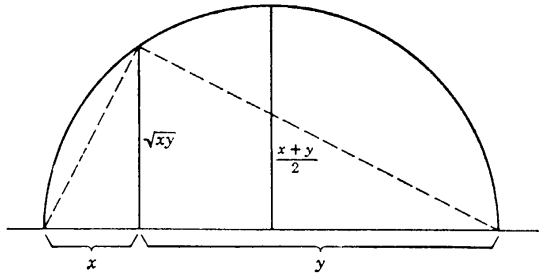


Figura 1.6 Medias geométrica y aritmética de x e y.

En el caso especial $n = 2$ podemos escoger

$$a_1 = \sqrt{x}, \quad a_2 = \sqrt{y}, \quad b_1 = \sqrt{y}, \quad b_2 = \sqrt{x},$$

donde x e y son números positivos. La desigualdad adquiere entonces la forma $(2\sqrt{xy})^2 \leq (x + y)^2$, o bien

$$\sqrt{xy} \leq \frac{x + y}{2}.$$

Esta desigualdad establece que la *media geométrica* $\sqrt{xy}$ de dos números positivos, x, y, no excede nunca su *media aritmética* $(x + y)/2$. La media geométrica de dos números, x, y, puede ser interpretada como la longitud de la altura en un triángulo rectángulo, cuya proyección sobre la hipotenusa la divide en segmentos de longitud x e y respectivamente. Esta desigualdad

establece entonces que en un triángulo rectángulo la altura no excede la mitad de la hipotenusa (ver Fig. 1.6).[1]

1.2 El concepto de función

Desde el comienzo de las matemáticas modernas en el siglo XVII el concepto de función ha estado en el centro mismo del pensamiento matemático. (Leibnitz parece haber sido el primero en utilizar la palabra "función".) Aunque la idea de las relaciones funcionales es significativa mucho más allá del dominio matemático, enfocaremos naturalmente nuestra atención sobre funciones en el sentido matemático, esto es, sobre la conexión entre cantidades matemáticas por relaciones matemáticas o prescripciones u "operaciones". Una gran parte de las matemáticas y las ciencias naturales es dominada por relaciones funcionales, pues aparecen en todo el Análisis, la Geometría, la Mecánica y otros campos. Por ejemplo, la presión en un gas ideal es una función de la densidad y de la temperatura; la posición de una molécula móvil es una función del tiempo; el volumen y la superficie de un cilindro son funciones de su radio y de su altura. Siempre que los valores de ciertas cantidades, $a, b, c, \ldots$, estén determinados por los de ciertas otras, $x, y, z, \ldots$, decimos que $a, b, c, \ldots$ *dependen* de $x, y, z, \ldots$, o bien que son *funciones* de $x, y, z, \ldots$. Ejemplos de relaciones funcionales están dados mediante expresiones formales tales como las siguientes:

a) La fórmula $A = a^2$ define A como función de a. Para $a > 0$, puede interpretarse A como el área de un cuadrado de lado a.

b) La fórmula

$$y = \sqrt{1 - x^2}$$

define y como función de x para todo x tal que $-1 \leq x \leq 1$. Para $x > 0$ esta función expresa el lado y de un triángulo rectángulo con hipotenusa 1, en términos del otro lado x.

c) Las ecuaciones

$$x = t, \qquad y = -t^2$$

asignan valores a x y a y para cada t, y definen por ello a x y a y como funciones de t. Si se interpretan x e y como las coordenadas rectangulares de un punto P en el plano, y t como el tiempo, entonces nuestras ecuaciones describen la localización de P en el tiempo t; en otras palabras, describen el *movimiento* del punto P.

d) Las ecuaciones

$$a = \frac{x}{x^2 + y^2}, \qquad b = \frac{y}{x^2 + y^2}$$

[1] El lector interesado encontrará más material en *An Introduction to Inequalities*, por E. F. Beckenbach y R. Bellman, Random House, 1961, y en *Geometric Inequalities*, por N. Kazarinoff, Random House, 1961.

definen a y b como funciones de x e y para $x^2 + y^2 \neq 0$. Interpretando las parejas de valores (x, y) y (a, b) como coordenadas rectangulares de dos puntos, vemos que las ecuaciones asignan a cada punto (x, y) [con excepción del origen $(0, 0)$] una "imagen" (a, b). El lector puede verificar fácilmente que la imagen (a, b) se encuentra siempre en el mismo rayo desde el origen que el "original" o "antecedente" (x, y) y tiene la distancia recíproca desde el origen. Se habla de *representar* o "mapear" (x, y) sobre (a, b) por medio de las ecuaciones que expresan a y b en términos de x e y.

En los ejemplos anteriores la ley funcional es expresada por fórmulas sencillas que determinan ciertas cantidades en términos de otras.[1] Las cantidades que aparecen en los miembros izquierdos, "las variables dependientes", son expresadas en términos de las "variables independientes", en el miembro derecho. La ley matemática que asigna valores únicos de las variables dependientes a valores dados de las variables independientes se llama una *función*. Esta no es afectada por los nombres x, y, etc., dados a estas variables. En el ejemplo *c)* se tiene una variable independiente t y dos variables dependientes x, y, mientras que en el ejemplo *d)* hay dos variables independientes, x e y y dos variables dependientes, a y b.

La dependencia de y con respecto de x, por una relación funcional, es indicada frecuentemente por la expresión breve: "y es función de x".[2]

a. Función-Gráfica

Dominio y rango de una función

Las variables independientes usualmente se interpretan geométricamente como coordenadas de un punto en una o más dimensiones. En el ejemplo *b)* éste sería un punto en el eje x; en el ejemplo *d)*, un punto en el plano x, y. En ocasiones, las variables independientes son libres de tomar todos los valores, como en los ejemplos *a)* y *c)*. Sin embargo, a menudo existe alguna restricción, inherente, o bien impuesta, y las funciones no están definidas para todos los valores. El conjunto de valores o los puntos para los cuales una función está definida constituye el "*dominio*" de la función. En el ejemplo *a)* el dominio es todo el eje a; en el *b)*, el intervalo $-1 \leq x \leq 1$; en el *c)*, todo el eje t; y en el ejemplo *d)*, los puntos del plano x, y distintos del origen.

[1] Más adelante se comprenderá gradualmente la necesidad de considerar funciones que no son susceptibles de tales representaciones mediante fórmulas sencillas. (Véase, por ejemplo, p. 49.)

[2] Esta locución es utilizada libremente en las ciencias, pero algunos de los textos más pedantes la evitan. No tiene objeto estorbarnos a nosotros mismos por un desmedido afán de tener una "precisión" quisquillosa cuando ésta no tiene relación con la esencia del asunto.

A cada punto P en el dominio las funciones asignan valores definidos para las variables dependientes. Estos valores pueden ser también interpretados como coordenadas de un punto Q, la *imagen* de P. Se dice que P es representado o transformado o "mapeado" por las funciones sobre el punto Q. Por lo tanto, en el ejemplo *d)* el punto $P = (1, 2)$ del plano x, y es transformado en el punto $Q = (\frac{1}{5}, \frac{2}{5})$ del plano a, b. Los puntos imagen Q constituyen el *rango* de la función.[1] Cada Q en el rango es la imagen de uno (o más) puntos en el dominio de la función.

En el ejemplo *c)*, puntos del eje t tienen como imágenes puntos en el plano x, y. El eje t es "mapeado" *en* el plano x, y. Pero no todo punto en el plano x, y aparece como imagen, sino sólo aquéllos para los cuales $y = -x^2$. Por lo tanto, el rango de la función es la parábola $y = -x^2$. Se dice que el eje t es transformado o "mapeado" *sobre* la parábola $y = -x^2$, en el sentido de que los puntos imagen llenan esta parábola.

En el ejemplo *d)*, el rango consiste de los puntos (a, b) en el plano a, b cuyas coordenadas pueden escribirse en la forma $a = x/(x^2 + y^2)$, $b = y/(x^2 + y^2)$, con x, y escogidos de tal manera que $x^2 + y^2 \neq 0$. En otras palabras, el rango consiste de aquellos puntos (a, b) para los cuales las ecuaciones anteriores tienen una solución (x, y). Como se ve inmediatamente, el rango consiste de los puntos (a, b) para los cuales a y b no se hacen cero simultáneamente; cada uno de tales puntos (a, b) es imagen del punto $x = a/(a^2 + b^2)$, $y = b/(a^2 + b^2)$. Toda figura geométrica en el plano x, y es transformada entonces sobre una figura correspondiente en el plano a, b, la cual consiste de las imágenes de los puntos de la primera figura. Por ejemplo, un círculo $x^2 + y^2 = r^2$ en torno al origen es transformado en el círculo $a^2 + b^2 = 1/r^2$ en el plano a, b.

En éste y en los siguientes capítulos trataremos casi exclusivamente con una sola variable independiente, digamos, x, y una sola variable dependiente, digamos, y, como se indica en el ejemplo *b)*.[2] Comúnmente, tal función se representa en la forma usual por su *gráfica* en el plano x, y, esto es, por la curva que consiste de aquellos puntos (x, y) cuya ordenada se encuentra en la relación funcional especificada con la abscisa x (ver Fig. 1.7). Para el ejemplo *b)* la gráfica es la mitad superior de un círculo de radio uno en torno al origen.

La interpretación de función como transformación o representación de un dominio en el eje x sobre un rango en el eje y conduce a una visualización diferente de las funciones. Interpretamos a x e y no como coordenadas del mismo punto en el plano x, y, sino como puntos en dos diferentes

[1] Es conveniente a menudo hablar del punto Q como de "una función" de P, no obstante que en la representación analítica aparecen varias funciones que expresan las diferentes coordenadas de Q.

[2] Sin embargo, se deberá recalcar desde el principio que funciones de varias variables aparecen de la misma natural manera en muchos casos. Estas serán discutidas sistemáticamente en el Volumen II.

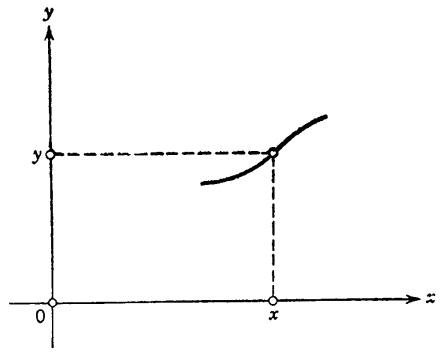


Figura 1.7 Gráfica de una función.

ejes numéricos independientes. Entonces la función transforma o representa un punto x del eje x en un punto y sobre el eje y. Tales transformaciones surgen frecuentemente en la geometría, tal como la transformación "afín", la cual se produce proyectando un punto x del eje x sobre un punto y en un eje y paralelo, desde un centro 0 localizado en el plano de los dos ejes (ver Fig. 1.8). Esta transformación puede ser expresada analíticamente, como es fácil comprobar, por la función lineal $y = ax + b$ con a y b cons-

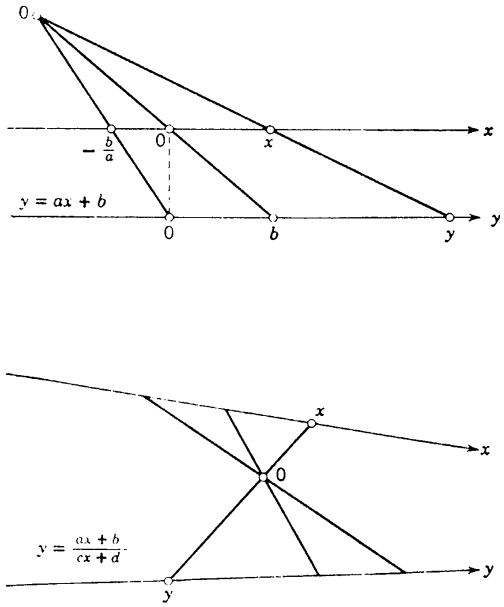


Figura 1.8 Mapeos.

tantes. Obviamente, es una función *biunívoca* o "uno a uno", en la cual inversamente al punto imagen y le corresponde un único punto x original.* Otro ejemplo más general es la "transformación *perspectiva*" definida por el mismo tipo de proyección, salvo que los dos ejes no son necesariamente paralelos. En este caso la expresión analítica está dada por la función racional lineal de la forma $y = (ax + b)/(cx + d)$, con a, b, c y d constantes.

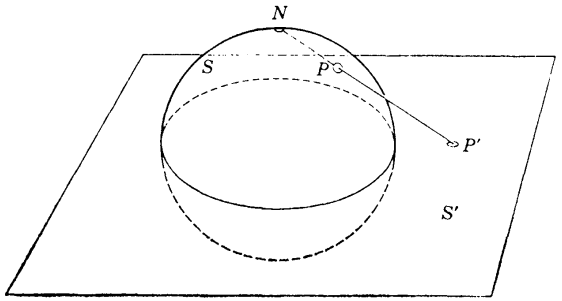


Figura 1.9 Proyección estereográfica.

Cualquier proyección de una superficie S en el espacio sobre otra superficie, S', desde algún centro N, puede ser vista como una transformación o función cuyo dominio es S y cuyo rango se encuentra en S'. Por ejemplo, podemos transformar o "representar" una esfera sobre un plano ecuatorial, proyectando cada punto P de la esfera sobre un punto P' del plano mediante rayos desde el Polo Norte (ver Fig. 1.9). Esta representación o "mapeo" es la "proyección estereográfica", que se utiliza frecuentemente para mapas de la Tierra. Ejemplos de este tipo sugieren la interpretación de funciones como "mapeos" o transformaciones.

Cuando intervienen más variables independientes o dependientes, la definición de funciones por transformaciones o "mapeos" proporciona una interpretación más flexible y adecuada que aquélla por gráficas. Este hecho resultará totalmente claro en el segundo volumen.

b. Definición del concepto de funciones de una variable continua. Dominio y rango de una función

Una función de una sola variable independiente x asigna valores y a valores x. El *dominio* de la función es la totalidad de valores x para los cuales la función está definida. En los casos que más nos interesan, el dominio de la función consiste de uno o varios intervalos (ver Fig. 1.10). Se

* (N. del T.) O sea, tanto la transformación dada como su *inversa* son "*unívocas*" (o univaluadas).

dice entonces que y es función de una *variable continua* (en contraste con otros casos donde, por ejemplo, la función puede estar definida sólo para valores racionales o enteros de x). Aquí, los "intervalos" que forman el dominio pueden o no contener sus puntos extremos, y pueden extenderse también al infinito en una o ambas direcciones.[1] Así, la función $y = \sqrt{1 - x^2}$ está definida en el intervalo cerrado $-1 \leq x \leq +1$; la función $y = 1/x$, en los dos intervalos abiertos semiinfinitos $x < 0$ y $x > 0$; la función $y = x^2$,

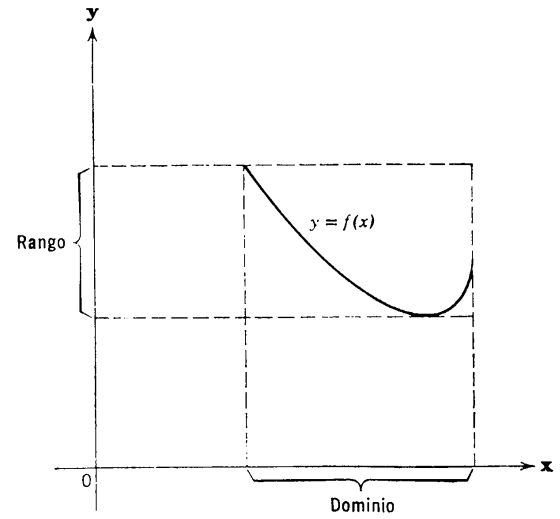


Figura 1.10 Dominio y rango de una función en representación gráfica.

en el 'intervalo" infinito $-\infty < x < +\infty$ que consiste de todo x; la función $y = \sqrt{(x^2 - 1)(4 - x^2)}$, en los dos intervalos separados $1 \leq x \leq 2$ y $-2 \leq x \leq -1$.

Las funciones se denotan por símbolos tales como f, F, g, etc. Las correspondientes relaciones entre x y los valores asociados y se escriben en la forma $y = f(x)$, o $y = F(x)$ o $y = g(x)$, etc., o también, en ocasiones, $y = y(x)$, para indicar[2] que y depende de x. Si, por ejemplo, $f(x)$ está definida por la expresión $x^2 + 1$, se tiene $f(3) = 3^2 + 1 = 10$, $f(-1) = (-1)^2 + 1 = 2$.

[1] Ordinariamente se reservará la palabra "intervalo" para el "acotado", esto es, para intervalos "finitos", que tienen puntos extremos que son números finitos definidos; entonces uno puede indicar el concepto más amplio, tal como se utiliza en el texto, por las palabras "conjuntos convexos", significando conjuntos que cuando contienen dos puntos deben contener todos los intermedios.

[2] En esta notación se trata de destacar las variables y no de indicar explícitamente la operación funcional por un símbolo tal como f. La notación

$$f: \quad x \to y$$

para la función f que mapea x en y se encuentra también ocasionalmente.

Naturaleza de la relación funcional

En la definición general de función, $f(x)$, no se dice nada respecto a la naturaleza de la relación mediante la cual se encuentra la variable dependiente cuando es dada la variable independiente. Como se dijo anteriormente, la función está dada a menudo en "forma cerrada" o explícita por una expresión sencilla como $f(x) = x^2 + 1$ o $f(x) = \sqrt{1 + \operatorname{sen}^2 x}$, y en los primeros días del cálculo tales expresiones explícitas fueron casi exclusivamente lo que los matemáticos entendieron por funciones. A menudo, los dispositivos mecánicos generan curvas geométricas o bien gráficas que entonces definen funciones. Un ejemplo notable es la cicloide, una curva descrita por un punto fijo sobre un círculo que rueda a lo largo del eje x (ver Fig. 1.11). Su expresión analítica funcional por fórmulas se dará posteriormente (ver Sec. 4.1-*c*).

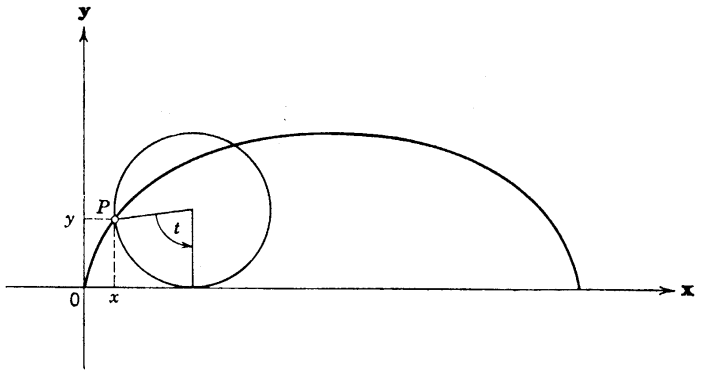


Figura 1.11.

Es claro que no estamos restringidos a tales funciones generadas geométrica o mecánicamente. *Cualquier regla o ley* mediante la cual un valor de y es asignado a valores de x constituye una función. En algunas investigaciones teóricas la amplia generalidad o bien la vaguedad del concepto de función es, de hecho, una ventaja. Sin embargo, para las aplicaciones, particularmente en el cálculo, el concepto general de función es innecesariamente amplio. Para hacer posibles desarrollos matemáticos significativos, las leyes "arbitrarias" de correspondencia mediante las cuales un valor de y es asignado a x deben estar sujetas a restricciones radicales. Durante el siglo pasado los matemáticos reconocieron y formularon en términos precisos las restricciones esenciales que había que imponer sobre tal concepto tan general para obtener funciones que verdaderamente tuvieran las propiedades útiles que uno intuitivamente podría esperar.

**Dominios de funciones extendidos o restringidos*

Aún para funciones dadas por fórmulas explícitas es importante darse cuenta de que cualquier descripción completa de una función debe incluir una definición del *dominio* de la función. Para nosotros la "función" f descrita por "$f(x) = x^2$ para $0 < x < 2$" no es estrictamente la misma función que la función g dada por "$g(x) = x^2$, en el dominio mayor $-2 < x < 2$", aun cuando $f(x)$ y $g(x)$ tengan los mismos valores en el intervalo $0 < x < 2$ donde ambas están definidas. Generalmente, una función f se llama una "restricción" de una función g (o bien g una "extensión" de f) si dondequiera que f está definida, g también está definida y toma los mismos valores. Por supuesto, la misma función f puede provenir por restricción de muchas funciones diferentes. En nuestro ejemplo anterior f es también una restricción de la función h definida por $h(x) = x^2$ para $0 < x < 2$, $h(x) = -x^2$ para $-2 < x \leq 0$. En efecto, este ejemplo ilustra el proceso inverso al de formar restricciones de una función, el cual puede denominarse "unión de secciones". Se pueden generar nuevas funciones definiéndolas simplemente por expresiones explícitas diferentes en distintas partes del dominio.

c. Representación gráfica. Funciones monótonas

La idea fundamental de la geometría analítica es dar una representación analítica a una curva definida originalmente por alguna propiedad geométrica. Esto se hace usualmente considerando una de las coordenadas rectangulares, digamos y, como una función $y = f(x)$ de la otra coordenada, x; por ejemplo, una parábola se representa por la función $y = x^2$; el círculo con radio 1 en torno del origen, por las dos funciones $y = \sqrt{1 - x^2}$ y $y = -\sqrt{1 - x^2}$. En el primer ejemplo puede pensarse de la función como definida en el intervalo infinito $-\infty < x < \infty$; en el segundo debemos restringirnos al intervalo $-1 \leq x \leq 1$, puesto que fuera de este intervalo la función no tiene sentido.[1]

Inversamente, si, en lugar de comenzar con una curva definida geométricamente, consideramos una función $y = f(x)$ dada analíticamente, podemos representar la dependencia funcional de y respecto de x gráficamente utilizando un sistema coordenado rectangular en la manera usual (ver Fig. 1.7). Si para cada abscisa x se toma la ordenada correspondiente $y = f(x)$, se obtiene la representación geométrica de la función. Las restricciones que han de imponerse al concepto de función deberán asegurar para su representación geométrica la forma de una curva geométrica "razonable". Esto, ciertamente, expresa una concepción intuitiva en lugar de una condición matemática estricta. Sin embargo, pronto se formularán condiciones, tales como continuidad, diferenciabilidad, etc., las cuales aseguran que la gráfica de una función es una curva capaz de ser visualizada geométricamente. Este no sería el caso si admitiésemos funciones "patológicas" tales como la siguiente: para todo valor racional de x la función y tiene el valor 1, y para

[1] Ordinariamente no se consideran valores imaginarios o complejos de x e y.

todo valor irracional de x el valor de y es 0.* Esta prescripción funcional asigna un valor definido de y a cada x; pero en todo intervalo de x, no importa cuán pequeño, el valor de y da saltos de 0 a 1 y de 1 a 0 una infinidad de veces. Este ejemplo demuestra que el concepto general no restringido de función da lugar a gráficas que no podríamos considerar como curvas.

Funciones multiformes

Consideraremos solamente funciones $y = f(x)$ que asignan un valor único de y a cada valor de x en el dominio, como, por ejemplo, $y = x^2$, o bien $y = \operatorname{sen} x$. Sin embargo, para una curva descrita geométricamente puede suceder lo que para el círculo $x^2 + y^2 = 1$: que una descripción completa de la curva no esté dada precisamente por una sola función (uniforme), sino que requiera varias funciones; para el caso del círculo, las dos funciones $y = \sqrt{1 - x^2}$ y $y = -\sqrt{1 - x^2}$. Lo mismo es cierto para la hipérbola $y^2 - x^2 = 1$, la cual está representada por las dos funciones $y = \sqrt{1 + x^2}$ e $y = -\sqrt{1 + x^2}$. Por eso, tales curvas no determinan sin ambigüedad las correspondientes funciones. Con frecuencia se dice que la curva está representada por una función *multiforme;* las funciones separadas que la representan se denominan entonces *ramas uniformes* de la función multiforme que representa la curva. Para mayor claridad se utilizará siempre la palabra "función" para indicar una función univaluada. Por ejemplo, el símbolo $\sqrt{x}$ (para $x \geq 0$) denotará siempre el número *no negativo* cuyo cuadrado es x.

Si una curva es la gráfica de *una* función, ésta es intersectada por cualquier paralela al eje y en a lo más un punto, puesto que a cada punto x en el intervalo de definición le corresponde justamente un valor de y. El círculo unitario representado por las dos funciones

$$y = \sqrt{1 - x^2} \qquad \text{e} \qquad y = -\sqrt{1 - x^2},$$

es intersectado por tales paralelas al eje y en más de un punto. Las porciones de una curva correspondientes a diferentes ramas uniformes son conectadas en ocasiones una con otra, de manera que la curva completa es una sola figura que puede ser dibujada con un trazo de la pluma; por ejemplo, el círculo (ver Fig. 1.12). Por otra parte, estas porciones pueden ser separadas completamente, como para la hipérbola (ver Fig. 1.13).

Ejemplos. Consideremos algunos ejemplos adicionales sobre la representación gráfica de funciones.

* (N. del T.) Esta es la llamada "función de Dirichlet".

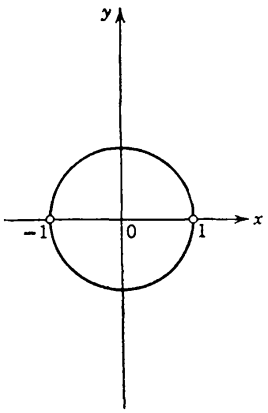


Figura 1.12.

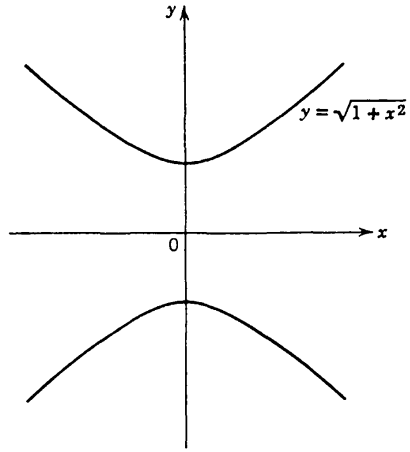


Figura 1.13.

a) y es proporcional a x,

$$y = ax.$$

La gráfica (ver Fig. 1.14) es una línea recta que pasa por el origen del sistema coordenado.

b) y es una "función lineal" de x,

$$y = ax + b.$$

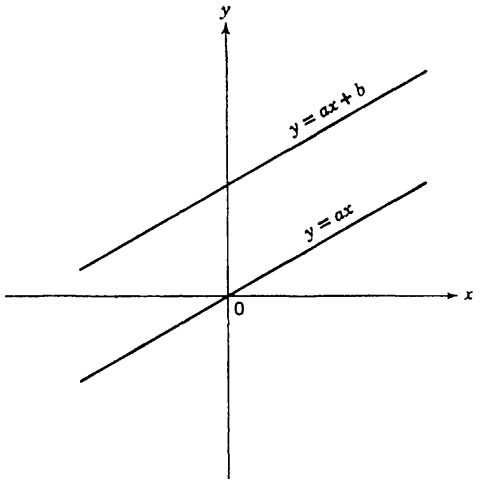


Figura 1.14 Funciones lineales.

La gráfica es una línea recta que pasa por el punto $x = 0$, $y = b$; la cual, si $a \neq 0$, pasa también por el punto $x = -b/a$, $y = 0$; y si $a = 0$ es una recta horizontal.

c) y es inversamente proporcional a x:

$$y = \frac{a}{x}.$$

En particular, para $a = 1$ es

$$y = \frac{1}{x},$$

de modo que

$y = 1$ para $x = 1$, $y = 2$ para $x = \frac{1}{2}$, $y = \frac{1}{2}$ para $x = 2$.

La gráfica (ver Fig. 1.15) es una *hipérbola rectangular,* una curva simétrica con respecto a las bisectrices de los ángulos entre los ejes coordenados.

Esta función obviamente no está definida para el valor $x = 0$, puesto que la división por cero no tiene significado. En la vecindad del punto excepcional $x = 0$ la función tiene valores arbitrariamente grandes, tanto positivos como negativos; este es el ejemplo más simple de una *discontinuidad infinita,* un concepto que se discutirá posteriormente (ver p. 59).

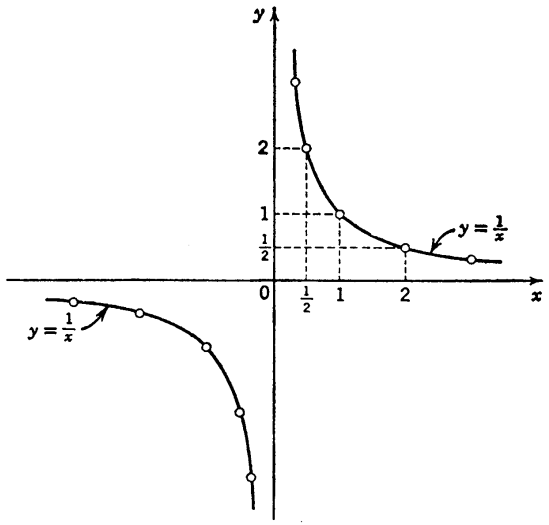


Figura 1.15 Discontinuidad infinita.

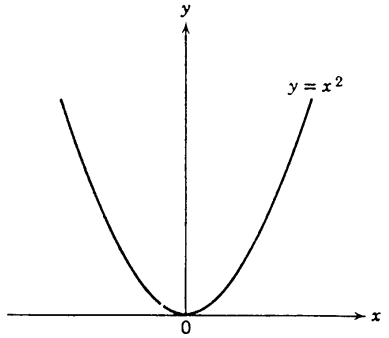


Figura 1.16 Parábola.

d) y es el cuadrado de x:

$$y = x^2.$$

Como es bien sabido, esta función es representada por una parábola (ver Fig. 1.16).

Análogamente, la función $y = x^3$ está representada por la así llamada *parábola cúbica* (ver Fig. 1.17).

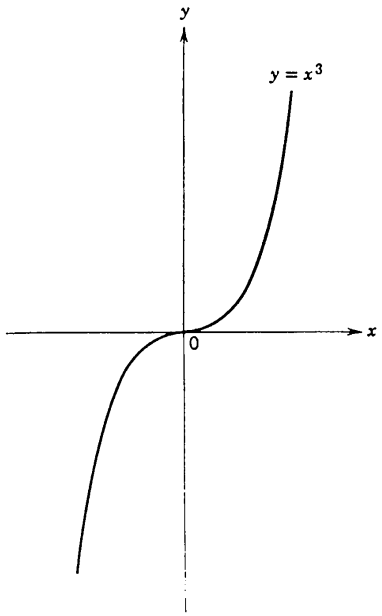


Figura 1 17 Parábola cúbica.

Funciones monótonas

Una función que para todos los valores de x en un intervalo tiene el mismo valor $y = a$ se denomina *función constante,* o simplemente *constante;* está representada gráficamente por una línea recta horizontal. Una función $y = f(x)$ para la cual un incremento en el valor de x resulta siempre en un incremento en el valor de y, esto es, para la cual $f(x) < f(x')$ siempre que $x < x'$, se denomina función *monótona creciente;* si, por otra parte, un incremento en el valor de x implica siempre un decremento en el valor de y, la función se denomina función *monótona decreciente.* Tales funciones son representadas gráficamente por curvas que siempre ascienden o bien siempre descienden conforme x recorre el intervalo de definición hacia valores crecientes (ver Fig. 1.18). Una función monótona transforma siempre valores distintos de x en diferentes y; esto es, la transformación es biunívoca o "uno a uno".

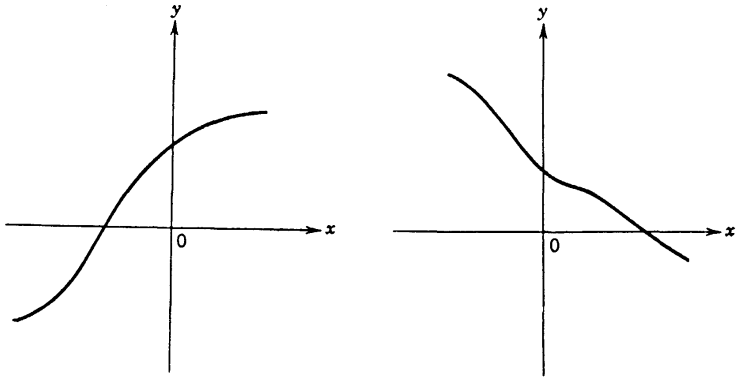


Figura 1.18 Funciones monótonas.

Funciones pares y funciones impares

Si la curva representada por $y = f(x)$ es simétrica con respecto al eje y, esto es, si $x = -a$ y $x = a$ dan el mismo valor de la función,

$$f(-x) = f(x),$$

la función se denomina función *par.* Por ejemplo, la función $y = x^2$ es par (ver Fig. 1.16). Si, por otra parte, la curva es simétrica con respecto al origen, esto es, si

$$f(-x) = -f(x),$$

se dice que la función es una función *impar;* entonces las funciones $y = x$, $y = x^3$ (ver Fig. 1.17), e $y = 1/x$ (ver Fig. 1.15) son impares.

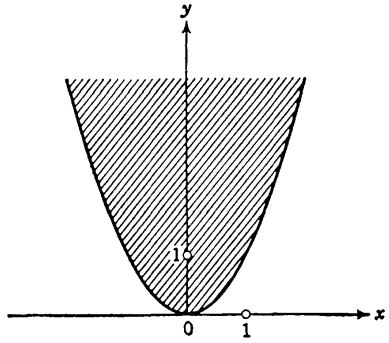


Figura 1.19 Gráfica de $y > x^2$.

Frecuentemente es útil considerar la representación geométrica de una desigualdad. Por ejemplo, la desigualdad $y > x^2$ está representada por el dominio sobre la parábola $y = x^2$ (Fig. 1.19). El interior del círculo unitario con centro en el origen (Fig. 1.20) se expresa por la desigualdad $x^2 + y^2 < 1$.

Frecuentemente, diversas desigualdades describen regiones más complicadas cuyas fronteras consisten de diferentes secciones. Así, el "primer" cuadrante del círculo unitario se describe por el sistema de desigualdades simultáneas:

$$x^2 + y^2 < 1, \qquad x > 0, \qquad y > 0.$$

(Ver Fig. 1.21.)

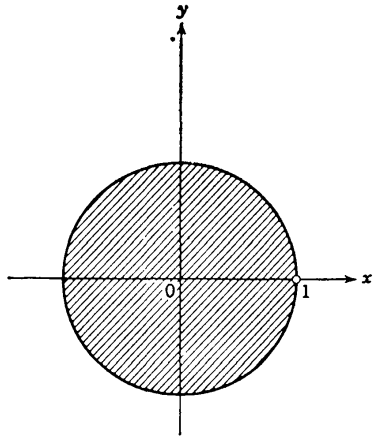


Figura 1.20 Gráfica de $x^2 + y^2 < 1$.

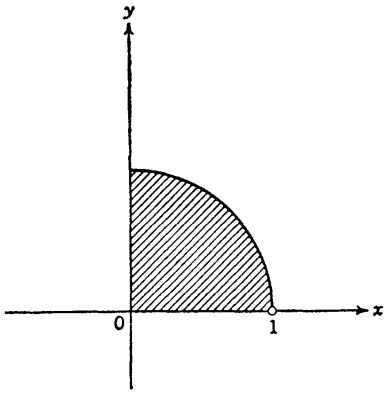


Figura 1.21 Gráfica de $x^2 + y^2 < 1$, $x > 0$, $y > 0$.

d. Continuidad

Explicaciones intuitiva y precisa

Las funciones y gráficas que se acaban de considerar exhiben una propiedad de gran importancia en el cálculo, a saber, la *continuidad*. Intuitivamente, la continuidad significa que un pequeño cambio en la variable independiente x implica sólo un pequeño cambio en la variable dependiente $y = f(x)$ y excluye saltos en el valor de y; así, la gráfica consiste de un solo pedazo de curva. En contraste, una gráfica $y = f(x)$ que consiste de pedazos de curva separados por un vacío en una abscisa x_0 exhibe allí una *discontinuidad de salto*. Por ejemplo, la función [1] $f(x) = \operatorname{sgn} x$, definida por $f(x) = +1$ para $x > 0$, por $f(x) = -1$ para $x < 0$ y $f(0) = 0$ tiene una "discontinuidad de salto" [2] en $x_0 = 0$ (ver Fig. 1.22).

La idea de la continuidad está implícita en el uso diario de las matemáticas elementales. Siempre que una función $y = f(x)$ esté descripta por tablas, tales como las logarítmicas o trigonométricas, el valor y puede ser registrado solamente para un conjunto "discreto" de valores de la variable independiente x; por ejemplo, a intervalos de 1/1000 o bien 1/100,000. Sin embargo, valores no registrados de la función pueden necesitarse para x intermedios. Entonces, tácitamente se supone que un valor no registrado $f(x_0)$ es aproximadamente el mismo que el de $f(x)$ para una x vecina que aparece en la tabla, y que $f(x_0)$ puede ser aproximado precisamente como

[1] Se pronuncia "signum" o bien "signo" de x.

[2] Técnicamente, la palabra "salto" se refiere solamente a la clase particular de discontinuidad en la cual la función aproxima desde la derecha e izquierda valores que no coinciden simultáneamente con $f(x_0)$. Una discontinuidad "infinita" es exhibida por la función $y = 1/x$ para $x \neq 0$ y $y = 0$ para $x = 0$. Aún otros tipos de discontinuidades se discutirán posteriormente.

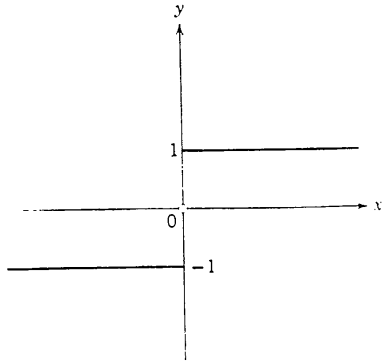


Figura 1.22 La función $f(x) = \operatorname{sgn} x$.

se quiere si sólo los valores de x en la tabla se espacian suficientemente cerca uno con otro.

La continuidad de la función $f(x)$ para un valor x_0 significa justamente que $f(x)$ difiere arbitrariamente poco del valor $f(x_0)$ cuando x está suficientemente cerca de x_0. Las expresiones "difiere arbitrariamente poco" y "suficientemente cerca" son un tanto vagas y deben ser explicadas precisamente en términos cuantitativos.

Prescríbase cualquier "margen de precisión" o "tolerancia", esto es, cualquier número real positivo ε (por pequeño que sea). Para la continuidad de f en x_0 se requiere que la diferencia entre $f(x)$ y $f(x_0)$ permanezca dentro de este margen, esto es, que $|f(x) - f(x_0)| < \varepsilon$, para todos los valores de x que se encuentran suficientemente cercanos a x_0 (o bien, para todos los valores x situados alguna distancia δ desde x_0).

Es más fácil visualizar lo que la continuidad significa si se interpreta a f como una transformación que asigna a puntos x en el eje x imágenes en

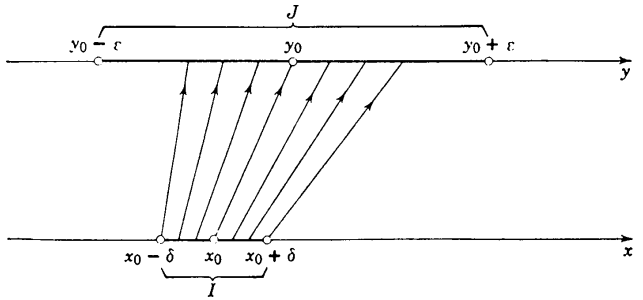


Figura 1.23 Continuidad del mapeo $y = f(x)$ en el punto x_0.

el eje y. Tómese cualquier punto x_0 en el eje x y su imagen $y_0 = f(x_0)$ (ver Fig. 1.23). Marcamos un intervalo abierto arbitrario J en el eje y, que tiene al punto y_0 como centro. Si la longitud de J es 2ε, entonces los puntos y de J son aquéllos cuya distancia a y_0 es menor que ε, o sea, para los cuales $|y - y_0| < \varepsilon$. La condición para la continuidad de $f(x)$ en x_0 es: Todos los puntos x suficientemente cercanos a x_0 tienen imágenes que se encuentran en J. O bien: Es posible marcar un intervalo I en el eje x con centro x_0, por ejemplo, el intervalo $x_0 - \delta < x < x_0 + \delta$, tal que todo punto x de I tiene una imagen $f(x)$ que se encuentra en J y por lo tanto $|f(x) - f(x_0)| < \varepsilon$. La continuidad de $f(x)$ en el punto x_0 significa que para una ε-*vecindad* arbitraria J del punto $y_0 = f(x_0)$ en el eje y, puede encontrarse una δ-vecindad I del punto x_0 en el eje x, cuya totalidad de puntos son transformados en puntos de J.[1] Por supuesto, esto solamente tiene sentido para puntos en el eje x en los cuales la función está definida, esto es, los que pertenecen al *dominio* de f. Lo anterior nos conduce a la siguiente *definición precisa de continuidad.*

La función $f(x)$ es continua en un punto x_0 de su dominio si para todo ε positivo puede encontrarse un número positivo δ tal que

$$|f(x) - f(x_0)| < \varepsilon$$

para todos los valores x en el dominio de f para los cuales $|x - x_0| < \delta$.

Muy útil es la interpretación geométrica de la continuidad cuando la función f se representa por su gráfica en el plano xy (ver Fig. 1.24). Sea $P_0 = (x_0, y_0)$ un punto sobre la gráfica. Los puntos (x, y) con $y_0 - \varepsilon < y < y_0 + \varepsilon$ forman ahora una "franja" horizontal J que contiene a P_0. La continuidad de f en x_0 significa que dada cualquiera de tales franjas horizontales J, no importa cuán angosta, es posible encontrar una franja vertical I dada por $x_0 - \delta < x < x_0 + \delta$ tan angosta que todo punto de la gráfica que se encuentre en I también caiga en J.

Como ilustración, considérese la función lineal $f(x) = 5x + 3$. Se tiene

$$|f(x) - f(x_0)| = |(5x + 3) - (5x_0 + 3)| = 5\,|x - x_0|,$$

la cual expresa que la transformación $y = 5x + 3$ *magnifica* distancias por el factor 5. Obviamente aquí $|f(x) - f(x_0)| < \varepsilon$ para todos los x para los

[1] En esta definición de continuidad I y J son intervalos que tienen sus *centros* en los puntos x_0 e y_0, respectivamente. Esto es conveniente para la definición analítica de continuidad en x_0 la cual recurre a las distancias $|x - x_0|$ e $|y - y_0|$; pero es algo artificial si se interpreta f geométricamente como un "mapeo" o transformación. En su lugar, se puede definir igualmente bien la continuidad de $y = f(x)$ en un punto x_0 por el requisito de que para todo intervalo *abierto* J en el eje y que contiene al punto $y_0 = f(x_0)$ puede encontrarse un intervalo abierto I en el eje x conteniendo al punto x_0 y tal que la imagen y de cualquier punto x en I para el cual la función esté definida está situada en J. La demostración de la equivalencia de las dos definiciones se deja al lector como un sencillo ejercicio.

cuales $|x - x_0| < \varepsilon/5$. Consecuentemente, la condición para la continuidad de $f(x)$ en el punto x_0 se satisface si se elige $\delta = \varepsilon/5$ (pero, por supuesto, cualquier número positivo $\delta < \varepsilon/5$ es también una posible elección); y la imagen de cualquier punto del intervalo $x_0 - \delta < x < x_0 + \delta$ se encontrará entonces en el intervalo $y_0 - \varepsilon < y < y_0 + \varepsilon$. En este ejemplo, a la aseveración de que la distancia $|y - y_0|$ es "arbitrariamente pequeña" para $|x - x_0|$ "suficientemente pequeño" puede dársele un significado específico en efecto, $|x - x_0|$ es *suficientemente pequeño* si no excede un quinto del valor de $|y - y_0|$.

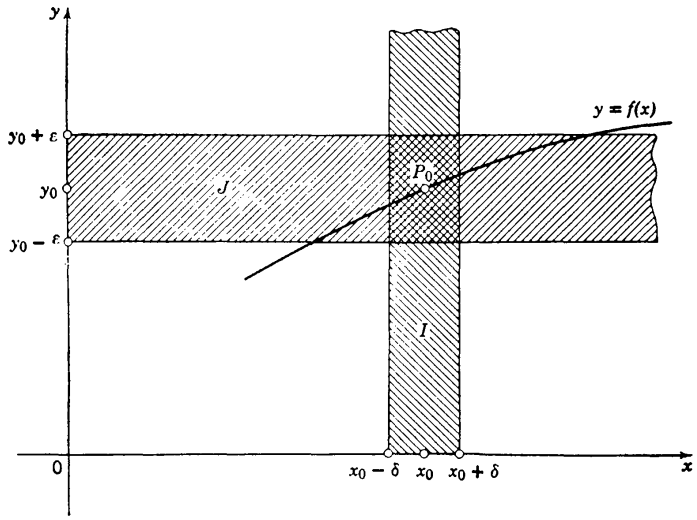


Figura 1.24 Continuidad de $y = f(x)$ en el punto x_0.

Otro ejemplo es proporcionado por la función $f(x) = x^2$. Aquí se tiene para $|x - x_0| < \delta$ que

$$|f(x) - f(x_0)| = |x^2 - x_0^2| = |x - x_0|\,|2x_0 + (x - x_0)|$$
$$\leq |x - x_0|(2\,|x_0| + |x - x_0|) \leq \delta(2\,|x_0| + \delta).$$

Se verifica inmediatamente que la condición $|f(x) - f(x_0)| < \varepsilon$ se satisface si se elige $\delta = -|x_0| + \sqrt{\varepsilon + |x_0|^2}$.

Intuitivamente, la idea de la continuidad parece obvia sin explicación, pero la formulación precisa puede ser en un principio algo difícil de entender debido a la vaguedad de palabras tales como "uno puede encontrar" o bien "elegido arbitrariamente". Sin embargo, el lector, que puede estar al principio muy satisfecho con alguna noción intuitiva de la continuidad, aprenderá gradualmente a apreciar la lógica precisión y generalidad de la

definición analítica, resultado de un largo y persistente esfuerzo para conciliar la necesidad del entendimiento intuitivo con el de la claridad lógica. Es indispensable un significado preciso para la palabra "continuidad". La definición analítica dada aquí es la formulación requerida y precisa de una importante propiedad de las funciones.

Para el principiante se deberá nuevamente insistir en que "pequeño" no es una designación absoluta de un número; más bien, el término "arbitrariamente pequeño" se refiere a un número que no es fijo al principio, sino que, cualquier valor positivo ε puede ser seleccionado, el cual está sujeto a una elección subsecuente menor para una aproximación refinada de $f(x_0)$. "Suficientemente pequeño" se refiere a un número δ que debe ser ajustado para satisfacer un margen de tolerancia establecido anteriormente por otro número ε.

Explicación de la continuidad y de la discontinuidad mediante ejemplos. Podemos aclarar la definición de continuidad haciendo contraste con ejemplos de discontinuidad, ejemplos que no satisfacen la definición anterior. Recuérdese el ejemplo sencillo de la función $f(x) = \operatorname{sgn} x$ en la p. 55. Obviamente, para todo $x_0 \neq 0$ esta función es continua de acuerdo con la definición en términos de ε y δ; de hecho, con un $\delta = |x_0|$ constante, no importa cuán pequeño se elija ε. Pero para $x_0 = 0$ no puede encontrarse ninguna δ si ε es menor que 1, puesto que $|f(x) - f(0)| = |f(x)| = 1 > \varepsilon$ para todo x distinto de cero, sin importar cuán próximo está x de cero.

La función $\operatorname{sgn} x$ ilustra el tipo más simple de discontinuidad en un punto ξ, conocida como *discontinuidad de salto,* en la cual $f(x)$ se aproxima a valores límite desde la derecha e izquierda, conforme x se aproxima a valores límite ξ, que difieren, ya sea el uno del otro o bien del valor de f en el punto ξ.[1] La gráfica en $x = \xi$ tiene entonces un vacío. Otras curvas con discontinuidades de salto están bosquejadas en la Fig. 1.25 *a* y *b*; la definición de estas funciones deberá resultar evidente de las figuras.[2]

En discontinuidades de este género existen tanto los límites por la derecha como por la izquierda. Volvamos ahora a discontinuidades en las cuales éste no es el caso. Las más importantes de éstas son las *discontinuidades infinitas* o infinidades. Estas son discontinuidades análogas a las exhibidas

[1] La definición precisa de *límite* se dará en la Sección 1.7; para los comentarios descriptivos que se hacen aquí, una idea intuitiva es suficiente.

[2] En todos estos ejemplos de discontinuidades de salto, los límites de la función por la derecha y por la izquierda en el punto de discontinuidad poseen valores diferentes. El ejemplo trivial de la función $f(x)$ definida por

$$f(x) = 0 \quad \text{para} \quad x \neq 0, \qquad f(x) = 1 \quad \text{para} \quad x = 0$$

ilustra una discontinuidad de salto en la cual los límites desde ambos lados son iguales entre sí pero difieren del valor de f en el punto mismo ξ de discontinuidad. Se tiene entonces una singularidad *eliminable.* Aquí f puede hacerse continua cambiando solamente el valor de f en ξ de modo que coincida con los límites desde ambos lados.

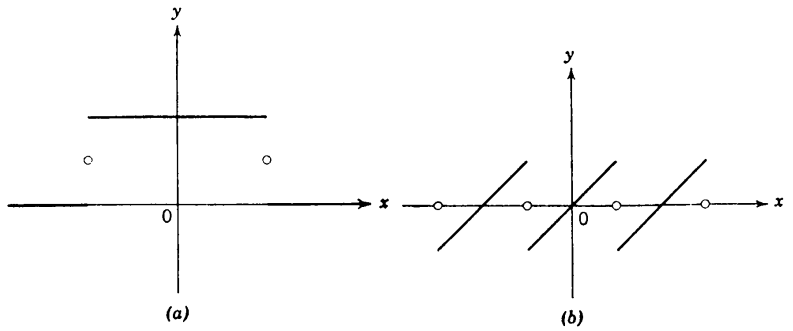


Figura 1.25.

por las funciones $1/x$ y $1/x^2$ en el punto $x = 0$; conforme $x \to 0$, el valor absoluto $|f(x)|$ de la función crece más allá de toda cota. La función $1/x$ crece numéricamente más allá de toda cota a través de valores positivos y negativos respectivamente, conforme x se aproxima al origen desde la izquierda y desde la derecha. Por otra parte, la función $1/x^2$ tiene para $x = 0$ una discontinuidad infinita en la cual el valor de la función crece

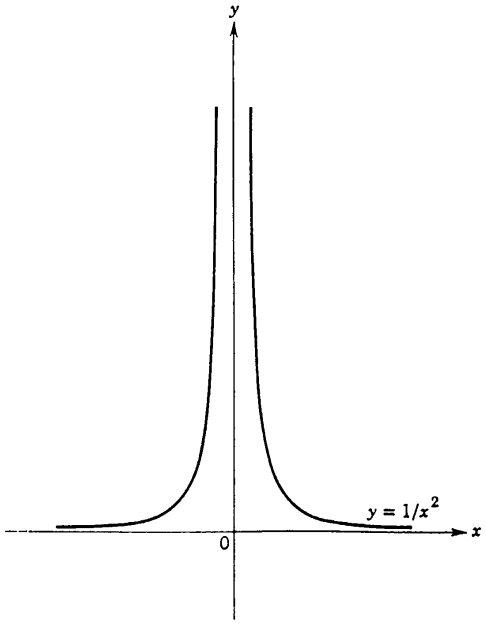


Figura 1.26 Gráfica de función con discontinuidad infinita.

más allá de cualquier cota positiva conforme x se aproxima al origen desde ambos lados (ver Fig. 1.26 y Fig. 1.27). La función $1/(x^2 - 1)$ mostrada en la Fig. 1.27 posee discontinuidades infinitas tanto en $x = 1$ cocmo en $x = -1$.

Un ejemplo de otro tipo de discontinuidad en el cual no existe ningún límite desde la derecha o desde la izquierda es la función par "lineal por pedazos", $y = f(x)$, ilustrada en la Fig. 1.28, la cual está definida para todos los valores de x distintos de cero como sigue. Esta función toma alternativamente los valores $+1$ y -1 para los valores $-x$ de la forma $\pm 1/2^n$,

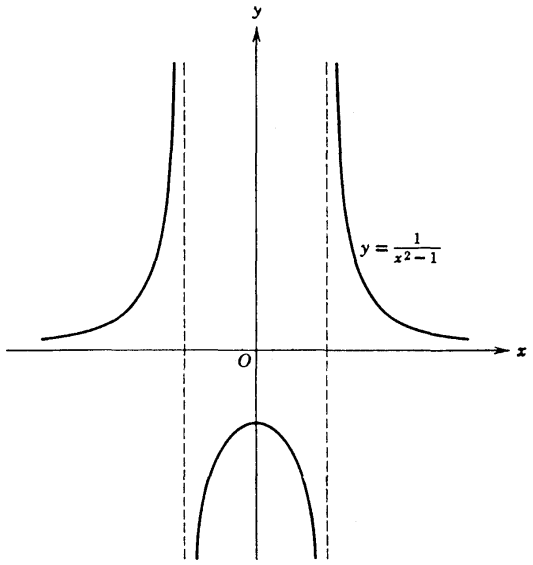


Figura 1.27 Función con discontinuidades infinitas.

donde n es cualquier entero: $f(\pm 1/2^n) = (-1)^n$. En todo intervalo $1/2^{n+1} < x < 1/2^n$, o bien $-1/2^n < x < -1/2^{n+1}$, la función $f(x)$ es lineal y recorre todos los valores entre -1 y $+1$. Por ello la función oscila hacia arriba y hacia abajo cada vez con más rapidez entre los valores -1 y $+1$ conforme x se aproxima más y más al punto $x = 0$, y en la vecindad inmediata de ese punto ocurren un número infinito de tales oscilaciones. Un comportamiento análogo es exhibido por la curva suave de la Fig. 1.29. [Aquí, de hecho $f(x)$ está dada por una expresión en forma exacta, a saber, $f(x) = \text{sen}\ (1/x)$, con la función seno definida apropiadamente como en la p. 74.]

Un contraste con este ejemplo es la función lineal por pedazos $y = f(x)$ que toma los valores $f(\pm 1/2^n) = (-\frac{1}{2})^n$ para todos los enteros n (ver

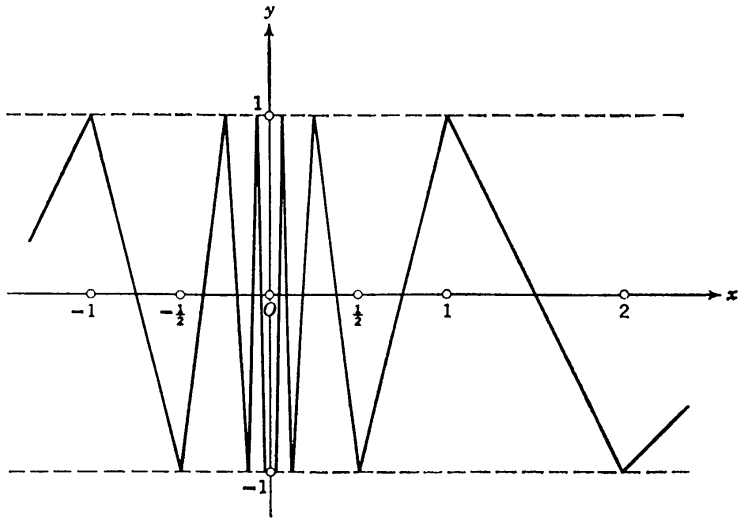


Figura 1.28 Función oscilante, lineal por pedazos, con discontinuidad.

Fig. 1.30) y es lineal para valores intermedios x. Aquí $f(x)$ permanece continua en el punto $x = 0$ si le asignamos el valor 0 en este punto. En la vecindad del origen la función oscila hacia arriba y hacia abajo un número infinito de veces, pero la magnitud de estas oscilaciones se hace arbitra-

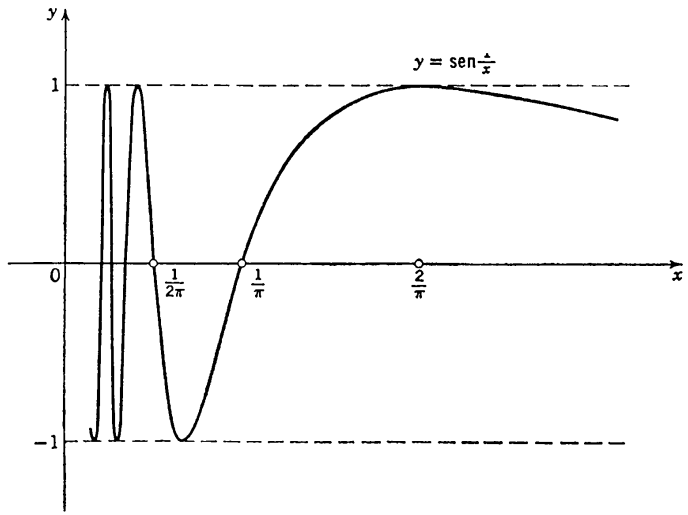


Figura 1.29 Función oscilante con discontinuidad.

riamente pequeña conforme se aproximan al origen. La situación es la misma para la función $y = x \operatorname{sen}(1/x)$ (ver Fig. 1.31).

Estos ejemplos muestran que la continuidad permite toda clase de posibilidades notables ajenas a nuestra cándida intuición.

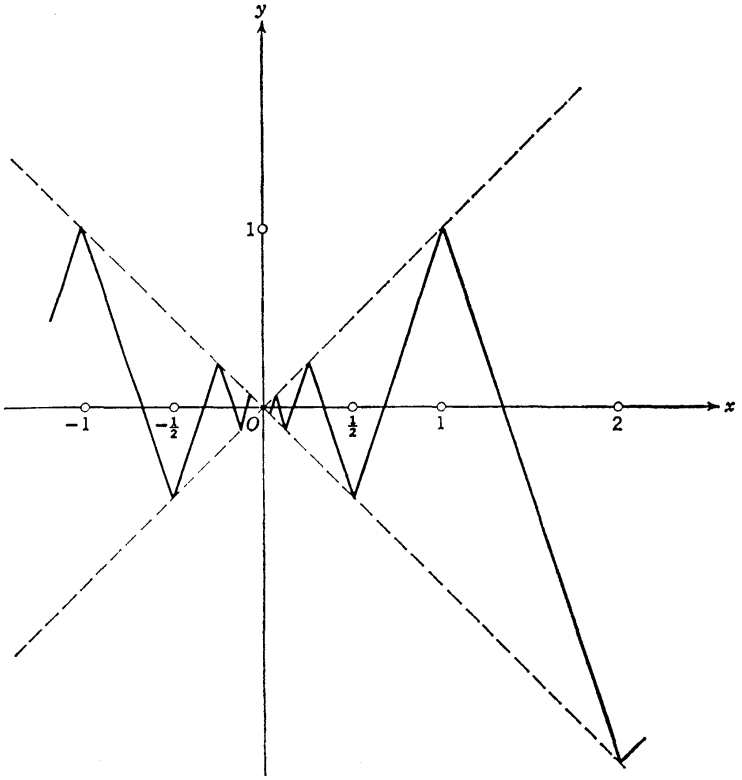


Figura 1.30 Función continua oscilante.

**Discontinuidades removibles*

Como se habrá notado, puede suceder que en un cierto punto, digamos $x = 0$, una función no esté definida por la ley original, tal como en los últimos ejemplos discutidos. Estamos entonces en libertad de extender la definición de la función, asignándole cualquier valor deseado en tal punto. En el último ejemplo podemos escoger la definición de tal manera que la función se torne también *continua* en ese punto, a saber, eligiendo $y = 0$ en $x = 0$. Una extensión continua análoga puede ser definida siempre que el límite por la izquierda así como el límite por la derecha existan

y sean iguales; bastará entonces con hacer el valor de la función en el punto en cuestión igual a estos límites de manera de hacer continua la función allí. Cualquiera que sea la discontinuidad que pueda ser impuesta por definición en $x = 0$, esta discontinuidad es "eliminable" asignando un valor adecuado a $f(0)$. Para la función $y = \operatorname{sen} 1/x$, como para la función de la Fig. 1.28, esto, sin embargo, no es posible: cualquiera que sea el

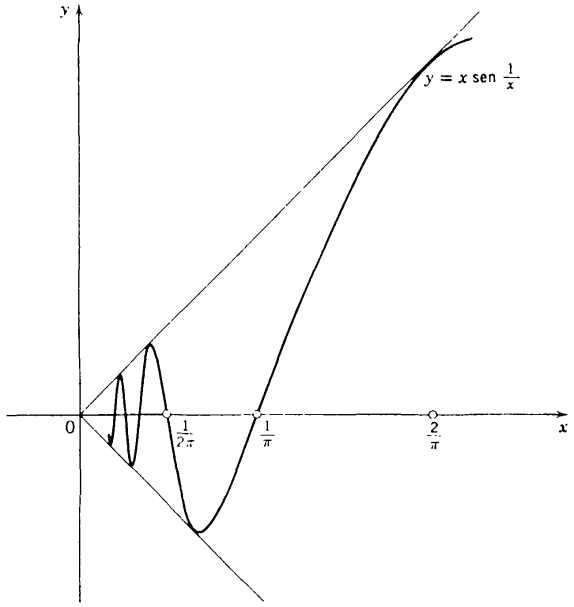


Figura 1.31 Función continua oscilante.

valor que se asigne a la función en $x = 0$, la función extendida es discontinua.

Módulo de continuidad. Continuidad uniforme. Nuestra definición de continuidad de la función $f(x)$ en x_0 requiere que para toda medida de precisión, $\varepsilon > 0$, *existan* cantidades $\delta > 0$ (llamadas módulos de continuidad) tales que $|f(x) - f(x_0)| < \varepsilon$ para todos los valores de x en el dominio de f para los cuales $|x - x_0| < \delta$. Un módulo de continuidad proporciona información sobre la sensibilidad de f a los cambios en x. Un módulo de continuidad δ nunca es único; puede ser siempre reemplazado (para las mismas x_0 y ε) por cualquier cantidad positiva menor δ', puesto que $|x - x_0| < \delta'$ implica $|x - x_0| < \delta$ y por ello $|f(x) - f(x_0)| < \varepsilon$. Para fines prácticos, como en los cálculos numéricos, podemos estar interesados

en una elección particular de δ; por ejemplo, en el máximo valor para δ. Por otra parte, si meramente deseamos establecer el hecho de que f es continua en x_0, entonces solamente se requiere exhibir un módulo cualquiera de continuidad para todo ε positivo.

En general, como lo muestran los ejemplos, este $\delta = \delta(\varepsilon)$ depende no sólo de ε sino que también del valor de x_0. Por supuesto, no se requiere considerar todos los valores positivos posibles de ε. Pueden restringirse siempre las consideraciones a ε suficientemente pequeños; digamos, a $\varepsilon \leq \varepsilon_0$ para un ε_0 elegido arbitrariamente, puesto que para $\varepsilon > \varepsilon_0$ puede utilizarse el mismo módulo de continuidad que para $\varepsilon = \varepsilon_0$. Análogamente, debemos tomar en cuenta solamente los puntos x en el dominio de f situados en una vecindad arbitraria de x_0; digamos, aquellos puntos con $|x - x_0| < \delta_0$, puesto que puede siempre reemplazarse cualquier módulo de continuidad δ por uno menor que no exceda a δ_0. La continuidad de f en x_0 es una *propiedad local,* es decir, una propiedad que depende solamente de los valores de f en alguna vecindad de x_0 arbitrariamente pequeña.

Como se ha visto, la función f puede ser continua para algún x_0 y discontinua para otros x. Una función se denomina *continua en un intervalo* si es continua en cada punto del intervalo. Para cada x_0 del intervalo se tiene entonces un módulo de continuidad $\delta = \delta(\varepsilon)$ el cual puede esperarse que varíe con x_0, reflejando las diferentes proporciones según las cuales y varía al variar x cerca de puntos x_0 diferentes.

Se dice que f es *uniformemente continua* en un intervalo si puede encontrarse un *módulo uniforme de continuidad* $\delta = \delta(\varepsilon)$ para ese intervalo, esto es, uno que no dependa del punto particular x_0 del intervalo. Por lo tanto, $f(x)$ es uniformemente continua en un intervalo [1] si para cada ε positivo existe un número positivo δ tal que $|f(x) - f(x_0)| < \varepsilon$ para *dos* puntos *cualesquiera* x y x_0 del intervalo para los cuales $|x - x_0| < \delta$.

Para una función uniformemente continua $y = f(x)$ los valores de y difieren "arbitrariamente poco" uno del otro para cualesquiera valores x que estén "suficientemente cercanos", *independientemente de su posición en el intervalo.* En ciertos aspectos la continuidad uniforme se acerca más a nociones intuitivas que la mera propiedad local de continuidad.

Por ejemplo, la función $f(x) = 5x + 3$ es uniformemente continua para todos los valores de la variable independiente puesto que aquí $|f(x) - f(x_0)| = 5|x - x_0| < \varepsilon$ para $|x - x_0| < \varepsilon/5$, y, por lo tanto, $\delta(\varepsilon) = \varepsilon/5$ representa un módulo uniforme de continuidad.

La función $f(x) = x^2$ para un intervalo x infinito, ciertamente no es uniformemente continua. Es evidente que pequeñas variaciones en x pueden producir cambios arbitrariamente grandes en x^2, bastando con que x sea suficientemente grande. Una ojeada a una tabla de cuadrados de enteros

[1] En esta definición la palabra *intervalo* puede referirse ya sea a intervalos cerrados o abiertos o bien infinitos.

x muestra cómo los cuadrados sucesivos se encuentran espaciados más y más conforme x crece. Pero, sin embargo, si se consideran solamente pares de valores x y x_0 pertenecientes a un intervalo fijo finito cerrado $[a, b]$, puede encontrarse un módulo uniforme de continuidad. En efecto, para $|x - x_0| < \delta$ se tiene

$$|f(x) - f(x_0)| = |x^2 - x_0^2| = |x - x_0|\,|x + x_0| \leq 2|x - x_0|(|b| + |a|)$$
$$< 2\delta(|b| + |a|) = \varepsilon$$

si se toma $\delta = \varepsilon/2(|b| + |a|)$.

Prevalece una situación análoga para la función $f(x) = 1/x$ para $x \neq 0$, con $f(0) = 0$. Considérese un intervalo cerrado acotado $a \leq x \leq b$ en todo el cual la función es continua. Tal intervalo no puede incluir al origen, el cual es un punto de discontinuidad, de modo que a y b deberán tener el mismo signo. Supóngase que a y b son ambos positivos. Entonces, para x y x_0 pertenecientes al intervalo y para $|x - x_0| < \delta$ se tiene

$$\left|\frac{1}{x} - \frac{1}{x_0}\right| = |x_0 - x|\,\frac{1}{|x_0|\,|x|} < \frac{\delta}{a^2} = \varepsilon$$

para $\delta = a^2\varepsilon$. Por lo tanto, la función es uniformemente continua en el intervalo $[a, b]$. Por supuesto, esto prueba también que la función $f(x) = 1/x$ es continua en todo punto $x_0 > 0$. Puesto que todo tal valor x_0 puede ser incluido en algún intervalo $a < x_0 < b$ con a y b positivos. La expresión $\delta = a^2\varepsilon$ es entonces un módulo de continuidad para la función en x_0, si se restringe x a una vecindad de x_0 que se encuentre totalmente en el intervalo.

Salta a la vista que las funciones continuas en los ejemplos anteriores son uniformemente continuas en cualquier intervalo cerrado acotado que pertenece a su dominio. Estas ilustran una propiedad general que será demostrada en el Suplemento, p. 123. Podemos enunciarla:

Toda función continua en un intervalo cerrado y acotado es por necesidad uniformemente continua en ese intervalo.

La restricción a intervalos *acotados* es esencial como lo muestra el ejemplo de la función x^2. Análogamente, se deberá estipular que el intervalo es *cerrado.* Por ejemplo, la función $y = 1/x$ es continua en el intervalo abierto $0 < x < 1$, pero no es uniformemente continua allí; pues cambios arbitrariamente grandes en y pueden ser producidos por cambios arbitrariamente pequeños en x, bastando sólo con que x se encuentre suficientemente cercano al origen. Si existiera un módulo uniforme de continuidad $\delta(\varepsilon)$ para el intervalo $(0, 1)$, podría tomarse, por ejemplo, $x_0 < \delta$, $x = \frac{1}{2}x_0$; obviamente entonces $|1/x - 1/x_0| = 1/x_0$ es mayor que cualquier ε preasignado siempre que x_0 sea suficientemente pequeño, de modo que la suposición de un $\delta(\varepsilon)$ uniforme conduce a una contradicción.

Continuidad de Lipschitz—Continuidad de Hölder. En los ejemplos anteriores de funciones uniformemente continuas en un intervalo $[a, b]$, se encontró un módulo de continuidad particularmente simple, a saber, $\delta(\varepsilon)$ proporcional a ε. Esta situación muy común se presenta en las así llamadas funciones *"continuas de Lipschitz"*,* esto es, las funciones $f(x)$ que satisfacen una desigualdad de la forma

$$|f(x_2) - f(x_1)| \leq L\,|x_2 - x_1|$$

(la llamada *condición de Lipschitz)* para todos los x_1 y x_2 en el intervalo, con un valor fijo L. La continuidad de Lipschitz significa que el "cociente incremental"

$$\frac{f(x_2) - f(x_1)}{x_2 - x_1}$$

formado para dos puntos distintos cualesquiera del intervalo nunca excede un valor finito L, fijo, en valor absoluto; o bien que la transformación $y = f(x)$ magnifica distancias de puntos en el eje x a lo más por el factor L. Evidentemente, para una función "continua de Lipschitz" la expresión $\delta(\varepsilon) = \varepsilon/L$ es un módulo de continuidad puesto que $|f(x_2) - f(x_1)| < \varepsilon$ para $|x_2 - x_1| < \varepsilon/L$. Recíprocamente, cualquier función con un módulo de continuidad proporcional a ε, digamos, $\delta(\varepsilon) = c\varepsilon$, es "continua de Lipschitz", con $L = 1/c$.

Como se verá en el capítulo 2, la mayoría de las funciones tratadas serán "continuas de Lipschitz" excepto en puntos aislados, como una consecuencia del hecho que sus derivadas son acotadas en cualquier intervalo cerrado que excluya a estos puntos. Sin embargo, la continuidad de Lipschitz es solamente suficiente pero no necesaria para la continuidad uniforme. El ejemplo más simple de una función que es continua sin ser "continua de Lipschitz" está dada por $f(x) = \sqrt{x}$ para $x \geq 0$, y hay que considerar $x_0 = 0$. Aquí el cociente incremental

$$\frac{f(x) - f(0)}{x - 0} = \frac{1}{\sqrt{x}}$$

se torna arbitrariamente grande para x suficientemente pequeños, y, por lo tanto, no puede ser acotado por una constante fija L. De aquí que no sea posible elegir $\delta(\varepsilon)$ proporcional a ε; pero existen otros módulos de continuidad no lineales para esta función, por ejemplo $\delta(\varepsilon) = \varepsilon^2$.

La función $\sqrt{x}$ pertenece a la clase general de funciones llamadas "continuas de Hölder",** que satisfacen una "condición de Hölder":

$$|f(x_2) - f(x_1)| \leq L\,|x_2 - x_1|^{\alpha}$$

para todos x_1, x_2 de un intervalo, donde L y α son constantes fijas, estando el "exponente de Hölder" α restringido a valores $0 < \alpha \leq 1$. Las funciones continuas Lipschitz se obtienen para el exponente Hölder especial $\alpha = 1$.

* (N. del T.) Decimos así para abreviar la expresión *"continuas en el sentido de Lipschitz"*. También se llaman "Lipschitzianas".

** (N. del T.) Brevemente, en lugar de "continuas en el sentido de Höolder". También llamadas "Hölderianas".

Es claro que $\delta = L^{-1/\alpha}\varepsilon^{1/\alpha}$ es un posible módulo de continuidad para una función "continua de Hölder" f; aquí δ es proporcional a $\varepsilon^{1/\alpha}$, y no a ε. La función $f(x) = \sqrt{x}$ es "continua de Hölder" con exponente $\alpha = \frac{1}{2}$. Esto se sigue de la desigualdad

$$|\sqrt{x_2} - \sqrt{x_1}| \leq |x_2 - x_1|^{1/2},$$

la que se obtiene observando que

$$|\sqrt{x_2} - \sqrt{x_1}| \leq |\sqrt{x_2} + \sqrt{x_1}|$$

y multiplicando por $|\sqrt{x_2} - \sqrt{x_1}|$. Esto proporciona el módulo de continuidad $\delta(\varepsilon) = \varepsilon^2$ para $\sqrt{x}$, como se mencionó anteriormente.

Con más generalidad, las potencias fraccionales $f(x) = x^{\alpha}$ para $0 < \alpha \leq 1$ son "continuas de Hölder" con exponente de Hölder α.

Las funciones hölderianas aún no agotan la clase de todas las funciones uniformemente continuas. No es difícil construir ejemplos de funciones continuas para las cuales las potencias de ε no sirven como módulos de continuidad. (Ver problema 13, p. 140.)

e. El teorema del valor intermedio. Funciones inversas

Intuitivamente, no existe duda de que una función que es continua, y, de aquí, que no tiene "saltos", no puede variar de un valor a otro sin pasar por todos los valores intermedios. Este hecho se expresa por el así llamado teorema del valor intermedio (su demostración precisa se da en el Suplemento, p. 123). Lo enunciamos así:

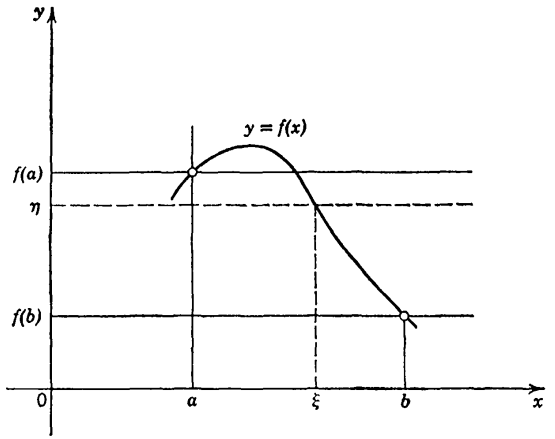


Figura 1.32 El teorema del valor intermedio.

TEOREMA DEL VALOR INTERMEDIO. *Considérese una función $f(x)$ continua en todo punto de un intervalo. Sean a y b dos puntos cualesquiera del intervalo y sea η cualquier número entre $f(a)$ y $f(b)$. Entonces, existe un valor ξ entre a y b para el cual $f(\xi) = \eta$.*

Interpretado geométricamente, el teorema establece que si dos puntos $(a, f(a))$ y $(b, f(b))$ de la gráfica de una función continua f están situados en diferentes lados de una paralela $y = \eta$ al eje x, entonces la paralela *intersecta* a la gráfica en algún punto intermedio (ver Fig. 1.32). Por supuesto que pueden haber varias intersecciones. En el importante caso en que la función $f(x)$ es monótona creciente o bien monótona decreciente en todo el intervalo, puede haber solamente una intersección, pues f no puede tener el mismo valor η para dos valores diferentes de ξ.

Como ejemplo tómese la función $f(x) = x^2$, la cual es monótona creciente y continua en el intervalo $1 \leq x \leq 2$. Aquí $f(1) = 1$ y $f(2) = 4$. Tomando para η el valor 2, intermedio entre 1 y 4, se encuentra que existe un único ξ entre 1 y 2 para el cual $\xi^2 = 2$. Este es, por supuesto, el número que se denota por $\sqrt{2}$.

Continuidad de la función inversa

Para cualquier función monótona creciente $f(x)$, definida en un intervalo $a \leq x \leq b$, se encuentra que para todo número real η tal que $f(a) \leq \eta \leq f(b)$, existe exactamente un ξ tal que $a \leq \xi \leq b$ para el cual $f(\xi) = \eta$.[1] Sean $\alpha = f(a)$ y $\beta = f(b)$. Puesto que ξ está determinado de manera única por η, se tiene una *función* $\xi = g(\eta)$ definida para argumentos η en el intervalo cerrado $[\alpha, \beta]$. Esta función g se denomina la *inversa* de f. Puesto que a valores más grandes de ξ corresponden valores más grandes de $\eta = f(\xi)$, la función g es también monótona creciente. Es fácil mostrar que la función inversa g también es continua.

En efecto, sea η cualquier valor entre α y β (ver Fig. 1.33). Entonces $\xi = g(\eta)$ debe estar situado entre $a = g(\alpha)$ y $b = g(\beta)$. Sea ε un número positivo dado, el cual puede suponerse tan pequeño que sea $a < \xi - \varepsilon < \xi + \varepsilon < b$. Se debe mostrar que $|g(y) - g(\eta)| < \varepsilon$ para toda y suficientemente próxima a η. Puesto que f es creciente, $\eta = f(\xi)$ está situado entre los valores $f(\xi - \varepsilon) = A$ y $f(\xi + \varepsilon) = B$, y puede encontrarse un δ tan pequeño que sea

$$A < \eta - \delta < \eta + \delta < B.$$

Si y es cualquier valor tal que $\eta - \delta < y < \eta + \delta$, y es $x = g(y)$, se tiene que $A < y < B$ y, por lo tanto, $g(A) < g(y) < g(B)$; esto es, $\xi - \varepsilon < g(y) < \xi + \varepsilon$, o bien $|g(y) - g(\eta)| < \varepsilon$. La misma demostración, ligeramente modificada, se aplica cuando η es uno de los puntos extremos α o β del intervalo de definición de g.

Las relaciones $y = f(x)$ y $x = g(y)$ son equivalentes y están representadas por la misma gráfica en el plano $\{x, y\}$; los puntos (x, y) del plano

[1] El teorema del valor intermedio, como se lo estableció, asigna ξ para η en el intervalo *abierto* $f(x) < \eta < f(b)$. Sin embargo, por supuesto que para $\eta = f(a)$, o bien para $\eta = f(b)$, basta sólo con tomar $\xi = a$, o bien $\xi = b$.

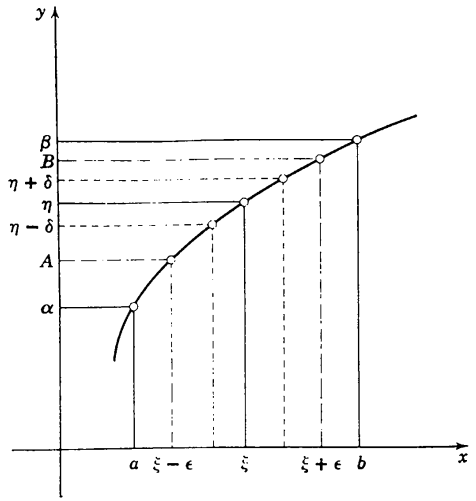


Figura 1.33 Continuidad de la inversa de una función monótona continua.

para los cuales $y = f(x)$ son los mismos que los puntos para los cuales $x = g(y)$. Si la función g se representa de la manera usual por $y = g(x)$, deben intercambiarse x e y; entonces la gráfica de $y = g(x)$ se obtiene de la gráfica $y = f(x)$ tomando la imagen especular respecto a la línea $y = x$.

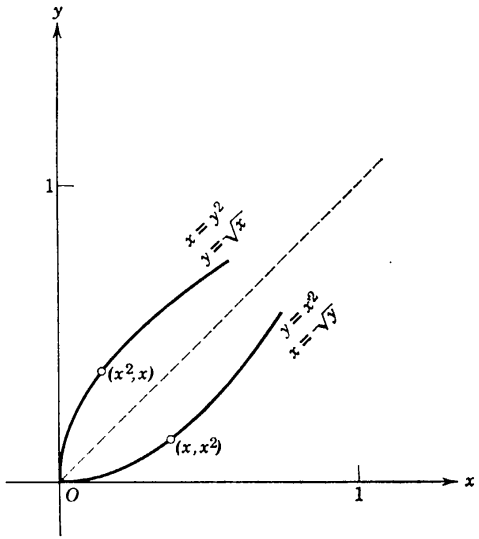


Figura 1.34 Funciones inversas.

Se tiene un ejemplo simple con las gráficas de la función $f(x) = x^2$ para $x \geq 0$ y de la función inversa $g(x) = \sqrt{x}$ para $x \geq 0$ (ver Fig. 1.34).

1.3 Las funciones elementales

a. Funciones racionales

Pasemos ahora a un breve repaso de las conocidas funciones elementales. Los tipos más simples de función se construyen mediante aplicación repetida de las operaciones elementales, adición y multiplicación. Si se aplican estas operaciones a una variable independiente x y a un conjunto de números reales $a_1, \ldots, a_n$ se obtienen los *polinomios*

$$y = a_0 + a_1 x + \cdots + a_n x^n.$$

Los polinomios son las funciones más simples del análisis y, en cierto sentido, las básicas.

Cocientes de tales polinomios, de la forma

$$y = \frac{a_0 + a_1 x + \cdots + a_n x^n}{b_0 + b_1 x + \cdots + b_m x^m},$$

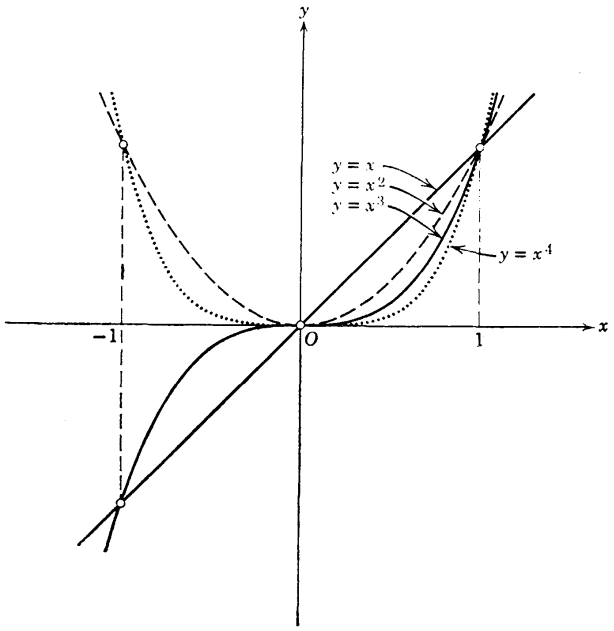


Figura 1.35 Potencias de x.

son las *funciones racionales* generales; éstas están definidas en todos los puntos donde el denominador es distinto de cero.

El polinomio más simple, la *función lineal*

$$y = ax + b,$$

se representa gráficamente por una línea recta. Toda *función cuadrática*

$$y = ax^2 + bx + c$$

se representa por una parábola. Las gráficas de polinomios de tercer grado

$$y = ax^3 + bx^2 + cx + d,$$

se llaman a veces "parábolas de tercer orden"; etc.

Las gráficas de la función $y = x^n$ para exponentes $n = 1, 2, 3, 4$ están dadas en la Fig. 1.35. Para valores pares de n la función $y = x^n$ satisface la ecuación $f(-x) = f(x)$, y es por ello una función par; mientras que para valores impares de n la función satisface la condición $f(-x) = f(x)$, y es por lo tanto impar.

El ejemplo más simple de función racional que no es un polinomio es la función $y = 1/x$ mencionada en la p. 51; su gráfica es una hipérbola equilátera. Otro ejemplo es la función $y = 1/x^2$ (ver Fig. 1.26, p. 60).

b. Funciones algebraicas

De inmediato nos vemos obligados a salir del conjunto de funciones racionales debido al problema de formar sus inversas. El ejemplo típico es la función $\sqrt[n]{x}$, la inversa de x^n. Se ve fácilmente que la función $y = x^n$ para $x \geq 0$ es monótona creciente y continua. Tiene por ello una inversa univaluada, que se denota por el símbolo $x = \sqrt[n]{y}$, o bien, intercambiando las letras utilizadas para las variables dependientes e independientes,

$$y = \sqrt[n]{x} = x^{1/n}.$$

Por definición, esta raíz es siempre no negativa. Para valores impares de n la función x^n es monótona para todos los valores de x, incluyendo valores negativos. Consecuentemente, para valores impares de n la definición de $\sqrt[n]{x}$ puede extenderse de manera única a todos los valores de x; y en este caso $\sqrt[n]{x}$ es negativa para valores negativos de x.

Con mayor generalidad, puede considerarse la función

$$y = \sqrt[n]{R(x)},$$

donde $R(x)$ es una función racional. Otras funciones de tipo semejante se forman aplicando las operaciones racionales a una o más de estas funciones especiales. Así, por ejemplo, pueden formarse las funciones

$$y = \sqrt[m]{x} + \sqrt[n]{x^2 + 1}, \qquad y = x + \sqrt{x^2 + 1}.$$

Estas funciones son casos especiales de *funciones algebraicas.* (El concepto general de una función algebraica se definirá en el Volumen II.)

c. Funciones trigonométricas

Las funciones racionales y las funciones algebraicas se definen directamente por medio de las operaciones elementales de cálculo, pero la geometría es la fuente de la cual obtenemos primero ejemplos de otras funciones, las así llamadas *funciones trascendentes.*[1] De éstas, se considerarán aquí las *funciones trascendentes elementales,* a saber, las funciones trigonométricas, la función exponencial y la función logaritmo.

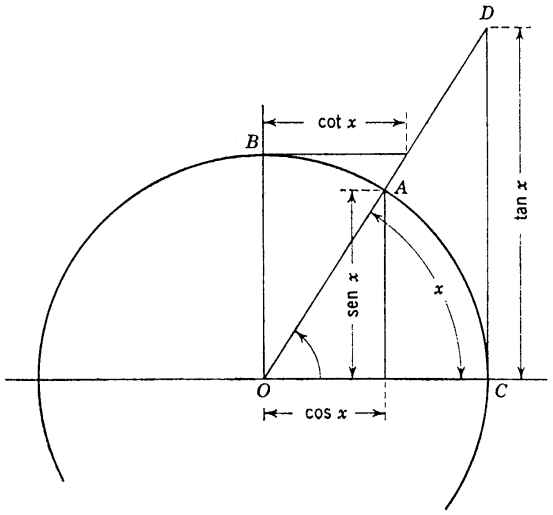


Figura 1.36 Las funciones trigonométricas.

En investigaciones analíticas los ángulos no se miden en grados, minutos y segundos, sino en *radianes.* El ángulo que se desea medir se coloca con su vértice en el centro de un círculo de radio 1 y el tamaño del ángulo se mide por la longitud del arco de circunferencia que es cortada por el ángulo.[2] Por lo tanto, un ángulo de 180° es el mismo que un ángulo de π

[1] La palabra "trascendente" no significa nada particularmente profundo o misterioso; sugiere simplemente que la definición de estas funciones trasciende las operaciones elementales de cálculos, "*quod algebrae vires transcendit*".

[2] La medida en radiantes de un ángulo puede ser definida también como el doble del *área* del sector correspondiente sobre el círculo de radio uno.

radianes (tiene medida π, en radianes). Un ángulo de 90° tiene medida, en radianes, $\pi/2$; un ángulo de 45° tiene a $\pi/4$ como su medida en radianes; un ángulo de 360° tiene medida, en radianes, 2π. Recíprocamente, un ángulo de 1 radián expresado en grados es

$$\frac{180°}{\pi}, \qquad \text{o bien, aproximadamente, } 57° \ 17' \ 45''.$$

En lo sucesivo, siempre que se hable de un ángulo x, se entenderá un ángulo cuya medida en radianes es x.

Recordemos brevemente el significado de las funciones trigonométricas $\operatorname{sen} x$, $\cos x$, $\tan x$, $\cot x$.[1] Estas están representadas en la Fig. 1.36, en la cual el ángulo x se mide a partir del segmento OC (de longitud 1), con-

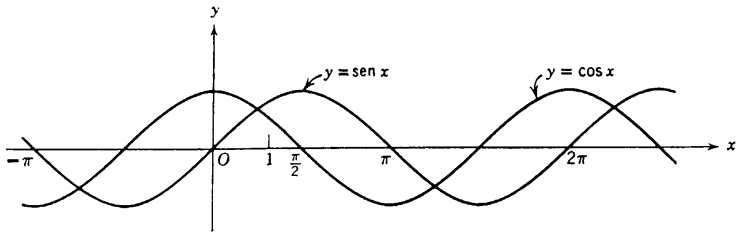


Figura 1.37.

siderando positivos a los ángulos en la dirección contraria a la de las manecillas del reloj. Las funciones $\cos x$ y $\operatorname{sen} x$ son las coordenadas rectangulares del punto A. Las gráficas de la función $\operatorname{sen} x$, $\cos x$, $\tan x$, $\cot x$ están dadas en las Figs. 1.37 y 1.38.

Más adelante (ver p. 237) estaremos en condiciones de reemplazar las definiciones geométricas por las analíticas correspondientes.

d. La función exponencial y el logaritmo

Además de las funciones trigonométricas, la función exponencial con base positiva a,

$$y = a^x,$$

y su inversa, el logaritmo de base a,

$$x = \log_a y,$$

se incluyen también entre las funciones trascendentes elementales. En matemáticas elementales es costumbre pasar por alto ciertas dificultades in-

[1] En ocasiones también es conveniente introducir las funciones $\sec x = 1/\cos x$, $\operatorname{cosec} x = 1/\operatorname{sen} x$.

herentes en su definición, y se pospondrá también aquí la discusión detallada de éstas hasta que dispongamos de mejores métodos (ver sección 2.5 p. 167). Se puede, sin embargo, indicar aquí al menos un camino "elemental" para definir estas funciones. Si $x = p/q$ es un número racional (donde p y q son enteros positivos), entonces —suponiendo positivo el número a— se define a^x como $\sqrt[q]{a^p} = a^{p/q}$, donde la raíz, de acuerdo a la convención, ha de ser tomada positiva. Puesto que los valores racionales de x son densos en todas partes, es natural extender esta función a^x a una función continua definida también para valores irracionales de x, dándole a a^x, cuando x es irracional, valores que son continuos con los valores ya definidos cuando

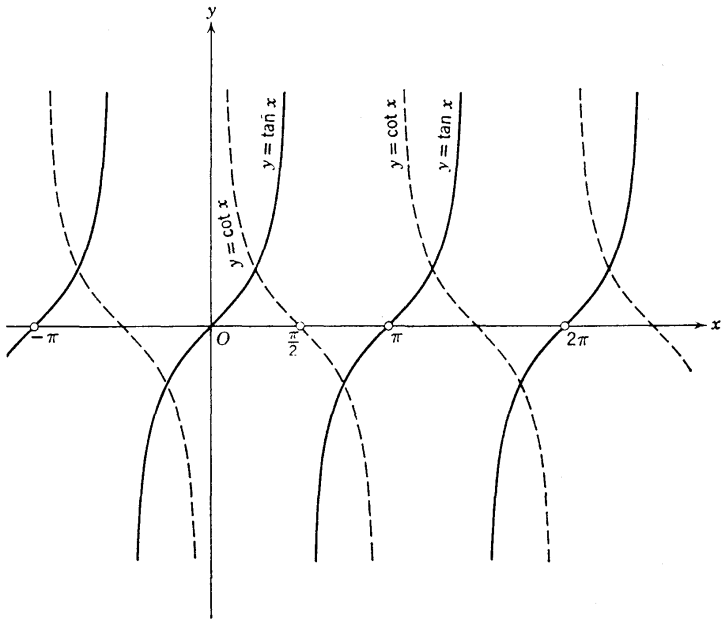


Figura 1.38.

x es irracional. Esto define una función continua $y = a^x$, la "función exponencial", la cual para todos los valores *racionales* de x dan el valor de a^x encontrado antes. Por el momento, podemos presuponer que esta extensión es realmente posible y puede ser llevada a cabo de una sola manera; pero deberá tenerse en mente que aún debemos *demostrar* que esto es cierto.[1]

La función

$$x = \log_a y$$

[1] Esto se hace en la p. 174.

puede ser entonces definida para $y > 0$ como la inversa de la función exponencial: $x = \log_a y$ es aquel número para el cual $y = a^x$.

e. Funciones compuestas, productos simbólicos, funciones inversas

Frecuentemente se forman nuevas funciones no sólo combinando funciones conocidas mediante operaciones racionales sino por el proceso más general y básico de formar *funciones de funciones* o funciones *compuestas*.

Sea $u = \phi(x)$ una función cuyo dominio está en el intervalo $a \leq x \leq b$ y cuyo rango se encuentra situado en el intervalo $\alpha \leq u \leq \beta$. Además, sea $y = g(u)$ una función definida para $\alpha \leq u \leq \beta$. Entonces $g(\phi(x)) = f(x)$ define una función f para $a \leq x \leq b$, la cual está "compuesta", o bien es la "composición" de g y ϕ. Por ejemplo, $f(x) = 1/(1 + x^{2n})$ es la composición de las funciones $\phi(x) = 1 + x^{2n}$ y $g(u) = 1/u$. Análogamente, la función $f(x) = \text{sen}\,(1/x)$ es la composición de $\phi(x) = 1/x$ y $g(u) = \text{sen}\,u$.

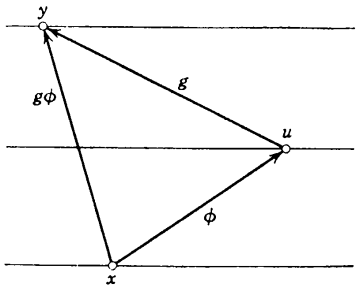


Figura 1.39 Producto simbólico $g\phi = f$ de dos funciones.

Es útil interpretar las funciones compuestas en términos de transformaciones o "*mapeos*". La transformación ϕ lleva cada punto x del intervalo $[a, b]$ a un punto u del intervalo $[\alpha, \beta]$. La transformación g lleva cualquier valor u en $[\alpha, \beta]$ a un punto y. La transformación f es el "producto simbólico" $g\phi$ de las transformaciones g y ϕ, esto es, la formada por ϕ y g sucesivamente, en ese orden; para cualquier x en $[a, b]$ se forma su imagen u por la transformación ϕ, y entonces se aplica g a la imagen $u = \phi(x)$, obteniéndose $g(\phi(x)) = f(x) = y$ (ver Fig. 1.39). Tal producto simbólico $g\phi$ es natural y tiene sentido para cualquier tipo de operación; significa que primero se realiza ϕ, y luego sobre el resultado se aplica g.[1] No debe confun-

[1] Que el producto $g\phi$ corresponde a realizar *primero* ϕ *y luego* g (en *este* orden) parece a primera vista no natural, pero en realidad corresponde a la convención siempre adoptada en matemáticas que consiste en escribir el argumento x de una función $f(x)$ *a la derecha* del símbolo f para la función. Así, por ejemplo, en sen(log x) se sobreentiende que primero se forma el logaritmo de x y después se toma el seno de esto, y no en sentido inverso.

dirse el producto simbólico $g\phi = g(\phi)$ de dos funciones con el producto *algebraico* ordinario $g(x) \cdot \phi(x)$ de funciones, en el cual tanto $g(x)$ como $\phi(x)$ se realizan para el mismo argumento x (transformaciones aplicadas al mismo punto) y se forma el *producto de los valores* de las funciones.

Naturalmente, no puede esperarse que los productos simbólicos sean conmutativos. En general, $g(\phi)$ y $\phi(g)$ no son los mismos, aún cuando ambos estén definidos; el orden en el cual las operaciones se realizan es de la mayor importancia. Si, por ejemplo, ϕ establece la operación de "sumar 1 a un número" y g la operación de "multiplicar un número por 2", entonces

$$g(\phi(x)) = 2(x + 1) = 2x + 2, \qquad \phi(g(x)) = (2x) + 1 = 2x + 1.$$

(Ver Fig. 1.40.)

Para estar en condiciones de formar el producto simbólico $g\phi$ de dos funciones, los "factores" g y ϕ deberán adaptarse mutuamente en el sentido de que el dominio de g deberá contener al rango de ϕ. Por lo tanto, no puede formarse $g\phi$ cuando

$$g(u) = \sqrt{u}, \qquad \text{y} \qquad \phi(x) = -1 - x^2.$$

Es útil considerar funciones que están compuestas más de una vez. Una de tales funciones es

$$f(x) = \sqrt{1 + \tan(x^2)},$$

que puede construirse mediante las composiciones sucesivas

$$\phi(x) = x^2, \qquad \psi(\phi) = 1 + \tan\phi, \qquad g(\psi) = \sqrt{\psi} = f(x).$$

Podría escribirse simbólicamente $f = g\psi\phi$.

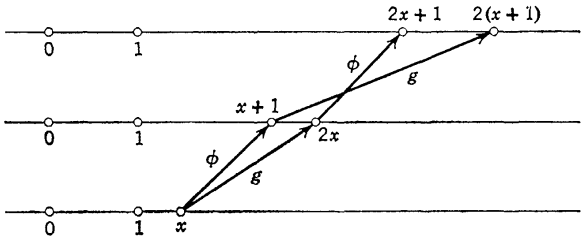


Figura 1.40 La no conmutatividad de transformaciones.

Funciones inversas

La noción de "función inversa" se aclara en el contexto del producto de transformaciones. Considérese la función o transformación ϕ que asocia a un punto x en el dominio de ϕ la imagen $u = \phi(x)$. Supóngase que la función es tal que x diferentes son transformados siempre en diferentes u,

y recíprocamente. La función se denomina entonces biunívoca. Ahora bien, un valor u es la imagen de a lo más un valor x. Puede asociarse con cada u en el rango de ϕ el valor $x = g(u)$ del cual u es la imagen bajo la transformación ϕ. De esta manera se ha definido una función g cuyo dominio es el rango de ϕ, y el cual, cuando es aplicado a una imagen $u = \phi(x)$, reproduce el valor original x; esto es, $g(\phi(x)) = x$. La función g se denomina la *inversa* de ϕ. Está caracterizada por la ecuación simbólica $g\phi x = x$.

La transformación identidad

La transformación identidad, I, se define como aquella que transforma a cada x en sí mismo. Para la inversa g de ϕ se tiene que $g\phi = I$.[1] La transformación I desempeña el mismo papel en lo que respecta a la multiplicación simbólica que el número 1 en la multiplicación ordinaria: la multiplicación por I no modifica a una transformación cualquiera dada. Por lo tanto, la ecuación $g\phi = I$ sugiere la notación $g = \phi^{-1}$ para la inversa de ϕ. Por ejemplo, la inversa $x = \text{arc sen}\, u$ de la función $u = \text{sen}\, x$ se denota a menudo por $x = \text{sen}^{-1} u$.[2]

A partir de la definición de la inversa g de ϕ, se sigue inmediatamente que ϕ es también la inversa de g, de modo que no sólo $g(\phi) = x$ sino que también $\phi(g(u)) = u$.

* Una función monótona $u = \phi(x)$ definida en un intervalo $a \le x \le b$ evidentemente define una transformación biunívoca de ese intervalo. Si, además, ϕ es continua, entonces, como se vio anteriormente, como consecuencia del teorema del valor intermedio (p. 68), es claro que el rango de ϕ, es el intervalo con puntos extremos $\phi(a)$ y $\phi(b)$. En este caso existe la inversa g de ϕ, que es también monótona y continua en ese último intervalo. En efecto, *las funciones monótonas continuas son las únicas funciones continuas que tienen inversas o que definen transformaciones biunívocas.* Así, sea $u = \phi(x)$ una función continua en el intervalo cerrado $[a, b]$ que transforma x diferentes del intervalo en diferentes u. Entonces, en particular, los valores $\phi(a) = \alpha$ y $\phi(b) = \beta$ son distintos. Supóngase que $\alpha < \beta$. Puede entonces mostrarse que $\phi(x)$ es monótona creciente en todo el intervalo. Pues, si no fuera éste el caso, podrían encontrarse dos valores c y d tales que $a \le c < d \le b$, para los cuales $\phi(d) < \phi(c)$. Si se tuviera también $\phi(d) \ge \phi(a)$, se seguiría del teorema del valor intermedio que existe un ξ en el intervalo $[a, c]$ para el cual $\phi(\xi) = \phi(d)$. Este ξ sería diferente de d, y la transformación no podría ser biunívoca. Si, por otra parte, $\phi(d) < \phi(a) = \alpha$, se seguiría que $\phi(a)$ es un punto intermedio entre $\phi(d)$ y $\phi(b)$; habría entonces un ξ intermedio entre d y b para el cual $\phi(\xi) = \phi(a)$, y esto contradice la naturaleza biunívoca de la transformación.

Una propiedad importante de las funciones compuestas, casi evidente, es que $g(\phi(x)) = f(x)$ es continua (donde esté definida) si g y ϕ lo son. Así, para un ε positivo dado se tiene

$$|f(x) - f(x_0)| = |g(\phi(x)) - g(\phi(x_0))| < \varepsilon \quad \text{para} \quad |\phi(x) - \phi(x_0)| < \delta,$$

[1] Más precisamente, $g\phi$ coincide con I en el dominio de ϕ.
[2] Esto no se deberá confundir con el *recíproco algebraico* $1/(\text{sen}\, u)$.

como consecuencia de la continuidad de la función g. Sin embargo, puesto que ϕ también es continua, ciertamente se tiene $|\phi(x) - \phi(x_0)| < \delta$ para todo x que satisface $|x - x_0| < \delta'$ con algún δ' positivo adecuado. Por tanto,

$$|f(x) - f(x_0)| < \varepsilon \qquad \text{para} \qquad |x - x_0| < \delta';$$

lo que muestra la continuidad de f.

Es mucho más fácil apelar a este teorema general al demostrar la continuidad de funciones compuestas como $\sqrt{1 - x^2}$ que tratar de construir directamente un módulo de continuidad para la función.

1.4 Sucesiones

Hasta ahora se han considerado funciones de una variable continua, o bien funciones cuyos dominios consisten de uno o más intervalos. Sin embargo, en matemáticas ocurren numerosos casos en los cuales una cantidad a depende de un entero positivo n. Tal función $a(n)$ asocia un valor a cada número natural n. La función $a(n)$ se denomina *sucesión,* más específicamente, *sucesión infinita,* si n recorre *todos* los enteros positivos. Usualmente se escribe a_n [1] en lugar de $a(n)$ para el "n-ésimo elemento" de la sucesión, y se consideran los elementos que forman una sucesión arreglados en orden de subíndices crecientes n:

$$a_1, a_2, a_3, \ldots.$$

Aquí la dependencia de los números a_n con respecto de n puede ser definida por cualquier ley que sea; y, en particular, los valores a_n no tienen que ser distintos uno del otro. La idea de sucesión será comprendida más fácilmente mediante ejemplos.

1. La suma de los primeros n enteros

$$S(n) = 1 + 2 + 3 + \cdots + n = \tfrac{1}{2}n(n + 1)$$

es una función de n, dando lugar a la sucesión

$$1, 3, 6, 10, 15, \ldots.$$

2. Otra función sencilla de n es la expresión "factorial de n", el producto de los primeros n enteros:

$$n! = 1 \cdot 2 \cdot 3 \cdots \cdots n.$$

3. Todo entero $n > 1$ que no sea un número primo es divisible por más de dos enteros positivos, mientras que los números primos son divisibles solamente por ellos mismos y por 1. Es claro que puede considerarse el

[1] Se pronuncia "*a*-sub-*n*".

número $T(n)$ de divisores de n como una función de n. Para los primeros números está dado por la tabla:

$n =$	1	2	3	4	5	6	7	8	9	10	11	12
$T(n) =$	1	2	2	3	2	4	2	4	3	4	2	6

4. Una sucesión de gran importancia en la Teoría de Números es $\pi(n)$, el número de primos menores que el número n. Su investigación detallada es uno de los problemas más fascinantes. El resultado principal es: "El número $\pi(n)$ está dado *asintóticamente,*[1] para valores grandes de n, por la función $n/\log n$". Aquí, por $\log n$ se entiende el logaritmo de "base natural" e, que se define posteriormente (pág. 100).

1.5 Inducción matemática

Se inserta aquí una discusión de un tipo muy importante de razonamiento que está difundido en la mayor parte del pensamiento matemático.

El hecho de que toda la sucesión de números naturales sea generada partiendo con el número 1 y pasando de n a $n + 1$, da lugar al fundamental "principio de inducción matemática". En las ciencias naturales se obtiene mediante "inducción empírica" de un gran número de muestras una ley que se espera se cumpla generalmente. El grado de certidumbre de la ley depende entonces del número de veces que una muestra o un "evento" haya sido observado y así confirmada la ley. Este tipo de inducción puede ser abrumadoramente convincente, aún cuando no lleve consigo la certidumbre lógica de una demostración matemática.

La *inducción matemática* es utilizada para establecer con certidumbre lógica la exactitud de un teorema para una sucesión infinita de casos. Denótese por A un aserto que se refiere a un número natural arbitrario n. Por ejemplo, la propiedad A puede ser "La suma de los ángulos interiores en un polígono simple de $n + 2$ lados es n veces 180°", o sea, $n\pi$. Para demostrar una propiedad de este tipo no es suficiente probarla para los primeros 10 o los primeros 100, o aún los primeros 1,000 valores de n. En lugar de eso debe aplicarse un método matemático que se explica primero para este ejemplo. Para $n = 1$ el polígono se reduce a un triángulo, para el cual es sabido que la suma de los ángulos es de 180°. Para un cuadrángulo, correspondiente a $n = 2$, trazamos una diagonal dividiendo el cuadrángulo en dos triángulos. Esto muestra que la suma de los ángulos del cuadrángulo es igual a la suma combinada de los ángulos de los dos triángulos, esto es, $180° + 180° = 2 \cdot 180°$. Procediendo al ejemplo de un pentágono, éste puede dividirse en un cuadrángulo y un triángulo mediante

[1] Esto es, el cociente del número $\pi(n)$ por el número $n/\log n$ difiere arbitrariamente poco de uno, bastando sólo con que n sea suficientemente grande.

el trazo de una diagonal adecuada. Esto da para la suma de los ángulos del pentágono el valor $2 \cdot 180° + 1 \cdot 180° = 3 \cdot 180°$. Puede continuarse de esta manera y probarse el teorema general sucesivamente para $n = 4, 5$, etcétera. La exactitud de la propiedad A para cualquier n se sigue de su exactitud para el n *precedente;* de esta manera se establece su validez general para todo n.

Formulación general

Lo que es esencial en la demostración de la propiedad A del ejemplo es que A se ha probado sucesivamente para los casos especiales A_1, A_2, ... A_n, ... La posibilidad de hacer esto depende de dos factores: (1) se ha de dar una demostración general mostrando que el aserto A_{r+1} es correcto siempre que A_r sea correcto; y (2) el aserto A_1 debe ser probado. Que estas condiciones son suficientes para probar la exactitud de todo aserto A_1, A_2, A_3, ..., constituye el *principio de inducción matemática.* En lo que sigue se acepta la validez de este principio como un hecho básico de lógica.

El principio puede ser formulado en una forma abstracta más general. "Sea S cualquier conjunto que consista de números naturales, que tenga las siguientes dos propiedades: (1) siempre que S contenga un número r, entonces contendrá también al número $r + 1$; y (2) S contiene al número 1. Entonces es cierto que S es el conjunto de todos los números naturales." La formulación previa del principio de inducción matemática se obtiene fácilmente si se toma como S el conjunto de los números naturales para los cuales la aseveración A es correcta.

El principio se aplica a menudo sin hacer mención específica del mismo; o bien, su uso se indica solamente mediante la expresión "etc.". Esto sucede a menudo, particularmente en matemáticas elementales. Sin embargo, en situaciones más complicadas es preferible apelar explícitamente al principio.

Ejemplos. Como ilustración se dan dos aplicaciones.

Primero se prueba una fórmula para la suma de los n primeros cuadrados. Se encuentra para n pequeños (digamos $n < 5$) que la siguiente fórmula,[1] denotada por A_n, se cumple:

$$1^2 + 2^2 + 3^2 + \cdots + n^2 = \frac{n(n+1)(2n+1)}{6}.$$

Se conjetura que esta fórmula es correcta para todo n. Para demostrarlo, se supone que r es cualquier número para el cual la fórmula A_r es correcta, esto es, que

$$1^2 + 2^2 + 3^2 + \cdots + r^2 = \frac{r(r+1)(2r+1)}{6};$$

sumando $(r + 1)^2$ a ambos miembros se obtiene

[1] Este resultado fue utilizado por el matemático griego Arquímedes en su trabajo sobre espirales.

$$1^2 + 2^2 + \cdots + r^2 + (r+1)^2 = \frac{r(r+1)(2r+1)}{6} + (r+1)^2$$

$$= \frac{(r+1)(r+2)[2(r+1)+1]}{6}$$

Esto, sin embargo, es precisamente la propiedad A_{r+1} obtenida substituyendo $r + 1$ por n en A_n. Por lo tanto, la veracidad de A_r implica la de A_{r+1}. Para completar la demostración de A_n para n general se requiere verificar solamente la exactitud de A_1, esto es, de

$$1^2 = \frac{1 \cdot 2 \cdot 3}{6}.$$

Puesto que esto es obviamente correcto, la fórmula A_n queda establecida para todo número natural n.

El lector deberá probar mediante un argumento análogo que

$$1^3 + 2^3 + 3^3 + \cdots + n^3 = \left[\frac{n(n+1)}{2}\right]^2.$$

Como ilustración adicional del principio de inducción se prueba

EL TEOREMA DEL BINOMIO. *El aserto A_n del teorema está representado por la fórmula*

$$(a+b)^n = a^n + \frac{n}{1}a^{n-1}b + \frac{n(n-1)}{1 \cdot 2}a^{n-2}b^2$$

$$+ \frac{n(n-1)(n-2)}{1 \cdot 2 \cdot 3}a^{n-3}b^3 + \cdots + \frac{n(n-1)(n-2) \cdot \cdots \cdot 2 \cdot 1}{1 \cdot 2 \cdot 3 \cdot \cdots \cdot (n-1) \cdot n}b^n.$$

Es costumbre escribir esta fórmula en la forma

$$(a+b)^n = \binom{n}{0}a^n + \binom{n}{1}a^{n-1}b + \binom{n}{2}a^{n-2}b^2 + \cdots + \binom{n}{n}b^n$$

donde el *coeficiente binomial* $\binom{n}{k}$ está definido por

$$\binom{n}{k} = \frac{n(n-1)(n-2)\cdots(n-k+1)}{k!} = \frac{n!}{k!(n-k)!}$$

para $k = 1, 2, \ldots, (n-1)$, y

$$\binom{n}{0} = \binom{n}{n} = 1.$$

(Si se define $0! = 1$, la fórmula general para $\binom{n}{k}$ se aplica también a los casos $k = 0$ y $k = n$.)

Si A_n se cumple para un cierto n, se encuentra, multiplicando ambos miembros por $(a + b)$, que

$$\begin{aligned}(a+b)^{n+1} &= (a+b)\left[\binom{n}{0}a^n + \binom{n}{1}a^{n-1}b + \cdots + \binom{n}{n}b^n\right] \\ &= \binom{n}{0}a^{n+1} + \left[\binom{n}{0} + \binom{n}{1}\right]a^n b + \left[\binom{n}{1} + \binom{n}{2}\right]a^{n-1}b^2 \\ &\quad + \cdots + \left[\binom{n}{n-1} + \binom{n}{n}\right]ab^n + \binom{n}{n}b^{n+1}.\end{aligned}$$

Ahora bien,

$$\begin{aligned}&\binom{n}{k} + \binom{n}{k+1} \\ &= \frac{n(n-1)\cdots(n-k+1)}{k!} + \frac{n(n-1)\cdots(n-k+1)(n-k)}{(k+1)!} \\ &= \frac{n(n-1)(n-2)\cdots(n-k+1)}{k!}\left(1 + \frac{n-k}{k+1}\right) \\ &= \frac{(n+1)n(n-1)\cdots(n-k+1)}{(k+1)!} = \binom{n+1}{k+1}.\end{aligned}$$

Puesto que $\binom{n}{0} = \binom{n+1}{0} = 1$ y $\binom{n}{n} = \binom{n+1}{n+1} = 1$, se tiene

$$\begin{aligned}(a+b)^{n+1} = \binom{n+1}{0}a^{n+1} + \binom{n+1}{1}a^n b + \binom{n+1}{2}a^{n-1}b^2 + \cdots \\ + \binom{n+1}{n}ab^n + \binom{n+1}{n+1}b^{n+1},\end{aligned}$$

que es la fórmula A_{n+1}. Puesto que también para $n = 1$ la fórmula vale; es decir,

$$(a+b)^1 = \binom{1}{0}a + \binom{1}{1}b = a + b,$$

resulta que el Teorema del Binomio se cumple para todos los números naturales n.

1.6 El límite de una sucesión

El concepto fundamental sobre el cual descansa la esencia toda del análisis matemático es el de *límite* de una sucesión infinita a_n. A menudo un número a se describe por medio de una sucesión infinita a_n de aproximaciones; esto es, el valor a está dado por el valor a_n con cualquier grado de precisión deseado si el índice n se elige suficientemente grande. Se han encontrado ya tales *representaciones de números a como "límites"* de sucesiones en su representación como fracciones decimales infinitas; los números reales aparecen entonces como límites, para n crecientes, de las sucesiones de fracciones decimales ordinarias con n dígitos. En la sección 1.7 se dará una discusión precisa general del concepto de límite; por el momento se ilustrará la idea de límite mediante algunos ejemplos significativos.

Las sucesiones $a_1, a_2, \ldots$ pueden ser convenientemente representadas por una sucesión de "bloques", correspondiendo el elemento a_n al rectángulo en el plano xy acotado por las rectas $x = n - 1$, $x = n$, $y = a_n$, $y = 0$, que tiene por área [1] el número $|a_n|$; o bien, equivalentemente, por la gráfica de una función constante por pedazos, $a(x)$, de una variable continua x con discontinuidades de salto en los puntos $x = n$.

a. $a_n = \dfrac{1}{n}$

Considérese la sucesión

$$1, \frac{1}{2}, \frac{1}{3}, \ldots, \frac{1}{n}, \ldots.$$

(Ver Fig. 1.41.) Ningún número de esta sucesión es cero; pero, conforme el número n crece, a_n se aproxima al cero. Además, si se toma cualquier

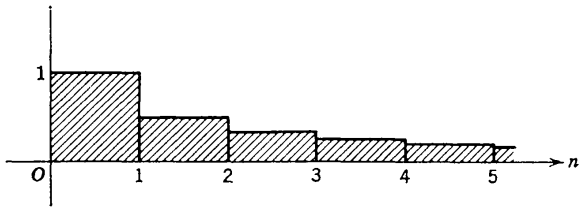


Figura 1.41 La sucesión $a_n = \dfrac{1}{n}$.

intervalo centrado en el origen, no importa cuán pequeño, entonces a partir de un índice definido hacia adelante todos los números a_n se encontrarán en ese intervalo. Esta situación se expresa diciendo que "si n crece los

[1] Pudo haberse escogido igualmente el rectángulo acotado por las rectas $x = n$, $x = n + 1$, $y = a_n$, $y = 0$ para representar a_n.

números a_n *tienden* a cero", o que "poseen el *límite* cero", o bien que "la sucesión $a_1, a_2, a_3, \ldots$ *converge* a cero".

Si los números son representados como puntos en una línea, esto significa que los puntos $1/n$ se apiñan cada vez más cerca del punto cero conforme n crece.

La situación es semejante para la sucesión

$$1, -\frac{1}{2}, \frac{1}{3}, -\frac{1}{4}, \ldots, \frac{(-1)^{n-1}}{n}, \ldots.$$

(Ver Fig. 1.42.) También aquí, los números a_n tienden a cero cuando n crece. La única diferencia es que los números a_n son a veces mayores y a veces menores que el límite cero; como se dice, la sucesión *oscila* en torno al límite.

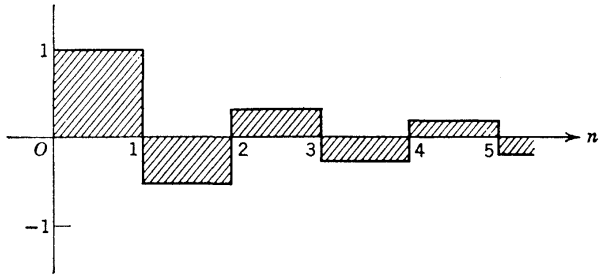


Figura 1.42 La sucesión $a_n = \dfrac{(-1)^{n-1}}{n}$.

La convergencia de la sucesión a cero usualmente es expresada simbólicamente por la ecuación

$$\lim_{n\to\infty} a_n = 0,$$

o bien, ocasionalmente, por la notación abreviada

$$a_n \to 0.$$

b. $a_{2m} = \dfrac{1}{m}$; $a_{2m-1} = \dfrac{1}{2m}$

En los ejemplos anteriores el valor absoluto de la diferencia entre a_n y el límite se tornaba progresivamente menor conforme n crecía. Esto no es necesariamente siempre el caso, como se muestra en la sucesión

$$\frac{1}{2}, 1, \frac{1}{4}, \frac{1}{2}, \frac{1}{6}, \frac{1}{3}, \ldots, \frac{1}{2m}, \frac{1}{m}, \ldots;$$

(ver Fig. 1.43) dada para valores pares $n = 2m$ por $a_n = a_{2m} = 1/m$; y para valores impares $n = 2m - 1$ por $a_n = a_{2m-1} = 1/2m$. Esta sucesión tiene también el límite cero; pues todo intervalo en torno al origen, por pequeño que sea, contiene a todos los números a_n desde un cierto valor

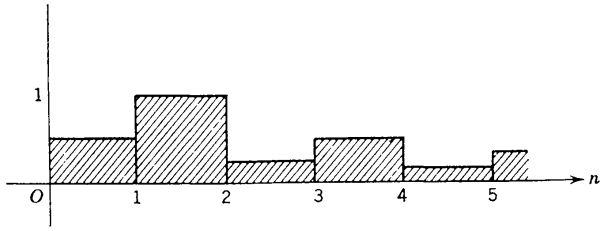


Figura 1.43 La sucesión $a_{2n} = \dfrac{1}{n}$, $a_{2n-1} = \dfrac{1}{2n}$.

de n hacia adelante; pero no es cierto que todo número a_n esté situado más cerca del límite cero que el número precedente.

c. $a_n = \dfrac{n}{n+1}$

Considérese la sucesión

$$a_1 = \frac{1}{2}, a_2 = \frac{2}{3}, \ldots, a_n = \frac{n}{n+1}, \ldots.$$

Escribiendo $a_n = 1 - 1/(n+1)$, se ve que al crecer n el número a_n se aproxima al número 1, en el sentido de que para todo intervalo definido en torno del punto 1 todos los números a_n sucesores a cierto a_N deberán encontrarse en ese intervalo. Se escribe

$$\lim_{n\to\infty} a_n = 1.$$

La sucesión

$$a_n = \frac{n^2 - 1}{n^2 + n + 1}$$

se comporta de una manera semejante. Esta sucesión tiende también a un límite al crecer n; de hecho, al límite uno: $\lim_{n\to\infty} a_n = 1$. Esto se ve más fácilmente si se escribe

$$a_n = 1 - \frac{n+2}{n^2+n+1} = 1 - r_n;$$

y se requiere mostrar solamente que los números r_n tienden a cero conforme n crece. Para todos los valores de n mayores que 2 se tiene $n + 2 < 2n$ y $n^2 + n + 1 > n^2$. Por lo tanto, para el residuo r_n se tiene

$$0 < r_n < \frac{2n}{n^2} = \frac{2}{n} \qquad (n > 2),$$

de donde se ve que r_n tiende a cero cuando n crece. Esta discusión da al mismo tiempo una estimación del máximo valor según el cual el número a_n (para $n > 2$) puede diferir del límite uno; esta diferencia no puede exceder a $2/n$.

Este ejemplo ilustra el hecho de que para valores grandes de n los términos con los mayores exponentes en el numerador y denominador de la fracción para a_n predominan y determinan el límite.

d. $a_n = \sqrt[n]{p}$

Sea p cualquier número positivo fijo. Considérese la sucesión $a_1, a_2, a_3, \ldots, a_n, \ldots$, donde

$$a_n = \sqrt[n]{p}.$$

Se afirma que

$$\lim_{n\to\infty} a_n = \lim_{n\to\infty} \sqrt[n]{p} = 1.$$

Se probará esto utilizando un lema que también será útil para otros propósitos.

LEMA. *Si h es un número positivo y n un entero positivo, entonces*

$$(1) \qquad (1 + h)^n \geq 1 + nh.$$

Esta desigualdad es una consecuencia trivial del teorema del binomio (ver p. 82), de acuerdo con el cual

$$(1 + h)^n = 1 + nh + \frac{n(n-1)}{2} h^2 + \cdots + h^n,$$

si se observa que todos los términos en el desarrollo de $(1 + h)^n$ son no negativos. El mismo argumento conduce a la desigualdad más fuerte

$$(1 + h)^n \geq 1 + nh + \frac{n(n-1)}{2} h^2.$$

Volviendo a la sucesión, se distinguirá entre los casos $p > 1$ y $p < 1$ (si $p = 1$, entonces $\sqrt[n]{p}$ es igual a 1 para todo n, y el aserto es ciertamente verdadero).

Si $p > 1$, entonces $\sqrt[n]{p}$ es también mayor que 1; hágase $\sqrt[n]{p} = 1 + h_n$, donde h_n es una cantidad positiva dependiente de n; y por la desigualdad (1) se tiene

$$p = (1 + h_n)^n \geq 1 + nh_n,$$

implicando

$$0 < h_n \leq \frac{p-1}{n}.$$

Cuando n crece el número h_n deberá tender a cero, lo que prueba que a_n converge al límite uno, como se afirmó. Al mismo tiempo se dispone de un medio para estimar cuán cerca está a_n del límite uno, puesto que la diferencia h_n entre a_n y uno no es mayor que $(p - 1)/n$.

Si $p < 1$, entonces $1/p > 1$, y $\sqrt[n]{1/p}$ converge al límite uno. Sin embargo,

$$\sqrt[n]{p} = \frac{1}{\sqrt[n]{1/p}}.$$

Como el recíproco de una cantidad que tiende a uno, el propio $\sqrt[n]{p}$ tiende a uno.

e. $a_n = \alpha^n$

Considérese la sucesión $a_n = \alpha^n$, donde α es fijo y n recorre la sucesión de enteros positivos.

Primero, sea α un número positivo menor que uno. Hágase entonces $\alpha = 1/(1 + h)$, donde h es positivo, y la desigualdad (1) da

$$a_n = \frac{1}{(1+h)^n} \leq \frac{1}{1 + nh} < \frac{1}{nh}$$

Puesto que h, y consecuentemente $1/h$, depende solamente de α y no cambia cuando n crece, se ve que α^n tiende a cero conforme n crece:

$$\lim_{n\to\infty} \alpha^n = 0 \qquad (0 < \alpha < 1).$$

Se satisface la misma relación cuando α es cero, o bien negativo pero mayor que -1. Esto es inmediatamente obvio, puesto que en cualquier caso $\lim_{n\to\infty} |\alpha|^n = 0$.

Si $\alpha = 1$, entonces α^n es siempre igual a uno y se tendrá que considerar al número uno como el límite de α^n.

Si $\alpha > 1$, hágase $\alpha = 1 + h$, donde h es positivo, y se verá de inmediato, de la desigualdad, que cuando n crece α^n no tiende hacia algún límite definido sino que crece más allá de toda cota. Se dice que α^n *tiende a infinito* cuando n crece indefinidamente, o bien que α^n *se torna infinito;* en símbolos,

$$\lim_{n\to\infty} \alpha^n = \infty \qquad (\alpha > 1).$$

Se insiste explícitamente en que *el símbolo ∞ no denota un número y que no puede calcularse con él de acuerdo a las reglas usuales;* asertos que afirmen que una cantidad es o se vuelve infinita jamás tendrán el mismo sentido que una aseveración que involucre cantidades definidas finitas. A pesar de esto, tales modos de expresión y el empleo del símbolo ∞ son extremadamente convenientes, como a menudo se verá en las siguientes páginas.

Si $\alpha = -1$, el valor α^n no tiende hacia límite alguno, pero cuando n crece y recorre la sucesión de enteros positivos α^n toma los valores $+1$ y -1 alternativamente. Análogamente, si $\alpha < -1$ el valor α^n crece numéricamente más allá de toda cota, pero el signo es alternativamente positivo y negativo.

f. Ilustración geométrica de los límites α^n y $\sqrt[n]{p}$

Si consideramos las gráficas de las funciones $y = x^n$ y $y = x^{1/n} = \sqrt[n]{x}$ y, por razones de conveniencia, nos restringimos a valores no negativos

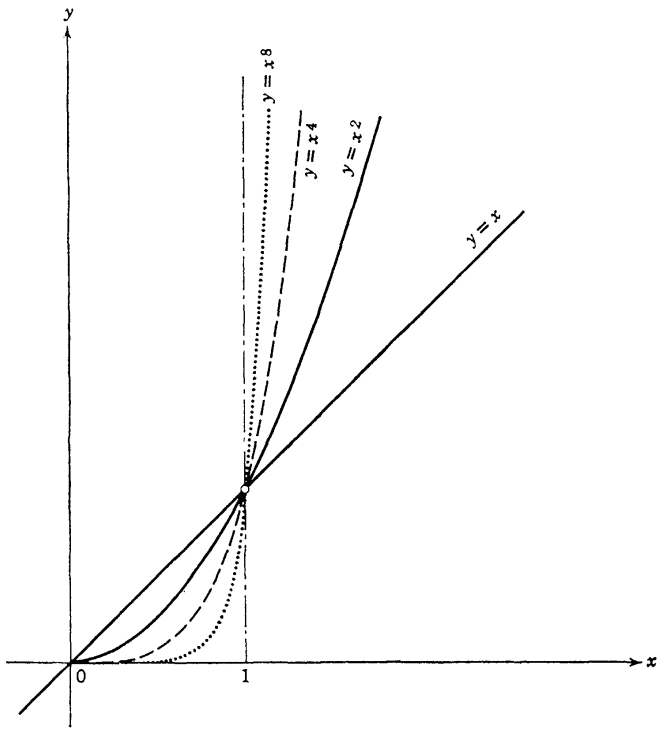


Figura 1.44 x^n conforme n crece.

de x, los límites anteriores están ilustrados en las Figs. 1.44 y 1.45, respectivamente. Se observa que en el intervalo de 0 a 1 las curvas $y = x^n$ se acercan más y más al eje x cuando n crece, mientras que fuera del intervalo se elevan cada vez con mayor pendiente y se aproximan a una

línea paralela al eje y. Todas las curvas pasan por el punto con coordenadas $x = 1$, $y = 1$ y por el origen.

Las gráficas de las funciones $y = x^{1/n} = +\sqrt[n]{x}$, se acercan cada vez más a la línea paralela al eje x y a una distancia 1 por encima de éste; y aquí también todas las curvas deberán pasar por el origen y el punto $(1, 1)$. De aquí que en el límite las curvas aproximan la línea quebrada que consiste de la parte del eje y entre los puntos $y = 0$ e $y = 1$ y de la paralela al eje x, $y = 1$. Además, es evidente que las dos figuras se encuentran muy relacionadas, como podría esperarse del hecho de que las funciones $y = \sqrt[n]{x}$ son las funciones inversas de las n-ésimas potencias, de lo cual se infiere de que para cada n la gráfica de $y = x^n$ es transformada en la de $y = \sqrt[n]{x}$ mediante una reflexión con respecto de la recta $y = x$.

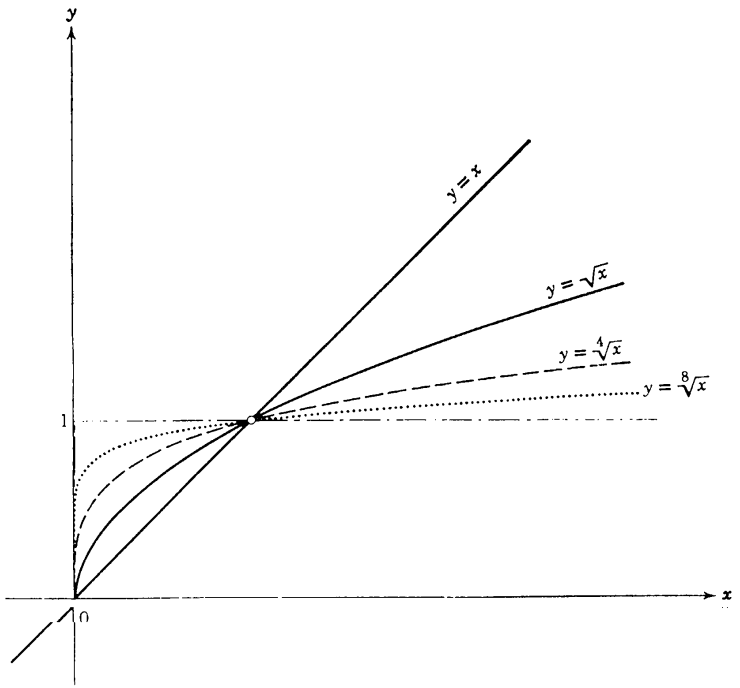


Figura 1.45 $x^{1/n}$ conforme n crece.

g. La serie geométrica

Un ejemplo de un límite muy conocido en las matemáticas elementales es proporcionado por la *serie geométrica:*

$$1 + q + q^2 + \cdots + q^{n-1} = S_n;$$

El número q se denomina la *razón común* o *cociente* de la serie. El valor de esta suma puede, como es bien sabido, expresarse en la forma

$$S_n = \frac{1 - q^n}{1 - q}$$

siempre que $q \neq 1$; y puede obtenerse esta expresión multiplicando la suma S_n por q y restando la ecuación así obtenida de la ecuación original, o bien puede verificarse la fórmula por división.

¿Qué sucede a la suma S_n cuando n crece indefinidamente? La respuesta es la siguiente: La sucesión de sumas S_n tiene un límite definido S si q está situado entre -1 y $+1$, excluyendo estos valores extremos, y

$$S = \lim_{n\to\infty} S_n = \frac{1}{1 - q}.$$

Para verificar esta propiedad se escribe S_n como $(1 - q^n)/(1 - q) = 1/(1 - q) - q^n/(1 - q)$. Se ha probado ya que siempre que $|q| < 1$ la cantidad q^n tiende a cero si n crece; de aquí que, bajo esta suposición, $q^n/(1 - q)$ tiende también a cero y S_n tiende al límite $1/(1 - q)$ cuando n crece.

El paso al límite, $\lim_{n\to\infty} (1 + q + q^2 + \cdots + q^{n-1}) = 1/(1 - q)$, se expresa usualmente diciendo que cuando $|q| < 1$ *la suma de la serie geométrica infinita es la expresión* $1/(1 - q)$.

Las sumas S_n de la serie geométrica finita también se denominan *sumas parciales* de la serie geométrica infinita $1 + q + q^2 + \ldots$ (Debe hacerse una distinción entre la *sucesión* de números q^n y las sumas parciales de la *serie* geométrica.)

El hecho de que las sumas parciales S_n de la serie geométrica tiendan al límite $S = 1/(1 - q)$ conforme n crece se expresa también diciendo que la serie geométrica infinita $1 + q + q^2 + \ldots$ *converge* a la suma $S = 1/(1 - q)$ cuando $|q| < 1$.

De paso, deberá notarse que si q es racional, por ejemplo, $q = \frac{1}{2}$, o bien $q = \frac{1}{3}$, entonces la suma de la serie geométrica infinita tiene un valor racional (en los casos mencionados los valores son 2 y $\frac{3}{2}$, respectivamente). Esta observación está detrás del hecho bien conocido de que las *fracciones decimales periódicas representan siempre números racionales.*[1] La demostración general de esta propiedad resultará evidente del ejemplo del número

$$x = 0.343434\cdots,$$

que puede ser evaluado escribiendo

[1] Véase Courant y Robbins, *What Is Mathematics?*, p. 66.

$$x = \frac{34}{10^2} + \frac{34}{10^4} + \frac{34}{10^6} + \cdots$$
$$= \frac{34}{10^2}\left(1 + \frac{1}{10^2} + \frac{1}{10^4} + \cdots\right)$$
$$= \frac{34}{100}\,\frac{1}{1 - 1/100} = \frac{34}{99}.$$

h. $a_n = \sqrt[n]{n}$

Mostraremos que la sucesión de números

$$a_1 = 1, \quad a_2 = \sqrt{2}, \quad a_3 = \sqrt[3]{3}, \ldots, \quad a_n = \sqrt[n]{n}, \ldots$$

tiende a 1 cuando n crece:

$$\lim_{n\to\infty} \sqrt[n]{n} = 1.$$

Puesto que a_n excede el valor 1, hacemos $a_n = 1 + h_n$, con h_n positivo. Entonces (ver p. 87)

$$n = (a_n)^n = (1 + h_n)^n$$
$$\geq 1 + nh_n + \frac{n(n-1)}{2} h_n^2 \geq \frac{n(n-1)}{2} h_n^2.$$

Se sigue para $n > 1$ que

$$h_n^2 \leq \frac{2}{n - 1};$$

y de aquí que

$$h_n \leq \frac{\sqrt{2}}{\sqrt{n - 1}}.$$

Se tiene ahora

$$1 \leq a_n = 1 + h_n \leq 1 + \frac{\sqrt{2}}{\sqrt{n - 1}}.$$

El miembro derecho de esta desigualdad obviamente tiende a uno, y por lo tanto también a_n.

i. $a_n = \sqrt{n + 1} - \sqrt{n}$

En este ejemplo los a_n son diferencias de dos términos, cada uno de los cuales crece más allá de toda cota. Intentando pasar al límite separadamente con cada uno de los dos términos, se obtiene la expresión simbólica sin significado $\infty - \infty$. En tal caso la existencia de un límite, y lo que pueda ser su valor, depende totalmente del caso especial. Se afirma que en este ejemplo es

$$\lim_{n\to\infty} (\sqrt{n+1} - \sqrt{n}) = 0.$$

Para la demostración se requiere escribir solamente la expresión en la forma

$$\sqrt{n+1} - \sqrt{n} = \frac{(\sqrt{n+1} - \sqrt{n})(\sqrt{n+1} + \sqrt{n})}{\sqrt{n+1} + \sqrt{n}} = \frac{1}{\sqrt{n+1} + \sqrt{n}};$$

y ver de inmediato que tiende a cero conforme n crece.

j. $a_n = \dfrac{n}{\alpha^n}$, **para** $\alpha > 1$

Formalmente, el límite de los a_n es del tipo indeterminado ∞/∞ encontrado ya en el ejemplo *c*. Se afirma que en este ejemplo la sucesión de números $a_n = n/\alpha^n$ tiende al límite cero.

Para la demostración hágase $\alpha = 1 + h$, donde $h > 0$, y hágase uso nuevamente de la desigualdad

$$(1+h)^n \geq 1 + nh + \frac{n(n-1)}{2} h^2 > \frac{n(n-1)}{2} h^2.$$

De aquí que para $n > 1$

$$a_n = \frac{n}{(1+h)^n} < \frac{2}{(n-1)h^2}.$$

Puesto que a_n es positivo y el miembro derecho de esta desigualdad tiende a cero, a_n deberá tender también a cero.

1.7 Discusión del concepto de límite

a. Definición de convergencia y divergencia

De los ejemplos discutidos en la sección 1.6 puede abstraerse el siguiente concepto general de límite:

Supóngase que para una dada sucesión infinita de puntos $a_1, a_2, a_3, \ldots$, existe un número finito l tal que todo intervalo abierto, por pequeño que sea, definido en torno del punto l, contiene todos los puntos a_n excepto, a lo más, un número finito. El número l se denomina entonces límite de la sucesión $a_1, a_2, \ldots$; o bien se dice que la sucesión $a_1, a_2, \ldots$ es convergente y converge a l. En símbolos $\lim_{n\to\infty} a_n = l$.

La siguiente definición de límite es equivalente:

A todo número positivo ε, por pequeño que sea, podemos asignarle un entero $N = N(\varepsilon)$ suficientemente grande tal que a partir del índice N en adelante [esto es, para $n > N(\varepsilon)$] se tiene siempre $|a_n - l| < \varepsilon$.

Por supuesto, es cierto como regla que $N(\varepsilon)$ deberá ser escogido cada vez mayor para valores cada vez menores de la tolerancia o error ε; en otras palabras, $N(\varepsilon)$ usualmente crecerá más allá de toda cota a medida que ε tienda a cero. La noción intuitivamente vaga de límite sugiere una representación de los a_n *moviéndose cada vez más hacia* l. Esta representación es reemplazada aquí por la definición "estática" precisa: Toda vecindad de l contiene a todos los a_n excepto, a lo más, un número finito de ellos.[1]

Obviamente, una sucesión $a_1, a_2, \ldots$, no puede tener más de un límite l. Si, por el contrario, dos números distintos l y l' fueran límites de la misma sucesión $a_1, a_2, \ldots$, podrían determinarse intervalos abiertos en torno de cada uno de los puntos l y l' que no se traslapen. Puesto que cada intervalo contiene todos los a_n excepto un número finito de ellos, la sucesión no podría ser infinita. El límite de una sucesión convergente está por ello determinado unívocamente.

Otra observación evidente pero útil es la siguiente: Si de una sucesión convergente se omite cualquier número de términos, la sucesión que resulta converge al mismo límite que la sucesión original.

Una sucesión que no converge se dice que es divergente. Si a medida que n crece los números a_n crecen más allá de toda cota positiva, se dice que la sucesión *diverge* a $+\infty$; como se ha hecho ya ocasionalmente, se escribe entonces $\lim_{n\to\infty} a_n = \infty$. De manera análoga, se escribe $\lim_{n\to\infty} a_n = -\infty$ si, cuando n crece, los números $-a_n$ crecen más allá de toda cota en la dirección positiva. Sin embargo, la divergencia puede manifestarse por sí misma de otras maneras, tal como para la sucesión $a_1 = -1$, $a_2 = +1$, $a_3 = -1$, $a_4 = +1, \ldots$, cuyos términos oscilan de un sitio a otro entre valores diferentes.

Evidentemente, ni la divergencia ni la convergencia de una sucesión son afectadas si se omite un número finito de términos.

Una sucesión $a_1, a_2, \ldots$ es *acotada* si existe un intervalo finito que contiene *todos* los puntos de la sucesión. Cualquier intervalo finito está contenido en algún intervalo finito que tiene el origen como centro. Por lo tanto, el requisito de que la sucesión sea acotada significa que existe un número M tal que $|a_n| \leq M$ para todo n.

Una sucesión convergente $a_1, a_2, \ldots$ también es necesariamente acotada. En efecto, sea l el límite de la sucesión. Tomando $\varepsilon = 1$ se encuentra a partir de la definición de convergencia que toda a_n a partir de cierta N en adelante está en el intervalo de longitud 2 con centro en l. Los únicos términos a_n de la sucesión que pueden estar fuera de ese intervalo son $a_1, \ldots, a_{N-1}$. Puede entonces, sin embargo, encontrarse un intervalo finito mayor que incluya también a $a_1, \ldots, a_{N-1}$.

[1] El lector notará la analogía con la definición de continuidad de una función $f(x)$ en un punto x_0. El papel que la cantidad suficientemente pequeña $\delta(\varepsilon)$ desempeña ahí es desempeñada aquí por el entero suficientemente grande $N(\varepsilon)$. En realidad, se verá en la p. 105 que la continuidad de una función es un punto puede ser formulada en términos de límites de sucesiones.

b. Operaciones racionales con límites

De la definición de límite se sigue de inmediato que se pueden efectuar las operaciones elementales de adición, multiplicación, substracción y división de límites de acuerdo con las siguientes reglas.

Si $a_1, a_2, \ldots$ es una sucesión con límite a y $b_1, b_2, \ldots$ es una sucesión con límite b, entonces la sucesión de números $c_n = a_n + b_n$ tiene también un límite c, y

$$c = \lim_{n\to\infty} c_n = a + b.$$

La sucesión de números $c_n = a_n b_n$ converge de la misma manera y

$$\lim_{n\to\infty} c_n = ab.$$

Análogamente, la sucesión $c_n = a_n - b_n$ converge y

$$\lim_{n\to\infty} c_n = a - b.$$

Siempre que el límite b difiera de cero, los números $c_n = a_n/b_n$ convergen de manera análoga y tienen el límite

$$\lim_{n\to\infty} c_n = \frac{a}{b}.$$

En palabras: Podemos *intercambiar* las operaciones racionales de cálculo con el proceso de formar límite; se obtiene el mismo resultado ya sea que se efectúe primero el paso al límite y después una operación racional o viceversa.

Las demostraciones de todas estas reglas resultan claras si una de ellas es llevada a cabo. Considérese la multiplicación de límites. Si las relaciones $a_n \to a$ y $b_n \to b$ valen, entonces para todo número positivo ε podemos asegurar ambas desigualdades

$$|a - a_n| < \varepsilon \qquad \text{y} \qquad |b - b_n| < \varepsilon$$

escogiendo n suficientemente grande, digamos $n > N(\varepsilon)$. Si se escribe

$$ab - a_n b_n = b(a - a_n) + a_n(b - b_n)$$

y se recuerda que existe una cota positiva M independiente de n tal que $|a_n| < M$, se obtiene

$$|ab - a_n b_n| \leq |b|\,|a - a_n| + |a_n|\,|b - b_n| < (|b| + M)\varepsilon.$$

Puesto que la cantidad $(|b| + M)\varepsilon$ puede hacerse arbitrariamente pequeña escogiendo ε suficientemente pequeño, la diferencia entre ab y $a_n b_n$ se hace en efecto tan pequeña como se desee para valores suficientemente grandes de n; y ésta es precisamente la aseveración expresada en la ecuación

$$ab = \lim_{n\to\infty} a_n b_n.$$

Usando este ejemplo como modelo, el lector puede demostrar las reglas para las operaciones racionales restantes.

Por medio de estas reglas muchos límites pueden ser evaluados fácilmente; en efecto se tiene:

$$\lim_{n\to\infty} \frac{n^2 - 1}{n^2 + n + 1} = \lim_{n\to\infty} \frac{1 - \frac{1}{n^2}}{1 + \frac{1}{n} + \frac{1}{n^2}} = 1,$$

puesto que en la segunda expresión puede pasarse directamente al límite en el numerador y en el denominador.

La siguiente regla simple es frecuentemente útil: *Si* $\lim a_n = a$ *y* $\lim b_n = b$, *y si, además,* $a_n > b_n$ *para todo* n, *entonces* $a > b$. Sin embargo, de ningún modo hay que esperar que a sea siempre *mayor* que b, como lo muestran las sucesiones $a_n = 1/n$, $b_n = 1/2n$, para las cuales $a = b = 0$.

c. Pruebas intrínsecas de convergencia. Sucesiones monótonas

En todos los ejemplos dados el límite de la sucesión considerada era un número conocido De hecho, para aplicar la definición anterior de límite de una sucesión debe conocerse el límite antes de que pueda verificarse la convergencia. Si del concepto de límite de una sucesión no se obtuviera otra cosa más que el reconocimiento de que algunos números conocidos pueden ser aproximados por ciertas sucesiones de otros números conocidos, se hubiera ganado muy poco con ello. La ventaja del concepto de límite en análisis radica esencialmente en el hecho de que problemas importantes tienen a menudo soluciones numéricas que posiblemente de otra forma no podrían ser conocidas o expresables directamente, pero que pueden ser descritas como límites. Mucho del análisis superior consiste de una sucesión de ejemplos de este hecho que se hará más claro progresivamente en los siguientes capítulos. La representación de los números irracionales como límites de números racionales puede ser interpretada como el primero y típico ejemplo.

Toda sucesión *convergente* de números conocidos $a_1, a_2, \ldots$ *define* un número l, su límite. Sin embargo, la única prueba de convergencia que surge de la definición de convergencia consiste en estimar las diferencias $|a_n - l|$, y esto es aplicable solamente si el número l es ya conocido. Es esencial disponer de criterios "*intrínsecos*" para la convergencia que no requieran de un conocimiento *a priori* del valor del límite sino que involucren solamente los términos de la propia sucesión. La más simple de tales pruebas o criterios se aplica a una clase especial de sucesiones, las sucesiones monótonas, e incluye muchos de los ejemplos importantes.

Límites de sucesiones monótonas

Una sucesión $a_1, a_2, \ldots$ se denomina *monótonamente creciente* si cada término a_n es mayor o no menor que el precedente; esto es,

$$a_n \geq a_{n-1}.$$

Análogamente, la sucesión es *monótonamente decreciente* si $a_n \leq a_{n-1}$ para todo n. Una sucesión *monótona* es aquélla que es monótonamente creciente o bien monótonamente decreciente. Con esta definición se tiene el principio básico:

Una sucesión que es tanto monótona como acotada converge.[1]

Este principio es sugerido convincentemente, mas no probado, por la intuición; está íntimamente relacionado con las propiedades de los números reales, y es de hecho equivalente al axioma de continuidad para los números reales.

El axioma (ver sección 1*b*) de que toda sucesión encaje de intervalos contiene un punto, se ve fácilmente como consecuencia de la convergencia de sucesiones acotadas monótonas. Así, sea $[a_1, b_1], [a_2, b_2], \ldots$ una sucesión de intervalos en encaje. Por la definición de encaje se tiene

$$a_1 \leq a_2 \leq \cdots \leq a_n < b_n \leq b_{n-1} \leq \cdots \leq b_1.$$

Obviamente, la sucesión $a_1, a_2, \ldots$ es monótonamente creciente. También es acotada puesto que $a_1 \leq a_n \leq b_1$ para todo n. De aquí que $l = \lim_{n\to\infty} a_n$ exista. Además, para todo m y para todo número $n > m$ se tiene

$$a_m \leq a_n \leq b_m.$$

Por lo tanto también

$$a_m \leq \lim_{n\to\infty} a_n = l \leq b_m.$$

De aquí que todos los intervalos del encaje anidado contengan uno y el mismo punto l. (El que no tengan otro punto en común se sigue de la propiedad adicional $\lim (b_n - a_n) = 0$ de los encajes de intervalos.)

Criterio de Cauchy para la convergencia

Una sucesión convergente es automáticamente acotada, pero no tiene por qué ser monótona (ver ejemplo *b*, p. 85). Por lo tanto, al tratar con sucesiones generales es deseable disponer de una prueba para la convergencia que sea también aplicable a sucesiones no monótonas. Esta necesidad se satisface mediante una simple condición, la *prueba o criterio de convergencia de Cauchy*. Este criterio caracteriza a las sucesiones de números reales que tienen un límite; y lo más importante es que no requiere un conocimiento *a priori* del valor del límite: *Una condición necesaria y suficiente para la convergencia de una sucesión $a_1, a_2, \ldots$ es que los elementos a_n de la sucesión con un índice n suficientemente grande difieran arbitrariamente poco unos de otros.* Formulado de manera precisa: Una sucesión a_n es convergente si para todo $\varepsilon > 0$ existe un número natural $N = N(\varepsilon)$ tal que $|a_n - a_m| < \varepsilon$ siempre que $n > N$ y $m > N$. Geométricamente, la condición de Cauchy establece que una sucesión converge si existen intervalos arbitrariamente pequeños fuera de los cuales se encuentran solamente un número finito de puntos de la sucesión. La exactitud del criterio de Cauchy para la convergencia será demostrada y discutido su significado en el suplemento.

[1] La hipótesis sobre la acotación es esencial puesto que ninguna sucesión no acotada puede converger. Obsérvese que una sucesión monótona creciente $a_1, a_2, \ldots$ está siempre "acotada desde abajo": $a_n \geq a_1$ para todo n. A fin de probar que una sucesión monótona creciente converge, es suficiente entonces encontrar un número M tal que $a_n \leq M$ para todo n.

d. Series infinitas y el símbolo de sumatoria

Una sucesión es simplemente un arreglo ordenado infinito de números $a_1, a_2, \ldots$ *Una serie infinita*

$$a_1 + a_2 + a_3 + \cdots$$

requiere que los términos sean sumados en el orden en el cual aparecen. Para llegar a un significado preciso de la *suma* de una serie infinita se considera la *n-ésima suma parcial,* esto es, la suma de los primeros n términos de la serie

$$s_n = a_1 + a_2 + \cdots + a_n.$$

Las sumas parciales s_n para n diferentes forman una sucesión

$$s_1 = a_1, \qquad s_2 = a_1 + a_2, \qquad s_3 = a_1 + a_2 + a_3,$$

y así sucesivamente. La suma s de la serie infinita está definida entonces como

$$s = \lim_{n\to\infty} s_n,$$

siempre que el límite exista. En ese caso la serie infinita se denomina *convergente*. Si la sucesión s_n diverge, la serie infinita se denomina *divergente*. Por ejemplo, la sucesión $1, q, q^2, q^3, \ldots$ da lugar a la serie geométrica infinita

$$1 + q + q^2 + q^3 + \cdots$$

cuyas sumas parciales son

$$s_n = 1 + q + q^2 + \cdots + q^{n-1}.$$

Para $|q| < 1$ la sucesión s_n converge hacia el límite

$$s = \frac{1}{1-q},$$

el cual representa entonces la suma de la serie infinita. Para $|q| \geq 1$ las sumas parciales s_n no poseen límite y la serie diverge (ver pp. 90-91).

Es costumbre utilizar para $a_1 + a_2 + \cdots + a_n$ el símbolo

$$\sum_{k=1}^{n} a_k$$

que indica que la suma de los a_k ha de tomarse con k recorriendo los enteros $k = 1$ a $k = n$. Por ejemplo,

$$\sum_{k=1}^{4} \frac{1}{k!} \text{ es el símbolo para } \frac{1}{1!} + \frac{1}{2!} + \frac{1}{3!} + \frac{1}{4!},$$

mientras que

$$\sum_{k=1}^{n} a^k b^{2k} \text{ es el símbolo para } a^1b^2 + a^2b^4 + a^3b^6 + \cdots + a^nb^{2n}.$$

Más generalmente, $\sum_{k=m}^{n} a_k$ significa la suma de todo a_k obtenido dando a k los valores m, $m + 1$, $m + 2$, ..., n. Por tanto

$$\sum_{k=3}^{5} \frac{1}{k!} = \frac{1}{3!} + \frac{1}{4!} + \frac{1}{5!}.$$

En estos ejemplos se ha utilizado la letra k para el índice de sumatoria. Por supuesto, la suma es independiente de la letra que denota este índice. Por lo tanto

$$s_n = \sum_{k=1}^{n} a_k = \sum_{i=1}^{n} a_i.$$

El símbolo

$$\sum_{k=1}^{\infty} a_k$$

se utiliza para denotar la suma de toda la serie infinita. Análogamente, $\sum_{k=0}^{\infty} a_k$ sería el símbolo para la suma de la serie infinita $a_0 + a_1 + a_2 + \ldots$, cuya n-ésima suma parcial es $s_n = a_0 + a_1 + a_2 + \cdots + a_{n-1}$.

Muchos de los resultados anteriores pueden ser escritos más concisamente en esta notación sumatoria. La fórmula de la p. 81 para la suma de los primeros n cuadrados se convierte en

$$\sum_{k=1}^{n} k^2 = \frac{n(n+1)(2n+1)}{6}$$

La fórmula para la suma de la serie geométrica es

$$\sum_{k=0}^{\infty} q^k = \frac{1}{1-q} \qquad \text{para } |q| < 1.$$

Finalmente, el teorema binomial está expresado por

$$(a+b)^n = \sum_{k=0}^{n} \binom{n}{k} a^{n-k}b^k.$$

Puesto que una serie infinita es simplemente el límite de una sucesión s_n, la convergencia puede ser decidida sobre la base de los criterios de convergencia para sucesiones. Por ejemplo, la convergencia de la serie

$$\sum_{k=1}^{\infty} \frac{1}{k^k} = \frac{1}{1^1} + \frac{1}{2^2} + \frac{1}{3^3} + \cdots$$

se sigue inmediatamente del hecho de que las sumas parciales

$$s_n = \sum_{k=1}^{n} \frac{1}{k^k} = \frac{1}{1^1} + \frac{1}{2^2} + \frac{1}{3^3} + \cdots + \frac{1}{n^n}$$

crecen monótonamente con n y están acotadas puesto que

$$1 \leq s_n \leq 1 + \frac{1}{2^2} + \frac{1}{2^3} + \frac{1}{2^4} + \cdots + \frac{1}{2^n}$$

$$= 1 + \frac{1}{4}\frac{1 - 1/2^{n-1}}{1 - \frac{1}{2}} = 1 + \frac{1}{2} - \frac{1}{2^n} < \frac{3}{2}.$$

Posteriormente, en el capítulo 7, se estudiarán las series infinitas más sistemáticamente.

e. El número e

Como primer ejemplo de un número que es generado como el límite de una sucesión, se considera

$$e = 1 + \frac{1}{1!} + \frac{1}{2!} + \frac{1}{3!} + \cdots.$$

Así e es la notación para $\lim_{n\to\infty} S_n$, donde

$$S_n = 1 + \frac{1}{1!} + \frac{1}{2!} + \cdots + \frac{1}{n!}$$ [1]

[1] Recordando la convención que define a 0! como 1, el primer término de la serie puede escribirse como 1/0! en concordancia con la ley de formación de los siguientes términos. Nótese que en nuestra notación S_n es realmente la $(n + 1)$-ésima suma parcial de la serie infinita, en vez de la n-ésima. Esto, sin embargo, no tiene ningún significado.

Los números e y π son las constantes trascendentes más exactamente utilizadas en análisis matemático. Para demostrar la existencia del límite e se requiere solamente probar que la sucesión S_n es acotada, puesto que los números S_n crecen monótonamente. Para todos los valores de n se tiene

$$\begin{aligned} S_n &= 1 + 1 + \frac{1}{2} + \frac{1}{2\cdot 3} + \frac{1}{2\cdot 3\cdot 4} + \cdots + \frac{1}{2\cdot 3\cdot 4\cdot \cdots \cdot n} \\ &\leq 1 + 1 + \frac{1}{2} + \frac{1}{2^2} + \frac{1}{2^3} + \cdots + \frac{1}{2^{n-1}} \\ &= 1 + \frac{1 - 1/2^n}{1 - \frac{1}{2}} < 3. \end{aligned}$$

Los números S_n tienen por lo tanto la cota superior 3, y puesto que forman una sucesión monótona creciente los números poseen un límite, que se denota por e.

La expresión de e como una serie permite rápidamente calcular e con gran precisión. El error que se comete al aproximar e por una suma parcial S_m puede ser estimado por el mismo método de comparación con una serie geométrica que proporcionó la cota superior 3 para e. Para toda $n > m$ se tiene

$$\begin{aligned} S_n &= S_m + \frac{1}{(m+1)!} + \frac{1}{(m+2)!} + \cdots + \frac{1}{n!} \\ &\leq S_m + \frac{1}{(m+1)!}\left[1 + \frac{1}{m+2} + \frac{1}{(m+2)(m+3)} + \cdots\right] \\ &\leq S_m + \frac{1}{(m+1)!}\left[1 + \frac{1}{m+1} + \frac{1}{(m+1)^2} + \cdots\right] \\ &= S_m + \frac{1}{(m+1)!}\,\frac{1}{1 - \frac{1}{m+1}} = S_m + \frac{1}{m}\,\frac{1}{m!}. \end{aligned}$$

Por tanto, para $n > m$

$$S_m < S_n \leq S_m + \frac{1}{m}\,\frac{1}{m!}.$$

Si se hace crecer n más allá de toda cota mientras a m se le mantiene fijo, se encuentra también que

$$S_m < e \leq S_m + \frac{1}{m}\,\frac{1}{m!}.$$

De aquí que e difiere de S_m en a lo más $(1/m)(1/m!)$. Puesto que $m!$ crece muy rápidamente con m, el número S_m es ya una buena aproximación para m bastante pequeño; por ejemplo, S_{10} difiere de e en menos que 10^{-7}. De esta manera se encuentra que $e = 2.718281\cdots$

e es un *número irracional*. La estimación para e en términos de S_m puede utilizarse también para establecer esta propiedad. Así, si e fuera racional, e podría escribirse en la forma p/m con p y m enteros positivos; aquí $m \geq 2$, puesto que e, situado entre 2 y 3, no puede ser un entero. Comparando e con la suma parcial S_m, se tendría

$$S_m < \frac{p}{m} \leq S_m + \frac{1}{m}\,\frac{1}{m!}.$$

Si se multiplican aquí los dos miembros por $m!$ se encuentra que

$$m!\,S_m < p(m-1)! \leq m!\,S_m + \frac{1}{m} < m!\,S_m + 1.$$

Pero

$$m!\,S_m = m! + m! + \frac{m!}{2!} + \frac{m!}{3!} + \cdots + \frac{m!}{m!}$$

es un entero puesto que cada término de la suma lo es. Por lo tanto, si e fuera racional el entero $p(m-1)$ estaría situado entre dos enteros sucesivos, lo cual es imposible.[1]

El número e como límite de $(1 + 1/n)^n$. El número e, que se definió aquí como la suma de una serie infinita, puede obtenerse también como límite de la sucesión

$$T_n = \left(1 + \frac{1}{n}\right)^n.$$

La demostración es sencilla y es al mismo tiempo un ejemplo instructivo de operaciones con límites. De acuerdo con el teorema del binomio,

$$\begin{aligned} T_n &= \left(1 + \frac{1}{n}\right)^n \\ &= 1 + n\,\frac{1}{n} + \frac{n(n-1)}{2!}\,\frac{1}{n^2} + \cdots + \frac{n(n-1)(n-2)\cdots 1}{n!}\,\frac{1}{n^n} \\ &= 1 + 1 + \frac{1}{2!}\left(1 - \frac{1}{n}\right) + \cdots \\ &\qquad + \frac{1}{n!}\left(1 - \frac{1}{n}\right)\left(1 - \frac{2}{n}\right)\cdots\left(1 - \frac{n-1}{n}\right) \end{aligned}$$

[1] La irracionalidad del número e significa que no existe ninguna ecuación lineal $ax + b = 0$ con coeficientes racionales a, b y $a \neq 0$ que tenga a e como solución. Una proposición mucho más fuerte ha sido probada (por Hermite), a saber: que no existe ninguna ecuación polinomial $a_0x^n + a_1x^{n-1} + \cdots + a_{n-1}x + a_n = 0$, de cualquier grado n y con coeficientes racionales $a_0, a_1, \ldots, a_n$ (con $a_0 \neq 0$), con $x = e$ como raíz. Se dice que e es un número *trascendente*, en contraste con los números "algebraicos" como $\sqrt{2}$ o bien $\sqrt[5]{10}$, que son raíces de ciertas ecuaciones polinomiales con coeficientes racionales.

De esto se ve de inmediato que $T_n \leq S_n < 3$. Más aún, puesto que T_{n+1} se obtiene de T_n reemplazando los factores $1 - 1/n$, $1 - 2/n$, ... por los factores mayores $1 - 1/(n+1)$, $1 - 2/(n+1)$, ..., y finalmente añadiendo un término positivo, se ve que los T_n forman también una sucesión monótona creciente de la cual se sigue la existencia del límite $\lim_{n\to\infty} T_n = T$.

Para probar que $T = e$, se observa que para $m > n$

$$T_m > 1 + 1 + \frac{1}{2!}\left(1 - \frac{1}{m}\right) + \cdots + \frac{1}{n!}\left(1 - \frac{1}{m}\right)\cdots\left(1 - \frac{n-1}{m}\right).$$

Si ahora se mantiene n fijo y se deja crecer m más allá de toda cota, se obtiene en el miembro izquierdo el número T y en el derecho la expresión S_n, de modo que $T \geq S_n$. Por lo tanto $T \geq S_n \geq T_n$ para todo valor de n. Se hace ahora crecer n, de modo que T_n tiende a T; y de la doble desigualdad se sigue que $T = \lim_{n\to\infty} S_n = e$. Esta era la propiedad por demostrar.

Posteriormente (sección 2.6, p. 171) este número e se obtendrá aún nuevamente desde otro punto de vista.

f. El número π como límite

Un proceso límite que en esencia se remonta a la antigüedad clásica (Arquímedes) es aquél mediante el cual se define el número π. Geométricamente π significa el área de un círculo de radio uno. Se considera como cosa obvia que esta área puede ser expresada por un número (racional o irracional) denotado por π. Sin embargo, esta definición no es de gran ayuda para nosotros si se desea calcular el número con toda precisión. Entonces no se tiene más alternativa que la de representar el número por medio de un proceso límite, a saber, como el límite de una sucesión de números conocidos y fácilmente calculados. Arquímedes utilizó ya este proceso en su método de exhaucción, el cual consiste en aproximar el círculo por medio de polígonos regulares con un número creciente de lados que se ajustan a él cada vez más. Si f_m denota el área de un m-ágono (polígono de m lados) regular inscrito en el círculo, el área del $2m$-ágono inscrito está dado por la fórmula [probada por geometría elemental o bien por la expresión $f_n = (n/2)\ \text{sen}\ (2\pi/n)$ (ver Fig. 1.46)]

$$f_{2m} = \frac{m}{2}\sqrt{2 - 2\sqrt{1 - (2f_m/m)^2}}.$$

Déjese recorrer ahora a m, no la sucesión de todos los enteros positivos, sino por la sucesión de potencias de 2, esto es, $m = 2^n$; en otras palabras, se forman aquellos polígonos regulares cuyos vértices se obtienen mediante

bisección repetida de la circunferencia. Es claro de la interpretación geométrica que los f_{2^n} forman una sucesión creciente y acotada, que por lo tanto *tiene* un límite que es el área del círculo:

$$\pi = \lim_{n\to\infty} f_{2^n}.$$

Esta representación de π como límite sirve en realidad como base para cálculos numéricos; así, partiendo con el valor $f_4 = 2$ pueden calcularse en orden los términos de la sucesión que tienden a π. Una estimación de la precisión con la cual cualquier término f_{2^n} representa a π puede obtenerse construyendo las rectas que tocan al círculo y son paralelas a los lados del 2^n-ágono inscrito. Estas rectas forman un polígono circunscrito, análogo al 2^n-ágono inscrito, que tiene dimensiones mayores en la razón $1 : \cos(\pi/2^n)$. Por lo tanto, el área F_{2^n} del polígono circunscrito puede encontrarse a partir de la razón dada por

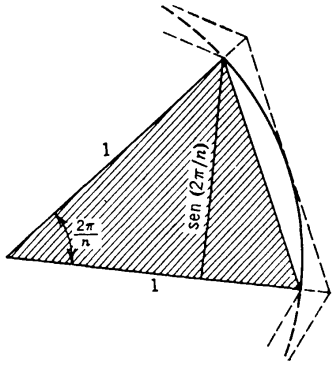


Figura 1.46.

$$\frac{f_{2^n}}{F_{2^n}} = \left(\cos\frac{\pi}{2^n}\right)^2.$$

Puesto que el área del polígono circunscrito es mayor que la del círculo, se tiene

$$f_{2^n} < \pi < F_{2^n} = \frac{f_{2^n}}{\left(\cos\frac{\pi}{2^n}\right)^2} = \frac{2f_{2^n}}{1 + \sqrt{1 - (f_{2^n}/2^{n-1})^2}}.$$

Por ejemplo, $f_8 = 2\sqrt{2}$, de modo que se tiene la estimación

$$2\sqrt{2} < \pi < \frac{4\sqrt{2}}{1 + \frac{1}{2}\sqrt{2}}.$$

Estas son cuestiones con las cuales el lector estará más o menos familiarizado. Sin embargo, lo que se desea hacer resaltar es que el cálculo de áreas por medio de exahucción por figuras rectilíneas cuyas áreas pueden ser calculadas fácilmente constituye la base para el concepto de integral, el que se introducirá en el capítulo 2. Para el cálculo numérico efectivo de π se dispone de métodos mucho más eficientes como se verá en la sección 6.26.

1.8 El concepto de límite para funciones de una variable continua

Hasta ahora se han considerado límites de sucesiones, esto es, de funciones de una variable entera n. La noción de límite, sin embargo, se encuentra frecuentemente en relación con una función $f(x)$ que está definida para toda x en algún intervalo.

Se dice que el valor de la función $f(x)$ tiende a un límite η cuando x tiende a ξ, o sea, en símbolos,

$$\lim_{x\to\xi} f(x) = \eta,$$

si $f(x)$ difiere arbitrariamente poco de η para toda x para la cual $f(x)$ está definida y que está situada suficientemente cerca de ξ.[1] Expresada más precisamente, la definición de $\lim f(x)$ es como sigue.

Si para una cantidad positiva arbitraria ε asignada puede definirse un intervalo $|x - \xi| < \delta$ suficientemente pequeño de modo que para toda x que pertenece tanto al dominio de f como al intervalo la desigualdad $|f(x) - \eta| < \varepsilon$ se satisface, entonces $\lim_{x\to\xi} f(x) = \eta$.

Existe una conexión muy estrecha entre los conceptos de límite de una función y continuidad. Si ξ pertenece al dominio de f, esto es, si $f(\xi)$ está definido, entonces $\lim_{x\to\xi} f(x)$, si existe, debe tener el valor $f(\xi)$. Realmente, la definición de $\eta = \lim_{x\to\xi} f(x)$ implica en particular $|f(\xi) - \eta| < \varepsilon$ para todo ε positivo, y por lo tanto $\eta = f(\xi)$. Ahora, comparando las definiciones de límite y de continuidad se ve que la relación $\lim_{x\to\xi} f(x) = f(\xi)$ justamente expresa la continuidad de la función f en el punto ξ. Luego, para ξ en el dominio de f, la existencia de $\lim_{x\to\xi} f(x)$ significa justamente que f es continua en ξ. Más generalmente, si $f(x)$ no está definida en ξ pero $\lim_{x\to\xi} f(x)$ existe y tiene el valor η, puede asignarse a f en el punto ξ el valor η, y la función f, completada de este modo, será continua en ξ. (Singularidad Removible. Ver p. 59.)

El *límite de una función* puede ser también completamente descrito en términos de *límites de sucesiones*. El aserto

$$\lim_{x\to\xi} f(x) = \eta$$

significa que

$$\lim_{n\to\infty} f(x_n) = \eta$$

para toda sucesión x_n con límite ξ (donde, claro está, se supone que los x_n pertenecen al dominio de f). Pues si $\lim_{x\to\xi} f(x) = \eta$ y si $\lim_{n\to\infty} x_n = \xi$, entonces $f(x)$ está

[1] Se supone aquí que arbitrariamente cerca de ξ existen puntos donde f está definida.

arbitrariamente cerca de η para x suficientemente cercano a ξ; pero x_n está suficientemente cercano a ξ sólo si n es suficientemente grande, y consecuentemente $\lim_{n\to\infty} f(x_n) = \eta$. Si, por otra parte, $\lim_{n\to\infty} f(x_n) = \eta$ para $x_n \to \xi$, debe tenerse también $\lim_{x\to\xi} f(x) = \eta$. Pues de lo contrario existiría un número positivo ε tal que $|f(x) - \eta| \geq \varepsilon$ para algún x arbitrariamente cercano a ξ; también existiría entonces una sucesión x_n convergente a ξ para la cual $|f(x_n) - \eta| \geq \varepsilon$; pero entonces $\lim_{n\to\infty} f(x_n)$ no podría ser igual a η.

La continuidad de la función $f(x)$ en el punto ξ implica entonces: $\lim_{n\to\infty} f(x_n) = f(\xi)$, para toda sucesión x_n en el dominio de f que converge a ξ. Más generalmente, para una función continua en el *intervalo* la relación

$$\lim_{n\to\infty} f(x_n) = f(\lim_{n\to\infty} x_n)$$

es válida para cualquier sucesión en el dominio de f que converge a un punto del intervalo. Se ve que para una función continua el símbolo de límite puede ser intercambiado (o bien, como se dice, "conmuta") con los símbolos para la función.

Límites de sumas, productos y cocientes de funciones se determinan por las mismas reglas que para las sucesiones (ver p. 94): Si $\lim_{x\to\xi} f(x) = \eta$ y $\lim_{x\to\xi} g(x) = \zeta$ existen, entonces

$$\lim_{x\to\xi} (f(x) + g(x)) = \eta + \zeta. \qquad \lim_{x\to\xi} (f(x)g(x)) = \eta\zeta;$$

y para $\zeta \neq 0$ también

$$\lim_{x\to\xi} \frac{f(x)}{g(x)} = \frac{\eta}{\zeta}$$

Las demostraciones son las mismas que para sucesiones. (Las reglas se seguirían también de aquéllas para sucesiones, escribiendo límites de funciones como límites de sucesiones.) Consecuentemente, cuando ξ pertenece al dominio de f y g, *la suma, producto y cociente de dos funciones* $f(x)$ *y* $g(x)$ *que son continuas en un punto* ξ *son también continuas* (donde para los cocientes se debe suponer que $g(\xi) \neq 0$).

Los casos en que ξ no pertenece al dominio de f resultarán ser de particular importancia para el cálculo diferencial. Como un primer ejemplo, considérese la relación

$$\lim_{x\to\xi} \frac{x^n - \xi^n}{x - \xi} = n\xi^{n-1},$$

donde n es un entero positivo. Por supuesto, $f(x) = (x^n - \xi^n)/(x - \xi)$ es una función definida solamente para $x \neq \xi$. Pero para $x \neq \xi$ la identidad algebraica

$$\frac{x^n - \xi^n}{x - \xi} = x^{n-1} + x^{n-2}\xi + x^{n-3}\xi^2 + \cdots + \xi^{n-1},$$

es válida como consecuencia de la fórmula sumatoria para la serie geométrica. Para encontrar el límite se requiere solamente hacer tender x a ξ y evaluar el límite del lado derecho por las reglas para límites de sumas y cocientes.

Menos obvia es la fórmula

$$\lim_{x\to 0} \frac{\operatorname{sen} x}{x} = 1$$

(donde, por supuesto, el ángulo x es medido en "radianes", como se explicó en las pp. 73-74). Aquí también, el cociente $(\operatorname{sen} x)/x$ está definido solamente para $x \neq 0$.

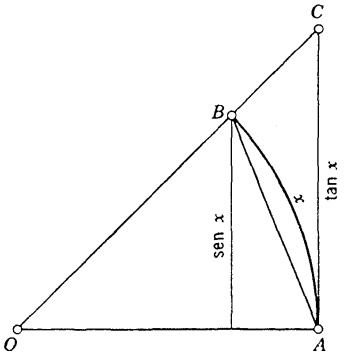


Figura 1.47.

Pero si se *define* $(\operatorname{sen} x)/x = 1$ para $x = 0$, el cociente se completa como una función que es continua también en $x = 0$. Para la demostración de la fórmula límite se apela aquí a un argumento geométrico.

De la Fig. 1.47 se encuentra, comparando las áreas de los triángulos OAB y OAC y el sector OAB [1] del círculo unitario, que, si $0 < x < \pi/2$, es

$$\tfrac{1}{2} \operatorname{sen} x < \tfrac{1}{2} x < \tfrac{1}{2} \tan x.$$

De aquí se sigue que si $0 < |x| < \pi/2$, es

$$1 < \frac{x}{\operatorname{sen} x} < \frac{1}{\cos x}.$$

Por lo tanto, el cociente $(\operatorname{sen} x)/x$ está situado entre los números 1 y $\cos x$. Se sabe que $\cos x$ tiende a 1 cuando $x \to 0$, y de esto se sigue que el cociente $(\operatorname{sen} x)/x$ puede sólo diferir arbitrariamente poco de 1 si x está

[1] Por supuesto, el ángulo x pudo haberse definido en primer término como el doble del área del sector OAB.

suficientemente cerca de 0. Esto es exactamente lo que se entiende por la ecuación que había que demostrar.

Del resultado acabado de probar se sigue que

$$\lim_{x\to 0} \frac{\tan x}{x} = \lim_{x\to 0} \frac{\operatorname{sen} x}{x} \lim_{x\to 0} \frac{1}{\cos x} = 1,$$

y también

$$\lim_{x\to 0} \frac{1 - \cos x}{x} = 0.$$

Esto último se sigue de la fórmula, válida para $0 < |x| < \pi/2$,

$$\frac{1 - \cos x}{x} = \frac{(1 - \cos x)(1 + \cos x)}{x(1 + \cos x)} = \frac{1 - \cos^2 x}{x(1 + \cos x)}$$

$$= \frac{\operatorname{sen} x}{x} \cdot \frac{1}{1 + \cos x} \cdot \operatorname{sen} x.$$

Para $x \to 0$ el primer factor en el lado derecho tiende a 1, el segundo a $\frac{1}{2}$, y el tercero a 0, como se estableció antes.

Dividiendo la misma fórmula por x, se obtiene

$$\frac{1 - \cos x}{x^2} = \left(\frac{\operatorname{sen} x}{x}\right)^2 \frac{1}{1 + \cos x},$$

de la cual surge

$$\lim_{x\to 0} \frac{1 - \cos x}{x^2} = \frac{1}{2}$$

Límites para $x \to \infty$. Finalmente se hace notar que igualmente es posible considerar procesos límite en los cuales la variable continua x crece más allá de toda cota. Por ejemplo, es claro el significado de la ecuación

$$\lim_{x\to\infty} \frac{x^2 + 1}{x^2 - 1} = \lim_{x\to\infty} \frac{1 + 1/x^2}{1 - 1/x^2} = 1.$$

Significa que la función en la izquierda difiere arbitrariamente poco de uno, bastando sólo con que x sea suficientemente grande. Las reglas para formar límites de este tipo para sumas, productos y cocientes son las mismas que antes.

* Existe un resultado adicional el cual es frecuentemente útil en el cálculo de límites, y es la regla para obtener el límite de una función compuesta. La función *compuesta* $f(g(z))$ está definida para aquellos valores de z para los cuales $x = g(z)$ está situado en el dominio de $f(x)$. La función $g(z)$ puede ser una función de una variable continua o bien una variable entera, pero $f(x)$ debe ser una función de una variable continua.

Si $\lim_{z\to\zeta} g(z) = \xi$, donde ξ está situado dentro de un intervalo abierto del dominio de f, y si $\lim_{x\to\xi} f(x) = \eta$, entonces $\lim_{z\to\zeta} f(g(z)) = \eta$. Como un corolario, se observa que una función continua de una función continua es asimismo continua (como se mencionó ya en la p. 78).

El resultado es obvio del hecho de que $f(x)$ puede hacerse arbitrariamente cercano a η tomando x suficientemente cercano a ξ, y para hacer $x = g(z)$ suficientemente cercano a ξ bastará solamente con tomar z suficientemente cercano a ζ. Con algunas ligeras modificaciones se aplican los mismos argumentos que valen en el caso en que cualquiera de las variables puede crecer más allá de toda cota.

a. Algunas observaciones sobre las funciones elementales

Hasta aquí tácitamente se ha supuesto que las funciones elementales son continuas. La demostración de este hecho es muy simple. En primer lugar, la función $f(x) = x$ es continnua; por ello, $x^2 = x \cdot x$ es continua, como producto de dos funciones continuas, y toda potencia de x es, de la misma manera, continua. Por lo tanto, todo polinomio es continuo, por ser la suma de funciones continuas. Toda función racional, como cociente de funciones continuas, es análogamente continua en todo intervalo en el cual el denominador no se hace cero.

La función x^n es continua y monótona para $x \geq 0$. Por lo tanto, la n-ésima raíz, que es la función inversa de la n-ésima potencia, es continua. De este hecho es fácil concluir que la n-ésima raíz de una función racional es continua (excepto donde el denominador se hace cero).

La continuidad de las funciones trigonométricas puede ser probada ahora, utilizando los conceptos ya desarrollados. Sin embargo, la discusión se omite aquí puesto que en el capítulo 2 (p. 188) se verá que la continuidad de todas estas funciones se sigue fácilmente como consecuencia de su diferenciabilidad.

Se harán solamente algunos comentarios en torno a la definición y continuidad de la función exponencial a^x, de la función potencia general, x^α, y del logaritmo. Se supone, como en la sección 1.3 (p. 74), que a es un número positivo, digamos, mayor que uno, y $r = p/q$ es un número racional positivo (siendo p y q enteros); entonces $a^r = a^{p/q}$ es el número positivo cuya q-ésima potencia es a^p. Si α es cualquier número irracional y $r_1, r_2, \ldots, r_m, \ldots$ es una sucesión de números racionales que se aproximan a α, se afirma que $\lim_{m\to\infty} a^{r_m}$ existe; y este límite se denomina entonces a^α.

Para probar la existencia de este límite por medio del criterio de Cauchy se requiere mostrar solamente que $|a^{r_n} - a^{r_m}|$ es arbitrariamente pequeño siempre que n y m sean suficientemente grandes. Se supone, por ejemplo, que $r_n > r_m$, o bien que $r_n - r_m = \delta$, donde $\delta > 0$. Entonces,

$$a^{r_n} - a^{r_m} = a^{r_m}(a^\delta - 1).$$

Puesto que los r_m convergen a α, éstos son acotados y también lo son los a^{r_m}; por lo tanto, es suficiente mostrar que

$$|a^\delta - 1| = a^\delta - 1$$

es arbitrariamente pequeño cuando los valores de n y m son suficientemente grandes. Sin embargo, ciertamente el número racional δ puede hacerse tan pequeño como se desee siempre que los valores de n y m sean suficientemente grandes. Por lo tanto, si l es un entero positivo arbitrariamente grande, $\delta < 1/l$ si n y m son suficientemente grandes. Ahora las relaciones $\delta < 1/l$ y $a > 1$ dan [1]

$$1 < a^\delta < a^{1/l},$$

y puesto que $a^{1/l}$ tiende a uno conforme l crece hacia infinito (ver p. 87), se concluye inmediatamente que vale el aserto y el límite existe.

Puede mostrarse que la función a^x extendida de esta manera a valores irracionales es también continua en todas partes, y, aún más, que es monótona. Para valores negativos de x esta función está naturalmente definida por la ecuación

$$a^x = \frac{1}{a^{-x}}.$$

Conforme x varía de $-\infty$ a $+\infty$, a^x toma valores entre cero y $+\infty$. Consecuentemente, posee una función inversa que es continua y monótona, la cual se denomina el *logaritmo de base a*. De manera parecida puede probarse que la potencia general x^α es una función continua de x, donde α es cualquier número fijo racional o irracional y x varía en el intervalo $0 < x < \infty$, la cual es monótona si $\alpha \neq 0$.

La discusión "elemental" de la función exponencial, del logaritmo y de la potencia x^α, delineada aquí será reemplazada posteriormente (p. 171) por otra discusión en principio mucho más simple.

Suplemento

Uno de los grandes logros de los matemáticos griegos fue la reducción de asertos matemáticos y teoremas, en forma lógica y coherente, a un pequeño número de postulados o axiomas muy simples, los bien conocidos axiomas de la geometría, o bien las reglas de la aritmética, que gobiernan relaciones entre unos cuantos objetos básicos, tales como los números enteros y los puntos geométricos. Los objetos matemáticos básicos se originaron como abstracciones o idealizaciones de la realidad física. Los axiomas, ya sea considerados como "evidentes" desde un punto de vista filosófico o bien meramente como abrumadoramente plausibles, se aceptan sin demostración; sobre ellos reposa la cristalizada estructura de las matemáticas. Por muchos siglos la axiomática matemática euclidiana fue aceptada como un modelo para el estilo matemático, y aun imitada para otras tareas intelectuales. (Por ejemplo, filósofos tales como Descartes y Spinoza trataron de hacer sus especulaciones más convincentes presentándolas axiomáticamente, o bien, como ellos decían, "more geometrico".)

[1] Esta propiedad surge del hecho de que para $a > 1$ la potencia $a^{m/n}$ es mayor que 1 si m/n es positivo. Pues $a = (a^{1/n})^n$ es el producto de m factores todos mayores que 1 y así es mayor que uno.

El método axiomático fue descartado cuando, después del estancamiento durante el medioevo, las matemáticas junto con la ciencia natural principiaron un desarrollo explosivamente vigoroso basado en el nuevo cálculo. Extendiendo vastamente el alcance de las matemáticas, los descubridores ingeniosos no podían ser estorbados teniendo que sujetar los nuevos descubrimientos a un análisis lógico riguroso, y por ello en el siglo XVII la invocación a la evidencia intuitiva se convirtió en un substituto ampliamente utilizado en cambio de la demostración deductiva. Matemáticos de primer orden operaron con los nuevos conceptos guiados por un infalible sentido para la exactitud de los resultados, aun, en ocasiones, con místicas asociaciones, como en las referencias a los "infinitesimales" o "cantidades infinitamente pequeñas". La confianza en la arrolladora potencia de las nuevas manipulaciones del cálculo transportó a los investigadores muy lejos, en trayectorias imposibles de seguir si hubiesen estado sujetos a las limitaciones del complejo rigor lógico. Solamente el instinto seguro de grandes maestros pudo prevenir en contra de graves errores.

El entusiasmo poco crítico pero enormemente fructífero del primer período se fue encontrando gradualmente con contracorrientes que se elevaron con gran fuerza en el siglo XIX pero que no impidieron el desarrollo del análisis constructivo iniciado anteriormente. Muchos de los grandes matemáticos del siglo XIX, en particular Cauchy y Weierstrass, desempeñaron un papel en el esfuerzo hacia la crítica revaluación. El resultado no fue sólo un nuevo y firme fundamento del análisis, sino que también crecieron la lucidez y la simplicidad como bases para notables progresos posteriores.

Una meta importante era la de reemplazar la confianza sin distinción sobre la "intuición" imprecisa por un razonamiento preciso basado en operaciones con números, pues el pensamiento geométrico intuitivo deja un margen indeseable de vaguedad, como se verá con frecuencia en los siguientes capítulos. Por ejemplo, el concepto general de curva continua escapa a la intuición geométrica. Una curva continua, que representa una función continua, como se definió anteriormente, puede no tener una dirección definida en cada punto; más aún, pueden construirse funciones continuas cuyas gráficas no tienen dirección en ninguna parte o bien a las cuales no puede asignárseles ninguna longitud.

Empero, jamás debe olvidarse que el razonamiento abstracto deductivo es meramente un aspecto de las matemáticas, mientras que la motivación impulsora y el gran alcance universal del análisis emanan de la realidad física y de la geometría intuitiva.

Este suplemento proporcionará un apoyo riguroso (con algunas repeticiones) para los conceptos básicos tratados intuitivamente antes en este capítulo.

S.1 Los límites y el concepto de número

Empezaremos con las ideas de la sección 1.1, analizando de lleno el concepto de número real y su conexión con el de límite. Se define el continuo de números por un procedimiento constructivo basado en los números naturales. Se probará entonces que el concepto extendido de número satisface las reglas de aritmética y los otros requisitos, haciendo de éste la herramienta adecuada para la medición.

Puesto que una exposición completa requeriría un libro separado,[1] se indicarán solamente los pasos principales. El estudiante, esforzándose a través del material algo tedioso, se maravillará del hecho de que sobre la base de los números naturales la mente humana puede erigir un sistema de números lógico y consistente, formidablemente adecuado para la labor de medición científica.[2]

a. Los números racionales

Límites definidos por intervalos racionales. Se comienza aceptando el sistema de números racionales con todas sus propiedades usuales, obtenidas de las propiedades básicas de números naturales. Así, los números racionales están *ordenados* por magnitud, permitiéndose definir intervalos "racionales" como conjuntos de números racionales situados entre dos números racionales dados (los intervalos que incluyen los puntos extremos se denominan *cerrados*). La longitud del intervalo con puntos extremos a, b es $|b - a|$. Como se observa en la sección 1*a*, los números racionales son *densos* y todo intervalo racional contiene una infinidad de números racionales. Por el momento todas las cantidades que ocurran se supondrán números racionales.

Dentro del dominio de los números racionales se definen sucesiones y límites (ver p. 93). Dada una sucesión infinita de números racionales $a_1, a_2, \ldots$ y un número *racional* r se dice que

$$\lim_{n\to\infty} a_n = r$$

[1] Véase por ejemplo, E. Landau, *Foundations of Analysis*, 2nd. Ed., Chelsea, New York, 1960.

[2] Los números reales pueden ser introducidos también de manera puramente axiomática, con todas sus propiedades básicas aceptadas como *axiomas*. En el tratamiento que se hará aquí se aceptan, en principio, solamente los axiomas para números naturales (incluyendo el principio de inducción matemática). Los números racionales y los números reales son *construidos* entonces sobre esta base. Los "axiomas" para números reales son entonces, en principio, simplemente *teoremas* sobre los números naturales, para los cuales se requieren demostraciones. En realidad, se comenzará ya con los números racionales como elementos conocidos, puesto que la construcción de los números racionales a partir de los números naturales y la deducción de las propiedades básicas de los números racionales no presentan ninguna dificultad.

si todo intervalo racional que contiene a r en su interior contiene también a *"casi todo"* a_n, esto es, todos los a_n excepto a lo más un número *finito*. Se sigue de inmediato que una sucesión de números racionales no puede tener más de un límite racional y que las reglas usuales para límites de sumas, diferencias, productos y cocientes (ver pp. 94-95) son válidas para sucesiones de números racionales con límites racionales.

Una consecuencia enteramente obvia de esta definición es que pasando al límite se preserva el orden: si $\lim a_n = a$, $\lim b_n = b$ y para todo n es $a_n \leq b_n$, entonces $a \leq b$. Nótese que aún suponiendo $a_n < b_n$ estrictamente, no puede decirse más que $a \leq b$, o bien excluir la posible igualdad de los límites (por ejemplo, las sucesiones $a_n = 1 - 2/n$ y $b_n = 1 - 1/n > a_n$ poseen el mismo límite 1).

Las aseveraciones en torno a límites pueden ser expresadas en términos de *sucesiones nulas* racionales, esto es, sucesiones $a_1, a_2, \ldots$ de números racionales para los cuales

$$\lim_{n\to\infty} a_n = 0.$$

Se dice a_n "se torna arbitrariamente pequeño conforme n tiende a infinito", dándose a entender con ello que para todo racional positivo ε, no importa cuán pequeño, se satisface la desigualdad $|a_n| < \varepsilon$ para casi todo n. Obviamente la sucesión $a_n = 1/n$ es una sucesión nula.

Así, una sucesión de números racionales a_n tiene el límite racional r si y sólo si los números $r - a_n$ forman una sucesión nula.

b. Números reales determinados por encaje de intervalos racionales

Se observó en la p. 29 que intuitivamente los puntos racionales son densos en el eje real y que existen siempre números racionales entre dos números reales cualesquiera. Esto sugiere la posibilidad de *definir* rigurosamente un número real totalmente en términos de relaciones de orden respecto a los racionales, procedimiento que se seguirá a continuación.

Una "sucesión anidada" o *encaje* de intervalos racionales (pp. 32-33) es una sucesión de intervalos *cerrados* J_n con puntos extremos *racionales* a_n, b_n, con cada intervalo contenido en el precedente, cuyas longitudes forman una sucesión nula

$$a_{n-1} \leq a_n \leq b_n \leq b_{n-1}$$

y

$$\lim_{n\to\infty} (b_n - a_n) = 0.$$

Puesto que cada intervalo $J_n = [a_n, b_n]$ de un encaje contiene todos los intervalos subsiguientes, un número racional r situado fuera de cualquier J_n está situado también fuera y sobre el mismo lado de todos los intervalos

subsiguientes. Así, un encaje de intervalos racionales da lugar a una separación de todos los números racionales en tres clases.[1] La primera clase consiste de los números racionales r situados a la izquierda de los intervalos J_n para n suficientemente grande, o sea, para los cuales $r < a_n$ para casi todo n. La segunda clase consiste de los números racionales r contenidos en todos los intervalos J_n. Esta clase contiene a lo más un número, puesto que la longitud del intervalo J_n se reduce a cero con n creciente. La tercera clase consiste de los números racionales r para los cuales $r > b_n$ para casi todo n. Es claro que todo número de la primera clase es menor que cualquiera de la segunda clase, y cualquier número de la segunda clase es menor que cualquiera de la tercera. Los propios puntos a_n son de la primera o bien de la segunda clase, y los números b_n están en la segunda o bien en la tercera.

Si la segunda clase no es vacía, consiste de un solo número racional r. En este caso la primera clase consiste de los números racionales menores que r y la tercera clase de los números racionales mayores que r. Se dice entonces que el encaje de intervalos J_n *representa* el número racional r. Por ejemplo, el encaje de intervalos $[r - 1/n, r + 1/n]$ representa el número r.

Si la segunda clase es vacía, el encaje no representa un número racional; estas sucesiones anidadas sirven entonces para representar números irracionales. Los intervalos individuales $[a_n, b_n]$ de la sucesión no son importantes para este propósito; solamente la separación de los números racionales en tres clases, generada por esta sucesión es esencial, diciéndonos dónde el número irracional encaja entre los racionales.

Así, dos encajes de intervalos racionales $[a_n, b_n]$ y $[a_n', b_n']$ se denominan *equivalentes* si dan lugar a la misma separación de los números racionales en tres clases. El lector deberá probar como ejercicio que es necesario y suficiente para la equivalencia que: $a_n' - a_n$ sea una sucesión nula, o también: que las desigualdades

$$a_n \leq b_n', \qquad a_n' \leq b_n$$

se cumplan para todo n.

A un *encaje de intervalos racionales* $[a_n, b_n]$ se le asigna un *número real. Los números reales determinados por dos encajes distintos se considerarán iguales si las sucesiones son equivalentes.* Un número real está representado entonces por la separación de los números racionales en tres clases generada por encajes equivalentes de intervalos racionales. Si la segunda clase consiste de un número racional r, se considera al número real representado por esta separación en clases como idéntico con el número racional r.

[1] Llamada "Cortadura de Dedekind".

*c. Orden, límites y operaciones aritméticas para números reales

Habiendo definido los números reales, pueden definirse ahora las nociones de orden, suma, diferencia, producto, límite, etc., para éstos y probar que poseen las propiedades usuales. Para ser lógicamente coherentes, cualquier definición concerniente a números reales deberá: (1) poseer el significado ordinario en caso de que los números reales sean racionales, y (2) ser independiente de los encajes individuales de intervalos utilizados para representar números reales.

Intervalos con puntos extremos reales

Si bien hasta ahora, aún para la definición de números irracionales, los puntos extremos de los intervalos anidados fueron supuestos racionales, deberán quitarse aquí tales restricciones y mostrar que se puede operar con números reales exactamente como se hace con números racionales.

Los números reales se denotarán por letras $x, y, \ldots$. Si el número real x está dado por el encaje de intervalos racionales $[a_n, b_n]$ se escribe $x \sim \{[a_n, b_n]\}$. De la definición de número real se infiere una definición natural de orden para un número real, $x \sim \{[a_n, b_n]\}$, relativa a un número racional r. Se dice que $r < x$, $r = x$, $r > x$, según que r pertenezca a la primera, segunda, o tercera clase de la separación de los números racionales generada por el encaje de intervalos. Esta definición es obviamente independiente del encaje especial $\{[a_n, b_n]\}$ que define a x, y posee el significado ordinario cuando x es racional. Equivalentemente, se dice que $r < x$ si $r < a_n$ para casi todo n, $r = x$ si $a_n \leq r \leq b_n$ para todo n, y $r > x$ si $r > b_n$ para casi todo n.

Mediante la comparación de números reales con números racionales pueden compararse los números reales el uno con el otro. Sean $x \sim \{[a_n, b_n]\}$, $y \sim \{[\alpha_n, \beta_n]\}$. Se dice que $x < y$ si existe un número racional r tal que $x < r < y$. Claramente, esta definición no depende de las representaciones particulares de x e y mediante encajes puesto que las comparaciones con r racional son independientes de tales representaciones. Así, se dice que $x < y$ si existe un número racional r tal que $b_n < r < \alpha_n$ para casi todo n, o sea, simplemente, si $b_n < \alpha_n$ para casi todo n. La relación $x < y$ impide la posibilidad de que $y < x$ o $x = y$. Obviamente, $x < y$ e $y < z$ implican $x < z$.

Para dos números reales cualesquiera, x, y, se debe cumplir una de las relaciones $x < y$, $x = y$, $y < x$. Pues si $x \neq y$ y uno de ellos, por ejemplo, y, es racional, entonces y deberá estar en la primera o bien en la tercera clase de la separación generada por x, esto es, o bien es $y < x$ o es $x < y$. Si ni x ni y son racionales, las segundas clases de las correspondientes subdivisiones son vacías, y deberá existir un número racional r en la primera clase respecto a uno de los números y en la tercera clase respecto al otro. Así se cumple una u otra de las relaciones $x < y$ o $y < x$.

Densidad. Una consecuencia inmediata de estas definiciones es la densidad de los números racionales, en el sentido de que entre dos números reales cualesquiera, x, y, existe siempre un número racional r. Se observa también que si un número real x está representado por un encaje de intervalos racionales $[a_n, b_n]$, entonces $a_n \leq x \leq b_n$ para todo n. Pues si $x < a_m$ para algún m, entonces $b_n < a_m$ para casi todo n, contradiciendo la desigualdad $a_m < b_n$, la cual se cumple para todo n. Por lo tanto, todo número real puede ser confinado en un intervalo racional $[a_n, b_n]$ de longitud arbitrariamente pequeña.

Una vez que los números reales están ordenados, puede hablarse de intervalos con *puntos extremos reales*. La densidad de los números racionales garantiza que todo tal intervalo incluye números racionales.

Límites. Un número real x se llama *límite* de una sucesión $x_1, x_2, \ldots$ de números reales si todo intervalo abierto con puntos extremos reales que contiene a x contiene también elementos x_n para casi todo n. Esta definición es compatible con la definición en términos de intervalos racionales dada anteriormente, en el sentido de que un límite racional de una sucesión racional es un límite de la misma sucesión de números en el sentido más general de límite real. Como consecuencia de la definición de límite se encuentra que para un número real x representado por un encaje de intervalos racionales $[a_n, b_n]$ es

$$x = \lim_{n\to\infty} a_n = \lim_{n\to\infty} b_n.$$

*Aritmética. En seguida se definen las *operaciones aritméticas* para números reales $x \sim \{[a_n, b_n]\}$ y $y \sim \{[\alpha_n, {}_n\beta]\}$: Esto se logra más fácilmente para las operaciones de adición y substracción. Se define

$$x + y \sim \{[a_n + \alpha_n, b_n + \beta_n]\}, \qquad x - y \sim \{[a_n - \beta_n, b_n - \alpha_n]\}.$$

La prueba de que estas definiciones tienen sentido es un ejercicio sencillo cuyos detalles se dejan al lector (ver problema 3, p. 139). Por ejemplo, para $x - y$ es necesario verificar solamente que los intervalos $[a_n - \beta_n, b_n - \alpha]$ forman un encaje cuyas longitudes tienden a cero, y por lo tanto que ellos representan un número real z. El hecho de que z no dependa de las representaciones especiales de x e y se demuestra caracterizando la separación de números racionales en tres clases generada por z directamente en términos de x e y; por ejemplo, la primera clase consiste de los números racionales $r < z$, o sea de aquellos r que son excedidos por $a_n - \beta_n$ para algún n; y se ve fácilmente que estos r son números racionales de la forma $s - t$, donde s y t son números racionales para los cuales $s < x$ y $t > y$.

El producto de los dos números reales, x, y, está definido para $y > 0$ por

$$x \cdot y \sim \{[a_n\alpha_n, b_n\beta_n]\},$$

donde se ha supuesto que todo $\alpha_n > 0$; es obvio que los encajes son adecuados para utilizarse con xy en los casos $y < 0$ e $y = 0$. Siempre que y sea un número *racional* positivo, el producto $x \cdot y$ es representable también en la forma

$$x \cdot y \sim \{[a_n y, b_n y]\}.$$

Para un número natural $y = m$ el producto $x \cdot y = mx$ puede ser obtenido también por adición repetida de x, esto es, $mx = x + (m-1)x = x + x + \cdots + x$.

Las operaciones aritméticas obedecen las reglas usuales. En particular, la relación $x < y$ es equivalente a $0 < y - x$. Se puede introducir el *valor absoluto* de un número real y probar la desigualdad del triángulo $|x + y| \leq |x| + |y|$. A la noción de límite de una sucesión de números reales, definida anteriormente en términos de relaciones de orden, puede dársele entonces la formulación equivalente: $x = \lim_{n\to\infty} x_n$ si para todo real positivo ε la relación $|x - x_n| < \varepsilon$ se cumple para casi todo n.

Verificaremos ahora el así llamado

AXIOMA DE ARQUÍMEDES. *Si x e y son números reales y x es positivo, existe entonces un número natural m tal que $mx > y$.*

En esencia, esto significa que un número real no puede ser "infinitamente pequeño" o bien 'infinitamente grande" comparado con otro (excepto si uno de ellos es cero). Para probar el axioma de Arquímedes (el cual en este contexto es realmente un *teorema)* se observa que para números racionales es una consecuencia de las propiedades comunes de los enteros. Si ahora $x \sim \{[a_n, b_n]\}$ e $y \sim \{[\alpha_n, \beta_n]\}$ son números reales y x es positivo, entonces $a_n > 0$ para casi todo n. Puesto que a_n y β_n son números racionales, puede entonces encontrarse un m tan grande que $ma_n > \beta_n$, de donde se sigue que $mx > \beta_n \geq y$.

d. Plenitud del continuo de números. Compacidad de intervalos cerrados. Criterios de convergencia

Los números reales hacen posible las operaciones límite con números racionales, pero serían de poco valor si las correspondientes operaciones de límite con los reales requirieran la introducción de algún tipo adicional de números "no reales" que habrían de ser intercalados entre los reales, y así sucesivamente *ad infinitum*. Afortunadamente, la definición de número real es tan amplia que no es posible ninguna extensión adicional del sistema numérico sin deponer algunas de sus propiedades esenciales (tal como el "orden", que debe ser descartado para números complejos).

Principio de continuidad

Esta completitud del continuo de números reales está expresada por el principio básico de continuidad (ver p. 32); todo encaje de intervalos con puntos extremos reales contiene un número real. Para probar esto considérense los intervalos cerrados $[x_n, y_n]$, siendo cada intervalo contenido en el precedente, cuyas longitudes $y_n - x_n$ forman una sucesión nula. Se afirma que existe un número real x contenido en todo $[x_n, y_n]$: Las sucesiones x_n e y_n tendrán entonces como límite a x. Para probar esto se reemplaza la sucesión anidada $[x_n, y_n]$ por un encaje de intervalos racionales $[a_n, b_n]$ que contiene a los $[x_n, y_n]$. Esta sucesión o encaje racional definirá entonces el número real x deseado. Para cada n, sea a_n el mayor número racional de la forma $p/2^n$ menor que x_n, y b_n el menor número racional de la forma $q/2^n$ mayor que y_n, donde p y q son enteros. Claramente, los intervalos $[a_n, b_n]$ forman un encaje que representa un número real x. Si x estuviera situado fuera de uno de los intervalos $[x_m, y_m]$, digamos, $x < x_m$, existiría un racional r con $x < r < x_m$, de donde, para todo n suficientemente grande, se tendría

$$y_n \leq b_n < r < x_m \leq x_n,$$

lo cual es imposible. Por lo tanto, todos los intervalos $[x_m, y_m]$ contienen al punto x.

Principio de Weierstrass—Compacidad

Son importantes otras diversas versiones de este principio de *continuidad*. La primera es el principio de Weierstrass de *existencia de puntos límite o puntos de acumulación de sucesiones acotadas*. Un punto x es un *punto límite* de una sucesión $x_1, x_2, \ldots$ si todo intervalo abierto que contiene a x contiene también puntos x_n para una *infinidad de números naturales n*. Nótese la diferencia entre esta definición y la definición de *límite*, donde los x_n deben estar situados en el intervalo abierto para *casi todo n*, o sea, para todo n excepto a lo más un número finito, o para todo n suficientemente grande. Si una sucesión tiene un límite, entonces este límite es también un punto límite de la sucesión y es, de hecho, el único. Puede no haber punto límite (como en el ejemplo de la sucesión 1, 2, 3, 4, ...) o bien un solo punto límite (como en una sucesión convergente) o diversos puntos límite (por ejemplo, la sucesión $1, -1, 1, -1, \ldots$ tiene los dos puntos límite $+1$ y -1). El principio de Weierstrass asegura que: *Toda sucesión acotada tiene al menos un punto límite.*

Para probar esto se observa que, como la sucesión $x_1, x_2, \ldots$ es *acotada*, existe un intervalo $[y_1, z_1]$ que contiene a todo x_n. Partiendo con $[y_1, z_1]$ se

construye por inducción sobre n un encaje de intervalos $[y_n, z_n]$ conteniendo cada uno puntos x_m para una infinidad de naturales m. Si $[y_n, z_n]$ contiene una infinidad de x_m, se divide $[y_n, z_n]$ en dos partes iguales por su punto medio. Por lo menos uno de los dos intervalos cerrados resultantes deberá contener, a su vez, una infinidad de x_m, y puede ser tomado como el intervalo $[y_{n+1}, z_{n+1}]$. Es claro que los $[y_n, z_n]$ forman un encaje que representa un número real x. Todo intervalo abierto que contiene a x contendrá los intervalos $[y_n, z_n]$ para n suficientemente grande y, por lo tanto, deberá contener una infinidad de x_m.

Los puntos límite pueden también ser definidos como *límites de subsucesiones* de la sucesión infinita $x_1, x_2, \ldots$ dada. Una subsucesión es cualquier sucesión infinita extraída de la sucesión dada, o sea, de la forma $x_{n_1}, x_{n_2}, x_{n_3}, \ldots$, donde $n_1 < n_2 < n_3 < \cdots$. Obviamente, un punto x es un punto límite de la sucesión $x_1, x_2, \ldots$ si es límite de alguna subsucesión. Recíprocamente, para todo punto límite x puede construirse por inducción una subsucesión $x_{n_1}, x_{n_2}, \ldots$ convergente a x. Si $x_{n_1}, \ldots, x_{n_{k-1}}$ están ya definidos, se toma para n_k un número de la infinidad de enteros n para los cuales $n > n_{k-1}$ y $|x_n - x| < 2^{-k}$.

Se reestablece el principio de Weierstrass en la forma:

TEOREMA. *Toda sucesión infinita acotada de números reales tiene una subsucesión convergente.*

Un conjunto se dice *compacto* si toda sucesión de sus elementos contiene una subsucesión convergente a un elemento del conjunto. En otras palabras, el teorema dice que *intervalos cerrados de números reales son conjuntos compactos.*

Sucesiones monótonas

Una consecuencia especial de este teorema es que *toda sucesión monótona acotada converge.* Así, sea la sucesión $x_1, x_2, \ldots$ monótona, por ejemplo, monótona creciente. Si la sucesión es también acotada, tiene un punto límite x. Arbitrariamente cerca de x deben existir puntos x_n de la sucesión, sin exceder ninguno a x puesto que los términos subsiguientes crecen y si $x_n > x$ entonces $x_m \geq x_n > x$ para $m > n$. Se sigue que todo intervalo que contiene a x contiene a casi todo x_n, o sea, que x es el límite de la sucesión.

Criterio de convergencia de Cauchy

La condición de que una sucesión sea acotada y monótona es *suficiente* para la convergencia. La importancia de este aserto es que permite a menudo probar la existencia del límite de una sucesión sin requerir un conocimiento *a priori* del valor del límite; además, la acotación y monotonía de una sucesión son propiedades que usualmente son fáciles de verificar en

aplicaciones concretas. Sin embargo, no toda sucesión convergente debe ser monótona (aunque debe ser acotada), y es importante disponer de un criterio de convergencia de aplicabilidad más general. Tal es la *prueba intrínseca o criterio de convergencia de Cauchy,* que es una condición necesaria y suficiente para la existencia del límite de una sucesión.

La sucesión $x_1, x_2, x_3, \ldots$ converge si y sólo si para todo ε positivo existe un N tal que $|x_n - x_m| < \varepsilon$ para todos n y m mayores que N.

En otras palabras, una sucesión converge si dos cualesquiera de sus elementos con índices suficientemente grandes difieren en menos de ε uno del otro.

Se afirma que la condición es necesaria para la convergencia. Si $x = \lim x_n$, entonces todo x_n con n suficientemente grande difiere de x en menos de $\varepsilon/2$ y, en consecuencia, por la desigualdad del triángulo, cualesquiera dos de tales valores x_n y x_m diferirán uno del otro en menos de ε. Recíprocamente, considérese una sucesión para la cual $|x_n - x_m| < \varepsilon$ para todo $\varepsilon > 0$, y para todos n, m suficientemente grandes. Existe entonces un valor N tal que casi todo x_n difiere de x_N en menos de 1. Esto significa que casi todo x_n puede ser encerrado en un intervalo de longitud 2. Puede entonces encontrarse un intervalo tan grande que contenga también el número finito de los x_n que pueden encontrarse fuera del intervalo en torno de x_N. Así, la sucesión está acotada y posee por lo tanto un punto límite x. Todo intervalo abierto que contiene a x contendrá también algunos puntos x_m con m arbitrariamente grande. Puesto que los puntos x_n difieren arbitrariamente poco uno del otro para n suficientemente grande, se sigue que el intervalo abierto en torno de x deberá contener a casi todo x_n, y así x resulta ser el límite de la sucesión.

e. Cotas. Extremo superior y extremo inferior *

Es de gran importancia que un conjunto acotado de números reales posea "las mejores cotas posibles" superiores e inferiores. Un conjunto S de números reales x es *acotado* si todos los números de S pueden ser encerrados en un mismo intervalo finito. Existen entonces *cotas superiores* de S, números B que no son excedidos por ningún número x de S:

$$x \leq B \qquad \text{para todo } x \text{ en } S.$$

Similarmente, existen *cotas inferiores* A de S:

$$A \leq x \qquad \text{para todo } x \text{ en } S.$$

* (N. del T.) En la América Latina algunos dicen "mínima cota superior" y "máxima cota inferior", por traducir del inglés "least upper bound" y "greatest lower bound", respectivamente.

Así, para el conjunto de recíprocos de números naturales, $1, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \ldots$, cualquier número $B \geq 1$ es una cota superior y todo número $A \leq 0$, una cota inferior. Aquí el número 1, un miembro del conjunto, es la menor de las cotas superiores; y el número 0, punto límite de los elementos del conjunto, aunque no un miembro de él, es la mayor de las cotas inferiores. La menor de las cotas superiores de un conjunto de números reales se denomina a menudo supremo; la mayor de las cotas inferiores, su *ínfimum*. En general el *supremum* y el *ínfimum* de un conjunto son o bien miembros del conjunto o al menos límites de sucesiones de miembros del conjunto. En efecto, si la menor de las cotas superiores, b de S, no pertenece a S, deberán existir algunos miembros de S situados arbitrariamente cerca de b, pues de otra forma podrían encontrarse cotas superiores de S menores que b; así, pueden seleccionarse sucesivamente una sucesión de números $x_1, x_2, \ldots$ de S situados cada vez más cerca de b y convergentes a b.

La *existencia de la menor de las cotas superiores* de un conjunto acotado S se sigue de inmediato de la convergencia de sucesiones monótonas acotadas. Para cualquier n defínase B_n como la menor cota racional superior de S con denominador 2^n. Claramente, para cualquier x en S y cualquier n

$$x \leq B_{n+1} \leq B_n \leq B_1.$$

Por lo tanto, los B_n forman una sucesión monótona decreciente y acotada la cual debe tener un límite b. Es fácil ver que b es una cota superior de S y que no existe ninguna cota superior menor. La existencia de la máxima cota inferior se prueba de la misma manera.

f. Numerabilidad de los números racionales

A fines del siglo XIX fue hecho un sorprendente descubrimiento relativo a los números racionales, que estimuló la creación, por Georg Cantor, de la Teoría de Conjuntos después de 1872. Aun cuando los números racionales son densos y no pueden ser ordenados por tamaño, pueden, sin embargo, ser arreglados como una sucesión infinita $r_1, r_2, \ldots, r_n, \ldots$ en la cual todo número racional aparece una vez. De esta forma los números racionales pueden ser numerados, o sea, ser contados como un primero, segundo, ..., n-ésimo, ... número racional, donde, por supuesto, el orden de los números en la sucesión no corresponde a su orden por magnitud. Este resultado, que se cumple también para números racionales en cualquier intervalo, es expresado por la propiedad: *Los números racionales son numerables, o sea, constituyen un conjunto numerable.*

Para probar este resultado se da simplemente una regla para arreglar los números racionales positivos como una sucesión. Cada uno de tales números puede ser escrito en la forma p/q, donde p y q son números naturales. Para cada entero positivo k existen exactamente $k - 1$ fracciones p/q para las cuales $p + q = k$. Estas son arregladas en orden creciente de p. Escribiendo los diferentes arreglos de números para $k = 2, 3, 4, \ldots$ sucesivamente, se obtiene una sucesión que contiene todos los números racionales positivos. Omitiendo las fracciones en las cuales el numerador y el denominador tienen un factor común mayor que 1, y por lo tanto representan el mismo número racional que una fracción anterior, se obtiene la sucesión

$$\frac{1}{1}, \frac{1}{2}, \frac{2}{1}, \frac{1}{3}, \frac{3}{1}, \frac{1}{4}, \frac{2}{3}, \frac{3}{2}, \frac{4}{1}, \frac{1}{5}, \frac{5}{1}, \frac{1}{6}, \frac{2}{5}, \ldots$$

en la cual todo número racional positivo aparece exactamente una vez. Una sucesión análoga conteniendo a todos los números racionales, o bien, todos los números racionales en algún intervalo particular, es construida fácilmente.

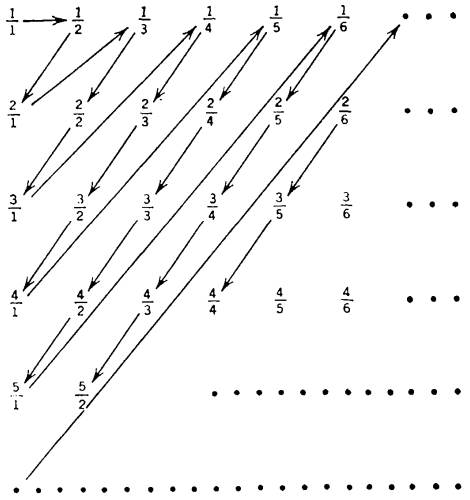


Figura 1.S.1. Numerabilidad de los racionales positivos.

Este resultado es visto en su debida perspectiva sólo a la luz de otro hecho básico: que *el conjunto de todos los* números *reales* no es numerable.[1] Esta es una indicación de que el conjunto de números reales contiene "muchos más" elementos que el de los números racionales, aun cuando ambos conjuntos son infinitos. Así, la numerabilidad es en realidad una propiedad altamente restrictiva de un conjunto.

[1] Para la demostración y una breve discusión general de los hechos básicos de la teoría de conjuntos, véase *What Is Mathematics?* por Courant y Robbins, p. 81.

La Teoría de Conjuntos desempeña un importante papel esclarecedor en matemáticas, aunque su uso con libre generalidad ha dado lugar a resultados paradójicos y controversias. Tales paradojas, sin embargo, no afectan la esencia de las matemáticas constructivas y están ausentes en la teoría de conjuntos para números reales.

S.2 Teoremas sobre funciones continuas

Propiedades importantes de las funciones continuas son establecidas sobre la base de la propiedad de completitud de los números reales. Recuérdese la definición de continuidad: la función $f(x)$ es continua en el punto ξ si para todo ε positivo dado la desigualdad $|f(x) - f(\xi)| < \varepsilon$ se cumple para todo x que difiere de ξ en menos de una cantidad adecuada δ, la cual depende generalmente de la elección de ε y ξ. Se sobreentiende en esta definición que se consideran solamente valores de x y ξ para los cuales f está definida.

Una definición más concisa de continuidad en términos de convergencia de sucesiones es la siguiente: $f(x)$ *es continua en el punto* ξ *si* $\lim_{n\to\infty} f(x_n) = f(\xi)$ *para toda sucesión* $x_1, x_2, \ldots$ *con límite* ξ (donde, nuevamente, los valores x_n y ξ están en el dominio de f). La equivalencia de las dos definiciones se probó en la sección 1.8, p. 105.

La función f se denomina *continua en un intervalo* si f es continua en cada punto del intervalo. La función $f(x)$ es *uniformemente continua* si para $\varepsilon > 0$ dado se tiene $|f(x) - f(\xi)| < \varepsilon$ siempre que x y ξ se encuentren suficientemente cercanos, independientemente de su posición en el intervalo; así, f es uniformemente continua si la cantidad δ que aparece en la definición de continuidad puede ser elegida independiente de ξ: Para todo $\varepsilon > 0$ existe un $\delta = \delta(\varepsilon) > 0$ tal que $|f(x) - f(\xi)| < \varepsilon$ siempre que $|x - \xi| < \delta$. Para propósitos prácticos, esto significa que si se subdivide el intervalo en el cual f está definida en un número suficientemente grande de intervalos iguales, entonces f variará en menos que una cantidad prescrita ε en cada subintervalo. En cualquier punto f diferirá entonces en menos de ε de su valor en cualquier otro punto del mismo subintervalo.

Se probará ahora lo siguiente: *Toda función continua en un intervalo cerrado* $[a, b]$ *es uniformemente continua en ese intervalo.*

Si f no fuese uniformemente continua en $[a, b]$ existiría un $\varepsilon > 0$ fijo y puntos, x, ξ, en $[a, b]$, arbitrariamente cercanos uno del otro, para los cuales $|f(x) - f(\xi)| \geq \varepsilon$. Entonces sería posible para cualquier n elegir puntos x_n, ξ_n en $[a, b]$ para los cuales $|f(x_n) - f(\xi_n)| \geq \varepsilon$ y $|x_n - \xi_n| < 1/n$. Puesto que los x_n forman una sucesión acotada de números, podría encontrarse una subsucesión convergente a un punto η del intervalo (utilizando la *compacidad* de intervalos cerrados). Los correspondientes valores ξ_n entonces convergerían también a η: puesto que f es continua en η se encontraría que $\eta = \lim f(x_n) = \lim f(\xi_n)$ para n tendiendo a infinito en la subsucesión, lo cual es imposible si $|f(x_n) - f(\xi_n)| \geq \varepsilon$ para todo n.

El *teorema del valor intermedio* afirma que: Si para una función $f(x)$ continua en un intervalo $a \leq x \leq b$, γ es cualquier valor entre $f(a)$ y $f(b)$, entonces $f(\xi) = \gamma$ para algún ξ adecuado entre a y b. Así, la *existencia de una solución* ξ de la ecuación $f(\xi) = \gamma$ es cierta si se exhiben dos valores a y b para los cuales $f(a) < \gamma$ y $f(b) > \gamma$ respectivamente. Esto implica inmediatamente la existencia de una *función inversa* determinada de manera única si f es continua y monótona, como se ha visto (pp. 67-68).

Para probar el teorema del valor intermedio, sean $a < b$, $f(a) = \alpha$, $f(b) = \beta$ y $\alpha < \gamma < \beta$. Sea S el conjunto de puntos x del intervalo $[a, b]$ para los cuales $f(x) < \gamma$. S es acotado y posee una misma cota superior ξ que pertenece también al intervalo cerrado $[a, b]$. Entonces $f(x) \geq \gamma$ para $\xi < x \leq b$. El punto ξ pertenece a S o bien es el límite de una sucesión de puntos x_n de S. En el primer caso, $f(\xi) < \gamma$; por lo tanto $\xi < b$, puesto que $f(b) > \gamma$, y existen puntos x entre ξ y b, arbitrariamente cerca de ξ, para los cuales $f(x) \geq \gamma$. Esto es imposible si f es continua en ξ y $f(\xi) < \gamma$. En el segundo caso, $f(\xi) \geq \gamma$, y se encuentra de $f(x_n) < \gamma$ y $\lim_{n\to\infty} x_n = \xi$ que $f(\xi) \leq \gamma$; y como se vio ya que $f(\xi) < \gamma$ es imposible, se debe tener $f(\xi) = \gamma$.

Una tercera propiedad básica de una función continua $f(x)$ en un intervalo cerrado $[a, b]$ es la *existencia de un valor máximo (máximo),* entendiéndose que existe un punto ξ en el intervalo $[a, b]$ tal que $f(x) \leq f(\xi)$ para todo x en el intervalo. Análogamente, f alcanzará su *valor mínimo (mínimo)* en algún punto η del intervalo: $f(x) \geq f(\eta)$ para todo x en el intervalo. Es esencial que el intervalo sea *cerrado:* por ejemplo, las funciones $f(x) = x$, $f(x) = 1/x$ son continuas pero no poseen un valor máximo en el intervalo abierto $0 < x < 1$; el máximo puede ocurrir justamente en uno de los puntos extremos o bien no existir si f no es continua en estos puntos.

Para probar este principio se observa que una función f continua en $[a, b]$ es necesariamente *acotada:* esto es, los valores $f(x)$ que constituyen el "rango" S de f están situados en algún intervalo finito. Realmente, por la continuidad uniforme de f pueden encontrarse un número finito de puntos $x_1, x_2, \ldots, x_n$ en el intervalo tales que $f(x)$ en cualquier x del intervalo difiera en menos de uno de alguno de los números $f(x_1), f(x_2), \ldots, f(x_n)$, los cuales pueden ser todos encajados en un intervalo finito. Puesto que entonces el conjunto S de valores $f(x)$ es acotado, posee una mínima cota superior M. Este M es el menor número tal que $f(x) \leq M$ para todo x en

$[a, b]$. M pertenece a S o bien es el límite de una sucesión de puntos de S. En el primer caso existe un ξ en $[a, b]$ tal que $f(\xi) = M$. En el segundo caso existe una sucesión de puntos x_n en $[a, b]$ tal que $\lim_{n\to\infty} f(x_n) = M$; así, puede encontrarse una subsucesión de los x_n que converge a un punto ξ de $[a, b]$, y entonces $f(\xi) = M$ por continuidad de f en ξ. Claramente, $f(\xi)$ es el máximo de f.

S.3 Coordenadas polares

En el capítulo 1 se han representado funciones geométricamente por curvas. La geometría analítica sigue el procedimiento inverso, principiando con una curva y representándola mediante una función, por ejemplo, mediante una función que expresa una de las coordenadas de un punto de la curva en términos de la otra. Este punto de vista nos conduce naturalmente a considerar, además de las coordenadas rectangulares a las cuales nos hemos restringido, otros sistemas de coordenadas posiblemente mejor situados para la representación de curvas dadas geométricamente. El ejemplo más importante es el de *coordenadas polares* r, θ conectadas con las coordenadas rectangulares x, y de un punto P por las ecuaciones

$$x = r\cos\theta, \quad y = r\,\mathrm{sen}\,\theta, \quad r^2 = x^2 + y^2, \quad \tan\theta = \frac{y}{x},$$

cuya interpretación geométrica se hace clara en la Fig. 1.S.2.[1]

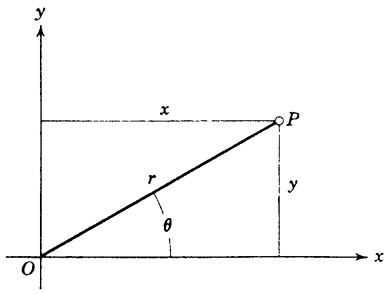


Figura 1.S.2 Coordenadas polares.

Considérese, por ejemplo, el *lemniscato*. Esta curva está geométricamente definida como el lugar geométrico de todos los puntos P para los cuales el producto de sus distancias r_1 y r_2 a los puntos fijos F_1 y F_2 cuyas

[1] Las coordenadas polares no están determinadas totalmente por el punto P. Además de θ, cualquiera de los ángulos $\theta \pm 2\pi$, $\theta \pm 4\pi$, ... puede ser considerado un ángulo polar de P.

coordenadas rectangulares son $x = a$, $y = 0$ y $x = -a$, $y = 0$ respectivamente, tiene el valor constante a^2 (ver Fig. 1.S.3). Puesto que

$$r_1^2 = (x - a)^2 + y^2, \quad r_2^2 = (x + a)^2 + y^2,$$

mediante un cálculo sencillo se obtiene la ecuación del lemniscato en la forma

$$(x^2 + y^2)^2 - 2a^2(x^2 - y^2) = 0.$$

Introduciendo coordenadas polares se obtiene

$$r^4 - 2a^2r^2(\cos^2\theta - \mathrm{sen}^2\,\theta) = 0;$$

dividiendo por r^2 y utilizando una fórmula trigonométrica sencilla obtenemos finalmente

$$r^2 = 2a^2\cos 2\theta.$$

Así, la ecuación del lemniscato es más sencilla en coordenadas polares que en coordenadas rectangulares.

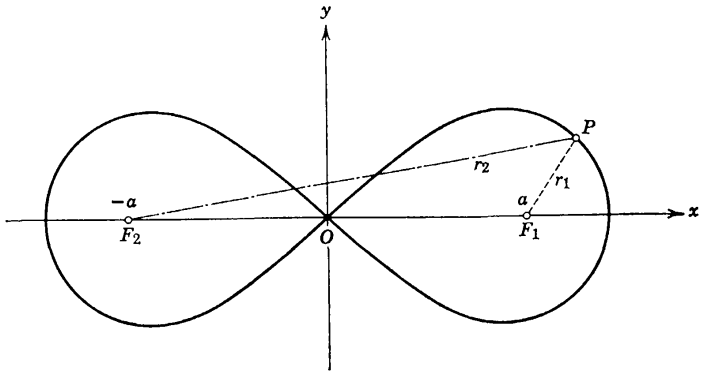


Figura 1.S.3 Lemniscato.

S.4 Observaciones sobre los números complejos

Nuestros estudios se basarán principalmente en el continuo de los números reales. Sin embargo, con miras hacia las discusiones de los capítulos 7, 8 y 9, se recuerda al lector que los problemas del álgebra han conducido hacia una extensión aún más amplia del concepto de número, los *números complejos*. El paso de los números naturales a los números reales surgió del deseo de eliminar fenómenos excepcionales y hacer siempre posible ciertas operaciones, tales como la substracción, la división y la correspondencia entre puntos y números. Análogamente, estamos precisados a introducir los números complejos, por el requisito de que toda ecuación cuadrática, y,

de hecho, toda ecuación algebraica, deberá tener solución. Si, por ejemplo, se desea que la ecuación

$$x^2 + 1 = 0$$

tenga raíces, estamos obligados a introducir nuevos símbolos i y $-i$ como sus raíces. (Como se muestra en la teoría de funciones de variable compleja, esto es suficiente para asegurar que *toda* ecuación algebraica tenga una solución.)[1]

Si a y b son números reales ordinarios, el *número complejo* $c = a + ib$ denota una pareja de números (a, b) con la cual se efectúan los cálculos de acuerdo a la siguiente regla general: Los números complejos (entre los cuales los números reales están incluidos como el caso especial $b = 0$) se suman, multiplican y dividen, tratándose al símbolo i como una cantidad indeterminada y simplificando todas las expresiones haciendo uso de la ecuación $i^2 = -1$ para eliminar todas las potencias de i mayores que la primera, dejando solamente una expresión de la forma $a + ib$.

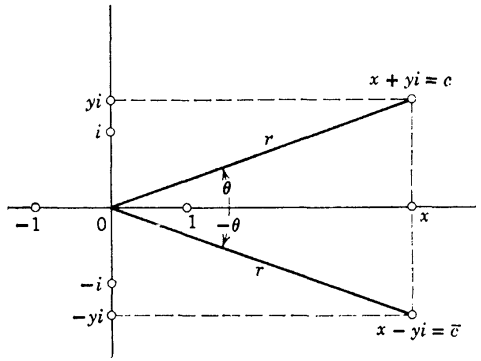


Figura 1.S.4 Representación geométrica de un número complejo $x + yi$ y de su conjugado.

Se supone que el lector ya posee un cierto grado de familiaridad con los números complejos. Se recalcará, sin embargo, una relación particularmente importante que será explicada en conexión con la representación geométrica o trigonométrica de los números complejos. Si $c = x + iy$ es uno de tales números, éste se representa en un sistema coordenado rectangular mediante el punto P con coordenadas x e y. Por medio de las ecuaciones $x = r\cos\theta$, $y = r\,\mathrm{sen}\,\theta$, se introducen las coordenadas r y θ (ver p. 125) en lugar de las coordenadas rectangulares x e y. Entonces $r = \sqrt{x^2 + y^2}$ es la distancia del punto P al origen y θ, el ángulo entre

[1] Una ecuación algebraica es de la forma $P(x) = 0$, donde P es un polinomio con coeficientes complejos.

el eje x positivo y el segmento OP. El número complejo c está representado en la forma

$$c = r(\cos\theta + i\,\mathrm{sen}\,\theta).$$

El ángulo θ se denomina la *amplitud* del número complejo c; la cantidad r, su *valor absoluto* o *módulo*, para el cual se escribe también $|c|$. Al número complejo "conjugado" $\bar{c} = x - iy$ obviamente le corresponde el mismo valor absoluto pero la amplitud $-\theta$ (Fig. 1.S.4). Claramente,

$$r^2 = |c|^2 = c\bar{c} = x^2 + y^2.$$

Si se utiliza esta representación trigonométrica, la multiplicación de números complejos toma una forma particularmente simple, pues entonces

$$c \cdot c' = r(\cos\theta + i\,\mathrm{sen}\,\theta) \cdot r'(\cos\theta' + i\,\mathrm{sen}\,\theta')$$

$$= rr'[(\cos\theta\cos\theta' - \mathrm{sen}\,\theta\,\mathrm{sen}\theta') + i(\cos\theta\,\mathrm{sen}\,\theta' + \mathrm{sen}\,\theta\cos\theta')].$$

Si se usan los teoremas de adición para las funciones trigonométricas, esto se transforma en

$$c \cdot c' = rr'(\cos(\theta + \theta') + i\,\mathrm{sen}\,(\theta + \theta')).$$

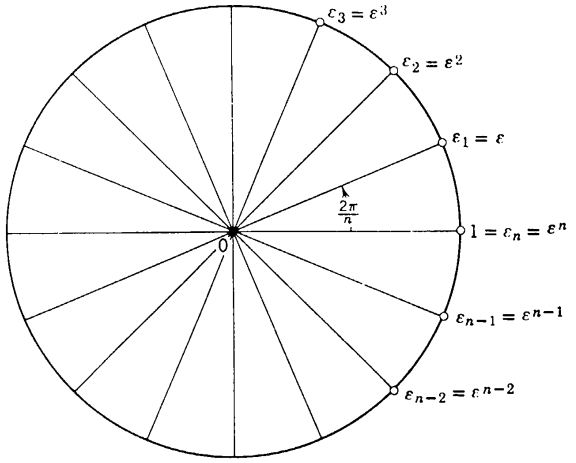


Figura 1.S.5 Las raíces n-ésimas de la unidad (para $n = 16$).

En consecuencia, los números complejos se multiplican multiplicando sus valores absolutos y sumando sus amplitudes. La notable fórmula

$$(\cos\theta + i\,\mathrm{sen}\,\theta)(\cos\theta' + i\,\mathrm{sen}\,\theta') = \cos(\theta + \theta') + i\,\mathrm{sen}\,(\theta + \theta')$$

se denomina usualmente *teorema de De Moivre*. Este nos conduce a la relación

$$(\cos\theta + i\,\mathrm{sen}\,\theta)^n = \cos n\theta + i\,\mathrm{sen}\,n\theta,$$

la cual, por ejemplo, permite de inmediato resolver la ecuación $x^n = 1$ para enteros positivos n; las raíces (las así llamadas raíces de la unidad) son

$$\varepsilon_1 = \varepsilon = \cos\frac{2\pi}{n} + i\,\mathrm{sen}\,\frac{2\pi}{n}, \qquad \varepsilon_2 = \varepsilon^2 = \cos\frac{4\pi}{n} + i\,\mathrm{sen}\,\frac{4\pi}{n}, \ldots,$$

$$\varepsilon_{n-1} = \varepsilon^{n-1} = \cos\frac{2(n-1)\pi}{n} + i\,\mathrm{sen}\,\frac{2(n-1)\pi}{n}, \qquad \varepsilon_n = \varepsilon^n = 1$$

(Fig. 1.S.5).

Geométricamente, los puntos que corresponden a las raíces de la unidad forman los vértices de un n-ágono regular inscrito en el círculo de radio 1 con centro en el origen.

Finalmente, si nos imaginamos la expresión del miembro izquierdo de la ecuación $(\cos\theta + i\,\mathrm{sen}\,\theta)^n = \cos n\theta + i\,\mathrm{sen}\,n\theta$ desarrollada por el teorema binomial, necesitaremos solamente separar las partes real e imaginaria de manera de obtener expresiones para $\cos n\theta$ y $\mathrm{sen}\,n\theta$ en términos de potencias y productos de potencias de $\mathrm{sen}\,\theta$ y $\cos\theta$:

$$\cos n\theta = \cos^n\theta - \binom{n}{2}\cos^{n-2}\theta\,\mathrm{sen}^2\theta + \binom{n}{4}\cos^{n-4}\theta\,\mathrm{sen}^4\theta + - \cdots,$$

$$\mathrm{sen}\,n\theta = \binom{n}{1}\cos^{n-1}\theta\,\mathrm{sen}\,\theta - \binom{n}{3}\cos^{n-3}\theta\,\mathrm{sen}^3\theta + \binom{n}{5}\cos^{n-5}\theta\,\mathrm{sen}^5\theta + - \cdots.$$

PROBLEMAS

SECCION 1.1a, página 26

1. (*a*) Si a es racional y si x es irracional, demuéstrese que $a + x$ es irracional, y si $a \neq 0$, que ax es irracional.

(*b*) Muéstrese que entre dos números racionales cualesquiera existe por lo menos un número irracional y, en consecuencia, una infinidad de ellos.

2. Probar que los números siguientes no son racionales: (*a*) $\sqrt{3}$. (*b*) $\sqrt{n}$, donde el entero n no es un cuadrado perfecto, es decir, no es el cuadrado de un entero. (*c*) $\sqrt[3]{2}$. (*d*) $\sqrt[p]{n}$, donde n no es una potencia p-ésima perfecta.

***3.** (*a*) Para cualquier raíz racional de un polinomio con coeficientes enteros,

$$a_n x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0, \qquad (a_n \neq 0),$$

expresada en sus menores términos como p/q, demuéstrese que el numerador p es factor de a_0 y que el denominador q es factor de a_n. (Por medio de este criterio podemos obtener todas las raíces reales racionales y, por lo tanto, demostrar la irracionalidad de cualesquiera otras raíces reales.)

(*b*) Demuéstrese la irracionalidad de $\sqrt{2} + \sqrt[3]{2}$ y $\sqrt{3} + \sqrt[3]{2}$.

SECCION 1.1c, página 33

1. Denotemos por $[x]$ la parte entera de x; es decir, $[x]$ es el entero para el que se cumple

$$x - 1 < [x] \leq x.$$

Sean $c_0 = [x]$, y $c_n = [10^n(x - c_0) - 10^{n-1}c_1 - 10^{n-2}c_2 - \cdots - 10c_{n-1}]$ para $n = 1, 2, 3 \ldots$. Verifíquese que la representación decimal de x es

$$x = c_0 + 0 \cdot c_1 c_2 c_3 \cdots$$

y que en esta construcción queda excluida la posibilidad de obtener una cadena infinita de nueves.

2. Defínase la desigualdad $x > y$ para dos números reales en términos de sus representaciones decimales (véase el suplemento, pág. 115).

***3.** Si p y q son enteros, $q > 0$, demuéstrese que el desarrollo decimal de p/q o bien termina (de manera que todos los dígitos posteriores al último sean ceros) o es periódico; es decir, hay un punto a partir del que el desarrollo decimal consiste en la repetición sucesional de una determinada cadena de dígitos. Por ejemplo, tenemos que $\frac{1}{4} = 0.25$ termina, y que $\frac{1}{11} = 0.090909\cdots$ es periódico. La longitud de la cadena que se repite se llama *período* del decimal; para $\frac{1}{11}$ el período es 2. En general, ¿qué tan grande puede ser el período de p/q?

SECCION 1.1e, página 36

1. Por medio de sólo signos de desigualdad (sin emplear los signos de valor absoluto) especifíquense los valores de x que satisfacen las relaciones siguientes. Discútanse todos los casos.

(*a*) $|x - a| < |x - b|$.
(*b*) $|x - a| < x - b$.
(*c*) $|x^2 - a| < b$.

2. Un intervalo (véanse las definiciones del texto) puede ser definido como cualquier porción conexa del continuo real. Se dice que un subconjunto S del continuo real es *conexo* si con todo par de puntos a, b en S, el conjunto S contiene todo el intervalo cerrado $[a, b]$. Al margen de los intervalos abiertos y cerrados que hemos ya mencionado, existen los intervalos "semiabiertos" $a \leq x < b$ y $a < x \leq b$ (que a veces se denotan respectivamente por $[a, b)$ y $(a, b]$) y los intervalos no acotados que pueden constituir o bien toda la recta real, o bien un rayo, es decir, una "semirrecta" $x \leq a$, $x < a$, $x > a$, $x \geq a$ (que a veces se denotan respectivamente $(-\infty, \infty)$] y $(-\infty, a]$, $(-\infty, a)$, (a, ∞), $[a, \infty)$), (véase también la nota al pie de la página, en la p. 46).

*(*a*) Demuéstrese que los casos de intervalos que se acaban de describir agotan todas las posibilidades de los subconjuntos conexos del eje numérico.

(*b*) Determínense los intervalos en los cuales se satisfacen las desigualdades siguientes.

(i) $x^2 - 3x + 2 < 0$.
(ii) $(x - a)(x - b)(x - c) > 0$, para $a < b < c$.
(iii) $|1 - x| - x \geq 0$.
(iv) $\dfrac{x - a}{x + a} \geq 0$.
(v) $\left|x + \dfrac{1}{x}\right| \geq 6$.

(vi) $[x] \leq x/2$. Véase el problema 1 de la sección 1.1c.

(vii) $\operatorname{sen} x \geq \sqrt{2}/2$.

(*c*) Demuéstrese que si $a \leq x \leq b$, entonces $|x| \leq |a| + |b|$.

3. Dedúzcanse las desigualdades

(*a*) $x + \dfrac{1}{x} \geq 2$, para $x > 0$,

(*b*) $x + \dfrac{1}{x} \leq -2$, para $x < 0$,

(*c*) $\left| + \dfrac{1}{x} \right| \geq 2$, para $x \neq 0$.

4. La media armónica ξ de dos números positivos a, b se define por

$$\frac{1}{\xi} = \frac{1}{2}\left(\frac{1}{a} + \frac{1}{b}\right).$$

Demuéstrese que la media armónica no excede la media geométrica; es decir, que $\xi \leq \sqrt{ab}$. ¿Cuándo son iguales las dos medias?

5. Dedúzcanse las desigualdades siguientes:

(*a*) $x^2 + xy + y^2 \geq 0$,

*(*b*) $x^{2n} + x^{2n-1}y + x^{2n-2}y^2 + \cdots + y^{2n} \geq 0$,

*(*c*) $x^4 - 3x^3 + 4x^2 - 3x + 1 \geq 0$.

¿Cuándo se cumple la igualdad?

***6.** ¿Cuál es la interpretación geométrica de la desigualdad de Cauchy para $n = 2, 3$?

7. Demuéstrese que la igualdad se cumple en la desigualdad de Cauchy si y sólo si las a_ν son proporcionales a las b_ν: es decir, $ca_\nu + db_\nu = 0$ para toda ν, donde c y d no dependen de ν y no son ambas cero.

8. (*a*) $|x - a_1| + |x - a_2| + |x - a_3| \geq a_3 - a_1$, para $a_1, < a_2 < a_3$.
¿Para qué valores de x se cumple la igualdad?

*(*b*) Encontrar el valor máximo de y para el que se tiene, para toda x, que

$$|x - a_1| + |x - a_2| + \cdots + |x - a_n| \geq y,$$

donde $a_1 < a_2 < \cdots < a_n$. ¿En qué condiciones se cumple la igualdad?

9. Demuéstrese que las desigualdades siguientes se cumplen cuando a, b, c son positivos.

(*a*) $a^2 + b^2 + c^2 \geq ab + bc + ca$.

(*b*) $(a + b)(b + c)(c + a) \geq 8abc$.

(*c*) $a^2b^2 + b^2c^2 + c^2a^2 \geq abc(a + b + c)$.

10. Supongamos que los números x_1, x_2, x_3 y a_{ik} $(i, k = 1, 2, 3)$ son todos positivos y además, $a_{ik} \leq M$ y $x_1^2 + x_2^2 + x_3^2 \leq 1$. Demuéstrese que

$$a_{11}x_1^2 + a_{12}x_1x_2 + \cdots + a_{33}x_3^2 \leq 3M.$$

***11.** Demuéstrese la desigualdad siguiente e interprétese geométricamente para $n \leq 3$.

$$\sqrt{(a_1 - b_1)^2 + \cdots + (a_n - b_n)^2} \leq \sqrt{(a_1^2 + \cdots + a_n^2)}\ \sqrt{(b_1^2 + \cdots + b_n^2)}.$$

12. Demuéstrese e interprétese geométricamente que para $n \leq 3$,

$$\sqrt{(a_1 + b_1 + \cdots + z_1)^2 + \cdots + (a_n + b_n + \cdots + z_n)^2}$$
$$\leq \sqrt{a_1^2 + \cdots + a_n^2} + \sqrt{b_1^2 + \cdots + b_n^2} + \cdots + \sqrt{z_1^2 + \cdots + z_n^2}.$$

13. Demuéstrese que la media geométrica de n números positivos no es mayor que la media aritmética; es decir, si $a_i > 0$ $(i = 1, \ldots, n)$, entonces

$$\sqrt[n]{a_1a_2 \cdots a_n} \leq \frac{1}{n}(a_1 + a_2 + \cdots + a_n).$$

(Indicación: supóngase que $a_1 \leq a_2 \leq \cdots \leq a_n$. Para el primer paso reemplácese a_n por la media geométrica y ajústese a_1 de manera que la media geométrica quede sin alteración.)

SECCION 1.2d, página 55

1. Si $f(x)$ es continua en $x = a$ y $f(a) > 0$, mostrar que el dominio de f contiene un intervalo abierto alrededor de a en el que $f(x) > 0$.

2. Por medio de la definición de continuidad muéstrese que los intervalos centrados

$$|f(x) - f(x_0)| < \varepsilon \quad \text{y} \quad |x - x_0| < \delta$$

se pueden substituir por un intervalo abierto arbitrario que contenga $f(x_0)$ y otro intervalo abierto suficientemente pequeño que contenga x_0, como se indica en la página 57.

3. Sea $f(x)$ continua para $0 \leq x \leq 1$. Supongamos además que $f(x)$ toma valores racionales y ningunos otros y que $f(x) = \frac{1}{2}$ cuando $x = \frac{1}{2}$. Demuéstrese que en todo punto $f(x) = \frac{1}{2}$.

4. (*a*) Definamos $f(x)$ para todos los valores de x de la manera siguiente:

$$f(x) = \begin{cases} 0, & x \text{ irracional} \\ 1, & x \text{ racional.} \end{cases}$$

Demuéstrese que $f(x)$ es discontinua en todo punto.

(*b*) Consideremos por otra parte

$$g(x) = \begin{cases} 0, & x \text{ irracional} \\ \dfrac{1}{q}, & x = \dfrac{p}{q} \text{ racional en sus menores términos.} \end{cases}$$

(Se dice que el número racional p/q está expresado en sus menores términos si los enteros p y q no tienen factores comunes mayores que 1, y $q > 0$. De esta manera, $f(16/29) = 1/29$). Demuéstrese que $g(x)$ es continua para todos los valores irracionales y discontinua para todos los valores racionales.

***5.** Si $f(x)$ satisface la ecuación funcional

$$f(x + y) = f(x) + f(y)$$

para todos los valores de x e y, encuéntrense los valores de $f(x)$ para los valores racionales de x y demuéstrese, si $f(x)$ es continua, que $f(x) = cx$, donde c es constante.

6. (*a*) Si $f(x) = x^n$, encuéntrese una δ que pueda depender de ξ tal que

$$|f(x) - f(\xi)| < \varepsilon$$

siempre que

$$|x - \xi| < \delta.$$

*(*b*) Hágase lo mismo si $f(x)$ es un polinomio cualquiera

$$f(x) = a_nx^n + a_{n-1}x^{n-1} + \cdots + a_1x + a_0,$$

donde $a_n \neq 0$.

SECCION 1.2e, página 68

1. Demuéstrese que si $f(x)$ es monótona en $[a, b]$ y satisface la propiedad del valor intermedio, entonces $f(x)$ es continua. ¿Se puede extraer la misma conclusión si f no es monótona?

2. (*a*) Demuéstrese que x^n es monótona para $x > 0$. En consecuencia, descríbase para $a > 0$ que $x^n = a$ posee una solución positiva única $\sqrt[n]{a}$.

(*b*) Sea $f(x)$ un polinomio

$$f(x) = a_n x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0, \qquad (a_n \neq 0).$$

Demuéstrese que (i) si n es impar entonces $f(x)$ tiene por lo menos una raíz real, (ii) si a_n y a_0 son de signos opuestos, entonces $f(x)$ tiene no menos de una raíz positiva y además si n es par, $n \neq 0$, entonces $f(x)$ tiene también una raíz negativa.

***3.** (*a*) Demuéstrese que existe una recta en cada dirección tal que bisecta cualquier triángulo dado, es decir, tal que lo divide en dos porciones de áreas iguales.

(*b*) Demuéstrese para cualquier par de triángulos que existe una recta que los bisecta simultáneamente.

SECCION 1.3b, página 72

1. (*a*) Demuéstrese que $\sqrt{x}$ no es una función racional. *(Indicación:* examínese la posibilidad de representar $\sqrt{x}$ como función racional para $x = y^2$. (Puede servir el hecho de que un polinomio diferente de cero no puede tener más que un número finito de raíces.)

(*b*) Demuéstrese que $\sqrt[3]{x}$ no es una función racional.

SECCION 1.3c, página 73

1. (*a*) Muéstrese que una recta no puede cortar la gráfica de una función polinomial de grado mayor que uno sino en un número finito de puntos.

(*b*) Obténgase el mismo resultado para las funciones racionales en general.

(*c*) Verifíquese que las funciones trigonométricas no son racionales.

SECCION 1.5, página 80

1. Demuéstrense las propiedades siguientes de los coeficientes binomiales.

(*a*) $1 + \binom{n}{1} + \binom{n}{2} + \cdots + \binom{n}{n-1} + \binom{n}{n} = 2^n.$

(*b*) $1 - \binom{n}{1} + \binom{n}{2} - \binom{n}{3} + \cdots + (-1)^n \binom{n}{n} = 0.$

(*c*) $\binom{n}{1} + 2\binom{n}{2} + 3\binom{n}{3} + \cdots + n\binom{n}{n} = n(2^{n-1}).$ *(Indicación:* represéntense los coeficientes binomiales en términos de factoriales.)

(*d*) $1 \cdot 2\binom{n}{2} + 2 \cdot 3\binom{n}{3} + \cdots + (n-1)n\binom{n}{n} = n(n-1)2^{n-2}.$

(*e*) $1 + \frac{1}{2}\binom{n}{1} + \frac{1}{3}\binom{n}{2} + \cdots + \frac{1}{n+1}\binom{n}{n} = \frac{2^{n+1}-1}{n+1}.$

*(*f*) $\binom{n}{0}^2 + \binom{n}{1}^2 + \cdots + \binom{n}{n}^2 = \binom{2n}{n}.$ *(Indicación:* considérese el coeficiente de x^n en $(1+x)^{2n}$.)

*(*g*) $S_n = -\binom{n}{0} - \frac{1}{3}\binom{n}{1} + \frac{1}{5}\binom{n}{2} - \frac{1}{7}\binom{n}{3} + \cdots + \frac{(-1)^n}{2n+1}\binom{n}{n}$

$$= \frac{4^n(n!)^2}{(2_n+1)!}.$$

(*Indicación:* demuéstrese que $\frac{2n+2}{2n+3}S_n = S_{n+1}$.)

2. Demuéstrese que $(1+x)^n \geq 1 + nx$, para $x > -1$.

3. Demuéstrese por inducción que $1 + 2 + \cdots + n = \frac{1}{2}n(n+1)$.

***4.** Demuéstrense por inducción las proposiciones siguientes:

(*a*) $1 + 2q + 3q^2 + \cdots + nq^{n-1} = \frac{1 - (n+1)q^n + nq^{n+1}}{(1-q)^2}.$

(*b*) $(1+q)(1+q^2)\cdots(1+q^{2^n}) = \frac{1-q^{2^{n+1}}}{1-q}.$

5. Demuéstrese que, para todos los números naturales n mayores que 1, o bien n es un número primo o bien se puede expresar como producto de primos. *(Indicación:* sea A_{n-1} la afirmación para todo entero k tal que $k \leq n$, que k es primo o se puede expresar como producto de primos.)

***6.** Considérese la sucesión de fracciones

$$\frac{1}{1}, \frac{3}{2}, \frac{7}{5}, \ldots, \frac{p_n}{q_n}, \ldots,$$

donde $P_{n+1} = p_n + 2q_n$ y $q_{n+1} = p_n + q_n$.

(*a*) Demuéstrese para toda n que p_n/q_n está en sus menores términos.

(*b*) Demuéstrese que el valor absoluto de la diferencia entre p_n/q_n y $\sqrt{2}$ se puede hacer arbitrariamente pequeño. Demuéstrese también que el error de aproximación a $\sqrt{2}$ alterna de signo.

7. Sean a, b, a_n y b_n enteros tales que

$$(a + b\sqrt{2})^n = a_n + b_n\sqrt{2},$$

donde a es el entero más cercano a $b\sqrt{2}$. Demuéstrese que a_n es el entero que más se aproxima a $b_n\sqrt{2}$.

***8.** Definamos a_n y b_n por

$$a_1 = 3, \qquad a_{n+1} = 3^{a_n}, \qquad \text{y} \qquad b_1 = 9, \qquad b_{n+1} = 9^{b_n}.$$

Determínese para cada valor de n el valor mínimo m tal que $a_m \geq b_n$.

9. Si n es un número natural, demuéstrese que

$$\frac{(1+\sqrt{5})^n - (1-\sqrt{5})^n}{2^n\sqrt{5}}$$

es un número natural.

10. Determínese el número máximo de regiones en las que se puede dividir un plano por n rectas. Demuéstrese que el máximo ocurre cuando no hay dos rectas paralelas entre sí y no hay tres que concurran al mismo punto, y determínese el número de regiones que se tiene al permitir concurrencias y paralelismos.

11. Demuéstrese, para cada número natural n, que existe un número natural k tal que

$$(\sqrt{2}-1)^n = \sqrt{k} - \sqrt{k-1}.$$

12. Demuéstrese inductivamente la desigualdad de Cauchy.

nSECCION 1.6, página 84

1. Demuéstrese que $\lim_{n\to\infty} (\sqrt{n+1} - \sqrt{n})(\sqrt{n+\frac{1}{2}}) = \frac{1}{2}$.

2. Demuéstrese que $\lim_{n\to\infty} (\sqrt[3]{n+1} - \sqrt[3]{n}) = 0$.

3. Sea $a_n = 10^n/n!$. (*a*) ¿A qué límite converge a_n? (*b*) ¿Es monótona la sucesión? (*c*) ¿Es monótona a partir de un cierto valor de n? (*d*) Dese una estimación de la diferencia entre a_n y el límite. (*e*) ¿A partir de qué valor de n es la diferencia menor que 1/100?

4. Demuéstrese que $\lim_{n\to\infty} \frac{n!}{n^n} = 0$.

5. (*a*) Demuéstrese que $\lim_{n\to\infty} \left(\frac{1}{n^2} + \frac{2}{n^2} + \cdots + \frac{n}{n^2} \right) = \frac{1}{2}$.

(*b*) Demuéstrese que $\lim_{n\to\infty} \left(\frac{1}{n^2} + \frac{1}{(n+1)^2} + \cdots + \frac{1}{(2n)^2} \right) = 0$.

(*Indicación:* compárese esta suma con su término mayor.)

(*c*) Demuéstrese que $\lim_{n\to\infty} \left(\frac{1}{\sqrt{n}} + \frac{1}{\sqrt{n+1}} + \cdots + \frac{1}{\sqrt{2n}} \right) = \infty$.

*(*d*) Demuéstrese que $\lim_{n\to\infty} \left(\frac{1}{\sqrt{n^2+1}} + \frac{1}{\sqrt{n^2+2}} + \cdots + \frac{1}{\sqrt{n^2+n}} \right) = 1$.

6. Demuéstrese que todo decimal periódico representa un número racional. (Compárese con sección 1.1c, problema 3.)

7. Demuéstrese que $\lim_{n\to\infty} \frac{n^{100}}{1.01^n}$ existe y determínese su valor.

8. Demuéstrese que si a y $b \leq a$ son positivos, la sucesión $\sqrt[n]{a^n + b^n}$ converge a a. De la misma manera, para cualesquiera k números positivos fijos a_1, a_2, ..., a_k demuéstrese que $\sqrt[n]{a_1^n + a_2^n + \cdots + a_k^n}$ converge y encuéntrese su límite.

9. Demuéstrese que la sucesión $\sqrt{2}, \sqrt{2\sqrt{2}}, \sqrt{2\sqrt{2\sqrt{2}}}, \ldots$, converge. Encuéntrese el límite.

10. Si $\nu(n)$ es el número de factores primos que tiene n, demuéstrese que

$$\lim_{n\to\infty} \frac{\nu(n)}{n} = 0.$$

11. Demuéstrese que si $\lim_{n\to\infty} a_n = \xi$, entonces $\lim_{n\to\infty} \sigma_n = \xi$, donde σ_n es la media aritmética $(a_1 + a_2 + \cdots + a_n)/n$.

12. Encuéntrense

(*a*) $\lim_{n\to\infty} \left(\frac{1}{1\cdot 2} + \frac{1}{2\cdot 3} + \cdots + \frac{1}{n(n+1)} \right)$

$\left(\textit{Indicación:} \ \frac{1}{k(k+1)} = \frac{1}{k} - \frac{1}{k+1} \right)$

(*b*) $\lim_{n\to\infty} \left(\frac{1}{1\cdot 2\cdot 3} + \frac{1}{2\cdot 3\cdot 4} + \cdots + \frac{1}{n(n+1)(n+2)} \right)$

13. Si $a_0 + a_1 + \cdots + a_p = 0$, demuéstrese que

$$\lim_{n\to\infty} (a_0\sqrt{n} + a_1\sqrt{n+1} + \cdots + a_n\sqrt{n+p}) = 0.$$

(*Indicación:* tómese $\sqrt{n}$ como factor común.)

14. Demuéstrese que $\lim_{n\to\infty} \sqrt[(2n+1)]{(n^2+n)} = 1$.

***15.** Sea a_n una sucesión tal que la sucesión $b_n = pa_n + qa_{n+1}$, donde $|p| < q$, es convergente. Probar que a_n converge. Si $|p| \geq q > 0$, mostrar que a_n no converge necesariamente.

16. Demuéstrese la validez de la relación

$$\lim_{n\to\infty} \frac{1}{n^{k+1}} \sum_{i=1}^{n} i^k = \frac{1}{k+1}$$

para cualquier entero k no negativo. (*Indicación:* aplíquese la inducción con respecto a k y empléese la relación

$$\sum_{i=1}^{n} [i^{k+1} - (i-1)^{k+1}] = n^{k+1},$$

con el desarrollo de $(i-1)^{k+1}$ en las potencias de i.)

SECCION 1.7, página 93

***1.** Sean a_1 y b_1 dos números positivos cualesquiera y sea $a_1 < b_1$. Definamos a_2 y b_2 por las ecuaciones

$$a_2 = \sqrt{a_1 b_1}, \qquad b_2 = \frac{a_1 + b_1}{2}.$$

De la misma manera, sean

$$a_3 = \sqrt{a_2 b_2}, \qquad b_3 = \frac{a_2 + b_2}{2},$$

y, en general,

$$a_n = \sqrt{a_{n-1} b_{n-1}}, \qquad b_n = \frac{a_{n-1} + b_{n-1}}{2}.$$

Demuéstrese (*a*) que la sucesión $a_1, a_2, \ldots$, converge, (*b*) que la sucesión $b_1, b_2, \ldots$, converge y (*c*) que las dos sucesiones tienen el mismo límite. (Este límite se llama *media aritmético-geométrica* de a_1 y b_1.)

***2.** Demuéstrese que el límite de la sucesión

$$\sqrt{2}, \sqrt{2+\sqrt{2}}, \sqrt{2+\sqrt{2+\sqrt{2}}}, \ldots$$

(*a*) existe y (*b*) es igual a 2.

***3.** Demuéstrese que el límite de la sucesión

$$a_n = \frac{1}{n} + \frac{1}{n+1} + \cdots + \frac{1}{2n}$$

existe. Dedúzcase que el límite es menor que 1 pero no menor que $\frac{1}{2}$.

4. Demuéstrese que el límite de la sucesión

$$b_n = \frac{1}{n+1} + \cdots + \frac{1}{2n}$$

existe y es igual al límite del ejemplo anterior.

5. Obténganse las cotas siguientes del límite L de los dos ejemplos anteriores: $37/60 < L < 57/60$.

***6.** Sean a_1, b_1 dos números positivos cualesquiera tales que $a_1 \leq b_1$. Sean

$$a_2 = \frac{2a_1b_1}{a_1 + b_1}, \qquad b_2 = \sqrt{a_1b_1},$$

y en general

$$a_n = \frac{2a_{n-1}b_{n-1}}{a_{n-1} + b_{n-1}}, \qquad b_n = \sqrt{a_{n-1}b_{n-1}}.$$

Demuéstrese que las sucesiones $a_1, a_2, \ldots$ y $b_1, b_2, \ldots$ convergen y tienen el mismo límite.

***7.** Demuéstrese que $1/e = 1 - 1 + \frac{1}{2!} - \frac{1}{3!} + \cdots + \frac{(-1)^n}{n!} + \cdots$. *(Indicación:* considérese el producto de las n-ésimas sumas parciales de los desarrollos de e y $1/e$.)

8. (a) Sin recurrir al teorema del binomio, demuéstrese que $a_n = (1 + 1/n)^n$ es monótona creciente y que $b_n = (1 + 1/n)^{n+1}$ es monótona decreciente. *(Indicación:* considérense a_{n+1}/a_n y b_n/b_{n+1}. Aplíquese el resultado de la sección 1.5, problema 2.)

(b) ¿Qué número es mayor $(1{,}000{,}000)^{1{,}000{,}000}$ ó $(1{,}000{,}001)^{999{,}999}$?

9. (a) A partir de los resultados del problema 8a demuéstrese que

$$\left(\frac{n}{e}\right)^n < n! < e(n + 1)\left(\frac{n}{e}\right)^n.$$

(b) Para $n > 6$ dedúzcase la desigualdad más fuerte

$$n! < n\left(\frac{n}{e}\right)^n.$$

***10.** Si $a_n > 0$ y $\lim_{n\to\infty} \frac{a_{n+1}}{a_n} = L$, entonces $\lim_{n\to\infty} \sqrt[n]{a_n} = L$.

11. Por medio del problema 10, evalúense los límites de las sucesiones siguientes:

(a) $\sqrt[n]{n}$, (b) $\sqrt[n]{n^5 + n^4}$, (c) $\sqrt[n]{\frac{n!}{n^n}}$

12. Por medio del problema 11c, demuéstrese que

$$n! = n^n e^{-n} a_n,$$

donde a_n es un número cuya n-ésima raíz tiende a 1. (Véase el Apéndice, capítulo 7.)

13. (a) Evalúese

$$\frac{1}{1 \cdot 3} + \frac{1}{2 \cdot 4} + \cdots + \frac{1}{n(n + 2)}.$$

(Indicación: compárese con la sección 1.6, problema 12a.)

(b) A partir del resultado anterior, demuéstrese que $\sum_{k=1}^{\infty} \frac{1}{2}$ converge.

14. Sean p y q números naturales arbitrarios. Evalúense

(a) $\sum_{k=1}^{n} \frac{1}{(k + p)(k + p + q)}$.

(b) $\sum_{k=1}^{n} \frac{1}{(k + p)(k + p + q)}$.

15. Evalúense

(a) $\frac{1}{1 \cdot 2 \cdot 3} + \frac{1}{2 \cdot 3 \cdot 4} + \cdots + \frac{1}{n(n + 1)(n + 2)}$

(b) $\sum_{k=1}^{n} \frac{1}{k(k + 1)(k + 3)}$.

(c) Evalúense los límites de las expresiones anteriores cuando $n \to \infty$.

***(d)** Sean $a_1, a_2, \ldots, a_m$ enteros no negativos donde $a_1 < a_2 < \cdots < a_m$. Señálese la manera de obtener una fórmula de

$$S_n = \sum_{k=1}^{n} \frac{1}{(k + a_1)(k + a_2) \cdots (k + a_m)}$$

y de encontrar $\lim_{n\to\infty} S_n$.

16. Si a_k es monótona y $\sum_{k=1}^{\infty} a_k$ converge, demuéstrese que $\lim_{k\to\infty} ka_k = 0$.

17. Si a_k es monótona decreciente con límite 0 se tiene que $b_k = a_k - 2a_{k+1} + a_{k+2} \geq 0$ para toda k, demuéstrese entonces que $\sum_{k=1}^{\infty} kb_k = a_1$.

SECCION 1.8, página 105

1. Demuéstrese que $\lim_{m\to\infty} (\cos \pi x)^{2m}$ existe para cada valor de x y es igual a 1 o 0 según sea x entero o no lo sea.

2. (a) Demuéstrese que $\lim_{n\to\infty} [\lim_{m\to\infty} (\cos n! \, \pi x)^{2m}]$ existe para cada valor de x y es igual a 1 o 0 según sea x racional o irracional.

(b) Examínese la cuestión de la continuidad de estas funciones límite.

3. Sea $f(x)$ continua para $0 \leq x \leq 1$. Supongamos además que $f(x)$ no toma sino valores racionales y que $f(x) = \frac{1}{2}$ cuando $x = \frac{1}{2}$. Demuéstrese que $f(x) = \frac{1}{2}$ en todas partes.

SECCION 1.S.1, página 112

1. Sean $r = p/q$, $s = m/n$ números racionales arbitrarios, en los que p, q, m, n son enteros y q, n son positivos. Defínanse, en términos de los enteros p, q, m, n,

(a) $r + s$, (b) $r - s$, (c) rs, (d) $\frac{r}{s}$, (e) $r < s$.

2. Demuéstrese para encajes de intervalos racionales, $[a_n, b_n]$ y $[a_n', b_n']$, que cada una de las condiciones siguientes es necesaria y suficiente para tener la equivalencia:

(a) $a_n' - a_n$ es una sucesión nula,

(b) $a_n \leq b_n'$ y $a_n' \leq b_n$.

3. Dados $x \sim \{[a_n, b_n]\}$, $y \sim \{[\alpha_n, \beta_n]\}$, (a) verifíquese que las definiciones de adición y substracción,

$$x + y = \{[a_n + \alpha_n, b_n + \beta_n]\}, \qquad x - y = \{[a_n - \beta_n, b_n - \alpha_n]\},$$

tienen sentido. En particular, verifíquese que

(i) las representaciones dadas constituyen, de hecho, encajes de conjuntos para $x + y$ y $x - y$ cuando x e y son racionales;

(ii) si $x < y$, entonces $x + z < y + z$, donde z es un número real arbitrario.

(b) Defínase el producto xy y verifíquese sobre todo que la definición de producto tiene sentido, y que

(i) el conjunto encajado que se da constituye, en efecto, un encaje para xy cuando x e y son racionales.

(ii) si $x < y$ y $z > 0$, entonces $xz < yz$.

4. Demuéstrese que los principios siguientes son equivalentes en el sentido de que uno cualquiera de ellos se puede deducir como consecuencia de cualquier otro.

(a) Todo encaje de intervalos con extremos reales contiene un número real.

(b) Toda sucesión monótona y acotada converge.

(c) Toda sucesión infinita y acotada tiene por lo menos un punto de acumulación o límite.

(d) Toda sucesión de Cauchy converge.

(e) Todo conjunto acotado de números reales tiene un ínfimo y un supremo.

Problemas misceláneos

1. Si $w_1, w_2, \ldots, w_n > 0$, demuéstrese que el promedio ponderado

$$\frac{w_1x_1 + w_2x_2 + \cdots + w_nx_n}{w_1 + w_2 + \cdots w_2}$$

queda entre la mayor y la menor de las x.

2. Demuéstrese que

$$2(\sqrt{n} + 1 - 1) < 1 + \frac{1}{\sqrt{2}} + \frac{1}{\sqrt{3}} + \cdots + \frac{1}{\sqrt{n}} < 2\sqrt{n}.$$

3. Demuéstrese para $x, y > 0$, que

$$\frac{x^n + y^n}{2} \geq \left(\frac{x + y}{2}\right)^n.$$

Interprétese geométricamente este resultado en términos de la gráfica de x^n.

4. Si $a_1 \geq a_2 \geq \cdots \geq a_n$ y $b_1 \geq b_2 \geq \cdots \geq b_n$, probar que

$$n\sum_1^n a_ib_i \geq \left(\sum_{i=1}^n a_i\right)\left(\sum_{i=1}^n b_i\right).$$

5. (a) Demuéstrese que la sucesión $a_1, a_2, a_3, \ldots$ se puede expresar como la sucesión de las sumas parciales de la serie $u_1, u_2, u_3, \ldots$ donde $u_n = a_n - a_{n-1}$ para $n \geq 1$ y $u_1 = a_1$.

(b) Exprésese la sucesión $a_n = n^3$ como sucesión de sumas parciales de una serie.

(c) *Del resultado anterior, obténgase una fórmula de la suma parcial n-ésima* de la serie

$$1 + 4 + 9 + \cdots + n^2 + \cdots.$$

(d) De la fórmula de $1^2 + 2^2 + \cdots + n^2$, obténgase una fórmula para

$$1^2 + 3^2 + 5^2 + \cdots + (2n + 1)^2.$$

6. Una sucesión se llama progresión aritmética de primer orden si las diferencias entre los términos sucesivos son constantes. Se llama progresión aritmética de

segundo orden si las diferencias de términos sucesivos constituyen una progresión aritmética de primer orden; y, en general, se llama progresión aritmética de orden k si las diferencias de términos sucesivos constituyen una progresión aritmética de orden $(k - 1)$.

Los números 4, 6, 13, 27, 50, 84 son los primeros seis términos de una progresión aritmética. ¿Cuál es el menor orden posible? ¿Cuál es el octavo término de la progresión del menor orden con estos términos iniciales?

7. Demuéstrese que el n-ésimo término de una progresión aritmética de segundo orden se puede expresar en la forma $an^2 + bn + c$, donde a, b, c son independientes de n.

***8.** Demuéstrese que el n-ésimo término de una progresión aritmética de orden k se puede expresar en la forma $an^k + bn^{k-1} + \cdots + pn + q$, donde $a, b, \ldots, p, q$ son independientes de n.

Encuéntrese el término n-ésimo de la progresión del menor orden en el problema 6.

9. Encuéntrese una fórmula del término n-ésimo de las progresiones aritméticas de menor orden de las que los términos iniciales son los siguientes:

(a) 1, 2, 4, 7, 11, 16, ...

(b) −7, −10, −9, 1, 25, 68, ...

***10.** Demuéstrese que la suma de los n primeros términos de una progresión aritmética de orden k es

$$a_kS_k + a_{k-1}S_{k-1} + \cdots + a_1S_1 + a_0n,$$

donde S_ν representa la suma de las n primeras potencias ν-ésimas y las a_i son independientes de n. Empléese este resultado para evaluar las sumas de las progresiones aritméticas del problema 9.

11. Al sumar

$$\nu(\nu + 1)(\nu + 2)\cdots(\nu + k + 1) - (\nu - 1)\nu(\nu + 1)\cdots(\nu + k)$$

desde $\nu = 1$ hasta $\nu = n$, demuéstrese que

$$\sum_{\nu=1}^{n} \nu(\nu + 1)(\nu + 2)\cdots(\nu + k) = \frac{n(n + 1)\cdots(n + k + 1)}{k + 2}.$$

12. Evalúese $1^3 + 2^3 + \cdots + n^3$ por medio de la relación

$$\nu^3 = \nu(\nu + 1)(\nu + 2) - 3\nu(\nu + 1) + \nu.$$

13. Demuéstrese que la función

$$f(x) = \begin{cases} \dfrac{1}{\log_2 |x|}, & x \neq 0 \\ 0, & x = 0 \end{cases}$$

es continua, aunque sin la continuidad de Hölder. (*Indicación:* mostrar que la continuidad de Hölder con exponente α no se cumple en el origen, al considerar los valores $x = 1/2^{n/\alpha}$.)

14. Sea a_n una sucesión monótona decreciente de números no negativos. Demuéstrese que $\sum_{n=1}^{\infty} a_n$ converge si y sólo si lo hace $\sum_{\nu=0}^{\infty} 2^\nu a_{2^\nu}$.

15. En lo que sigue, investíguese la convergencia y determínese el límite cuando ello sea posible.

(a) $n!\, e - [n!\, e]$

(b) a_n/a_{n+1}, donde $a_1 = 0$, $a_2 = 1$, y $a_{k+2} = a_{k+1} + a_k$.

2

Las ideas fundamentales del cálculo integral y diferencial

Los procesos de límite fundamentales del cálculo son la integración y la derivación o diferenciación. Casos aislados de estos procesos del cálculo (que culminaron con el trabajo de Arquímedes) fueron considerados aún en la antigüedad, y con frecuencia creciente en los siglos XVI y XVII. Sin embargo, el desarrollo sistemático del cálculo, no iniciado sino hasta en el siglo XVII, es usualmente atribuido a los dos grandes pioneros de la ciencia, Newton y Leibnitz. La clave para este desarrollo sistemático es la comprensión de que los procesos de derivación e integración, que fueron tratados separadamente, están íntimamente relacionados al ser recíprocos uno del otro.[1]

Una evaluación histórica imparcial de los méritos no puede atribuir la invención del cálculo a repentinos destellos inexplicables de genio por parte de uno o dos individuos. Muchas gentes, tales como Fermat, Galileo y Kepler, estimulados por las nuevas ideas revolucionarias en la ciencia, contribuyeron a los fundamentos del cálculo. De hecho, el profesor de Newton, Barrow, estuvo casi en posesión completa de la comprensión básica de la reciprocidad entre derivación e integración, que es la piedra angular del cálculo sistemático de Newton y Leibnitz. Newton expuso los conceptos de manera un poco más clara; por otra parte, la ingeniosa notación de Leibnitz y sus métodos de cálculo son altamente sugestivos y siguen siendo indispensables. El trabajo de estos dos hombres estimuló inmediatamente las ramas superiores del análisis, incluyendo el cálculo de variaciones y la teoría de ecuaciones diferenciales, y condujo a innumerables aplicaciones en las ciencias naturales. Muy curioso resulta que, no obstante que Newton, Leibnitz y sus sucesores inmediatos hicieron tales variados usos de la poderosa herramienta puesta en sus manos, ninguno consiguió clarificar completamente los conceptos básicos involucrados en sus trabajos. Sus argumentos emplearon "cantidades infinitamente pequeñas" en formas que son lógicamente insostenibles y no convincentes. La clarificación aconteció a fines del siglo XIX con la cuidadosa formulación del concepto de límite y con el análisis del continuo de números, como se explica en el capítulo 1.[1]

[1] Este enunciado constituye el "teorema fundamental del cálculo".

Comenzaremos con una discusión de los conceptos fundamentales. Estos pueden apreciarse totalmente sólo a través de ilustraciones concretas y ejemplos; es, por lo tanto, que se recomienda aquí, como en muchas partes en este libro, que las secciones teóricas y generales sean cuidadosamente estudiadas nuevamente después de que el lector haya absorbido el material más específico y concreto en secciones subsiguientes.

2.1 La integral

a. Introducción

Sólo después de un prolongado desarrollo los procedimientos sistemáticos de integración y derivación alcanzaron la necesidad de descripciones matemáticas precisas para nociones intuitivas que surgen en la geometría y en las ciencias naturales. La derivación es el concepto requerido para describir las nociones de *tangentes* a curvas y el de velocidad de partículas en movimiento, o bien, más generalmente, el concepto de *razón de variación o cambio*. El concepto intuitivo de *área* de una región con fronteras curvas encuentra su formulación matemática precisa en el proceso de integración. Muchos otros conceptos relacionados en la geometría y en la física requieren también de la integración, como se verá posteriormente. En esta sección se introduce el concepto de integral en conexión con el problema de medir el área de una región plana acotada por curvas.

Areas. Se tiene una percepción intuitiva de que una región contenida dentro de una curva cerrada posee un "área" la cual mide el número de unidades cuadradas dentro de la curva. Empero, la pregunta de *cómo* esta medida para el área puede ser descrita en términos precisos exige una cadena de pasos matemáticos. Las propiedades básicas del área que la intuición sugiere son: el área es un número (positivo, dependiente de la elección de la unidad de longitud); este número es el mismo para figuras congruentes; para todos los rectángulos es el producto de las longitudes de

[1] El surgimiento del cálculo, que se prolonga por más de 2,000 años, representa uno de los capítulos más fascinantes en la historia del descubrimiento científico. A los lectores interesados se les recomienda el libro de Carl B. Boyer, *Concept of the Calculus*, Hafner Publishing Company, 1949. Véase también O. Toeplitz, *Calculus*, A Genetic Approach, University of Chicago, 1963.

los lados adyacentes; y, finalmente, para una región descompuesta en secciones el área del todo es igual a la suma de las áreas de las secciones.

Una consecuencia inmediata es el hecho de que: para una región A que es parte de una región B, el área de A no puede exceder el área de B.

Estas propiedades permiten el cálculo directo del área de cualquier figura que puede ser descompuesta en un número finito de rectángulos. Más generalmente, para asignar un valor F al área de una región R, consideramos otras dos regiones R' (inscripta) y R'' (circunscripta) que se puedan descomponer en rectángulos, donde R'' contiene a R y R' está contenida en R

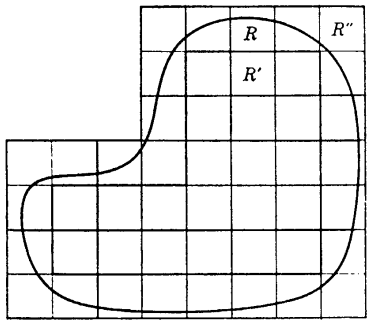


Figura 2.1 Aproximación de un área.

(ver Fig. 2.1). Se sabe entonces, por lo menos, que F debe encontrarse entre las áreas de R' y R''. El valor de F está completamente determinado si se encuentran sucesiones de regiones circunscriptas R_n'' y regiones inscriptas R_n' que puedan ambas descomponerse en rectángulos y tales que las áreas de R_n'' y R_n' tengan el mismo límite cuando n tiende a infinito. Este es, remontándonos a la antigüedad, el método de "exhaucción", el cual es usado en geometría elemental para describir el área de un círculo.[1] La formulación precisa de esta idea intuitiva conduce ahora a la noción de integración.

b. La integral como un área

Area bajo una curva

La noción analítica de integral surge cuando se asocian áreas con *funciones:* Considérese el área de una región acotada por la izquierda y por

[1] Por supuesto, puede utilizarse cualquier clase de *polígonos* inscriptos y circunscritos, puesto que un polígono puede ser descompuesto en triángulos rectángulos y el área de un triángulo rectángulo es claramente la mitad de la de un rectángulo con los mismos lados.

la derecha mediante líneas verticales $x = a$ y $x = b$, debajo por el eje x y arriba por la gráfica de una función positiva continua $f(x)$ (Fig. 2.2). En breve, nos referiremos a esto como el área "bajo la curva". Por el mbmento se acepta como intuitiva la idea de que el área de tal región es un número definido. Llamamos esta área F_a^b la *integral de la función* f entre los lí-

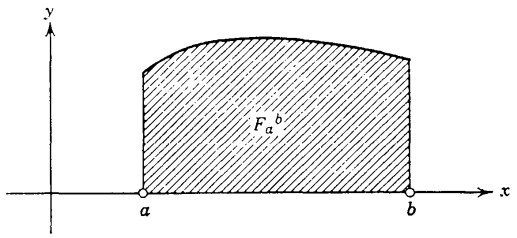


Figura 2.2.

mites[1] a y b. En la investigación del valor numérico de F_a^b se hace uso de aproximaciones mediante sumas de áreas de rectángulos. Para este propósito se divide el intervalo (a, b) del eje x en n partes (pequeñas), no necesariamente del mismo tamaño, las cuales se denominarán *celdas.* En cada punto de división se traza la línea perpendicular al eje x, hacia arriba hasta la curva. La región con área F_a^b es así dividida en n franjas, cada una

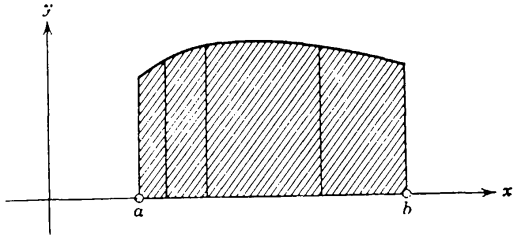


Figura 2.3.

acotada por una porción de la gráfica de la función $f(x)$ y por tres segmentos de recta (Fig. 2.3).

Area o integral como límite de una suma. El cálculo preciso del área de tales franjas no es más fácil que el cálculo del área de la región original. Es, sin embargo, dar un paso hacia adelante el aproximar el área de cada franja desde arriba y desde abajo mediante las áreas de los rectángulos circunscritos e inscritos con la misma base, donde la frontera curva de la

[1] No debe surgir confusión del uso de la palabra "límite" para puntos frontera del intervalo de integración.

franja se reemplaza por una línea horizontal a una distancia del eje x que es o bien el máximo o bien el mínimo valor de $f(x)$ en la celda (Fig. 2.4). Más generalmente, se obtiene una aproximación intermedia si se reemplaza la franja por un rectángulo de la misma base y acotado en la parte superior por una línea horizontal que intersecta la frontera curva de la franja (véase Fig. 2.5). Analíticamente, esto equivale a reemplazar la fun-

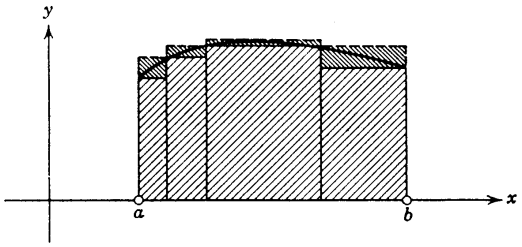


Figura 2.4.

ción $f(x)$ en cada una de las celdas por algún valor intermedio constante. Denótese por F_n la suma de las n áreas rectangualres. La intuición nos dice que los valores F_n tienden a F_a^b si la subdivisión se hace cada vez más fina, esto es, si se hace crecer n sin límite mientras que la máxima longitud de las celdas individuales tiende a cero. De esta forma F_a^b está representado como un límite de áreas que consisten de rectángulos.

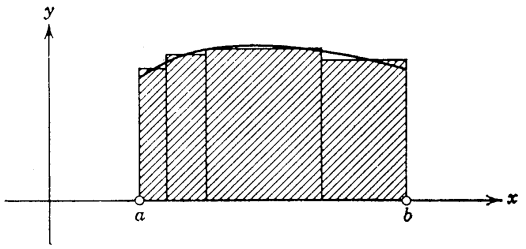


Figura 2.5.

c. Definición analítica de la integral. Notaciones

Definición y existencia de integrales

En el párrafo anterior aceptamos al área bajo una curva como una cantidad dada intuitivamente, y a continuación ésta se representó como un valor límite. Ahora se invertirá el procedimiento. No se invocará más a la intuición para asignar un área a la región bajo una curva continua; por

el contrario, se comenzará de manera puramente analítica con las sumas F_n definidas anteriormente y se *probará* que estas sumas tienden a un límite definido. Este límite es entonces la *definición* precisa de la integral y del área.

Sea la función $f(x)$ continua (pero no necesariamente positiva) en el intervalo cerrado $a \leq x \leq b$. Se divide el intervalo por $(n-1)$ puntos $x_1, x_2, \ldots, x_{n-1}$ en n celdas iguales o distintas, con las longitudes

$$x_i - x_{i-1} = \Delta x_i, \qquad (i = 1, 2, \ldots, n)^1,$$

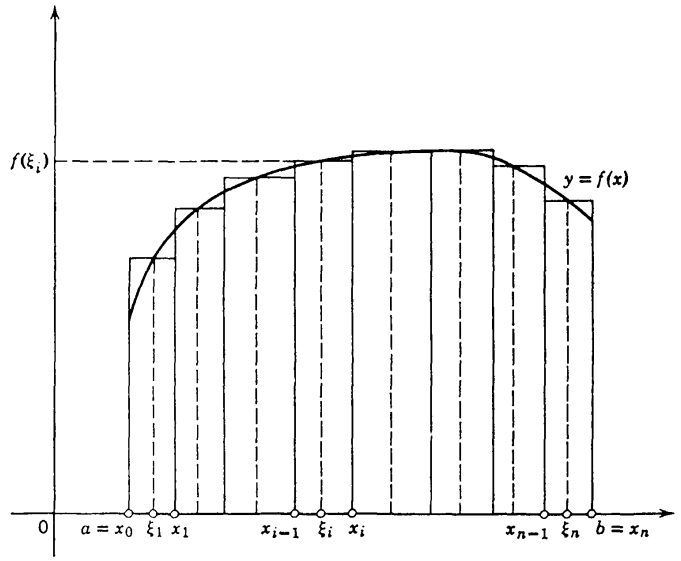


Figura 2.6 Para ilustrar la definición analítica de integral.

donde, además, se hace $x_0 = a$, $x_n = b$ (ver Fig. 2.6). En cada subintervalo cerrado $[x_{i-1}, x_i]$ o celda se elige cualquier punto ξ_i. Se forma la suma

$$\begin{aligned} F_n &= f(\xi_1)(x_1 - x_0) + f(\xi_2)(x_2 - x_1) + \cdots + f(\xi_n)(x_n - x_{n-1}) \\ &= f(\xi_1)\Delta x_1 + f(\xi_2)\Delta x_2 + \cdots + f(\xi_n)\Delta x_n. \end{aligned}$$

Utilizando el símbolo de sumatoria se escribe más concisamente

$$F_n = \sum_{i=1}^{n} f(\xi_i)(x_i - x_{i-1})$$

[1] El símbolo Δ no debe ser interpretado como un factor sino solamente como indicador de una diferencia en valores de la variable que le sigue. Así, el símbolo Δx_i significa la diferencia $x_i - x_{i-1}$ de valores consecutivos de x.

o bien

$$F_n = \sum_{i=1}^{n} f(\xi_i)\Delta x_i.$$

Si $f(x)$ es positivo, el valor F_n representa el área bajo la curva, obtenida substituyendo f en cada subintervalo por el valor constante $f(\xi_i)$. Por supuesto, las sumas F_n pueden ser formadas sin suponer que f es positivo. Parece plausible intuitivamente que las sumas F_n deben tender a un límite F_a^b conforme el número n de intervalos crece indefinidamente y al mismo tiempo la longitud del máximo subintervalo tiende a cero. Esto implicaría que el valor del límite F_a^b es independiente de la manera particular según la cual los puntos de subdivisión $x_1, x_2, \ldots, x_{n-1}$ y los puntos intermedios $\xi_1, \xi_2, \ldots, \xi_n$ son elegidos. Denominamos a F_a^b la integral de $f(x)$ entre los límites a y b.

La intuición geométrica, no importa cuán convincente, puede servir solamente como guía a los procesos de límite analíticos; por lo tanto, una justificación analítica es necesaria y se debe proporcionar una demostración para la existencia de la integral como el límite descripto arriba. Además, como se ha dicho ya, no debe insistirse en la hipótesis de que la función f es positiva en el intervalo.

Así se afirma:

TEOREMA DE EXISTENCIA. *Para toda función continua $f(x)$ en un intervalo cerrado $[a, b]$ la integral sobre este intervalo existe como el límite de las sumas F_n descritas arriba (independientemente de la elección de los puntos de subdivisión $x_1, \ldots, x_{n-1}$ y de los puntos intermedios $\xi_1, \ldots, \xi_n$, siempre que el máximo de las longitudes Δx_i tienda a cero).*

Se adquirirá primero cierta experiencia y comprensión antes de considerar la demostración de la existencia de la integral, lo que se hará en el Suplemento (p. 213).

Notación de Leibnitz para la integral

La definición de la integral como el límite de una suma condujo a Leibnitz a expresar la integral mediante el siguiente símbolo:

$$\int_a^b f(x)\,dx.$$

El signo de integral es una modificación del signo de sumatoria en forma de una S grande que se usó en la época de Leibnitz. El paso al límite a partir de una subdivisión finita en porciones Δx_i es indicada mediante el uso de la letra d en lugar de Δ. Sin embargo, al utilizarse esta notación no debe tolerarse el misticismo del siglo XVIII de considerar dx como un

"infinitamente pequeño" o "cantidad infinitesimal", o de considerar la integral como una "suma de un número infinito de cantidades infinitamente pequeñas". Tal concepción está desprovista de significado claro y oscurece lo que anteriormente se ha formulado con precisión. Desde el punto de vista presente, el símbolo individual dx no ha sido definido. La sugestiva combinación de símbolos $\int_a^b f(x)\,dx$ está definida para una función $f(x)$ en el intervalo $[a, b]$ mediante la formación de las sumas ordinarias F_n y pasando al límite con $n \to \infty$.

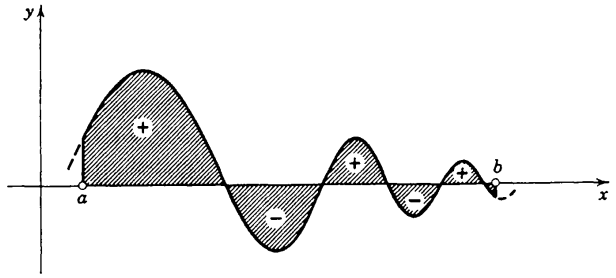


Figura 2.7.

El símbolo particular que se utiliza para la variable de integración es una cuestión sin importancia alguna (así como en la notación para sumas no importa cómo se denominó al índice de la sumatoria); en lugar de $\int_a^b f(x)\,dx$ puede igualmente escribirse $\int_a^b f(t)\,dt$ o bien $\int_a^b f(u)\,du$. El *integrando* denotado por f es una función de una variable independiente sobre el intervalo $[a, b]$ y el nombre de la variable es indiferente. Solamente los puntos extremos del intervalo de integración a y b afectan el valor de la integral para una f dada. Expresiones como $\int_a^b f(x)\,dx$ o bien $\int_a^b f(a)\,da$, en las cuales se utiliza la misma letra para la variable de integración y un punto extremo, son engañosas bajo nuestra definición y deberán, al principio, evitarse.

Si el *integrando* $f(x)$ es positivo en el intervalo $[a, b]$, puede inmediatamente identificarse $\int_a^b f(x)\,dx$ con el área acotada por la gráfica de f y las rectas $x = a$, $x = b$ e $y = 0$. Sin embargo, la integral de f está definida analíticamente como el límite de las sumas F_n independientemente de cualquier suposición acerca del signo de f. Si $f(x)$ es negativa en todo o parte del intervalo, el único efecto es el de hacer los correspondientes factores

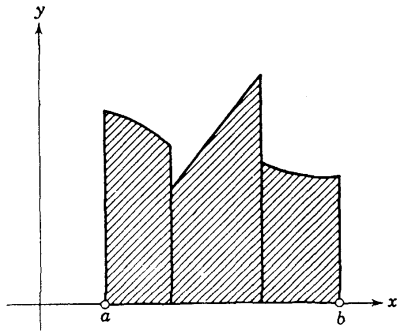


Figura 2.8.

$f(\xi_i)$ en la suma negativos en vez de positivos. A la región acotada por la parte de la curva debajo del eje x se asignará entonces naturalmente un área negativa. La integral será así la suma de términos positivos y negativos, correspondiendo respectivamente a porciones de la curva encima y debajo del eje x [1] (véase Fig. 2.7).

Es intuitivamente convincente afirmar que nuestro proceso de límite converge, aún si la función $f(x)$ no es continua en todas partes sino que

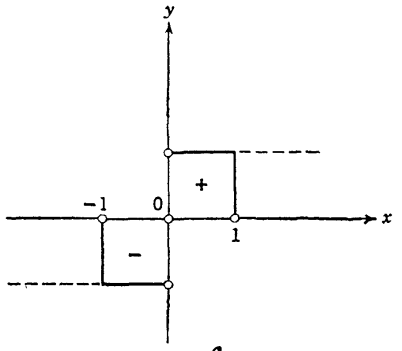


Figura 2.9 $\int_{-1}^{1} \operatorname{sgn} x\, dx = 0.$

posee discontinuidades de salto en uno o varios puntos, como la función indicada por la curva en la Fig. 2.8, donde claramente existe un área bajo la curva.[2]

[1] Las áreas de regiones acotadas por curvas cerradas arbitrarias se considerarán en el capítulo 4.

[2] Como otro ejemplo considérese $f(x) = \operatorname{sgn} x$ en $[-1, 1]$. Se tiene $f(x) = -1$ para $x < 0$ y $f(x) = +1$ para $x > 0$ (véase Fig. 2.9). Entonces $\int_{-1}^{+1} f(x)\, dx = 0$.

Así, el proceso de límite anterior puede muy bien resultar en un límite definido de la suma F_n para funciones que poseen algunas discontinuidades; y esta posibilidad se indica denominando a tales funciones *integrables*. A mediados del siglo XIX el gran Bernhard Riemann analizó por primera vez la aplicabilidad del proceso de integración a funciones generales. Más recientemente se han introducido varias extensiones del concepto mismo de integración. Empero, tales refinamientos poseen una menor importancia inmediata para el cálculo relativo a fenómenos intuitivamente accesibles, y no será necesario para nosotros destacar siempre la integrabilidad de nuestras funciones como mero recordatorio de que pueden definirse funciones no integrables.

En cursos avanzados la integral que se ha definido aquí se denomina "integral de Riemann", para distinguirla de varios conceptos generalizados de integral; las sumas de aproximación F_n se denominan sumas de *Riemann*.

2.2 Ejemplos elementales de integración

En un buen número de casos significativos estamos ya en condiciones de calcular la integral de una función llevando a cabo los procesos de límite prescriptos. Esto se hará mediante una evaluación explícita de las sumas F_n para una elección adecuada de los puntos intermedios ξ_i (usualmente, el punto extremo izquierdo o derecho de las celdas). El teorema sobre la existencia de la integral de una función continua asegura que el límite de los F_n es el mismo para cualquier otra elección de los puntos intermedios ξ_i y para cualquier método de subdivisión.

a. Integración de una función lineal

Primero se verificará que la integral realmente da el valor correcto del área para algunas figuras sencillas que conocemos de la geometría.

Sea $f(x) = \text{constante} = \gamma$. Para calcular la integral de $f(x)$ entre los límites a y b se forman las sumas F_n (véase Fig. 2.10). Puesto que aquí $f(\xi_i) = \gamma$, se encuentra

$$F_n = \sum_{i=1}^{n} \gamma\, \Delta x_i = \gamma \sum_{i=1}^{n} \Delta x_i = \gamma(b - a).$$

Por lo tanto, de manera análoga

$$\lim_{n\to\infty} F_n = \int_a^b \gamma\, dx = \gamma(b - a).$$

Esta es justamente la fórmula para el área de un rectángulo de altura γ y base $b - a$.

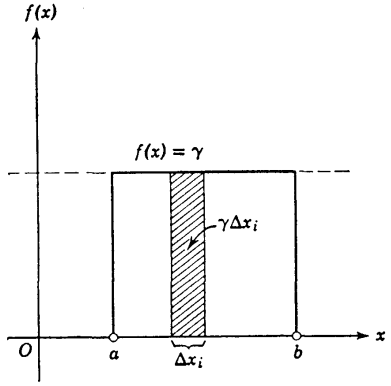


Figura 2.10 Integral de una constante.

La integral de la función $f(x) = x$,

$$\int_a^b x\,dx,$$

(Fig. 2.11), como se sabe de la geometría elemental, posee el valor

$$\tfrac{1}{2}(b-a)(b+a) = \tfrac{1}{2}(b^2 - a^2).$$

Para confirmar que el proceso límite conduce analíticamente al mismo resultado, se subdivide el intervalo de a a b en n partes iguales por medio de los puntos de división

$$a + h, a + 2h, \ldots, a + (n-1)h,$$

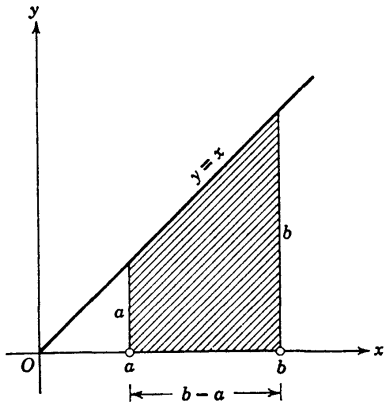


Figura 2.11.

donde $h = (b-a)/n$. Tomando para ξ_i el punto extremo derecho de cada intervalo se encuentra la integral como el límite, para $n \to \infty$, de la suma

$$F_n = (a+h)h + (a+2h)h + \cdots + (a+nh)h$$
$$= nah + (1 + 2 + 3 + \cdots + n)h^2 = nah + \tfrac{1}{2}n(n+1)h^2,$$

donde se ha utilizado la bien conocida fórmula para la suma de una progresión aritmética (véase p. 134, problema 3). Substituyendo $h = (b-a)/n$, se ve que

$$F_n = a(b-a) + \frac{1}{2}\left(1 + \frac{1}{n}\right)(b-a)^2,$$

de lo que se sigue inmediatamente que

$$\lim_{n\to\infty} F_n = a(b-a) + \tfrac{1}{2}(b-a)^2 = \tfrac{1}{2}(b^2 - a^2).$$

b. Integración de x^2

La geometría elemental no conduce tan fácilmente a la integración de la función $f(x) = x^2$, esto es, a la determinación del área de la región[1] acotada por un segmento de una parábola, un segmento del eje x y dos coordenadas. Es necesario un verdadero proceso de límite. Suponiendo $a < b$, se eligen los mismos puntos de división y los mismos puntos intermedios que en el ejemplo anterior (véase Fig. 2.12). Se sigue entonces que la integral de x^2 entre los límites a y b es el límite de las sumas

$$F_n = (a+h)^2h + (a+2h)^2h + \cdots + (a+nh)^2h$$
$$= na^2h + 2ah^2(1 + 2 + 3 + \cdots + n)$$
$$+ h^3(1^2 + 2^2 + 3^2 + \cdots + n^2);$$

usando los valores conocidos de las sumas encerradas entre paréntesis se encuentra (véase pp. 81-82)

$$F_n = na^2h + n(n+1)ah^2 + \frac{1}{6}[n(n+1)(2n+1)]h^3$$
$$= a^2(b-a) + \left(1 + \frac{1}{n}\right)a(b-a)^2 + \frac{1}{6}\left(1 + \frac{1}{n}\right)\left(2 + \frac{1}{n}\right)(b-a)^3.$$

Puesto que $\lim_{n\to\infty} \frac{1}{n} = 0$, se tiene

$$\lim_{n\to\infty} F_n = a^2(b-a) + a(b-a)^2 + \frac{1}{3}(b-a)^3 = \frac{1}{3}(b^3 - a^3).$$

[1] Algunas veces se expresa esto diciendo que se está "cuadrando" la región.

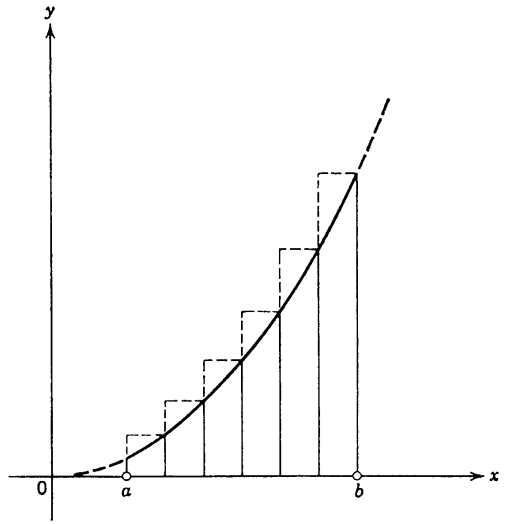


Figura 2.12 Area bajo un arco parabólico mediante subdivisión aritmética.

Así, para $a < b$,

$$\int_a^b x^2\,dx = \frac{1}{3}(b^3 - a^3).$$

*c. Integración de x^α para enteros $\alpha \neq -1$

Los siguientes ejemplos de esta sección son ilustraciones instructivas que muestran que en algunos casos la integración puede llevarse a cabo mediante recursos especiales elementales. Posteriormente, en la Sección 2.9d (p. 212) se obtendrán los mismos resultados de manera más simple utilizando métodos generales.

El mismo tipo de argumento que el utilizado para x y x^2, aplicado a las funciones x^3, x^4, . . . , da como resultado la relación

$$\int_a^b x^\alpha\,dx = \frac{1}{\alpha+1}(b^{\alpha+1} - a^{\alpha+1}), \tag{1}$$

donde α es cualquier entero positivo; esto puede ser demostrado encontrando fórmulas apropiadas para las sumas $1^\alpha + 2^\alpha + \cdots + n^\alpha$, tal como la relación

$$\lim_{n\to\infty}\left[(1^\alpha + 2^\alpha + \cdots + n^\alpha)\frac{1}{n^{\alpha+1}}\right] = \frac{1}{\alpha+1}$$

que puede ser probada por inducción sobre α (véase el Problema 16, p. 136). En la siguiente sección la fórmula (1) será probada de una forma

diferente, con mayor generalidad y simplicidad, indicándose con ello la potencia de los métodos que desarrollaremos. Su validez será extendida a todos los valores de α excepto $\alpha = -1$.

Afortunadamente, la definición de integral nos deja bastante latitud en la elección de las subdivisiones y proporciona un camino más sencillo para evaluar la integral. No se tienen que utilizar sumas basadas en puntos equidistantes de división. En vez de eso, con el "cociente" $\sqrt[n]{b/a} = q$ se

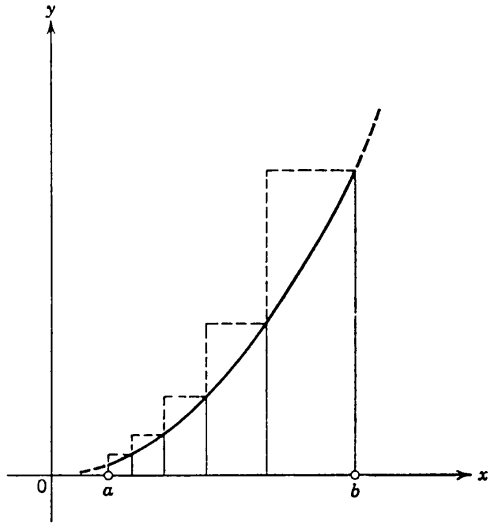


Figura 2.13 Area bajo un arco parabólico mediante subdivisión geométrica.

subdivide el intervalo $[a, b]$ mediante los puntos de una progresión geométrica (Fig. 2.13),

$$a, aq, aq^2, \ldots, aq^{n-1}, aq^n = b;$$

se requiere entonces solamente evaluar la suma de una serie geométrica. Dados los puntos de división $x_i = aq^i$ la longitud de la i-ésima celda está dada por

$$\Delta x_i = aq^i - aq^{i-1} = \frac{aq^i(q-1)}{q}.$$

El máximo Δx_i es el último:

$$\Delta x_n = \frac{b(q-1)}{q}.$$

Para $n \to \infty$ el número q tiende hacia el valor uno (véase Ejemplo *d*, p. 87), y por lo tanto la longitud Δx_n de la máxima celda, y entonces también

la longitud de todas las celdas, tiende a cero. Para los puntos intermedios ξ_i se eligen nuevamente los puntos extremos derechos x_i de cada celda. La suma

$$F_n = \sum_{i=1}^{n} (\xi_i)^\alpha \Delta x_i = \sum_{i=1}^{n} (aq^i)^\alpha aq^i \frac{q-1}{q} \tag{2}$$

$$= a^{\alpha+1} \frac{q-1}{q} \sum_{i=1}^{n} (q^{1+\alpha})^i$$

es conocida explícitamente a partir de la suma de la progresión geométrica con razón $q^{1+\alpha}$. Aplicando la bien conocida fórmula (p. 91), se encuentra

$$F_n = a^{\alpha+1} \frac{q-1}{q} q^{\alpha+1} \frac{q^{n(\alpha+1)}-1}{q^{\alpha+1}-1}$$

$$= a^{\alpha+1}(q-1)q^\alpha \frac{(b/a)^{\alpha+1}-1}{q^{\alpha+1}-1} = (b^{\alpha+1}-a^{\alpha+1})q^\alpha \frac{q-1}{q^{\alpha+1}-1}.$$

Puesto que $q \neq 1$, puede utilizarse una vez más la fórmula para la suma de una progresión geométrica y escribir

$$\frac{q-1}{q^{1+\alpha}-1} = \frac{1}{q^\alpha + q^{\alpha-1} + \cdots + 1}.$$

Para $n \to \infty$ todas las potencias de q tienden a uno y se sigue que

$$\lim_{n\to\infty} F_n = \frac{1}{1+\alpha}(b^{1+\alpha} - a^{1+\alpha}).$$

De esta forma se ha verificado la fórmula (1) para la integral de x^α para $0 < a < b$ y cualquier entero positivo α.

El mismo método se aplica también para enteros negativos α, siempre que $\alpha \neq -1$. Para la suma F_n se obtiene como antes

$$F_n = (b^{\alpha+1} - a^{\alpha+1})q^\alpha \frac{q-1}{q^{\alpha+1}-1}$$

$$= (b^{\alpha+1} - a^{\alpha+1}) \frac{q-1}{q(1-q^{-\alpha-1})},$$

donde se recuerda que $-\alpha$ es positivo y mayor que uno. Aplicando la fórmula para una progresión geométrica, se obtiene

$$\frac{1}{q}\left(\frac{q-1}{q^{-\alpha-1}-1}\right) = \frac{1}{q^{-\alpha-1} + q^{-\alpha-2} + \cdots + q}$$

que tiende a $1/(-\alpha-1)$ conforme $n \to \infty$. Consecuentemente, como antes,

$$\lim_{n\to\infty} F_n = \frac{1}{\alpha+1}(b^{\alpha+1} - a^{\alpha+1}),$$

La fórmula integral no tiene significado para $\alpha = -1$, puesto que tanto el numerador como el denominador en el miembro derecho serían entonces cero. En vez de esto se encuentra, de la expresión original (2) para F_n para el caso $\alpha = -1$, que $F_n = n(q-1)/q$. Consecuentemente, observando que $q = \sqrt[n]{b/a}$ tiende a uno conforme $n \to \infty$, se encuentra

$$\int_a^b \frac{1}{x}\,dx = \lim_{n\to\infty} n(\sqrt[n]{b/a} - 1). \tag{3}$$

Aquí el límite en el miembro derecho no puede ser expresado en términos de potencias de a y de b pero puede ser expresado en términos de logaritmos de esas cantidades como se verá posteriormente (p. 167).

*d. Integración de x^α para α racional distinto de -1

El resultado obtenido anteriormente puede ser considerablemente generalizado sin complicar esencialmente la demostración. Sea $\alpha = r/s$ un número racional positivo, siendo r y s enteros positivos; entonces en la evaluación de la integral dada arriba nada se cambia excepto la evaluación del límite $(q-1)/(q^{\alpha+1}-1)$ cuando q se aproxima a uno. Esta expresión es ahora simplemente $(q-1)/(q^{(r+s)/s}-1)$. Hagamos $q^{1/s} = \tau (\tau \neq 1)$: Entonces, si q tiende a uno τ también tiende a uno. Tenemos por ello que encontrar el valor límite de $(\tau^s - 1)/(\tau^{r+s} - 1)$ cuando τ se aproxima a uno. Si se dividen tanto el numerador como el denominador por $\tau - 1$ y se transforman como antes mediante la fórmula para progresiones geométricas, el límite se torna simplemente

$$\lim_{\tau\to 1} \frac{\tau^{s-1} + \tau^{s-2} + \cdots + 1}{\tau^{r+s-1} + \tau^{r+s-2} + \cdots + 1}.$$

Puesto que tanto el numerador como el denominador son continuos en τ, este límite se obtiene de inmediato substituyendo $\tau = 1$, y por lo tanto se hace igual a $s/(r+s) = 1/(\alpha+1)$; luego, para todo valor racional positivo de α se obtiene la fórmula integral

$$\int_a^b x^\alpha\,dx = \frac{1}{\alpha+1}(b^{\alpha+1} - a^{\alpha+1}),$$

tal como con enteros positivos.

Esta fórmula permanece válida para valores negativos racionales de $\alpha = -r/s$, siempre que se excluya el valor $\alpha = -1$ (para el cual la fórmula utilizada arriba para la suma de la progresión geométrica pierde su significado).

Para α negativo se evalúa nuevamente el límite de $(q-1)/(q^{\alpha+1}-1)$ haciendo $q^{-1/s} = \tau$ para $\alpha = -r/s$; esto se deja como un ejercicio para el lector.

Es natural conjeturar que el campo de validez de la última fórmula se extiende también a valores irracionales de α. En realidad, se establecerá la fórmula integral para todos los valores reales de α (excepto $\alpha = -1$) en la sección 2.7 (p. 176) de una manera enteramente simple como una consecuencia de la teoría general.

*e. Integración de sen x y cos x

El último ejemplo elemental que será tratado aquí mediante un artificio especial es la integral de $f(x) = \operatorname{sen} x$. La integral

$$\int_a^b \operatorname{sen} x\, dx$$

es claramente el límite de la suma

$$S_h = h[\operatorname{sen}(a + h) + \operatorname{sen}(a + 2h) + \cdots + \operatorname{sen}(a + nh)],$$

que surge de la división del intervalo de integración en celdas de longitud $h = (b - a)/n$. Multiplíquese la expresión en el miembro derecho por $2 \operatorname{sen} h/2$ y recuérdese la bien conocida fórmula trigonométrica

$$2 \operatorname{sen} u \operatorname{sen} v = \cos(u - v) - \cos(u + v).$$

Siempre que h no sea un múltiplo de 2π se obtiene la fórmula

$$S_h = \frac{h}{2\operatorname{sen}\frac{h}{2}}\left[\cos\left(a + \frac{h}{2}\right) - \cos\left(a + \frac{3}{2}h\right) + \cos\left(a + \frac{3}{2}h\right)\right.$$

$$\left. - \cos\left(a + \frac{5}{2}h\right) + \cdots + \cos\left(a + \frac{2n-1}{2}h\right) - \cos\left(a + \frac{2n+1}{2}h\right)\right]$$

$$= \frac{h}{2\operatorname{sen}\frac{h}{2}}\left[\cos\left(a + \frac{h}{2}\right) - \cos\left(a + \frac{2n+1}{2}h\right)\right].$$

Puesto que $a + nh = b$, la integral se convierte en el límite de

$$\frac{h}{2\operatorname{sen}\frac{h}{2}}\left[\cos\left(a + \frac{h}{2}\right) - \cos\left(b + \frac{h}{2}\right)\right] \text{ conforme } h \to 0.$$

Se sabe ahora, del capítulo 1 (p. 107), que para $h \to 0$ la expresión $(h/2)/(\operatorname{sen} h/2)$ se aproxima al límite uno. El límite deseado es entonces simplemente $\cos a - \cos b$, y se obtiene la integral

$$\int_a^b \operatorname{sen} x\, dx = -(\cos b - \cos a).$$

Análogamente,

$$\int_a^b \cos x\, dx = \operatorname{sen} b - \operatorname{sen} a \qquad \text{(véase problema 3, p. 217)}.$$

Cada uno de los ejemplos anteriores fue tratado con un artificio especial. Empero, el punto esencial del cálculo integral y diferencial sistemáticos es el simple hecho de que en vez de tales artificios especiales se usan consideraciones generales que conducen directamente al resultado. Llegaremos a estos métodos discutiendo primero algunas reglas generales concernientes a integrales, introduciendo luego el concepto de la derivada y finalmente estableciendo la conexión entre integral y derivada.

2.3 Reglas fundamentales de integración

Las propiedades básicas de la integral se siguen directamente de su definición como el límite de una suma:

$$\int_a^b f(x)\, dx = \lim_{n\to\infty} \sum_{i=1}^{n} f(\xi_i)\Delta x_i,$$

donde el intervalo $[a, b]$ es cortado en subintervalos o celdas de longitud Δx_i, el número ξ_i es cualquier valor en el i-ésimo subintervalo y se requiere que el máximo Δx_i tienda a cero conforme $n \to \infty$.

a. Aditividad

Sea c cualquier valor entre a y b. Si las integrales se interpretan como áreas y se recuerda que el área de una región que consiste de varias secciones es la suma de las áreas de las secciones (Fig. 2.14), esto nos conduce a la regla

$$\int_a^b f(x)\, dx = \int_a^c f(x)\, dx + \int_c^b f(x)\, dx. \tag{4}$$

Para una demostración analítica las subdivisiones se eligen de tal manera que el punto c aparezca como un punto de división, digamos $c = x_m$ (donde m varía con n). Entonces,

$$\sum_{i=1}^{n} f(\xi_i)\Delta x_i = \sum_{i=1}^{m} f(\xi_i)\Delta x_i + \sum_{i=m+1}^{n} f(\xi_i)\Delta x_i,$$

donde la primera suma en el miembro derecho corresponde a una subdivisión del intervalo $[a, c]$ en m celdas, y la segunda suma corresponde a una subdivisión del intervalo $[c, b]$. Ahora para $n \to \infty$ se obtiene nuestra regla para integrales.

Hasta ahora solamente hemos definido $\int_a^b f(x)\, dx$ cuando $a < b$. Para $a = b$, o bien para $a > b$, definimos la integral de tal manera que la regla de aditividad sea preservada. Por eso, para $c = a$ debe definirse

$$\int_a^a f(x)\, dx = 0, \tag{5}$$

y entonces para $b = a$ se sigue que

$$\int_a^c f(x)\,dx + \int_c^a f(x)\,dx = \int_a^a f(x)\,dx = 0.$$

Esto nos conduce a *definir* $\int_a^c f(x)\,dx$ para $c < a$ por la fórmula

$$\int_a^c f(x)\,dx = -\int_c^a f(x)\,dx, \tag{6}$$

donde el miembro derecho posee el significado establecido originalmente. Su *significado geométrico es que el área* bajo la curva $y = f(x)$ ha de ser

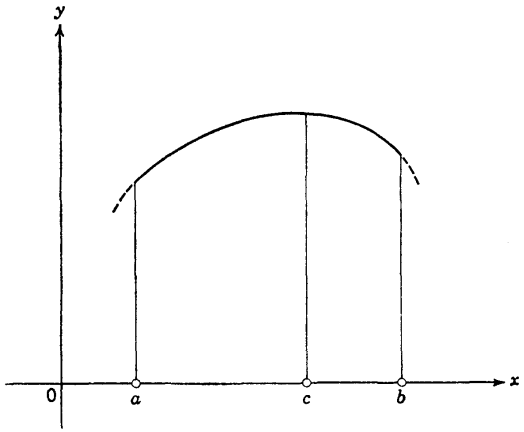


Figura 2.14.

considerada *negativa* si la dirección de movimiento desde el límite inferior de integración hacia el límite superior es el de las x decrecientes. Una ojeada a los ejemplos anteriores de integrales confirma que, en efecto, un intercambio en los límites de integración a y b da como resultado el cambio de signo en el valor de la integral.

b. Integral de una suma y de un producto con una constante

Si $f(x)$ y $g(x)$ son dos funciones cualesquiera (integrables), las leyes básicas de operar con límites implican

$$\begin{aligned}\int_a^b f(x)\,dx + \int_a^b g(x)\,dx &= \lim_{n\to\infty}\left[\sum_{i=1}^n f(\xi_i)\,\Delta x_i\right] + \lim_{n\to\infty}\left[\sum_{i=1}^n g(\xi_i)\,\Delta x_i\right] \\ &= \lim_{n\to\infty}\left[\sum_{i=1}^n f(\xi_i)\,\Delta x_i + \sum_{i=1}^n g(\xi_i)\,\Delta x_i\right] \\ &= \lim_{n\to\infty}\left\{\sum_{i=1}^n [f(\xi_i) + g(\xi_i)]\,\Delta x_i\right\};\end{aligned}$$

y, por lo tanto, la importante regla para la suma de dos funciones

$$\int_a^b f(x)\,dx + \int_a^b g(x)\,dx = \int_a^b [f(x) + g(x)]\,dx; \tag{7}$$

y análogamente para la diferencia:

$$\int_a^b f(x)\,dx - \int_a^b g(x)\,dx = \int_a^b [f(x) - g(x)]\,dx.$$

Además, con cualquier constante α es

$$\begin{aligned}\int_a^b \alpha f(x)\,dx &= \lim_{n\to\infty}\sum_{i=1}^n \alpha f(\xi_i)\,\Delta x_i \\ &= \alpha \lim_{n\to\infty}\sum_{i=1}^n f(\xi_i)\,\Delta x_i,\end{aligned}$$

y así

$$\int_a^b \alpha f(x)\,dx = \alpha\int_a^b f(x)\,dx.$$

Las últimas dos reglas permiten integrar "*combinaciones lineales*" de dos o más funciones que pueden ser integradas individualmente. Así, para cualquier función cuadrática $y = Ax^2 + Bx + C$ con cualesquiera constantes A, B, C, se tiene

$$\begin{aligned}\int_a^b (Ax^2 + Bx + C)\,dx &= \int_a^b Ax^2\,dx + \int_a^b Bx\,dx + \int_a^b C\,dx \\ &= A\int_a^b x^2\,dx + B\int_a^b x\,dx + C\int_a^b 1\,dx \\ &= \frac{A}{3}(b^3 - a^3) + \frac{B}{2}(b^2 - a^2) + C(b - a).\end{aligned}$$

De la misma manera se integra el *polinomio* general

$$y = A_0x^n + A_1x^{n-1} + \cdots + A_{n-1}x + A_n:$$

$$\begin{aligned}\int_a^b y\,dx = \frac{1}{n+1}A_0(b^{n+1} - a^{n+1}) + \frac{1}{n}A_1(b^n - a^n) + \cdots \\ + \tfrac{1}{2}A_{n-1}(b^2 - a^2) + A_n(b - a).\end{aligned}$$

c. Estimación de integrales

Otra observación obvia concerniente a integrales es básica. Considérese para $a < b$ una función $f(x)$ la cual es positiva o cero en cada punto del intervalo $[a, b]$. Entonces,

$$\int_a^b f(x)\,dx \geq 0. \tag{9}$$

Esto surge inmediatamente si se escribe la integral como el límite de una suma y se observa que las sumas contienen solamente términos no negativos.

Más generalmente, si se tienen dos funciones f y g con la propiedad $f(x) \geq g(x)$ para todo x en el intervalo $[a, b]$ entonces

$$\int_a^b f(x)\,dx \geq \int_a^b g(x)\,dx. \tag{10}$$

Pues se tiene

$$\int_a^b f(x)\,dx - \int_a^b g(x)\,dx = \int_a^b [f(x) - g(x)]\,dx \geq 0,$$

puesto que $f(x) - g(x)$ jamás es negativa.

Aplíquese este resultado a una función $f(x)$ la cual es continua en el intervalo $[a, b]$. Sea M el máximo valor y m el mínimo valor de f en ese intervalo. Puesto que

$$m \leq f(x) \leq M$$

para todo x en $[a, b]$, se tiene

$$\int_a^b m\,dx \leq \int_a^b f(x)\,dx \leq \int_a^b M\,dx.$$

Recordando que para cualquier constante C

$$\int_a^b C\,dx = C\int_a^b 1\,dx = C(b-a),$$

se obtiene la desigualdad

$$m(b-a) \leq \int_a^b f(x)\,dx \leq M(b-a), \tag{11}$$

la cual da cotas superiores o inferiores sencillas para la integral definida de cualquier función continua.

Nuevamente, esta estimación es intuitivamente obvia. Si se considera la integral interpretada como un área, las cantidades $M(b-a)$ y $m(b-a)$ representan áreas de un rectángulo circunscripto y un rectángulo inscripto sobre la base común de longitud $b-a$ (véase Fig. 2.15).

d. El teorema del valor medio para integrales

La integral como un valor medio

Es significativa una interpretación ligeramente diferente de nuestras desigualdades en términos del *promedio de la función f en un intervalo* $[a, b]$. Para un número finito de cantidades $f_1, f_2, \ldots, f_n$, el *promedio* o *media aritmética* es el número

$$\frac{f_1 + f_2 + \cdots + f_n}{n}$$

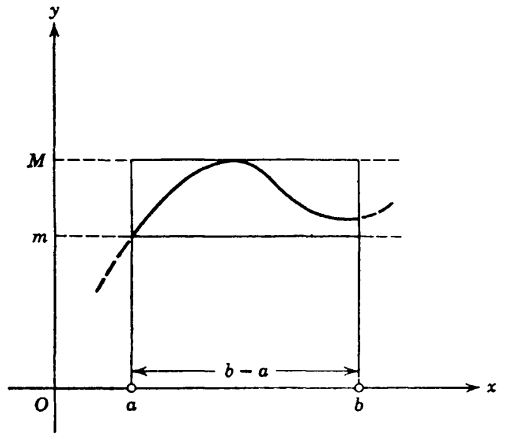


Figura 2.15.

Si se desea asignar un significado al valor promedio de un número infinito de cantidades $f(x)$ correspondientes a x arbitrarios en el intervalo $[a, b]$, es natural escoger primero un número finito n de valores de f, digamos $f(x_1), f(x_2), \ldots, f(x_n)$, para formar su promedio

$$\frac{f(x_1) + \cdots + f(x_n)}{n},$$

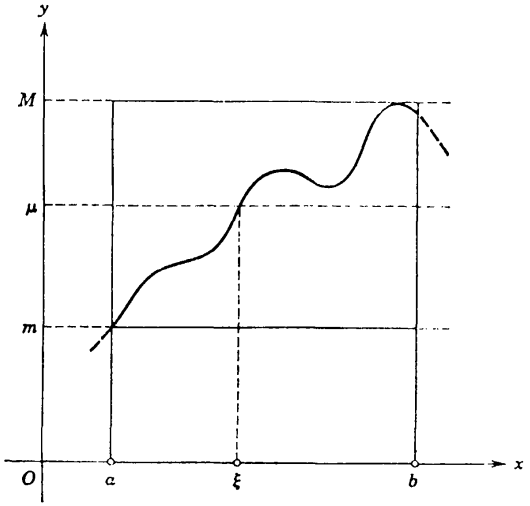


Figura 2.16. El valor medio μ de una función.

y entonces tomar el límite cuando n crece más allá de toda cota. El valor de este límite, si existe, dependerá mucho de cómo los puntos x_i se encuentren espaciados en el intervalo $[a, b]$. Un valor definido para el promedio de f se logra si se toman para los x_i los puntos que se obtienen cuando se divide el intervalo $[a, b]$ en n partes *iguales* de longitud $\Delta x_i = (b - a)/n$. Se tiene entonces

$$\frac{f(x_1) + \cdots + f(x_n)}{n} = \frac{1}{b - a} \sum_{i=1}^{n} f(x_i)\Delta x_i,$$

y es claro que en el límite, para $n \to \infty$, los n-ésimos promedios convergen hacia el valor

$$\mu = \frac{1}{b - a}\int_a^b f(x)\,dx = \frac{\displaystyle\int_a^b f(x)\,dx}{\displaystyle\int_a^b dx}.$$

A μ se le llamará la "media aritmética" o bien el *valor medio* de f en el intervalo $[a, b]$. Nuestras desigualdades entonces establecen simplemente que el valor medio de una función continua no puede ser mayor que el valor máximo ni menor que el valor mínimo de la función (Fig. 2.16).

Puesto que la función $f(x)$ es continua en el intervalo $[a, b]$, deben existir puntos en el intervalo donde f posee el valor M o el valor m. Por el teorema del valor intermedio para funciones continuas, deberá entonces existir también un punto ξ en el intervalo donde f tome, en efecto, el valor intermedio μ. Se ha probado entonces:

TEOREMA DEL VALOR MEDIO. *Para una función continua $f(x)$ en el intervalo $[a, b]$, existe un valor ξ en el intervalo tal que*

$$\int_a^b f(x)\,dx = f(\xi)\,(b - a). \tag{12}$$

Este es el sencillo pero muy importante *teorema del valor medio del cálculo integral.* En palabras, afirma que el valor medio de una función continua en un intervalo pertenece al rango de la función.

El teorema asegura solamente la *existencia* de por lo menos un ξ en el intervalo, para el cual $f(\xi)$ es igual al valor promedio de f, pero no da ninguna información adicional sobre la ubicación de ξ.

Nótese que la fórmula que expresa al teorema del valor medio permanece válida si los límites a y b son intercambiados; por lo tanto, el teorema del valor medio es correcto también cuando $a > b$.

El teorema del valor medio generalizado. En vez de la sencilla media aritmética es necesario considerar a menudo "medias pesadas" de n cantidades $f_1, \ldots, f_n$, dadas por

$$\frac{p_1f_1 + p_2f_2 + \cdots + p_nf_n}{p_1 + p_2 + \cdots + p_n} = \mu,$$

donde "los factores peso" p_i son cualesquiera cantidades positivas. Si, por ejemplo, $f_1, f_2, \ldots, f_n$ son, realmente, los pesos de partículas localizadas en los puntos respectivos $f_1, f_2, \ldots, f_n$ del eje x, entonces μ representará la posición del centro de gravedad. Si todos los pesos p_i son iguales, la cantidad μ es justamente la media *aritmética* definida anteriormente.

Para una función $f(x)$ puede formarse análogamente la media pesada

$$\mu = \frac{\displaystyle\int_a^b f(x)p(x)\,dx}{\displaystyle\int_a^b p(x)\,dx} \tag{13}$$

sobre el intervalo $[a, b]$ donde $p(x)$, la *función peso,* es cualquier función positiva en el intervalo. La hipótesis de que p es positiva garantiza que el denominador no se anula.

La media pesada μ está situada también entre el máximo valor M y el mínimo valor m de la función f en el intervalo.

En efecto, multiplicando la desigualdad

$$m \le f(x) \le M,$$

por el número *positivo* $p(x)$, se encuentra que

$$mp(x) \le f(x)p(x) \le Mp(x).$$

La integración conduce entonces a

$$m\int_a^b p(x)\,dx \le \int_a^b f(x)p(x)\,dx \le M\int_a^b p(x)\,dx.$$

Dividiendo por la cantidad positiva $\displaystyle\int_a^b p(x)\,dx$, indudablemente se obitene el resultado

$$m \le \mu \le M.$$

Si aquí $f(x)$ es continua, se concluye del teorema del valor intermedio (p. 68) que $\mu = f(\xi)$, donde ξ es un valor apropiado en el intervalo $a \le \xi \le b$. Esto conduce al siguiente *teorema del valor medio generalizado del cálculo integral:*

Si $f(x)$ y $p(x)$ son continuas en el intervalo $[a, b]$ y además $p(x)$ es positiva en este intervalo, entonces existe un valor ξ en el intervalo tal que

$$\int_a^b f(x)p(x)\,dx = f(\xi)\int_a^b p(x)\,dx. \tag{14}$$

El caso especial $p(x) = 1$ conduce al teorema del valor medio anterior.

2.4 La integral como función del límite superior (integral indefinida)

Definición y fórmula básica

El valor de la integral de una función $f(x)$ depende de los límites de integración a y b: La integral es una función de los dos límites a y b. Para estudiar más de cerca esta dependencia sobre los límites imaginemos que el límite inferior es un número fijo, digamos α, denotemos la variable de integración ya no por x sino por u (véase p. 148) y denótese el límite superior por x en vez de denotarlo por b, de manera de indicar que se considerará al límite superior como la variable y que se desea investigar el valor de la integral como función de su límite superior. En efecto, escribimos

$$\phi(x) = \int_a^x f(u)\, du.$$

La función $\phi(x)$ se denomina una *integral indefinida* de la función $f(x)$. Cuando se habla de *una* y no de *la* integral indefinida, se sugiere que en vez del límite inferior α pudo haberse elegido cualquier otro, en cuyo caso se obtendría ordinariamente un valor diferente para la integral. Geométrica-

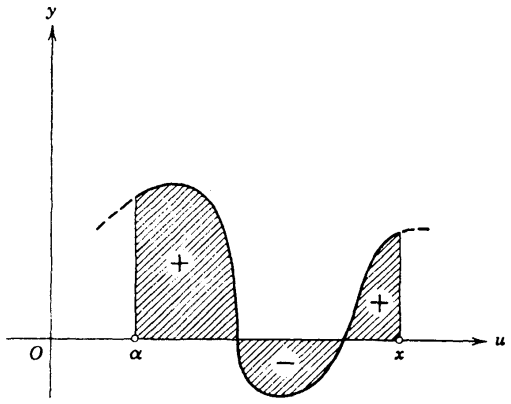


Figura 2.17 La integral indefinida como un área.

mente, la integral indefinida $\phi(x)$ está dada por el área (mostrada por el sombreado en la Fig. 2.17) bajo la curva $y = f(u)$ y acotada por el eje u, la ordenada $u = \alpha$ y la variable ordenada $u = x$, estando el signo determinado por las reglas discutidas anteriormente (p. 148).

Cualquier integral *definida* particular se encuentra a partir de la integral indefinida $\phi(x)$. Sin duda, por las reglas básicas para integrales,

$$\int_a^b f(u)\, du = \int_a^\alpha f(u)\, du + \int_\alpha^b f(u)\, du$$
$$= -\int_\alpha^a f(u)\, du + \int_\alpha^b f(u)\, du = \phi(b) - \phi(a).$$

En particular, puede expresarse cualquier otra integral indefinida con límite inferior α' en términos de $\phi(x)$:

$$\int_{\alpha'}^x f(u)\, du = \phi(x) - \phi(\alpha').$$

Como se ve, *cualquier integral indefinida difiere de la integral indefinida especial* $\phi(x)$ *sólo por una constante.*

Continuidad de la integral indefinida

Si la función $f(x)$ es continua en el intervalo $[a, b]$ y α es un punto de este intervalo, entonces la integral indefinida

$$\phi(x) = \int_\alpha^x f(u)\, du$$

representa una función de x la cual está definida nuevamente en el mismo intervalo. Como se ve fácilmente: *La integral indefinida* $\phi(x)$ *de una función continua* $f(x)$ *es asimismo continua.* En efecto, si x e y son dos valores cualesquiera en el intervalo, se tiene por el teorema del valor medio que

$$(15) \qquad \phi(y) - \phi(x) = \int_x^y f(u)\, du = f(\xi)\,(y - x)$$

donde ξ es algún valor en el intervalo con puntos extremos x e y. De la continuidad de f se tiene entonces

$$\lim_{y \to x} \phi(y) = \lim_{y \to x} [\phi(x) + f(\xi)\,(y - x)] = \phi(x) + f(x) \cdot 0 = \phi(x),$$

lo que muestra que ϕ es continua. Más específicamente, en cualquier intervalo cerrado se tiene $|\phi(y) - \phi(x)| \leq M\,|y - x|$, donde M es el máximo de $|f|$ en el intervalo, de modo que ϕ es aún Lipschitz-continua.

La fórmula (15) para $\phi(y) - \phi(x)$ muestra: que $\phi(x)$ es una función creciente de x en el caso en que f sea positiva en todo el intervalo, o sea, para $y > x$

$$\phi(y) = \phi(x) + f(\xi)\,(y - x) > \phi(x).$$

El formar la integral indefinida de una función es un camino importante para *general nuevos tipos de funciones.* En la sección 2.5 se aplicará este método para introducir la función logaritmo. Esto nos hará también vislumbrar el hecho de que teoremas generales del análisis matemático conducen a las más notables fórmulas específicas.

Como se verá en la sección 3.14a (p. 298), la definición de nuevas funciones por medio de integrales de funciones ya definidas es un procedimiento satisfactorio si se desean establecer las definiciones (por ejemplo, de las funciones trigonométricas) sobre una base puramente analítica en vez de atenerse a explicaciones geométricas intuitivas.

2.5 El logaritmo definido mediante una integral

a. Definición de la función logaritmo

En la sección 2.2 se tuvo buen éxito en expresar $\int_a^b x^\alpha\,dx$ para cualquier racional $\alpha \neq -1$ en términos de potencias de a y de b. Para $\alpha = -1$ se pudo solamente representar la integral como el límite de la sucesión

$$\int_a^b \frac{1}{u}\,du = \lim_{n\to\infty} n(\sqrt[n]{b/a} - 1).$$

Independientemente de las discusiones de la sección 2.2 se *introduce* ahora la función representada por la integral indefinida

$$\int_1^x \frac{1}{u}\,du,^1$$

o bien, geométricamente, mediante el área bajo una hipérbola como se indica en la Fig. 2.18. La denominamos el *logaritmo* de x, o bien, más precisamente, el *logaritmo natural* de x, y se escribe

$$\log x = \int_1^x \frac{1}{u}\,du. \tag{16}$$

Puesto que $y = 1/u$ es una función continua y positiva para todo $u > 0$, la función $\log x$ está definida para todo $x > 0$, siendo además continua y también monótona creciente. La elección de 1 como el límite inferior en la integral indefinida para $\log x$ es cuestión de conveniencia. Implica que

$$\log 1 = 0, \tag{17}$$

[1] En esta sección otra vez se utilizará libremente el hecho de que la integral de una función continua (en este caso la función $1/u$) existe; la demostración general está dada en el Suplemento.

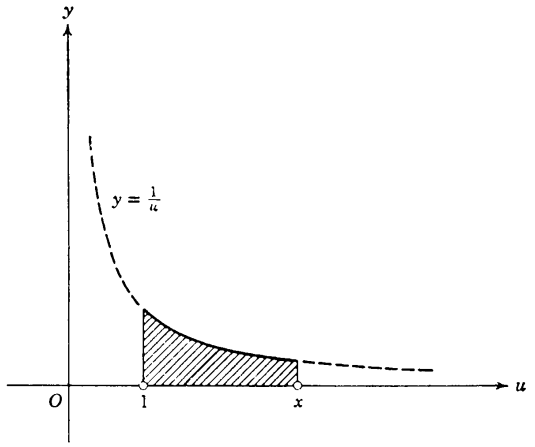


Figura 2.18 Log x representado por un área.

y que $\log x$ es positivo para $x > 1$ y negativo para x entre cero y 1 (Fig. 2.19). Cualquier integral definida de $1/u$ entre límites positivos a y b puede ser expresada en términos de logaritmos mediante la fórmula (véase pp. 165-166)

$$\int_a^b \frac{1}{u}\,du = \log b - \log a. \tag{18}$$

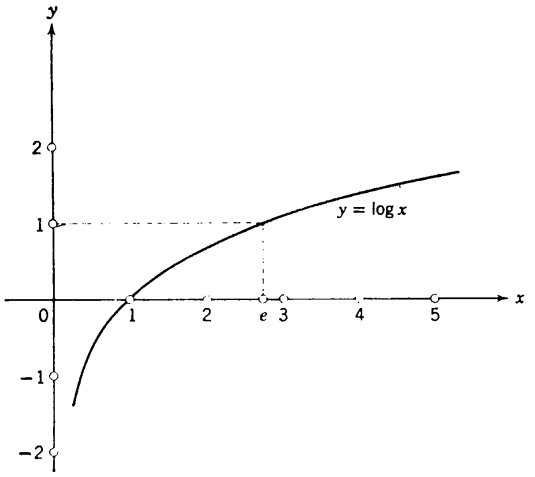


Figura 2.19 El logaritmo natural.

Geométricamente, esta integral representa el área bajo la hipérbola $y = 1/x$ entre las ordenadas $x = a$ y $x = b$.

b. El teorema de adición para logaritmos

La propiedad fundamental que justifica el nombre tradicional para $\log x$ está expresada por el

TEOREMA DE ADICIÓN. *Para todo x e y positivos*

$$\log(xy) = \log x + \log y. \tag{19}$$

DEMOSTRACIÓN. El teorema de adición se escribe en la forma

$$\log(xy) - \log y = \log x,$$

o bien

$$\int_y^{xy} \frac{1}{v}\,dv = \int_1^x \frac{1}{u}\,du,$$

donde deliberadamente se han elegido diferentes letras para las variables de integración en las dos integrales. La igualdad de las dos integrales se deducirá del hecho de que las sumas de aproximación poseen el mismo valor para elecciones adecuadas de subdivisiones y de puntos intermedios. Supóngase primero que $x > 1$. Entonces

$$\int_1^x \frac{1}{u}\,du = \lim_{n\to\infty} \sum_{i=1}^{n} \frac{1}{\xi_i}\,\Delta u_i,$$

donde $u_0 = 1, u_1, u_2, \ldots, u_n = x$ representan los puntos que aparecen en una subdivisión del intervalo $[1, x]$ y ξ_i está situado en la i-ésima celda. Haciendo $v_i = yu_i$, $\eta_i = y\xi_i$, se ve que los puntos $v_0, v_1, \ldots, v_n$ corresponden a una subdivisión del intervalo $[y, xy]$ con puntos intermedios $\eta_i = \xi_i y$. Obviamente,

$$\Delta v_i = y\,\Delta u_i,$$

de modo que

$$\sum_{i=1}^{n} \frac{1}{\eta_i}\,\Delta v_i = \sum_{i=1}^{n} \frac{1}{\xi_i}\,\Delta u_i.$$

Para n tendiendo a infinito se obtiene la identidad deseada entre integrales para el caso $x > 1$.

Para $x = 1$ el teorema de adición se cumple trivialmente, puesto que $\log 1 = 0$. Para probar el teorema también para el caso $0 < x < 1$, se observa que entonces $1/x > 1$, y por tanto

$$\begin{aligned}
\log x + \log y &= \log x + \log\left(\frac{1}{x}\,xy\right) \\
&= \log x + \log\frac{1}{x} + \log(xy) \\
&= \log\frac{1}{x} + \log x + \log(xy) \\
&= \log\left(\frac{1}{x}\,x\right) + \log(xy) \\
&= \log 1 + \log(xy) = \log(xy).
\end{aligned}$$

Esto completa la demostración del teorema de adición.

Una demostración del teorema de adición puede también basarse en la fórmula (3) (p. 156), de acuerdo con la cual

$$\log x = \lim_{n\to\infty} n(\sqrt[n]{x} - 1).$$

Entonces

$$\begin{aligned}
\log(xy) &= \lim_{n\to\infty} n(\sqrt[n]{xy} - 1) \\
&= \lim_{n\to\infty} [n(\sqrt[n]{x} - 1)\sqrt[n]{y} + n(\sqrt[n]{y} - 1)] \\
&= [\lim_{n\to\infty} n(\sqrt[n]{x} - 1)](\lim_{n\to\infty} \sqrt[n]{y}) + \lim_{n\to\infty} n(\sqrt[n]{y} - 1) \\
&= \log x + \log y,
\end{aligned}$$

puesto que $\lim_{n\to\infty} \sqrt[n]{y} = 1$ (véase p. 87).

Aplicando el teorema de adición para el caso especial $y = 1/x$, éste conduce a

$$\log 1 = \log x + \log\frac{1}{x},$$

o bien

$$\log\frac{1}{x} = -\log x. \tag{20}$$

Entonces, más generalmente,

$$\log\frac{y}{x} = \log y + \log\frac{1}{x} = \log y - \log x. \tag{21}$$

Aplicando repetidamente el teorema de adición a un producto de n factores se tiene

$$\log(x_1 x_2 \cdots x_n) = \log x_1 + \log x_2 + \cdots + \log x_n.$$

En particular, se encuentra que para cualquier entero positivo n

$$\log(x^n) = n \log x. \tag{22}$$

Esta identidad se cumple también para $n = 0$, puesto que $x^0 = 1$, y puede ser extendida a enteros negativos n observando que

$$\log (x^n) = \log \left(\frac{1}{x^{-n}} \right) = -\log (x^{-n}) = -(-n) \log x = n \log x.$$

Para cualquier racional $\alpha = m/n$ y cualquier a positivo puede formarse $a^\alpha = a^{m/n} = x$. Se tiene entonces

$$\log x = \frac{1}{n} \log x^n = \frac{1}{n} \log a^m = \frac{m}{n} \log a = \alpha \log a.$$

Así, la identidad

$$\log (a^\alpha) = \alpha \log a \tag{23}$$

se cumple para cualquier real positivo a y cualquier racional α.

2.6 Función exponencial y potencias

a. El logaritmo del número e

La constante e obtenida en la p. 102 como el límite de $(1 + 1/n)^n$ desempeña un marcado papel para la función $\log x$. Sin duda, el número e está caracterizado por la ecuación [1]

$$\log e = 1.$$

Para la demostración, se observa que la continuidad de la función $\log x$ implica

$$\log e = \log \left[\lim_{n \to \infty} \left(1 + \frac{1}{n} \right)^n \right] = \lim_{n \to \infty} \log \left[\left(1 + \frac{1}{n} \right)^n \right]$$

$$= \lim_{n \to \infty} n \log \left(1 + \frac{1}{n} \right).$$

Ahora, por el teorema del valor medio del cálculo integral,

$$\log \left(1 + \frac{1}{n} \right) = \int_1^{1+1/n} \frac{1}{u} \, du = \frac{1}{\xi} \frac{1}{n},$$

donde ξ es algún número entre 1 y $1 + 1/n$ que depende de la elección de n. Obviamente, $\lim_{n \to \infty} \xi = 1$, de modo que

$$\log e = \lim_{n \to \infty} \frac{1}{\xi} = 1. \tag{24}$$

[1] Esto significa geométricamente que el área acotada por la hipérbola $y = 1/x$ y las rectas $y = 0$, $x = 1$ y $x = e$, tiene el valor uno (véase Fig. 2.18).

b. La función inversa del logaritmo. La función exponencial

De la relación $\log e = 1$ se sigue que para todo racional α es

$$\log (e^\alpha) = \alpha \log e = \alpha.$$

Esto muestra que todo número racional α aparece como un valor del $\log x$ para algún x positivo. Puesto que $\log x$ es continuo, toma entonces cualquier valor entre dos números racionales, esto es, todos los valores reales contenidos entre tales números. Se sigue que cuando x varía sobre todos los valores positivos, los valores de $y = \log x$ recorren todos los números y.

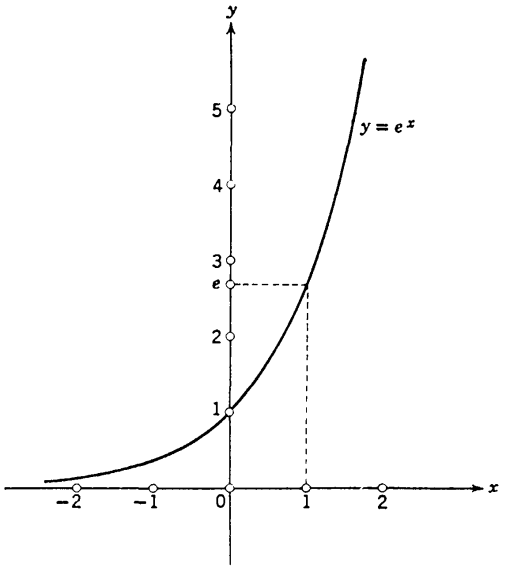


Figura 2.20 La función exponencial.

Puesto que $\log x$ es monótono creciente, existe para todo número real y exactamente un real x positivo tal que $\log x = y$. La solución x de la ecuación $y = \log x$ está dada por la función inversa del logaritmo, la cual se denotará por $x = E(y)$. Se sabe entonces que $E(y)$ (Fig. 2.20) está definida y es positiva para todo y, y que es también continua y creciente (véase pp. 65-66).

Puesto que las ecuaciones $y = \log x$ y $x = E(y)$ se cumplen para la misma relación entre x e y, la ecuación $\alpha = \log (e^\alpha)$, la cual es válida para α racional, puede escribirse también en la forma

$$E(\alpha) = e^\alpha.$$

Se ve que: para todo racional α el valor de $E(\alpha)$ es la α-ésima potencia del número e. Para $\alpha = m/n$ racional la potencia e^α está definida directamente como $\sqrt[n]{e^m}$. Para α irracional la expresión e^α está definida más naturalmente representando a α como el límite de una sucesión de números racionales α_n y haciendo $e^\alpha = \lim (e^{\alpha_n})$. Puesto que $e^{\alpha_n} = E(\alpha_n)$, y puesto que la función $E(y)$ depende continuamente de y, se puede estar seguro de que el límite de e^{α_n} existe y que posee el valor $E(\alpha)$, independientemente de la sucesión especial utilizada para aproximar a α. Esto prueba que la ecuación $E(\alpha) = e^\alpha$ se cumple también para α irracional. Para todo α real puede escribirse ahora e^α en vez de $E(\alpha)$. Denominamos a e^x *función exponencial.* Esta función está definida y es continua para todo x, y es creciente y positiva en todas partes.

Puesto que las ecuaciones $y = \log x$ y $x = e^y$ son dos formas de expresar la misma relación entre los números x e y, se ve que la designación $\log x$, el "logaritmo natural" de x (como se define aquí mediante una integral), se mantiene para el *logaritmo de base e,* con la interpretación que a ese término se le daría en matemáticas elementales; esto es, $\log x$ es el exponente de aquella potencia de e que es igual a x, o sea,

$$e^{\log x} = x. \tag{25}$$

Puede escribirse [1] $\log x = \log_e x$.

Análogamente, $x = e^y$ es aquel número cuyo logaritmo es y, o sea,

$$\log e^y = y. \tag{26}$$

Desde el punto de vista del cálculo, es realmente más fácil introducir en primer término logaritmos naturales como integrales de la función sencilla $y = 1/x$, como se hizo aquí, y definir después potencias de e tomando la inversa de la función logaritmo. De esta forma la continuidad y monotonía de las funciones $\log x$ y e^x aparecen justamente como consecuencia de teoremas generales y no requieren argumentos especiales.

c. La función exponencial como límite de potencias

Originalmente se obtuvo el número e como el límite

$$e = \lim_{n\to\infty} \left(1 + \frac{1}{n}\right)^n.$$

Una fórmula más general representa e^x para cualquier x como un límite, a saber:

[1] El lector puede pensar que el nombre "logaritmo natural" debió haber sido reservado más bien para logaritmos de base 10. Sin embargo, históricamente, la primera tabla de logaritmos, publicada por Napier en 1614, dio esencialmente logaritmos de base e. Los logaritmos de base 10 fueron introducidos sólo subsecuentemente por Briggs debido a sus obvias ventajas de cálculo.

$$e^x = \lim_{n\to\infty} \left(1 + \frac{x}{n}\right)^n. \tag{27}$$

Para la demostración, es suficiente mostrar que la sucesión

$$s_n = \log \left(1 + \frac{x}{n}\right)^n$$

posee el límite x. Pues entonces la sucesión de valores

$$e^{s_n} = \left(1 + \frac{x}{n}\right)^n$$

debe tender a e^x puesto que la función exponencial es continua. Ahora bien,

$$s_n = n \log \left(1 + \frac{x}{n}\right) = n \int_1^{1+x/n} \frac{1}{\xi}\, d\xi.$$

Por el teorema del valor medio del cálculo integral se tiene

$$s_n = n \frac{1}{\xi_n} \left[\left(1 + \frac{x}{n}\right) - 1\right] = \frac{x}{\xi_n},$$

donde ξ_n es algún valor entre uno y $1 + x/n$. Puesto que obviamente ξ_n tiende a uno para n tendiendo a ∞, se tiene en efecto $\lim_{n\to\infty} s_n = x$.

d. Definición de potencias arbitrarias de números positivos

Las potencias arbitrarias de cualesquiera números positivos pueden ahora ser expresadas en términos de las funciones exponencial y logaritmo.[1]

Se encuentra para α racional y cualquier x positivo que la relación

$$\log (x^\alpha) = \alpha \log x$$

se cumple. Se escribe esta ecuación en la forma

$$x^\alpha = e^{\alpha \log x}.$$

Para α irracional se representa a α nuevamente como el límite de una sucesión de números racionales α_n, y se define

$$x^\alpha = \lim_{n\to\infty} x^{\alpha_n} = \lim_{n\to\infty} e^{\alpha_n \log x}.$$

La continuidad de la función exponencial implica nuevamente que el límite existe y que posee el valor $e^{\alpha \log x}$, puesto que

$$e^{\alpha \log x} = e^{\lim (\alpha_n \log x)} = \lim e^{\alpha_n \log x}.$$

[1] Esto evita la definición "elemental", más incómoda, y la justificación de estos procesos mediante el paso al límite a partir de exponentes racionales, indicado en la p. 109.

Por lo tanto, la ecuación

$$x^\alpha = e^{\alpha \log x} \tag{28}$$

se cumple de manera completamente general para todo α y cualquier x positivo. Haciendo $\log x = \beta$, o bien, lo que es lo mismo, $x = e^\beta$, se infiere que

$$(e^\beta)^\alpha = e^{\alpha\beta}, \tag{29}$$

y, más generalmente, para todo x positivo:

$$(x^\alpha)^\beta = (e^{\alpha \log x})^\beta = e^{\alpha\beta \log x} = x^{\alpha\beta}.$$

Otra regla para operar con potencias, que es fácilmente establecida en forma completamente general, es la *ley de multiplicación,* o sea,

$$x^\alpha x^\beta = x^{\alpha+\beta},$$

donde x es un número positivo y α y β son arbitrarios. Es suficiente probar la correspondiente fórmula obtenida tomando los logaritmos de ambos lados:

$$\log (x^\alpha x^\beta) = \log (x^{\alpha+\beta}).$$

Ahora bien, por las reglas (19), (26) y (28) ya establecidas, se sigue que

$$\begin{aligned} \log (x^\alpha x^\beta) &= \log x^\alpha + \log x^\beta = \log (e^{\alpha \log x}) + \log (e^{\beta \log x}) \\ &= \alpha \log x + \beta \log x = (\alpha + \beta) \log x \\ &= \log (e^{(\alpha+\beta) \log x}) = \log (x^{\alpha+\beta}). \end{aligned}$$

e. Logaritmos de base arbitraria

Es fácil expresar logaritmos relativos a una base diferente de e en términos de logaritmos naturales. Si, para un número positivo, a, la ecuación $x = a^y$ se satisface, se escribe entonces

$$y = \log_a x.$$

Ahora bien, $a^y = e^{y \log a}$, de modo que $x = e^{y \log a}$, o sea que $y \log a = \log x$. Se sigue que

$$\log_a x = \frac{\log x}{\log a}, \tag{30}$$

donde $\log x$ es el logaritmo natural de base e. En particular, los logaritmos comunes de base 10 están dados por

$$\log_{10} x = \frac{\log x}{\log 10}.$$

Puesto que los logaritmos de cualquier base a son proporcionales a los logaritmos naturales, satisfacen el mismo teorema de adición:

$$\log_a x + \log_a y = \log_a (xy).$$

2.7 La integral de una potencia arbitraria de x

En la sección 2.2 se obtuvo la fórmula

$$\int_a^b u^\alpha \, du = \frac{b^{\alpha+1} - a^{\alpha+1}}{\alpha + 1},$$

para todo racional $\alpha \neq -1$. (El caso $\alpha = -1$ se vio que conduce al logaritmo.) Para evaluar la integral cuando α es un número irracional es suficiente discutir la integral indefinida

$$\phi(x) = \int_1^x u^\alpha \, du.$$

de la cual todas las integrales definidas con límites positivos a y b pueden ser obtenidas. Supóngase $x > 1$ (el caso $x < 1$ puede ser tratado de la misma manera después de intercambiar los límites). Se tiene entonces, por (28), que

$$u^\alpha = e^{\alpha \log u},$$

donde $\log u \geq 0$ para u en el intervalo de integración. Sean β y γ dos números racionales cualesquiera diferentes de -1 para los cuales

$$\beta \leq \alpha \leq \gamma.$$

Entonces también

$$\beta \log u \leq \alpha \log u \leq \gamma \log u.$$

Puesto que la función exponencial es creciente, esto implica

$$e^{\beta \log u} \leq e^{\alpha \log u} \leq e^{\gamma \log u};$$

esto es,

$$u^\beta \leq u^\alpha \leq u^\gamma.$$

Se tiene entonces

$$\int_1^x u^\beta \, du \leq \phi(u) \leq \int_1^x u^\gamma \, du.$$

Las integrales de u^β y u^γ se evaluaron anteriormente, conduciendo a

$$\frac{1}{\beta + 1} (x^{\beta+1} - 1) \leq \phi(x) \leq \frac{1}{\gamma + 1} (x^{\gamma+1} - 1).$$

Si se hacen ahora converger los números racionales β y γ a α, se obtiene en el límite

$$\phi(x) = \frac{1}{\alpha + 1}(x^{\alpha+1} - 1),$$

puesto que $x^{\beta+1} = e^{(\beta+1)\log x}$ y $x^{\gamma+1} = e^{(\gamma+1)\log x}$ tienden a $e^{(\alpha+1)\log x} = x^{\alpha+1}$ debido a la continuidad de la función exponencial. El mismo resultado se sigue para x entre cero y uno. Así, de manera general, para a, b positivos es

$$\int_a^b u^\alpha \, du = \phi(b) - \phi(a) = \frac{1}{\alpha + 1}(b^{\alpha+1} - a^{\alpha+1}),$$

tal como para α racional.

Cuando α es un entero positivo la fórmula permanece válida aun cuando los límites a o b se tornen cero o bien negativos, y es fácil extender la fórmula directamente a esos casos.

2.8 La derivada

El concepto de derivada, análogamente al de integral, posee un inmediato origen intuitivo y es fácil de entender. Además, abre la puerta hacia un caudal enorme de hechos matemáticos e ideas. Y el estudiante sólo gradualmente se hará conocedor de la variedad de aplicaciones significativas y de la potencia de las técnicas que se desarrollarán en este libro.

El concepto de derivada es sugerido primero mediante la noción intuitiva de la *tangente a una curva suave* $y = f(x)$ en un punto P de coordenadas x e y. Esta tangente es caracterizada por el ángulo α entre su dirección y el eje x positivo. Pero, ¿cómo obtiene uno este ángulo a partir de la descripción analítica de la función $f(x)$? El conocimiento de los valores de x e y en el punto P no basta para determinar el ángulo α, puesto que hay un número infinito de rectas, aparte de la tangente que pasa por P. Por otra parte, para determinar α no se requiere conocer la función $f(x)$ en su comportamiento global; el conocimiento de la función en una vecindad arbitraria del punto P debe ser suficiente para determinar la dirección α, no importa cuán minúscula sea elegida esa vecindad. Esto indica que se debería definir la dirección de la tangente a una curva $y = f(x)$ mediante un proceso de límite, como se hará en breve.

El problema de calcular la dirección de tangentes, o de "derivación", fue sugerido a los matemáticos alrededor del siglo XVI por problemas de optimización, esto es, cuestiones de máximos y mínimos que aparecen en la geometría, mecánica y óptica. (Véase la discusión en la sección 3.6.)

Otro problema de capital importancia que conduce a la derivación es el de dar un significado matemático preciso a la noción intuitiva de *velocidad* en un movimiento arbitrario no uniforme (véase p. 183).

Se principiará con el problema de describir analíticamente la tangente a una curva mediante un proceso de límite.

a. La derivada y la tangente

Definición geométrica. De acuerdo con la intuición ingenua, se define la tangente a la curva dada, $y = f(x)$, en uno de sus puntos, P, por medio de los siguientes procesos de límite geométricos (Fig. 2.21). Considérese un segundo punto P_1 sobre la curva cercano a P. Por los dos puntos P y P_1 se traza una línea recta secante de la curva. Si ahora el punto P_1 se mueve a lo largo de la curva hacia el punto P, entonces es de esperarse que la secante se aproxime a una posición límite la cual es independiente del lado del cual P_1 tiende a P. Esta *posición límite de la secante es la tangente;* la aseveración de que existe tal posición límite de la secante es equivalente a la suposición de que la curva posee una tangente definida o bien una dirección definida en el punto P. Se ha utilizado la palabra "suposición" debido a que en realidad se ha hecho una. La hipótesis de que la tangente existe en cada punto no es de ninguna manera cierta para todas las curvas que representan funciones sencillas. Por ejemplo, cualquier curva con una esquina o vértice en un punto P no posee ahí una dirección determinada de manera única; por ejemplo, la curva definida por $y = |x|$, en $(0, 0)$. (Véase la discusión en las pp. 187-188.)

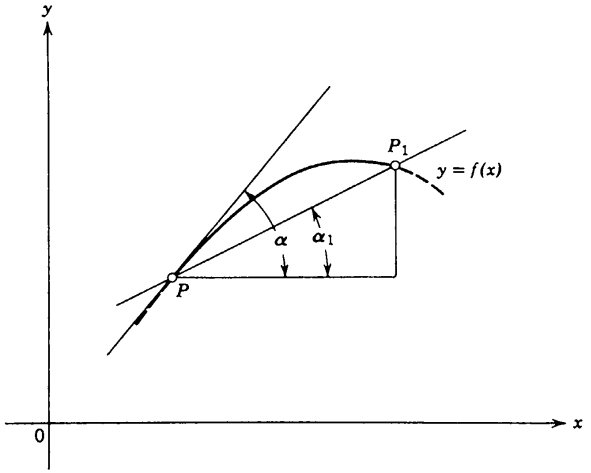


Figura 2.21 Secante y tangente.

Puesto que nuestra curva está representada por medio de una función $y = f(x)$, debe formularse el proceso de límite geométrico analíticamente con referencia a $f(x)$. Este proceso de límite analítico se denomina *derivación* de $f(x)$.

Considérese al ángulo que una línea recta forma con el eje x como aquél según el cual el eje x positivo debe ser girado en la dirección positiva o en el sentido contrario al de las manecillas del reloj [1] de modo que quede por primera vez paralelo a la recta. (Este sería un ángulo α en el intervalo $0 \leq \alpha \leq \pi$.) Sea α_1 el ángulo que la secante PP_1 forma con el eje x positivo (ver Fig. 2.22) y α el ángulo que la tangente forma con el eje x positivo. Entonces

$$\lim_{P_1 \to P} \alpha_1 = \alpha,$$

donde el significado de los símbolos es obvio. Sean x, y y x_1, y_1 las coordenadas de los puntos P y P_1 respectivamente. Entonces se tiene de inmediato.[2]

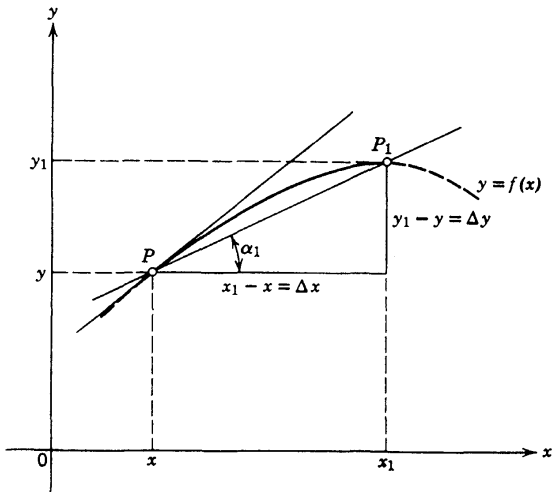


Figura 2.22.

$$\tan \alpha_1 = \frac{y_1 - y}{x_1 - x} = \frac{f(x_1) - f(x)}{x_1 - x};$$

así, nuestro proceso de límite (omitiendo el caso $\alpha = \pi/2$ de una tangente perpendicular) está representado por la ecuación

$$\lim_{x_1 \to x} \frac{f(x_1) - f(x)}{x_1 - x} = \lim_{x_1 \to x} \tan \alpha_1 = \tan \alpha.$$

[1] Esto es, en una dirección tal que una rotación de $\pi/2$ lo hace coincidir con el eje positivo y.

[2] Para que esta ecuación pueda tener significado debe suponerse que tanto x como x_1 están situadas en el dominio de f. En lo que sigue, tácitamente se harán las hipótesis correspondientes en los pasos que dan lugar a procesos de límite.

Notación. La expresión

$$\frac{f(x_1) - f(x)}{x_1 - x} = \frac{y_1 - y}{x_1 - x} = \frac{\Delta y}{\Delta x}$$

se denomina *cociente incremental* de la función $y = f(x)$, y los símbolos Δy y Δx denotan las diferencias de la función $y = f(x)$ y de la variable independiente x. (Aquí, como en la p. 146, el símbolo Δ es una abreviación para diferencia y *no* es un factor.) La tangente trigonométrica de α, la "pendiente" de la curva,[1] es por eso igual al límite al cual el cociente incremental de nuestra función tiende cuando x_1 tiende a x.

A este límite del cociente incremental se le denomina *derivada* [2] de la función $y = f(x)$ en el punto x. Generalmente se utilizará la notación de Lagrange, $y' = f'(x)$, para denotar la derivada, o bien, como lo hiciera Leibnitz, el símbolo [3] dy/dx, o $df(x)/dx$, o $(d/dx)f(x)$. En la p. 192 se discutirá el significado de la notación de Leibnitz con más detalle. Aquí se puntualiza que: La notación $f'(x)$ indica el hecho de que la *derivada es, ella misma, una función de x,* puesto que un valor de $f'(x)$ corresponde a cada valor de x en el intervalo considerado. Este hecho es destacado en ocasiones mediante el uso de los términos *función derivada, curva derivada.* La definición de la derivada aparece en varias formas diferentes:

$$f'(x) = \lim_{x_1 \to x} \frac{f(x_1) - f(x)}{x_1 - x} = \lim_{h \to 0} \frac{f(x + h) - f(x)}{h},$$

donde en la segunda expresión x_1 se reemplaza por $x + h$; o bien en la notación de Leibnitz,

$$\frac{dy}{dx} = \frac{df(x)}{dx} = f'(x) = \lim_{x_1 \to x} \frac{f(x_1) - f(x)}{x_1 - x} = \lim_{\Delta x \to 0} \frac{\Delta y}{\Delta x}.$$

Si f está definida en una vecindad del punto x, entonces el cociente $[f(x + h) - f(x)]/h$ está definido como una función de h para todos los valores $h \neq 0$ para los cuales $|h|$ es lo suficientemente pequeño para asegurar que $x + h$ está en el intervalo bajo consideración. La definición de $f'(x)$ como límite requiere que $\left| \dfrac{f(x + h) - f(x)}{h} - f'(x) \right|$ sea arbitrariamente pequeño para todo $h \neq 0$ (positivo *o* negativo) tal que $|h|$ sea suficientemente pequeño.

Cálculo analítico de derivadas. El concepto intuitivo y la noción analítica general de derivada son simples y naturales. Menos obvio es el procedimiento de llevar a cabo verdaderamente tales procesos de límite.

[1] Las palabras *gradiente* o bien *coeficiente de dirección* son utilizadas ocasionalmente.

[2] El término *coeficiente diferencial* es también utilizado en textos antiguos.

[3] La notación de Cauchy $Df(x)$ y la notación de Newton $\dot{y}$ son tambien utilizadas.

Es imposible encontrar la derivada haciendo simplemente $x_1 = x$ en la expresión para el cociente incremental, pues entonces tanto el numerador como el denominador serían iguales a cero y esto nos conduciría a la expresión sin significado 0/0. Así, el paso al límite en cada caso depende de ciertos pasos preliminares (la transformación del cociente incremental).

Por ejemplo, para la función $f(x) = x^2$ se tiene

$$\frac{f(x_1) - f(x)}{x_1 - x} = \frac{x_1^2 - x^2}{x_1 - x} = x_1 + x \quad \text{siempre que } x \neq x_1.$$

Esta función $x_1 + x$ no posee exactamente el mismo dominio que $(x_1^2 - x^2)/(x_1 - x)$: La función $x_1 + x$ está definida en el punto $x_1 = x$, donde el cociente $(x_1^2 - x^2)/(x_1 - x)$ no está definido. Para cualesquiera otros valores de x_1 las dos funciones son iguales; por lo tanto, al pasar al límite, para el cual se requiere específicamente que $x_1 \neq x$, se obtiene el mismo valor para $\lim_{x_1 \to x} (x_1^2 - x^2)/(x_1 - x)$ que para $\lim_{x_1 \to x} (x_1 + x)$. Sin embargo, puesto que la función $x_1 + x$ está definida y es continua en el punto $x_1 = x$, puede hacerse con ésta lo que no se pudo hacer con el cociente, a saber, pasar al límite haciendo simplemente $x_1 = x$. Para la derivada se obtiene entonces

$$f'(x) = \frac{d(x^2)}{dx} = 2x.$$

Como otro ejemplo, derivamos, esto es, calculamos la derivada de la función $y = \sqrt{x}$ para $x > 0$. Para $x_1 \neq x$ se tiene

$$\frac{f(x_1) - f(x)}{x_1 - x} = \frac{\sqrt{x_1} - \sqrt{x}}{x_1 - x} = \frac{(\sqrt{x_1} - \sqrt{x})(\sqrt{x_1} + \sqrt{x})}{(x_1 - x)(\sqrt{x_1} + \sqrt{x})}$$

$$= \frac{x_1 - x}{(x_1 - x)(\sqrt{x_1} + \sqrt{x})} = \frac{1}{\sqrt{x_1} + \sqrt{x}}.$$

Por lo tanto (para $x > 0$)

$$\frac{d\sqrt{x}}{dx} = \lim_{x_1 \to x} \frac{1}{\sqrt{x_1} + \sqrt{x}} = \frac{1}{2\sqrt{x}}.$$

Para $x = 0$ se tiene una singularidad: La *derivada es infinita,* puesto que $(\sqrt{x_1} - 0)/(x_1 - 0) = 1/\sqrt{x_1} \to \infty$ para $x_1 \to 0$.

Definición analítica

Es sumamente significativo que el proceso de derivar una función posea un significado analítico definido completamente separado de la concepción intuitiva geométrica de la tangente. La definición analítica de la

integral, libre de la visualización geométrica de área, nos permitió basar la noción de área sobre la de integral. Con igual espíritu, independientemente de la representación geométrica de una función $y = f(x)$ por medio de una curva, *definimos* la derivada de la función $y = f(x)$ como la nueva función $y' = f'(x)$ dada mediante el límite del cociente incremental $\Delta y/\Delta x$, siempre que el límite exista.

Aquí las diferencias $\Delta y = y_1 - y = f(x_1) - f(x)$ y $\Delta x = x_1 - x$ son "cambios o variaciones correspondientes" en las variables y y x. La razón $\Delta y/\Delta x$ puede denominarse "razón promedio de variación" de y con respecto a x en el intervalo $(x, x + \Delta x)$. El límite $f'(x) = dy/dx$ representa entonces la "razón instantánea de variación" o simplemente la "razón de variación" de y con respecto a x.

Si este límite existe se dice que la función $f(x)$ es *derivable* o *diferenciable.* Se supondrá siempre que cada función que se considere es derivable a menos que se especifique lo contrario.[1] Se destaca que si la función $f(x)$ ha de ser derivable en el punto x, el límite para $h \to 0$ del cociente $[f(x + h) - f(x)]/h$ debe existir, donde h puede poseer cualquier valor $\neq 0$ para el cual $x + h$ pertenece al dominio de f. Si, en particular, f está definida en todo un intervalo que contiene al punto x en su interior, entonces el límite debe existir *independientemente de la manera en la cual h tiende a* cero, ya sea a través de valores positivos o a través de valores negativos, sin restricción sobre el signo.

Disponiendo ahora de una definición analítica para la derivada $f'(x)$, se toma el ángulo de dirección α con el eje x positivo, dado por la ecuación $\tan \alpha = f'(x)$, como la dirección de la tangente a la curva en el punto (x, y).[2] De este modo, basando la definición geométrica en la analítica, se evitan dificultades que pueden aparecer por la vaguedad de la visualización geométrica. De hecho, se ha *definido* ahora precisamente lo que se entiende por una tangente a la gráfica de $y = f(x)$ en el punto (x, y) y se dispone de un criterio analítico para decidir si una curva posee o no una tangente en un punto dado (x, y).

Funciones monótonas

No obstante, la interpretación objetiva de la derivada como la pendiente de la tangente a la curva es una ayuda altamente provechosa para el entendimiento, aun en discusiones puramente analíticas. Este es el caso de la siguiente proposición basada en la intuición geométrica:

[1] Ejemplos en los cuales esta hipótesis no se satisface se darán posteriormente (ver p. 188). Tales ejemplos justifican el mencionar la derivabilidad como una hipótesis si el contexto lo asegura.

[2] El ángulo α no está determinado absolutamente de manera única sino que puede ser substituido por $\alpha \pm \pi$, $\alpha \pm 2\pi$, etc., a menos que se especifique como arriba que $0 \leq \alpha < \pi$.

La función $f(x)$ es monótona creciente cuando $f'(x) > 0$ y monótona decreciente cuando $f'(x) < 0$.

En verdad, si $f'(x)$ es positiva y la curva es recorrida en la dirección de x creciente, entonces la tangente se inclina hacia arriba, esto es, hacia y creciente (α es un "ángulo agudo"); por lo tanto, en el punto en cuestión la curva se eleva conforme x crece; y si, por otra parte, $f'(x)$ es negativa, la tangente se inclina hacia abajo (α es un "ángulo obtuso") y la curva cae conforme x crece (véase Fig. 2.23). Esto será probado analíticamente en las pp. 198, 199.

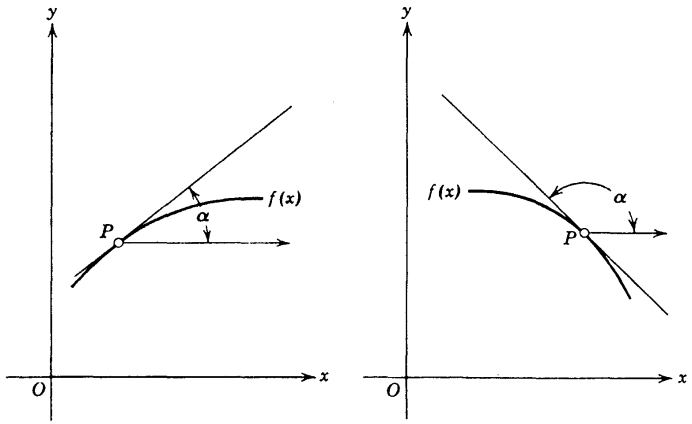


Figura 2.23 Tangentes a gráficas de funciones crecientes y decrecientes.

b. La derivada como una velocidad

La necesidad de substituir el concepto intuitivo de velocidad o rapidez por una definición precisa conduce de nuevo a exactamente el mismo proceso de límite que ya hemos denominado diferenciación.

Considérese el ejemplo de un punto que se mueve sobre una línea recta, el eje y dirigido, estando determinada la posición del punto por una sola coordenada, y. Esta coordenada y es la distancia, con su signo adecuado, de nuestro punto móvil medida desde un punto inicial fijo sobre la línea. El movimiento está dado si se conoce y como función del tiempo t: $y = f(t)$. Si esta función es una función lineal $f(t) = ct + b$, se habla de un *movimiento uniforme con velocidad c*, y para cualquier par de valores distintos, t y t_1, puede obtenerse la velocidad dividiendo la distancia recorrida en un intervalo de tiempo por la longitud de dicho intervalo de tiempo:

$$c = \frac{f(t_1) - f(t)}{t_1 - t}.$$

La velocidad es, por lo tanto, el cociente incremental de la función $ct + b$, y este cociente de incrementos es independiente del par particular de instantes que se fijen. Pero, ¿qué es lo que se ha de entender por la velocidad de movimiento en un instante t si el movimiento no es ya uniforme?

Para responder a esta pregunta, se considera el cociente de incrementos

$$[f(t_1) - f(t)]/(t_1 - t),$$

el cual se denominará *"velocidad promedio"* en el intervalo de tiempo entre t_1 y t. Ahora bien, si esta velocidad promedio tiende a un límite definido cuando se hace tender t_1 a t, este límite se definirá como la velocidad en el tiempo t. En otras palabras: *la velocidad,* esto es, la razón instantánea de variación de la distancia con respecto al tiempo *en el tiempo t, es la derivada*

$$f'(t) = \lim_{t_1 \to t} \frac{f(t_1) - f(t)}{t_1 - t}.$$

Newton destacó la interpretación de derivadas [1] como velocidades, y escribió $\dot{y}$, o bien $\dot{f}(x)$, en vez de $f'(t)$, una notación que se utilizará ocasionalmente. Nuevamente, la derivabilidad de la función es una suposición necesaria si la noción de velocidad ha de tener un significado.

Un ejemplo sencillo es el del movimiento de cuerpos en caída libre. Se parte de la ley establecida experimentalmente según la cual la distancia recorrida en un tiempo t por un cuerpo en caída libre, que parte del reposo en $t = 0$, es proporcional a t^2; es por lo tanto representada mediante una función de la forma

$$y = f(t) = at^2$$

con a constante. Como en las pp. 180-181, la velocidad entonces está dada mediante la expresión $f'(t) = 2at$; así: *la velocidad de un cuerpo en caída libre crece en proporción al tiempo.*

c. Ejemplos de diferenciación

Se ilustrará ahora la técnica de la derivación mediante un número de ejemplos típicos.

Funciones lineales

Para la función $y = f(x) = c$ con c constante, se ve para todo x, es $f(x + h) - f(x) = c - c = 0$, de modo que $\lim_{h \to 0} [f(x + h) - f(x)]/h = 0$; esto es, *la derivada de una función constante es cero.*

[1] Denominadas por él "fluxiones".

Para una función lineal $y = f(x) = cx + b$, se encuentra

$$f'(x) = \lim_{h\to 0} \frac{f(x+h) - f(x)}{h} = \lim_{h\to 0} \frac{ch}{h} = c.$$

La derivada de una función lineal es constante.

Potencias de x

En seguida se deriva la función potencia

$$y = f(x) = x^\alpha,$$

suponiendo primero que α es un entero positivo. Siempre que $x_1 \neq x$ se tiene

$$\frac{f(x_1) - f(x)}{x_1 - x} = \frac{x_1^\alpha - x^\alpha}{x_1 - x} = x_1^{\alpha-1} + x_1^{\alpha-2}x + \cdots + x^{\alpha-1},$$

donde se divide directamente o bien se utiliza la fórmula para la suma de una progresión geométrica. Esta sencilla manipulación algebraica es la clave para pasar al límite; pues la última expresión en el miembro derecho de la ecuación es una función continua de x_1, en particular para $x_1 = x$, y así puede efectuarse el paso al límite con $x_1 \to x$ para esta expresión simplemente reemplazando x_1 en todas partes por x. Cada término toma entonces el valor $x^{\alpha-1}$, y, puesto que el número de términos es exactamente α, se obtiene

$$y' = f'(x) = \frac{d(x^\alpha)}{dx} = \alpha x^{\alpha-1}.$$

Se llega al mismo resultado si α es un entero negativo $-\beta$; se debe, sin embargo, suponer que x no es cero. Entonces se encuentra

$$\frac{f(x_1) - f(x)}{x_1 - x} = \frac{\dfrac{1}{x_1^\beta} - \dfrac{1}{x^\beta}}{x_1 - x} = -\frac{x^\beta - x_1^\beta}{x - x_1}\,\frac{1}{x^\beta x_1^\beta}$$

$$= -\frac{x^{\beta-1} + x^{\beta-2}x_1 + \cdots + x_1^{\beta-1}}{x_1^\beta x^\beta}.$$

Una vez más, puede efectuarse el paso al límite simplemente substituyendo x por x_1. Entonces, tal como se hizo antes se obtiene para el límite

$$y' = -\beta\frac{x^{\beta-1}}{x^{2\beta}} = -\beta x^{-\beta-1}.$$

Por lo tanto, para valores enteros *negativos* $\alpha = -\beta$ la derivada está dada nuevamente por la fórmula

$$y' = \alpha x^{\alpha-1}.$$

Finalmente, se probará la misma fórmula cuando x es positivo y α cualquier número racional. Se supone que $\alpha = p/q$, donde p y q son ambos enteros y además positivos. (Si uno de ellos fuera negativo no se requerirían cambios esenciales en la demostración; para $\alpha = 0$ el resultado es ya conocido, puesto que x^α es entonces constante.) Se tiene ahora

$$\frac{f(x_1) - f(x)}{x_1 - x} = \frac{x_1^{p/q} - x^{p/q}}{x_1 - x}.$$

Si se hace ahora $x^{1/q} = \xi$ y $x_1^{1/q} = \xi_1$, se obtiene

$$\frac{f(x_1) - f(x)}{x_1 - x} = \frac{\xi_1^p - \xi^p}{\xi_1^q - \xi^q} = \frac{\xi_1^{p-1} + \xi_1^{p-2}\xi + \cdots + \xi^{p-1}}{\xi_1^{q-1} + \xi_1^{q-2}\xi + \cdots + \xi^{q-1}}.$$

Después de esta última transformación puede efectuarse inmediatamente el paso al límite con $x_1 \to x$ (o, lo que equivale a lo mismo, $\xi_1 \to \xi$) y obtener así para el valor límite la expresión

$$y = \frac{p}{q}\frac{\xi^{p-1}}{\xi^{q-1}} = \frac{p}{q}\xi^{p-q} = \frac{p}{q}x^{(p-q)/q} = \frac{p}{q}x^{(p/q)-1}$$

o, finalmente,

$$f'(x) = y' = \alpha x^{\alpha-1},$$

que es formalmente el mismo resultado obtenido antes. Se deja al lector probar que la misma fórmula de derivación vale también para exponentes racionales negativos.

Se volverá (p. 207) a la derivación de potencias y se probará la validez general de la fórmula anterior para exponentes arbitrarios α.

Funciones trigonométricas

Como un último ejemplo se considera la *derivación de las funciones trigonométricas* sen x y cos x. Se utiliza la elemental fórmula de adición trigonométrica para transformar el cociente incremental

$$\frac{\operatorname{sen}(x+h) - \operatorname{sen} x}{h} = \frac{\operatorname{sen} x \cos h + \cos x \operatorname{sen} h - \operatorname{sen} x}{h}$$

$$= \operatorname{sen} x\frac{\cos h - 1}{h} + \cos x\frac{\operatorname{sen} h}{h}.$$

Recordando las relaciones de la sección 1.8, pp. 107-108,

$$\lim_{h\to 0}\frac{\operatorname{sen} h}{h} = 1, \qquad \lim_{h\to 0}\frac{\cos h - 1}{h} = 0,$$

y se obtiene inmediatamente

$$y' = \frac{d(\operatorname{sen} x)}{dx} = \cos x.$$

La función $y = \cos x$ puede ser derivada exactamente de la misma forma. Partiendo con

$$\frac{\cos(x+h) - \cos x}{h} = \cos x \frac{\cos h - 1}{h} - \operatorname{sen} x \frac{\operatorname{sen} h}{h}$$

y tomando el límite para $h \to 0$, se obtiene la derivada [1]

$$y' = \frac{d(\cos x)}{dx} = -\operatorname{sen} x.$$

d. Algunas reglas fundamentales para la derivación

Tal como en el caso de la integral, existen ciertas reglas básicas para la derivación que se obtienen inmediatamente a partir de la definición y bastan para formar la derivada de muchas funciones.

1. Si $\phi(x) = f(x) + g(x)$, entonces $\phi'(x) = f'(x) + g'(x)$.
2. Si $\psi(x) = cf(x)$ (donde c es una constante), entonces $\psi'(x) = cf'(x)$. Se tiene

$$\frac{\phi(x+h) - \phi(x)}{h} = \frac{f(x+h) - f(x)}{h} + \frac{g(x+h) - g(x)}{h}$$

y

$$\frac{\psi(x+h) - \psi(x)}{h} = c\frac{f(x+h) - f(x)}{h},$$

y nuestras proposiciones se obtienen directamente mediante el paso al límite.

Así, por ejemplo, la derivada de la función $\phi(x) = f(x) + ax + b$ (donde a y b son constantes) está dada por la ecuación

$$\phi'(x) = f'(x) + a.$$

Con la ayuda de estas reglas y de la fórmula para la derivada de una potencia se puede inmediatamente derivar cualquier polinomio $y = a_0x^n + a_1x^{n-1} + \cdots + a_n$ y encontrar

$$y' = na_0x^{n-1} + (n-1)a_1x^{n-2} + \cdots + 2a_{n-2}x + a_{n-1}.$$

e. Derivabilidad y continuidad de funciones

Es útil saber que la derivabilidad es una condición más fuerte que la continuidad:

[1] Si x es interpretado como un ángulo, entonces estas fórmulas sencillas para las derivadas de sen x y cos x presuponen, por supuesto, que el ángulo x está medido en radianes.

Si una función es derivable, entonces es también continua.

Pues si el cociente de incrementos $[f(x+h) - f(x)]/h$ se aproxima a un límite definido cuando h tiende a cero, el numerador de la fracción, esto es, $f(x+h) - f(x)$ debe [1] tender a cero con h; esto precisamente expresa la continuidad de la función $f(x)$ en el punto x. Por lo tanto, son innecesarias otras demostraciones engorrosas de la continuidad de funciones que puede mostrarse que son derivables (esto es, para la mayoría de las funciones que se encontrarán).

Discontinuidades de la derivada-Esquinas

La recíproca, sin embargo, es falsa; *no* es cierto que toda función continua posea una derivada en cada punto. El contraejemplo más sencillo es la función $f(x) = |x|$, esto es, $f(x) = -x$ para $x \leq 0$ y $f(x) = x$ para $x \geq 0$; su gráfica está mostrada en la Fig. 2.24. En el punto $x = 0$ esta

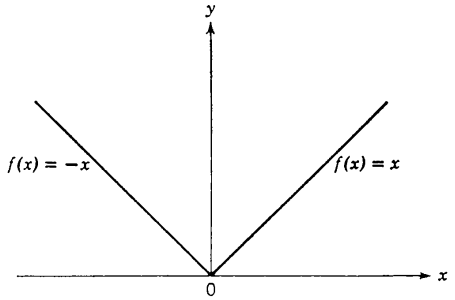


Figura 2.24 $f(x) = |x|$.

función es continua pero no posee derivada. El límite de $[f(x+h) - f(x)]/h$ es igual a 1 si h tiende a cero a través de valores positivos, y es igual a -1 si h tiende a cero por valores negativos; y si el signo de h no se restringe, ningún límite existe. Se dice que nuestra función posee diferentes *derivadas anterior* y *posterior* en el punto $x = 0$, donde por derivada anterior y derivada posterior se entienden respectivamente los valores límite de $[f(x+h) - f(x)]/h$ conforme h se aproxima a cero sólo a través de valores positivos y sólo a través de valores negativos respectivamente. La *derivabilidad* de una función definida en un intervalo en torno al punto considerado requiere así no solamente que existan las derivadas anterior y posterior, sino que éstas sean iguales. Geométricamente, la desigualdad

[1] Puesto que entonces

$$\lim_{h\to 0}[f(x+h) - f(x)] = \left[\lim_{h\to 0}\frac{f(x+h) - f(x)}{h}\right](\lim_{h\to 0} h) = f'(x)\cdot 0 = 0.$$

de las dos derivadas significa que la curva posee una *esquina*. La derivabilidad expresa de una manera precisa lo que intuitivamente podría denominarse *suavidad* de la gráfica de la función.

Discontinuidades infinitas

Como un ejemplo más de puntos donde una función continua no es derivable, se consideran los puntos donde la derivada se torna infinita, esto es, los puntos en los cuales ni existe una derivada anterior ni una derivada posterior, y el cociente incremental $[f(x+h)-f(x)]/h$ crece más allá de toda cota conforme $h \to 0$. Por ejemplo, la función $y = f(x) = \sqrt[3]{x} = x^{1/3}$ está definida y es continua para todos los valores de x. Para todos los valores de x distintos de cero su derivada está dada (p. 185) por la fórmula $y' = \frac{1}{3}x^{-2/3}$. En el punto $x = 0$ se tiene $[f(x+h)-f(x)]/h = h^{1/3}/h = h^{-2/3}$, y se ve de inmediato que cuando $h \to 0$ la expresión no posee valor límite sino que, por el contrario, tiende a infinito. Este estado de cosas es a menudo descrito brevemente diciendo que la función posee una derivada infinita en el punto en cuestión; como se podrá recordar, sin embargo, esto significa solamente que conforme h tiende a cero el cociente diferencia crece más allá de toda cota y que la derivada en el sentido en que se ha definido realmente *no existe*. El significado geométrico de una *derivada infinita* es que la tangente a la curva es vertical (ver Fig. 2.25).

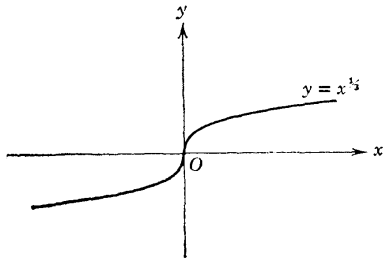


Figura 2.25.

La función $y = f(x) = \sqrt{x}$, la cual está definida y es continua para $x \geq 0$, tampoco es derivable en el punto $x = 0$. Puesto que y no está definida para valores negativos de x, se considera aquí solamente la derivada lateral derecha. La ecuación $[f(h)-f(0)]/h = 1/\sqrt{h}$ muestra que esta derivada es infinita; la curva toca al eje y en el origen (Fig. 2.26).

Finalmente, en la función $y = \sqrt[3]{x^2} = x^{2/3}$ se tiene un caso en el cual la derivada lateral derecha en el punto $x = 0$ es positiva e infinita, en tanto que la derivada lateral izquierda es negativa e infinita, como se sigue de la relación

$$\frac{f(h)-f(0)}{h} = \frac{1}{\sqrt[3]{h}}$$

En realidad, la curva continua $y = x^{2/3}$, llamada *parábola semicúbica* o *parábola de Neil*, posee en el origen una *cúspide* con una tangente perpendicular al eje x (ver Fig. 2.27).

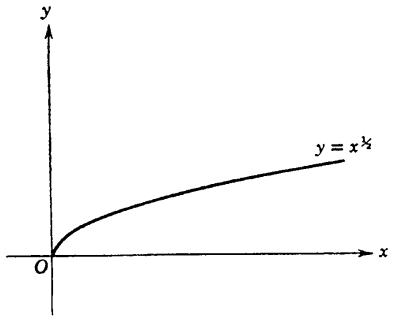


Figura 2.26.

f. Derivadas superiores y su significado

La gráfica de la derivada $f'(x)$ de una función se denomina la *curva derivada* de la gráfica de $f(x)$. Por ejemplo, la curva derivada de la parábola $y = x^2$ es una línea recta representada por la función $y = 2x$. La curva derivada de la curva seno $y = \operatorname{sen} x$ es la curva coseno $y = \cos x$; y, análogamente, la curva derivada de la curva $y = \cos x$ es la curva $y = -\operatorname{sen} x$. (Estas últimas curvas pueden obtenerse una de la otra mediante

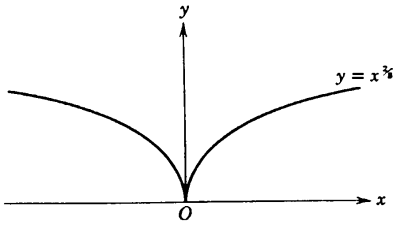


Figura 2.27.

una translación en la dirección del eje x, como se muestra en la Fig. 2.28.)

Es bastante natural el formar curvas derivadas de las curvas derivadas, esto es, formar la derivada de la función $f'(x) = \phi(x)$. Esta derivada

$$\phi'(x) = \lim_{h \to 0} \frac{f'(x+h)-f'(x)}{h}$$

suponiendo que existe, se denomina *segunda derivada* de la función $f(x)$; y se la denotará por $f''(x)$.

Análogamente, puede intentarse formar la derivada de $f''(x)$, llamada *tercera derivada* de $f(x)$, la cual entonces se denota por $f'''(x)$. Para la mayoría de las funciones que nos conciernen no hay nada que nos impida repetir el proceso de derivación tantas veces como queramos, definiendo así una n-ésima *derivada* $f^{(n)}(x)$.[1] Ocasionalmente, será conveniente denominar a la función $f(x)$ su propia 0-ésima derivada.

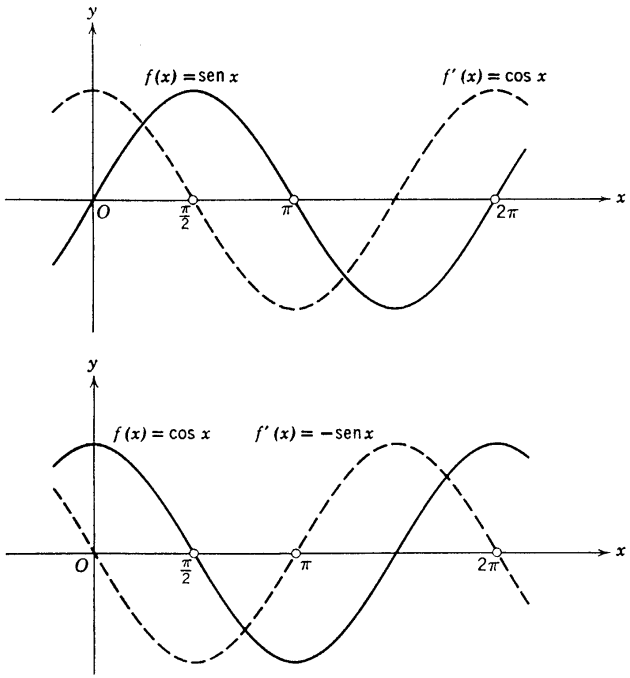


Figura 2.28 Curvas derivadas de sen x y cos x.

Si la variable independiente es interpretada como el tiempo t y el movimiento de un punto está representado como antes por la función $f(t)$, el significado *físico* de la segunda derivada es la razón de variación de la velocidad $f'(t)$ con respecto al tiempo, o sea, como se le denomina usualmente, la *aceleración*. En el ejemplo del cuerpo en caída libre la distancia recorrida en el tiempo t estaba dada mediante la función $y = f(t) = at^2$. Se encontró $f'(t) = 2at$ para la velocidad en el tiempo t. La aceleración

[1] Los términos *segundo, tercero, . . . , n-ésimo coeficiente diferencial* son también utilizados, o bien $D^2f, \ldots, D^nf$ (ver nota 3, p. 180).

posee entonces el valor constante $f''(t) = 2a$ (el cual usualmente es identificado con la constante gravitacional g). Posteriormente (p. 257), se discutirá la interpretación geométrica de la segunda derivada en detalle. Aquí, sin embargo, se toma nota de los siguientes hechos: En un punto donde $f''(x)$ es positiva, $f'(x)$ crece conforme x crece; si aquí $f'(x)$ es positiva, la curva se torna más empinada para x creciente. Si por otra parte, $f''(x)$ es negativa, $f'(x)$ decrece conforme x crece, y si $f'(x)$ es positiva, la curva se torna menos empinada conforme x crece.

Finalmente, se observa que las derivadas superiores pueden utilizarse para definir una función. Así, las funciones trigonométricas se pueden caracterizar por una así llamada *ecuación diferencial* que involucra la función y su segunda derivada. De las fórmulas $(d\cos x)/dx = -\operatorname{sen} x$, $(d\operatorname{sen} x)/dx = \cos x$ se obtiene inmediatamente al derivar nuevamente,

$$\frac{d^2}{dx^2}\cos x = -\cos x, \qquad \frac{d^2}{dx^2}\operatorname{sen} x = -\operatorname{sen} x.$$

Por tanto, si el símbolo u es representativo de cualquiera de las funciones sen x o cos x, se tiene la relación *(ecuación diferencial)*

$$u'' = -u.$$

Esta ecuación diferencial claramente la satisface también cualquier combinación lineal $u = a\cos x + b\operatorname{sen} x$ con coeficientes constantes a, b. Se verá en la Sec. 3.16 b que tales combinaciones lineales, con constantes arbitrarias a y b, son las únicas funciones u para las cuales $u'' = -u$.

En todos los tipos de aplicaciones que involucran oscilaciones o fenómenos ondulatorios, tales como movimientos de resortes o bien ondas sobre la superficie del agua, pasaremos directamente de consideraciones físicas hacia una ecuación diferencial del tipo $u'' = -u$ para la variable física significativa u (usualmente la variable independiente es el *tiempo)*. Es por ello importante reconocer que u puede ser representada sencillamente en términos de funciones trigonométricas (véase capítulo 9).

g. Derivada y cociente de incrementos. Notación de Leibnitz

En la notación de Leibnitz el paso al límite en el proceso de derivación es simbólicamente expresado reemplazando el símbolo Δ por el símbolo d, motivando el símbolo de Leibnitz para la derivada *definida* mediante la ecuación

$$\frac{dy}{dx} = \lim_{\Delta x \to 0} \frac{\Delta y}{\Delta x}.$$

Si se desea obtener un entendimiento claro del significado del cálculo diferencial, debemos precavernos contra la vieja falacia de imaginar la derivada como el cociente de dos "cantidades" dy y dx que son de hecho

"infinitamente pequeñas". El *cociente de incrementos* $\Delta y/\Delta x$ posee un significado sólo para diferencias Δx que no son iguales a cero. *Después* de formar este genuino cociente incremental debe efectuarse el paso al límite por medio de una transformación o algún otro plan el cual también en el límite evita la división por cero. No tiene sentido suponer que *primero* Δx y Δy atraviesan por algo parecido a un proceso de límite y alcanzan valores que son infinitesimalmente pequeños pero aún diferentes de cero, de modo que Δx y Δy se reemplazan por "cantidades infinitamente pequeñas" o "infinitesimales" dx y dy, y que el cociente de estas cantidades es entonces formado. Tal concepción de la derivada es incompatible con la claridad matemática; de hecho, no tiene significado alguno. Para mucha gente indudablemente posee un cierto encanto de misterio, siempre asociado con la palabra "infinito"; y en los primeros días del cálculo diferencial aun el mismo Leibnitz fue capaz de combinar estas vagas ideas místicas con un tratamiento enteramente claro de los procesos de límite. Pero ahora el misticismo de las cantidades "infinitamente pequeñas" no tiene lugar en el cálculo.

La notación de Leibnitz, sin embargo, no es meramente sugestiva por sí misma, sino que es en realidad sumamente flexible y útil. La razón es que en muchos cálculos y transformaciones formales podemos tratar con los símbolos dy y dx exactamente *como si* fueran números ordinarios. Tratando a dx y dy como números es posible dar una expresión más clara a muchos cálculos que pueden llevarse a cabo con toda validez sin su uso. En los siguientes capítulos se verá este hecho verificado una y otra vez, y nos habremos justificado al hacer un uso libre y repetido de ellos, siempre que no se pierda de vista el carácter simbólico de los signos dy y dx.

*También para las derivadas segunda y superiores Leibnitz ideó una notación sugestiva. Consideró la segunda derivada como el límite del "segundo cociente de incrementos" de la siguiente manera: Además de la variable x se considera $x_1 = x + h$ y $x_2 = x + 2h$. Entonces se toma el segundo cociente de incrementos, entendiéndose por éste el primer cociente de incrementos del primer cociente, esto es, la expresión

$$\frac{1}{h}\left(\frac{y_2 - y_1}{h} - \frac{y_1 - y}{h}\right) = \frac{1}{h^2}(y_2 - 2y_1 + y),$$

donde $y = f(x)$, $y_1 = f(x_1)$ y $y_2 = f(x_2)$. Escribiendo $h = \Delta x$, $y_2 - y_1 = \Delta y_1$ y $y_1 - y = \Delta y$, se puede denominar apropiadamente la expresión en el último paréntesis "diferencia de la diferencia" de y, o bien, *segunda diferencia* de y, y escribir simbólicamente [1]

$$y_2 - 2y_1 + y = \Delta y_1 - \Delta y = \Delta(\Delta y) = \Delta^2 y.$$

[1] Aquí $\Delta\Delta = \Delta^2$ es simplemente un símbolo para "diferencia de diferencia" o bien "segunda diferencia".

En esta notación simbólica el segundo cociente incremental es escrito entonces $\Delta^2 y/(\Delta x)^2$, donde el denominador es realmente el cuadrado de Δx, mientras que en el numerador el índice superior 2 denota simbólicamente la repetición del proceso de diferencia. La segunda derivada es expresada entonces mediante

$$f''(x) = \lim_{x_1 \to x} \frac{\Delta^2 f}{(\Delta x)^2}.$$

Este simbolismo para el cociente incremental [1] condujo a Leibnitz a introducir la notación

$$y'' = f''(x) = \frac{d^2y}{dx^2}, \qquad y''' = f'''(x) = \frac{d^3y}{dx^3}, \quad \text{etc.},$$

para las derivadas segundas y superiores, y se verá que esta notación es también de utilidad.[2]

h. El teorema del valor medio del cálculo diferencial

El cociente incremental involucra los valores de una función para distintos valores de x, mientras que la derivada en un punto no nos dice nada acerca de la función en cualquier otro punto; el cociente incremental refleja propiedades de la función "globalmente", mientras la derivada refleja una propiedad *local*, o sea, una propiedad "en lo pequeño". A menudo se requerirá deducir propiedades del todo, o propiedades "globales", de una función a partir de las propiedades locales dadas por su derivada. Para este propósito se utiliza una relación fundamental entre el cociente de incrementos y la derivada, conocido como "el teorema del valor medio del cálculo diferencial".

El teorema del valor medio es fácil de apreciar intuitivamente. Considérese el cociente incremental

$$\frac{f(x_1) - f(x_2)}{x_1 - x_2} = \frac{\Delta f}{\Delta x}$$

de una función $f(x)$ y supóngase que la derivada existe en todas partes en el intervalo cerrado $x_1 \leq x \leq x_2$, de modo que la gráfica de la curva posea una tangente en todas partes. El cociente de incrementos es la tan-

[1] Como debe destacarse, la proposición de que la segunda derivada puede ser representada como el límite del segundo cociente de incrementos requiere, una demostración. Previamente, la segunda derivada no se definió de esta forma sino como el límite del primer cociente de incrementos de la primera derivada. Las dos definiciones son equivalentes siempre que la segunda derivada sea *continua;* la demostración, sin embargo, se dará sólo posteriormente (véase el capítulo 5, apéndice II), puesto que no tenemos aún ninguna necesidad particular de esta propiedad.

[2] Esta es la notación que se acostumbra. Escribiendo $y'' = d^2y/(dx)^2$, $y''' = d^3y/(dx)^3$, con paréntesis, sería un poco más clara, pero de ordinario no se hace.

gente del ángulo de inclinación α de la secante, mostrada en la Fig. 2.29. Imagínese esta secante trasladada paralelamente a sí misma. Al menos una vez alcanzará una posición en la cual es una tangente a la curva en un punto entre x_1 y x_2; esto ocurre en un punto $x = \xi$ de la curva el cual se encuentra a la máxima distancia de la secante. Por lo tanto existe un valor intermedio ξ en el intervalo, tal que

$$\frac{f(x_1) - f(x_2)}{x_1 - x_2} = f'(\xi).$$

Esta proposición se denomina *teorema del valor medio del cálculo diferencial.*[1] Puede expresarse también en forma algo diferente observando que el número ξ puede ser escrito en la forma

$$\xi = x_1 + \Theta(x_2 - x_1),$$

donde todo lo que se sabe en torno a Θ es que se encuentra situado entre 0 y 1. Aún cuando Θ (o bien ξ) en general no puede ser especificado más exactamente, el teorema es sumamente poderoso en las aplicaciones.

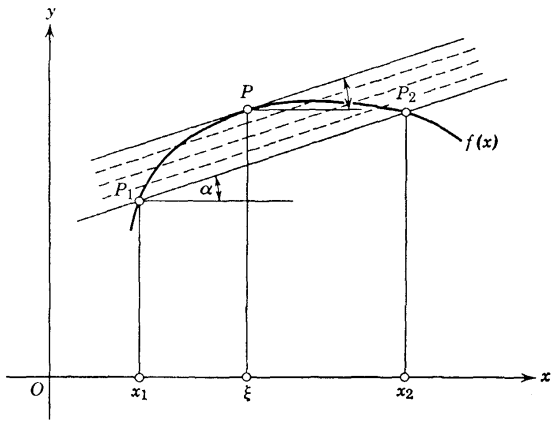


Figura 2.29.

Considérese, por ejemplo, el caso donde x es el tiempo y $y = f(x)$ la distancia de un automóvil desde su punto de partida a lo largo de cierto camino. Entonces $f'(x)$ es la velocidad del automóvil en el tiempo x. Si, digamos, durante las dos primeras horas ($\Delta x = 2$) el automóvil ha cubierto una distancia $\Delta f = 120$ millas, puede concluirse, a partir del teorema del valor medio, que al menos en un momento ξ durante estas dos horas el automóvil tuvo una velocidad de exactamente 60 millas por hora (siempre que la velocidad exista en todo momento). El conductor no puede sostener, por ejemplo, que ha viajado todo el tiempo a menos de 50 millas por hora. Por otra parte, no hay nada que indique cuál fue el tiempo ξ en el cual se alcanzó la velocidad precisa de 60 millas por hora; pudo haber sido en algún tiempo durante la primera hora, o bien durante la segunda hora o bien en diversas ocasiones.

[1] Un nombre más apropiado sería el de *teorema del valor intermedio* del cálculo diferencial.

Una formulación precisa del teorema del valor medio es la siguiente:

Si $f(x)$ es continua en el intervalo cerrado $x_1 \leq x \leq x_2$ y derivable en todo punto del intervalo abierto $x_1 < x < x_2$, entonces existe al menos un valor θ, con $0 < \theta < 1$, tal que

$$\frac{f(x_2) - f(x_1)}{x_2 - x_1} = f'[x_1 + \theta(x_2 - x_1)].$$

Si se substituye x_1 por x y x_2 por $x + h$, el teorema del valor medio puede ser expresado mediante la fórmula

$$\frac{f(x+h) - f(x)}{h} = f'(\xi) = f'(x + \theta h), \qquad x < \xi < x + h.$$

Aun cuando es esencial que $f(x)$ sea continua para todos los puntos del intervalo, incluyendo los puntos extremos, no se tiene que suponer que la derivada exista en los puntos extremos.

Si en cualquier punto en el interior del intervalo la derivada deja de existir, el teorema del valor medio no es necesariamente cierto. Es fácil ver esto en el ejemplo de $f(x) = |x|$.

i. Demostración del teorema

El teorema del valor medio se deduce usualmente mediante la reducción a un caso especial que será establecido primero.

TEOREMA DE ROLLE. *Si una función $\phi(x)$ es continua en el intervalo cerrado $x_1 \leq x \leq x_2$ y derivable en el intervalo abierto $x_1 < x < x_2$, y si, además, $\phi(x_1) = 0$ y $\phi(x_2) = 0$, entonces existe al menos un punto ξ en el interior del intervalo en el cual $\phi'(\xi) = 0$.*

Interpretado geométricamente, esto significa que si una curva alcanza el eje x en dos puntos, entonces debe poseer una tangente horizontal en algún punto intermedio (Fig. 2.30).

Sin duda, puesto que $\phi(x)$ es continua en el intervalo cerrado $[x_1, x_2]$, existe un valor máximo M de $\phi(x)$ y un valor mínimo m en este intervalo (véase p. 124). Puesto que ϕ se hace cero en los puntos extremos, se debe tener $m \leq 0 \leq M$. Si estos valores máximo y mínimo deben ser iguales, entonces necesariamente $m = M = 0$ y así $\phi(x) = 0$ en todos los puntos

del intervalo; entonces también $\phi'(x) = 0$ en el intervalo, y, por lo tanto, $\phi'(\xi) = 0$ para todo ξ en el intervalo. Entonces, debe considerarse solamente el caso donde m y M no son cero ambos. Si, en particular M no es cero, entonces M debe ser positivo. Existe un punto en el intervalo $[x_1, x_2]$ donde $\phi(\xi) = M$. Puesto que ϕ se hace cero en los puntos extremos del intervalo, el punto ξ debe ser un punto interior. Además, $\phi(x) \leq \phi(\xi) = M$

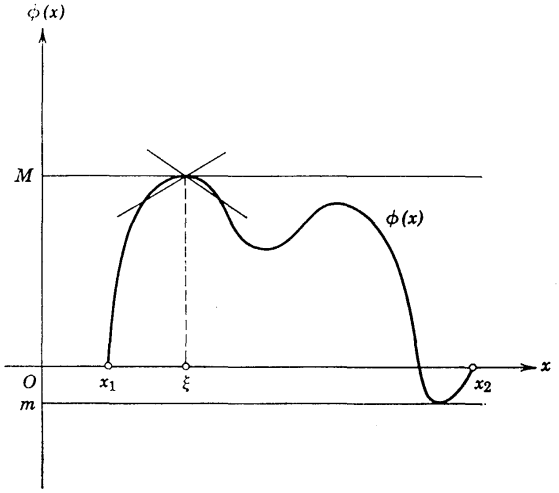


Figura 2.30.

para todo x en $[x_1, x_2]$. Consecuentemente, para todo número h cuyo valor absoluto $|h|$ es suficientemente pequeño, la desigualdad $\phi(\xi + h) - \phi(\xi) \leq 0$ se cumple. Esto implica que el cociente

$$\frac{\phi(\xi + h) - \phi(\xi)}{h}$$

es negativo o bien cero para $h > 0$ y positivo o bien cero para $h < 0$. Si se hace h tender a cero a través de valores positivos se encuentra que $\phi'(\xi) \leq 0$, mientras que para h tendiendo a cero a través de valores negativos se sigue que $\phi'(\xi) \geq 0$. Por lo tanto, $\phi'(\xi) = 0$; y así el teorema de Rolle ha sido probado en el caso $M \neq 0$. El mismo argumento se cumple para $m \neq 0$.

Para probar el teorema del valor medio se aplica el teorema de Rolle a una función que representa la distancia vertical entre el punto $(x, f(x))$ de la gráfica y su secante:

$$\phi(x) = f(x) - f(x_1) - \frac{f(x_2) - f(x_1)}{x_2 - x_1}(x - x_1).$$

Esta función[1] obviamente satisface la condición $\phi(x_1) = \phi(x_2) = 0$, y es de la forma $\phi(x) = f(x) + ax + b$ con coeficientes constantes $a = -[f(x_2) - f(x_1)]/(x_2 - x_1)$ y b. De la p. 187 se sabe que

$$\phi'(x) = f'(x) + a,$$

y así, por el teorema de Rolle,

$$0 = \phi'(\xi) = f'(\xi) + a$$

para un valor intermedio ξ elegido adecuadamente; por lo tanto,

$$f'(\xi) = -a = \frac{f(x_2) - f(x_1)}{x_2 - x_1};$$

y así el teorema del valor medio está probado.

Significado del teorema

La derivada de una función fue definida como el límite de cocientes de incrementos para un intervalo cuando los puntos extremos se aproximan entre sí. El teorema del valor medio establece una conexión entre cocientes de incrementos y derivadas de una función derivable, que no involucra la contracción del intervalo. Cada cociente incremental es igual a la derivada en un punto intermedio adecuado ξ.

Ejemplos. Tal como en el teorema del valor medio del cálculo integral, no hay nada afirmado específicamente en el teorema del valor intermedio respecto a la ubicación de ξ además del hecho de que ξ está situado en el interior del intervalo. Para el ejemplo de la función cuadrática $y = f(x) = x^2$ con derivada $f'(x) = 2x$, se encuentra

$$\frac{f(x_2) - f(x_1)}{x_2 - x_1} = x_1 + x_2 = f'(\xi),$$

donde $\xi = (x_1 + x_2)/2$ es el punto medio del intervalo $[x_1, x_2]$. En general, sin embargo, ξ puede estar situado en cualquier otra parte entre x_1 y x_2. Por ejemplo, si $f(x) = x^3$, se tiene $[f(1) - f(0)]/(1 - 0) = 1 = f'(\xi) = 3\xi^2$, donde $\xi = 1/\sqrt{3}$.

Funciones monótonas. Como una de las muchas aplicaciones del teorema del valor medio del cálculo diferencial, se prueba que si la derivada de

[1] Esta función es también proporcional a la distancia del punto $(x, f(x))$, de la curva, a la secante. El lector puede verificar esto por sí mismo fácilmente; por ejemplo, utilizando la propiedad de la geometría analítica elemental de que la expresión $(y - mx - b)/\sqrt{1 + m^2}$ representa la distancia (con su signo) del punto (x, y) a la recta con ecuación $y - mx - b = 0$. De esta forma se encuentra que, efectivamente, en los puntos de la curva que están a una distancia mayor de la secante, la tangente es paralela a la secante.

$f(x)$ posee un signo constante, entonces f es monótona. Específicamente, se supone $f(x)$ continua en el intervalo cerrado $[a, b]$ y derivable en cada punto del intervalo abierto (a, b). *Si en estas condiciones es* $f'(x) > 0$ *para* x *en* (a, b), *entonces la función* $f(x)$ *es monótona creciente; similarmente, si* $f'(x) < 0$, *la función es monótona decreciente.* La demostración es obvia: Sean x_1 y x_2 dos valores cualesquiera en el intervalo cerrado $[a, b]$. Entonces existe un ξ entre x_1 y x_2, y, por lo tanto, también entre a y b, tal que

$$f(x_2) - f(x_1) = f'(\xi)\,(x_2 - x_1).$$

Si $f'(x) > 0$ en todas partes en (a, b), se tiene en particular $f'(\xi) > 0$. Por lo tanto, $f(x_2) - f(x_1)$ es positivo para $x_2 > x_1$; esto es, $f(x)$ es creciente. Análogamente, f es decreciente si $f'(x) < 0$ en (a, b).

De la misma manera se muestra que *una función* $f(x)$ *continua en* $[a, b]$ *y derivable en el intervalo abierto* (a, b) *debe ser constante si* $f'(x) = 0$ *en todas partes en* (a, b). Pues entonces

$$f(x_2) - f(x_1) = f'(\xi)\,(x_2 - x_1) = 0.$$

Esta importante propiedad corresponde al hecho intuitivamente obvio de que una curva cuya tangente en cada punto es paralela al eje x debe ser una línea recta paralela al eje x.

Continuidad de Lipschitz de funciones diferenciables. Se mencionó anteriormente que una función $f(x)$ que posee una derivada es necesariamente continua. El teorema del valor medio del cálculo diferencial proporciona una información cuantitativa mucho más precisa, a saber, un *módulo de continuidad.* Considérese una función $f(x)$ la cual está definida en el intervalo cerrado $[a, b]$ y posee una derivada $f'(x)$ en cada punto de ese intervalo. Supóngase que $f'(x)$ es *acotada* en el intervalo (éste es ciertamente el caso siempre que $f'(x)$ sea definida y continua en el intervalo cerrado $[a, b]$); existe un número M tal que $|f'(x)| \leq M$. Para dos valores cualesquiera x_1, x_2 en (a, b), se infiere del teorema del valor medio que

$$|f(x_2) - f(x_1)| = |f'(\xi)\,(x_2 - x_1)| \leq M\,|x_2 - x_1|.$$

Para $\varepsilon > 0$ dado hemos así producido un módulo sencillo de continuidad, $\delta = \varepsilon/M$, tal que

$$|f(x_2) - f(x_1)| \leq \varepsilon \qquad \text{para} \quad |x_2 - x_1| \leq \delta.$$

Tómese, por ejemplo, la función $f(x) = x^2$ en el intervalo $-a \leq x \leq +a$. Puesto que

$$|f'(x)| = |2x| \leq 2a,$$

se ve que aquí

$$|f(x_2) - f(x_1)| \leq \varepsilon \qquad \text{para} \quad |x_2 - x_1| \leq \varepsilon/2a.$$

Se dijo que una función $f(x)$ "satisface una condición de Lipschitz" o bien que es "continua en el sentido de Lipschitz" (también, que es "lipschitziana") si existe una constante M tal que

$$|f(x_2) - f(x_1)| \leq M\,|x_2 - x_1|$$

para todos x_1, x_2 en cuestión. Esto significa que todos los *cocientes de incrementos*

$$\frac{f(x_2) - f(x_1)}{x_2 - x_1}$$

poseen la misma cota superior M para su valor absoluto. Se ve que toda función f con derivada continua f' en un intervalo cerrado es continua en el sentido de Lipschitz. Sin embargo, aun funciones que no poseen derivada en cada punto pueden ser "continuas de Lipschitz", como lo muestra el ejemplo $f(x) = |x|$. El lector puede verificar por él mismo que para esta función siempre es $|f(x_2) - f(x_1)| \leq |x_2 - x_1|$.

Por otra parte, no toda función es continua en el sentido de Lipschitz. Esto se muestra mediante el ejemplo de $f(x) = x^{1/3}$; aquí el cociente

$$\frac{f(x) - f(0)}{x - 0} = x^{-2/3}$$

no está acotado para x pequeño; por lo tanto, $f(x)$ no es continua en el sentido de Lipschitz en $x = 0$. Esto es compatible con el hecho de que la derivada $f'(x) = 1/3x^{2/3}$ no permanece acotada cuando x tiende a cero. Las funciones que son continuas en el sentido de Lipschitz constituyen una clase importante intermedia entre aquéllas que son simplemente continuas y aquéllas que poseen una derivada continua.

j. La aproximación de funciones mediante funciones lineales. Definición de diferenciales

Definición. La derivada de una función $y = f(x)$ se definió mediante

$$f'(x) = \lim_{h \to 0} \frac{f(x + h) - f(x)}{h} = \lim_{\Delta x \to 0} \frac{\Delta y}{\Delta x},$$

donde $\Delta x = h$. Si para un x fijo y una h variable se define una cantidad ε mediante

$$\varepsilon(h) = \frac{f(x + h) - f(x)}{h} - f'(x) = \frac{\Delta y}{\Delta x} - f'(x),$$

entonces el hecho de que $f'(x)$ es la derivada de f en el punto x equivale a la ecuación

$$\lim_{h \to 0} \varepsilon(h) = 0.$$

La cantidad $\Delta y = f(x + h) - f(x)$ representa la variación o *incremento* en el valor de la variable dependiente y que resulta cuando el valor x de la variable independiente es variado en la cantidad $\Delta x = h$. Puesto que

$$\Delta y = f'(x)\,\Delta x + \varepsilon\,\Delta x,$$

la cantidad Δy aparece como la suma de dos partes, a saber, una parte $f'(x)\,\Delta x$ que es proporcional a Δx y una parte $\varepsilon\,\Delta x$ que puede hacerse tan pequeña como se desee comparada con Δx haciendo a Δx mismo suficientemente pequeño. La parte lineal, dominante, en la expresión para Δy la denominaremos la *diferencial* dy de y, y escribimos

$$dy = df(x) = f'(x)\,\Delta x.$$

Para cualquier función derivable f y para un x fijo esta diferencial es una función lineal bien definida de $h = \Delta x$. Por ejemplo, para la función $y = x^2$ se tiene $dy = d(x^2) = 2x\,\Delta x = 2xh$. Para la función particular $y = x$, cuya derivada posee el valor constante uno, se tiene simplemente $dx = \Delta x$. Es entonces compatible con nuestra definición el escribir dx para Δx cuando x es la variable independiente; por lo tanto, la diferencial de cualquier función $y = f(x)$ puede ser escrita también como

$$dy = df(x) = f'(x)\,dx.$$

El incremento de la variable dependiente

$$\Delta y = f'(x)\,dx + \varepsilon\,dx = dy + \varepsilon\,dx$$

difiere de la diferencial dy en la cantidad $\varepsilon\,dx$, que, en general, no es cero. En el ejemplo de la función $y = x^2$ se tiene $dy = 2x dx$, mientras que

$$\Delta y = (x + dx)^2 - x^2 = 2x\,dx + (dx)^2 = dy + \varepsilon\,dx,$$

donde $\varepsilon = dx$.

Anteriormente se utilizó el símbolo dy/dx de manera puramente simbólica para denotar el límite del cociente $\Delta y/\Delta x$ para Δx tendiendo a cero. Con la presente definición de las diferenciales dy y dx, la derivada dy/dx puede ser considerada en realidad como el cociente ordinario de dy y dx. Aquí, sin embargo, dy y dx no son ahora en ningún sentido cantidades "infinitamente pequeñas" o "infinitesimales"; tal interpretación estaría desprovista de sentido. En vez de eso, dy y dx son funciones lineales bien definidas de $h = \Delta x$, las cuales para Δx grande pueden poseer valores numéricos grandes. No hay nada sobresaliente en el hecho de que el cociente dy/dx de estas cantidades posea el mismo valor que la derivada $f'(x)$. Esta es solamente una tautología reestableciendo la definición de dy como $f'(x)\,dx$.[1]

[1] Análogamente, las diferenciales de orden superior podrían ser definidas por $d^2y = f''(x)h^2 = f''(x)\,(dx)^2$, $d^3y = f'''(x)\,(dx)^3$, etc., en concordancia con la notación de Leibnitz para las derivadas superiores.

Reescribiendo la relación entre incremento y diferencial de f en la forma

$$f(x + h) = f(x) + hf'(x) + \varepsilon h,$$

se ve que la expresión $f(x + h)$ considerada como una función de h está representada por la función lineal $f(x) + hf'(x)$, con un error εh que es arbitrariamente pequeño comparado con h si h es suficientemente pequeño. Esta representación aproximada de $f(x + h)$ mediante la función lineal $f(x) + hf'(x)$ significa geométricamente que se reemplaza la curva por su tangente en el punto x (véase Fig. 2.31).

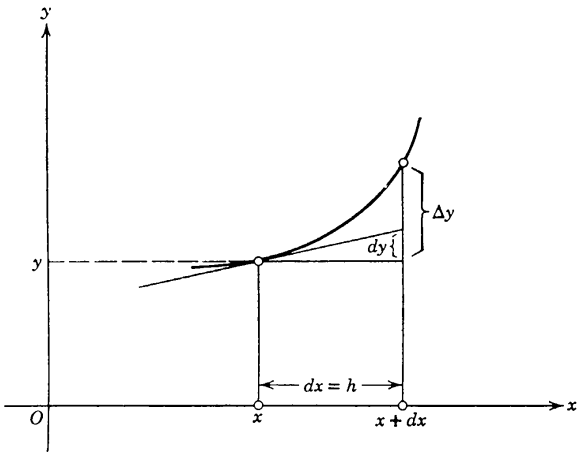


Figura 2.31 Incremento Δy y diferencial dy.

Aproximación lineal

Una estimación más precisa de la magnitud del *"error"*, esto es, de la desviación de la función $f(x)$ respecto de la función lineal que representa a la tangente, está dada mediante el teorema del valor medio del cálculo diferencial. Para una ξ adecuada entre x y $x + h$ se tiene

$$f(x + h) - f(x) = hf'(\xi),$$

de modo que

$$\varepsilon = \frac{f(x + h) - f(x)}{h} - f'(x) = f'(\xi) - f'(x).$$

Si, como ocurre usualmente en las aplicaciones, la función $f'(x)$ misma posee una derivada $f''(x)$, se encuentra, aplicando el teorema del valor medio por segunda vez, que

$$f'(\xi) - f'(x) = (\xi - x)f''(\eta),$$

donde η es un valor intermedio entre x y ξ y por lo tanto también entre x y $x + h$. Se sigue que

$$|\varepsilon| = |(\xi - x)f''(\eta)| = |\xi - x|\,|f''(\eta)| \leq hM,$$

donde M es una cota superior para el valor absoluto de la segunda derivada de f en el intervalo $[x, x + h]$. Entonces $|\varepsilon h|$, que mide la desviación de $f(x + h)$ respecto a la función lineal $f(x) + hf'(x)$, es a lo más igual a Mh^2. Para h suficientemente pequeño la expresión Mh^2 es, por supuesto, mucho menor que $f'(x)h$, a menos que suceda que $f'(x)$ posee el valor cero. Esta *aproximación de una función en un intervalo pequeño mediante una función lineal* es de gran significado tanto para las aplicaciones prácticas como para el análisis matemático avanzado. En capítulos posteriores volveremos a este tópico, y se deducirá entonces la mejor estimación de $|\varepsilon h| \leq \frac{1}{2}Mh^2$.

Interpolación

*Cuando una función $f(x)$ está descripta numéricamente mediante una tabla de valores, f está determinada ordinariamente mediante *interpolación lineal* para argumentos x intermedios entre aquéllos para los cuales f está registrada. Este procedimiento corresponde también a reemplazar la función f mediante una función lineal en un intervalo. En este caso la gráfica de la función lineal está dada mediante una *secante* en vez de una *tangente* a la curva que representa f. Si, digamos, los valores de f son conocidos en los puntos a y b, reemplazamos a $f(x)$ para un x intermedio por la expresión

$$\phi(x) = f(a) + (x - a)\frac{f(b) - f(a)}{b - a},$$

la cual es lineal en x y da los valores correctos de f en los puntos extremos $x = a$ y $x = b$ del intervalo (véase Fig. 2.32). Aquí, nuevamente mediante el uso del teorema del valor medio, puede estimarse el error en esta aproximación. Se tiene

$$f(x) - \phi(x) = (x - a)\left[\frac{f(x) - f(a)}{x - a} - \frac{f(b) - f(a)}{b - a}\right]$$
$$= (x - a)[f'(\xi_1) - f'(\xi_2)].$$

Puesto que ξ_1 está situado entre a y x, estará también situado entre a y b, como lo está ξ_2. Una segunda aplicación del teorema del valor medio del cálculo diferencial conduce entonces a

$$f(x) - \phi(x) = (x - a)f''(\eta)(\xi_1 - \xi_2),$$

donde η está entre ξ_1 y ξ_2 y, por lo tanto, también entre a y b. Consecuentemente, denotando por M a una cota superior para $|f''|$ en el intervalo $[a, b]$, se encuentra que

$$|f(x) - \phi(x)| \leq |x - a|\,|\xi_1 - \xi_2|\,|f''(\eta)| \leq M(b - a)^2.$$

Una vez más, la desviación de f de su aproximación lineal puede ser estimada mediante el cuadrado de la longitud del intervalo.

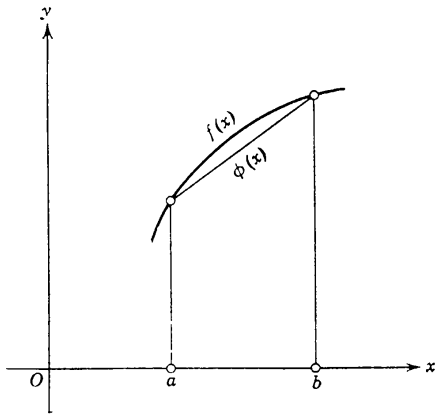


Figura 2.32 Interpolación lineal.

Como ejemplo numérico se toman de una tabla de funciones trigonométricas con medida en radianes los valores

$$\text{sen } 0.75 = 0.6816, \qquad \text{sen } 0.76 = 0.6889,$$

donde los errores no exceden de 0.00005. Si se desea deducir el valor de la función seno para el argumento intermedio 0.754, se encuentra mediante interpolación lineal que

$$\text{sen } 0.754 \approx 0.6816 + \tfrac{4}{10}(0.6889 - 0.6816) \approx 0.6845.$$

Para la función $f(x) = \text{sen } x$ la primera derivada es $f'(x) = \cos x$, y la segunda derivada $f''(x) = -\text{sen } x$. Obviamente, $|f''(x)| \leq 1$, de modo que el error en el valor que se encontró para sen 0.754 como resultado del procedimiento de interpolación lineal no excede de $1 \times (0.01)^2 = 0.0001$. A esta estimación del error debemos añadir posiblemente errores debidos a redondeos en los valores tabulados y en la interpolación.

Puede compararse este valor obtenido mediante interpolación lineal con el valor que se obtendría reemplazando la curva seno por su tangente en el punto $x = 0.75$. Tomando $f'(0.75) = \cos 0.75 = 0.7317$ de la tabla, se encuentra

$$\text{sen } 0.754 \approx f(0.75) + f'(0.75)\,(0.004) = \text{sen } 0.75 + 0.004 \cos 0.75$$
$$\approx 0.6845.$$

Nótese que el valor verdadero de sen 0.754 correcto para seis cifras significativas es 0.684560.

k. Observaciones sobre aplicaciones a las ciencias naturales

Al aplicar las matemáticas a fenómenos naturales nunca se opera con cantidades *conocidas en forma precisa*. El que una longitud sea *exactamente* un metro es una cuestión que no puede ser decidida mediante ningún experimento y que en consecuencia no posee significado físico. Además, no existe significado físico al decir que la longitud de una varilla de material es racional o irracional; puede ser medida siempre con cualquier grado deseado de precisión mediante números racionales, y el único problema significativo es el de si se puede manipular para efectuar tal medición utilizando números racionales con denominadores relativamente pequeños. Tal como la cuestión de racionalidad o de irracionalidad en el sentido riguroso de "matemáticas exactas" no posee significado físico, efectuar procesos de límite en las aplicaciones no es usualmente más que una idealización matemática.

El significado práctico —y sobresaliente— de tales idealizaciones estriba en el hecho de que a través de idealizaciones las expresiones analíticas se tornan esencialmente más sencillas y más manejables. Por ejemplo, es mucho más simple y más conveniente operar con la noción de velocidad instantánea, la cual es función de *un* solo instante definido de tiempo, que con la noción de velocidad promedio entre dos instantes diferentes. Sin tal idealización toda investigación científica de la Naturaleza estaría condenada a complicaciones irremediables y se atollaría desde el principio.

No se tiene la intención de entrar en una discusión filosófica de la relación de las matemáticas con la realidad. Para un mejor entendimiento de la teoría se deberá destacar que en las aplicaciones se tiene el derecho de reemplazar una derivada por un cociente incremental y viceversa, siempre que las diferencias sean suficientemente pequeñas para garantizar una aproximación suficientemente cercana. El físico, el biólogo, el ingeniero, o cualquier otro que tenga que tratar con estas ideas en la práctica, tendrá por ello el derecho de identificar el cociente incremental con la derivada, dentro de sus límites de precisión. Mientras menor sea el incremento $h = dx$ de la variable independiente, con mayor precisión podrá representarse el incremento $\Delta y = f(x + h) - f(x)$ mediante la diferencial $dy = hf'(x)$. Mientras se mantenga, a sabiendas, dentro de los límites de precisión requeridos por el problema, puede estarle permitido hablar de las cantidades $dx = h$ y $dy = hf'(x)$ como "infinitesimales". Estas cantidades "físicamente

infinitesimales" poseen un significado preciso. Dichas cantidades son variables con valores finitos, diferentes de cero y elegidos suficientemente pequeños para la investigación dada; por ejemplo, menores que una parte fraccional de una longitud de onda, o menores que la distancia entre dos electrones en un átomo; en general, menores que el grado de precisión requerido.

2.9 La integral, la función primitiva y los teoremas fundamentales del cálculo

a. La derivada de la integral

Como se afirmó ya, la conexión entre integración y derivación es la piedra angular del cálculo diferencial e integral.

Recordamos de la sección 2.4 que una integral indefinida de una función continua $f(x)$ está definida como una función $\phi(x)$ del punto extremo superior de integración mediante la fórmula

$$\phi(x) = \int_a^x f(u)\,du,$$

donde a era cualquier punto en el dominio de f. Se probará ahora:

El Teorema Fundamental del Cálculo (Primera Parte). *La integral indefinida $\phi(x)$ de una función continua $f(x)$ posee siempre una derivada $\phi'(x)$ y, además,*

$$\phi'(x) = f(x).$$

Esto es, la derivación de la integral indefinida de una función continua reproduce siempre el integrando:

$$\frac{d}{dx}\int_a^x f(u)\,du = f(x).$$

Este carácter inverso de las operaciones de derivación e integración es el hecho básico del cálculo. La demostración es una consecuencia inmediata del teorema del valor medio del cálculo integral. De acuerdo con ese teorema, para cualesquiera valores de x y $x + h$ del dominio de f se tiene

$$\phi(x + h) - \phi(x) = \int_a^{x+h} f(u)\,du = hf(\xi),$$

donde ξ es algún valor en el intervalo con puntos extremos x y $x + h$. Para h tendiendo a cero el valor ξ debe tender a x, de modo que

$$\lim_{h\to 0} \frac{\phi(x + h) - \phi(x)}{h} = \lim_{h\to 0} f(\xi) = f(x),$$

puesto que f es continua. Por lo tanto, $\phi'(x) = f(x)$, como afirma el teorema.

Aplicaciones. (*a*) Puede utilizarse el teorema para encontrar derivadas de algunas funciones introducidas anteriormente. El logaritmo natural se definió para $x > 0$ mediante la integral indefinida

$$\log x = \int_1^x \frac{1}{u}\,du.$$

Se sigue inmediatamente que

$$\frac{d \log x}{dx} = \frac{1}{x}.$$

(*b*) Logaritmos más generales de una base arbitraria a fueron expresados en la forma

$$\log_a x = \frac{\log x}{\log a}.$$

Aplicando la regla para la derivada del producto de una constante y de una función se encuentra que

$$\frac{d}{dx}\log_a x = \frac{1}{x \log a}.$$

(*c*) Se encontró que

$$\frac{d}{dx}x^\alpha = \alpha x^{\alpha-1}$$

en el caso en que el exponente α es un entero o, más generalmente, un número racional. Puede ahora extenderse esta fórmula para α arbitrario. Para este propósito se recuerda la fórmula de integración

$$\int_a^b u^\beta\,du = \frac{1}{\beta+1}(b^{\beta+1} - a^{\beta+1}),$$

la cual se probó para cualquier pareja de números positivos a, b y cualquier $\beta \neq -1$. Si se remplaza aquí el límite superior b por la variable x y se derivan ambos miembros respecto a x, se sigue, para $x > 0$, que

$$x^\beta = \frac{d}{dx}\frac{1}{\beta+1}(x^{\beta+1} - a^{\beta+1}).$$

Utilizando las reglas para la derivada de una suma y el producto de una constante por una función, podemos escribir este resultado en la forma

$$x^\beta = \frac{1}{\beta+1}\frac{d}{dx}x^{\beta+1}.$$

Sustituyendo α por $\beta + 1$, se obtiene la fórmula

$$\frac{d}{dx}x^\alpha = \alpha x^{\alpha-1}$$

para cualquier $\beta \neq -1$, esto es, para $\alpha \neq 0$. Sin embargo, la fórmula también se cumple trivialmente para $\alpha = 0$, pues entonces $x^\alpha = 1$, y la derivada de una constante es cero.

b. La función primitiva y su relación con la integral

Inversión de la derivación

El teorema fundamental muestra que la integral indefinida $\phi(x)$, esto es, la integral con un límite superior variable x, de una función $f(x)$, es una solución del siguiente problema: *Dada $f(x)$ determinar una función $F(x)$ tal que*

$$F'(x) = f(x).$$

Este problema requiere *invertir el proceso de derivación.* Es típico de los problemas inversos que ocurren en muchas partes de las matemáticas y que hemos encontrado ya como un método matemático fructífero para generar nuevos conceptos. (Por ejemplo, la primera extensión de la idea de los números naturales está sugerida por el deseo de invertir ciertos procesos elementales de la aritmética. Nuevamente se obtuvieron nuevos tipos de funciones de las inversas de funciones conocidas.)

Cualquier función $F(x)$ tal que $F'(x) = f(x)$, se denomina *función primitiva de $f(x)$*, o, simplemente, una *primitiva de $f(x)$*; esta terminología sugiere que la función $f(x)$ es deducida de $F(x)$.

Este problema de la inversión de la derivación, o de encontrar una función primitiva, es a primera vista de un carácter bastante diferente al problema de la integración. La primera parte del teorema fundamental del cálculo afirma, sin embargo:

Toda integral indefinida $\phi(x)$ de la función $f(x)$ es una primitiva de $f(x)$.

Empero, este resultado no resuelve completamente el problema de encontrar las funciones primitivas. Pues no se sabe aún si se han encontrado *todas* las soluciones del problema. La pregunta acerca del conjunto de todas las funciones primitivas es contestada por el siguiente teorema, mencionado en ocasiones como *segunda parte del teorema fundamental* del cálculo diferencial e integral

La diferencia de dos funciones primitivas $F_1(x)$ y $F_2(x)$ de la misma función $f(x)$ es siempre una constante

$$F_1(x) - F_2(x) = c.$$

Así, a partir de cualquier función primitiva $F(x)$ pueden obtenerse todas las demás en la forma

$$F(x) + c$$

mediante una elección adecuada de la constante c. Recíprocamente, para cada valor de la constante c la expresión $F_1(x) = F(x) + c$ *representa una función primitiva de* $f(x)$.

Es claro que para cualquier valor de la constante c la función $F(x) + c$ es una función primitiva, siempre que la propia $F(x)$ lo sea. Pues se tiene (ver p. 187)

$$\frac{d}{dx}[F(x) + c] = \frac{d}{dx}F(x) + \frac{d}{dx}c = F'(x) = f(x).$$

Por tanto, para completar la demostración del teorema queda sólo por mostrar que la diferencia de dos funciones primitivas $F_1(x)$ y $F_2(x)$ es siempre constante. Con este fin, considérese la diferencia

$$F_1(x) - F_2(x) = G(x).$$

Claramente,

$$G'(x) = F_1'(x) - F_2'(x) = f(x) - f(x) = 0.$$

Sin embargo, se probó en la p. 199, a partir del teorema del valor medio del cálculo diferencial, que una función cuya derivada se hace cero en todas partes de un intervalo es una constante. Por lo tanto, $G(x)$ es una constante c, y el teorema queda demostrado.

Combinando las dos partes que se acaban de probar, puede formularse ahora el

TEOREMA FUNDAMENTAL DEL CÁLCULO. *Toda función primitiva* $F(x)$ *de una función dada* $f(x)$, *continua en un intervalo, puede ser representada en la forma*

$$F(x) = c + \phi(x) = c + \int_a^x f(u)\,du,$$

donde c y a son constantes; y, recíprocamente, para cualesquiera valores constantes de a y c, elegidos arbitrariamente,[1] *esta expresión representa siempre una función primitiva.*

Observaciones

Es posible conjeturar que la constante c puede, en general, ser omitida debido a que variando el límite inferior a se varía la función primitiva mediante una constante aditiva; esto es, que todas las funciones primitivas son integrales indefinidas. Frecuentemente, sin embargo, no pueden obtenerse todas las funciones primitivas si se omite la c, como el ejemplo $f(x) = 0$ muestra. Para esta función la integral indefinida será siempre cero, independientemente del límite inferior; así, cualquier constante arbitraria es una función primitiva de $f(x) = 0$. Un segundo ejemplo es la función $f(x) = \sqrt{x}$, la cual está definida sólo para valores no negativos de x. La integral indefinida es

$$\phi(x) = \tfrac{2}{3}x^{3/2} - \tfrac{2}{3}a^{3/2},$$

y se ve que no importa cómo se elija el límite inferior a; la integral indefinida $\phi(x)$ se obtiene siempre de $\frac{2}{3}(x)^{3/2}$ añadiendo una constante $-\frac{2}{3}a^{3/2}$ que es menor que o igual a cero; sin embargo, una función tal como $\frac{2}{3}x^{3/2} + 1$ es también una primitiva de $\sqrt{x}$. Así, en la expresión general para la función primitiva no puede hacerse caso omiso de la constante aditiva arbitraria.

La relación que se ha encontrado sugiere extender la noción de integral indefinida de modo que incluya todas las funciones primitivas. En lo sucesivo toda expresión de la forma $c + \phi(x) = c + \int_a^x f(u)\,du$ se denominará una *integral indefinida de* $f(x)$, *y no se distinguirá más entre la función primitiva y la integral indefinida.* Sin embargo, si el lector ha de poseer un correcto entendimiento de las interrelaciones de estos conceptos, es absolutamente necesario que tenga en mente que en primera instancia la integración y la inversión de la derivación son dos cosas diferentes, y que es sólo el conocimiento de la relación entre ellas lo que da el derecho de aplicar el término "integral indefinida" también a la función primitiva.

Es bastante usual utilizar una notación la cual no es perfectamente clara sin comentario: se escribe

$$F(x) = \int f(x)\,dx,$$

cuando se entiende que la función $F(x)$ es de la forma

$$F(x) = c + \int_a^x f(u)\,du$$

para constantes apropiadas c y a; esto es, se omite el límite superior x, el límite inferior a y la constante aditiva c, y se utiliza la letra x para la variable de integración. Estrictamente hablando, por supuesto, hay una pequeña incompatibilidad en utilizar la misma letra para la variable de integración y el límite superior de integración x que es la variable independiente en $F(x)$. Al usar la notación $\int f(x)\,dx$ no debe jamás perderse de vista la indeterminación asociada a ella, esto es, el hecho de que el símbolo denota siempre *una* de las funciones primitivas de f solamente. La fórmula $F(x) = \int f(x)\,dx$ es justamente una forma simbólica de escribir la relación

$$\frac{d}{dx}F(x) = f(x).$$

[1] Siempre que a esté situado en el dominio de f.

c. El uso de la función primitiva para la evaluación de integrales definidas

Supóngase que se conoce cualquier función primitiva $F(x)$ para la función $f(x)$ y que se desea evaluar la integral definida $\int_a^b f(u)\,du$. Se sabe que la integral indefinida

$$\phi(x) = \int_a^x f(u)\,du,$$

siendo también una primitiva de $f(x)$, puede sólo diferir de $F(x)$ por en constante aditiva. Por ello

$$\phi(x) = F(x) + c,$$

y la constante aditiva c está determinada de inmediato debido a que la integral indefinida $\phi(x) = \int_a^x f(u)\,du$ debe tomar el valor cero cuando $x = a$. Se obtiene así $0 = \phi(a) = F(a) + c$, de lo cual se sigue que $c = -F(a)$ y $\phi(x) = F(x) - F(a)$. En particular, para el valor $x = b$ se tiene la fórmula básica

$$\int_a^b f(u)\,du = F(b) - F(a),$$

si

$$F'(u) = f(u).$$

Por ello,

Si $F(x)$ es cualquier función primitiva de la función continua $f(x)$, la integral definida de $f(x)$ entre los límites a y b es igual a la diferencia $F(b) - F(a)$.

Si se utiliza la relación $F'(x) = f(x)$, esta consecuencia del teorema fundamental puede ser escrita en la forma

$$(31)\qquad F(b) - F(a) = \int_a^b F'(x)\,dx = \int_a^b \frac{dF(x)}{dx}\,dx = \int_a^b dF(x),$$

donde ahora $F(x)$ puede ser cualquier función con una derivada continua $F'(x)$, y donde se utiliza la sugestiva notación simbólica $dF(x) = F'(x)\,dx$ de Leibnitz.

Al aplicar nuestra regla, a menudo se utiliza una barra vertical para denotar la diferencia de valores en los puntos extremos, escribiéndose

$$\int_a^b \frac{dF(x)}{dx}\,dx = F(b) - F(a) = F(x)\Big|_a^b.$$

Puede escribirse (31) en la forma

$$(32)\qquad \frac{F(b) - F(a)}{b - a} = \frac{1}{b - a}\int_a^b F'(x)\,dx.$$

Recordando la definición de la p. 163 relativa al promedio de una función en un intervalo, la regla afirma entonces que *el cociente incremental de la función $F(x)$ formado para los puntos a y b es igual a la media aritmética o promedio de la derivada de $F(x)$ en el intervalo con puntos extremos a y b.* Cuando se consideró el movimiento de una partícula sobre una recta, a la variación en distancia s dividida por la variación en tiempo t se le denominó "velocidad promedio". Se ve ahora que en realidad $\Delta s/\Delta t$ es precisamente el promedio de las velocidades ds/dt para el intervalo de tiempo dado, si t es la variable independiente utilizada al formar el promedio.

RELACIÓN ENTRE LOS TEOREMAS DE VALOR MEDIO

La fórmula

$$(33)\qquad F(b) - F(a) = \int_a^b f(x)\,dx$$

que se cumple para cualquier función continua f y una de sus primitivas, F, hace también evidente la relación entre los teoremas del valor medio del cálculo integral (p. 163) y del cálculo diferencial (p. 194). Mediante el teorema del valor medio del cálculo integral se concluye de (33) que

$$F(b) - F(a) = (b - a)f(\xi).$$

Puesto que F es una primitiva de f, puede reemplazarse $f(\xi)$ por $F'(\xi)$ y obtenerse el teorema del valor medio del cálculo diferencial para la función F. Por supuesto, el requisito de que F posea derivada continua es más fuerte que el requisito del teorema del valor medio del cálculo diferencial, o sea, que la derivada simplemente exista.

d. Ejemplos

En el capítulo 3 se hará un uso extensivo del teorema fundamental al evaluar integrales. Por el momento se ilustra el método que está basado en la utilización de la fórmula

$$\int_a^b \frac{dF(x)}{dx} = F(b) - F(a)$$

mediante algunos ejemplos.

En la p. 185 se dedujo la fórmula

$$\frac{d}{dx}x^n = nx^{n-1}$$

para enteros positivos n. Esta fórmula es, en realidad, una consecuencia trivial del *teorema del binomio,* puesto que

$$\frac{d}{dx}x^n = \lim_{h\to 0}\frac{1}{h}[(x+h)^n - x^n]$$
$$= \lim_{h\to 0}\frac{1}{h}\left(x^n + nhx^{n-1} + \frac{n(n-1)}{2}h^2x^{n-2} + \cdots + h^n - x^n\right)$$
$$= \lim_{h\to 0}\left(nx^{n-1} + \frac{n(n-1)}{2}hx^{n-2} + \cdots + h^{n-1}\right) = nx^{n-1}.$$

Integrando entre los límites a y b se encuentra que

$$\int_a^b nx^{n-1}\,dx = b^n - a^n.$$

Escribiendo m para $n-1$ se obtiene la fórmula

$$\int_b^a x^m\,dx = \frac{1}{m+1}(b^{m+1} - a^{m+1})$$

para enteros $m \geq 0$. Esta deducción de la expresión para la integral de x^m es mucho más sencilla que la dada en la p. 153, la cual estuvo basada en una subdivisión geométrica del intervalo $[a, b]$; además, el resultado es ahora, en verdad, más general puesto que podemos excluir de la hipótesis el que a y b sean positivos.

Las fórmulas

$$\frac{d\,\text{sen}\,x}{dx} = \cos x, \qquad \frac{d\cos x}{dx} = -\text{sen}\,x$$

fueron obtenidas en la p. 186 aplicando los teoremas de adición para funciones trigonométricas y utilizando $\lim_{h\to 0}\left(\frac{\text{sen}\,h}{h}\right) = 1$. Integrando se obtiene inmediatamente

$$\int_a^b \cos x\,dx = \text{sen}\,b - \text{sen}\,a, \qquad \int_a^b \text{sen}\,x\,dx = \cos a - \cos b.$$

Nuevamente, esta deducción de las fórmulas de integración a partir del teorema fundamental es más sencilla que aquélla basada en la definición de la integral definida considerando ésta como límite de una suma.

Suplemento. La existencia de la integral definida de una función continua

Se tiene todavía que demostrar el hecho de que la integral de una función $f(x)$ entre los límites a y b $(a < b)$ existe siempre que $f(x)$ sea continua en el intervalo cerrado $[a, b]$. La demostración estará basada principalmente en la continuidad uniforme de $f(x)$ (véase p. 65): para cualquier ε positivo dado los valores de f en dos puntos cualesquiera ξ y η del intervalo difieren en menos que ε si ξ y η están suficientemente próximos entre sí; y el grado de proximidad depende solamente de ε y es independiente de ξ, η; en otras palabras, existe un módulo uniforme de continuidad $\delta(\varepsilon)$ tal que $|f(\xi) - f(\eta)| < \varepsilon$ para cualesquiera valores ξ, η en $[a, b]$ para los cuales $|\xi - \eta| < \delta$.

La definición de integral como un límite de sumas requiere que se subdivida el intervalo $[a, b]$ en n partes mediante puntos sucesivos $x_0, x_1, \ldots, x_n$, donde $x_0 = a$, $x_n = b$ y $x_0 < x_1 < \cdots < x_n$. Sea S_n un nombre para una particular subdivisión de $[a, b]$, de este tipo, en n celdas. La *fineza* de la subdivisión será medida mediante la longitud de la mayor de las celdas resultantes, esto es, el máximo de las cantidades $\Delta x_i = x_i - x_{i-1}$, la cual se denominará la "norma" de S_n. Debido a la continuidad uniforme de f, los valores de f en dos puntos cualesquiera de la misma celda difieren en menos que ε tan pronto como la norma de S_n sea menor que $\delta = \delta(\varepsilon)$. Una suma de aproximación basada en la subdivisión S_n se obtiene eligiendo un valor ξ_i en cada celda $[x_{i-1}, x_i]$ y formando

$$F_n = \sum_{i=1}^{n} f(\xi_i)\Delta x_i.$$

Debemos probar que para la sucesión de subdivisiones S_n con norma tendiendo a cero, las sumas F_n convergen hacia un límite, el que se denotará por $\int_a^b f(x)\,dx$, y que el valor de este límite no depende de la particular elección de las subdivisiones y de puntos intermedios ξ_i. Para llevar a cabo la demostración se comparan primero los valores F_n y F_N pertenecientes a dos subdivisiones S_n y S_N, donde la norma de S_n es menor que δ y donde la subdivisión S_N es un "refinamiento" de S_n; esto es, todos los puntos de subdivisión de S_n aparecen entre aquéllos de S_N. En una notación modificada apropiadamente se tiene

$$F_N = \sum_{j=1}^{N} f(\eta_j)\Delta y_j,$$

donde los valores y_j son los puntos de subdivisión de S_N, donde $\Delta y_j = y_j - y_{j-1}$, y η_j está situado en el intervalo $[y_{j-1}, y_j]$. Dos puntos de subdivisión sucesivos x_{i-1} y x_i de S_n aparecen también entre los valores y_j, por ejemplo, $x_{i-1} = y_{r-1}$, $x_i = y_s$. En S_N la celda $[x_{i-1}, x_i]$ es cortada en intervalos, digamos, $[y_{r-1}, y_r]$, $[y_r, y_{r+1}]$, $\ldots$, $[y_{s-1}, y_s]$, siendo

$$\sum_{j=r}^{s} f(\eta_j)\,(y_j - y_{j-1})$$

la contribución total a F_N. Se compara ésta con la contribución a F_n de la celda $[x_{i-1}, x_i]$, dada por $f(\xi_i)\,(x_i - x_{i-1})$, la cual puede ser escrita como

$$\sum_{j=r}^{s} f(\xi_i)\,(y_j - y_{j-1})$$

(véase Fig. 2.33), y se encuentra para el valor absoluto de la diferencia de las contribuciones:

$$\left|\sum_{j=r}^{s} [f(\eta_j) - f(\xi_i)](y_j - y_{j-1})\right| \leq \sum_{j=r}^{s} \varepsilon \cdot (y_j - y_{j-1}) = \varepsilon(x_i - x_{i-1}).$$

Por lo tanto, sumando las diferencias de las contribuciones a F_n y F_N para todas las celdas $[x_{i-1}, x_i]$ de S_n, se encuentra la estimación

$$|F_N - F_n| \leq \sum_{i=1}^{n} \varepsilon(x_i - x_{i-1}) = \varepsilon(b - a),$$

siempre que S_n posea una norma menor que $\delta(\varepsilon)$ y S_N sea un refinamiento de S_n.

Si ahora S_n y S_m son dos subdivisiones cualesquiera, puede considerarse la subdivisión S_N formada por todos los puntos de subdivisión de S_n junto con los de S_m. Entonces, S_N será un refinamiento tanto de S_n como de S_m.

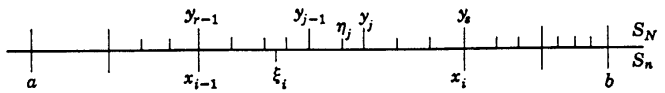


Figura 2.33.

Supóngase que S_n y S_m poseen norma menor que $\delta(\varepsilon)$. Eligiendo cualesquiera puntos intermedios η_j de las celdas de S_N para definir F_N, se encuentra

$$|F_n - F_m| = |(F_n - F_N) + (F_N - F_m)| \leq |F_n - F_N| + |F_m - F_N| \leq 2\varepsilon(b - a).$$

Se ve entonces que dos sumas de aproximación cualesquiera difieren arbitrariamente poco una de la otra, si las normas de las correspondientes subdivisiones son suficientemente pequeñas. Considérese ahora cualquier sucesión de subdivisiones S_n cuyas normas tiendan a cero para $n \to \infty$. Sean F_n las correspondientes sumas de aproximación. Para cualquier $\varepsilon > 0$ la norma de S_n es menor que $\delta(\varepsilon)$ para todo n suficientemente grande,. Por lo tanto

$$|F_n - F_m| < 2\varepsilon(b - a)$$

para n y m suficientemente grandes. Se sigue que la sucesión F_n satisface el criterio de convergencia de Cauchy (véase p. 120); consecuentemente,

$$\lim_{n\to\infty} F_n = F$$

existe.

Queda por probar que el valor de $\lim_{n\to\infty} F_n$ no depende de las particulares subdivisiones y de los puntos intermedios. Si S_n' denota cualquier otra sucesión de subdivisiones cuyas normas tienden a cero, entonces la suma correspondiente F_n' posee un límite F'. Puesto que

$$|F_n' - F_n| < 2\varepsilon(b - a)$$

siempre que las normas de S_n y S_n' sean menores que $\delta(\varepsilon)$, se encuentra para $n \to \infty$ que también $|F - F'| \leq 2\varepsilon(b - a)$. Puesto que aquí ε es un número positivo arbitrario, se sigue que $F' = F$. Por lo tanto, el límite F, que se denota por $\int_a^b f(x)\,dx$, está determinado de manera única.

La demostración de la existencia de la integral definida de una función continua queda así completamente terminada.

Sumas de aproximación más generales. Nuestra demostración indica más claramente lo que es esencial en la aproximación de una integral mediante una suma. Se hace evidente el hecho de que se podría formular un proceso de límite algo más general que conduzca también a la integral, y que la siguiente forma, más general, del teorema es cierta: f_i no tiene que ser un valor de la función para que las sumas $F_n = \Sigma f_i \Delta x_i$ converjan a la integral; en vez de esto es suficiente que $|f_i - f(\xi_i)| < \delta(\varepsilon)$ para algún punto ξ_i en el intervalo $[x_{i-1}, x_i]$, donde $\delta(\varepsilon) \to 0$ para $\varepsilon \to 0$.

Este aserto general es útil a menudo. Si, por ejemplo, $f(x) = \phi(x)\psi(x)$, entonces en vez de la suma $\Sigma f(\xi_\nu)\,\Delta x_\nu$, puede considerarse la suma más general

$$\Sigma \phi(\xi_\nu')\psi(\xi_\nu'')\,\Delta x_\nu,$$

donde ξ_ν' y ξ_ν'' son dos puntos de la celda no necesariamente coincidentes. Esta suma también tiende a la integral

$$\int_a^b f(x)\,dx = \int_a^b \phi(x)\psi(x)\,dx$$

conforme n crece, siempre que la longitud de la máxima celda tienda a cero.

Una propiedad similar vale para otras sumas formadas de manera análoga; por ejemplo, la suma

$$\sum_{\nu=1}^{n} \sqrt{\phi(\xi_\nu')^2 + \psi(\xi_\nu'')^2}\,\Delta x_\nu$$

tiende a la integral

$$\int_a^b \sqrt{\phi(x)^2 + \psi(x)^2}\,dx.$$

Para probar estas propiedades se tiene solamente que probar que la variación D en las sumas de aproximación debida a la desviación de ξ_ν'' con respecto a ξ_ν' tiende a cero en el límite. Esto es obvio en el primer ejemplo, donde la variación en la suma de aproximación es

$$D = \sum_{\nu=1}^{n} \phi(\xi_\nu')[\psi(\xi_\nu'') - \psi(\xi_\nu')]\,\Delta x_\nu.$$

Puesto que ϕ es acotada y ψ uniformemente continua, D puede hacerse arbitrariamente pequeña eligiendo celdas suficientemente pequeñas.

La variación en la segunda suma está representada por

$$D = \sum_{\nu=1}^{n} (\sqrt{\phi(\xi_\nu')^2 + \psi(\xi_\nu'')^2} - \sqrt{\phi(\xi_\nu')^2 + \psi(\xi_\nu')^2})\,\Delta x_\nu.$$

Utilizando la desigualdad del triángulo aplicada al triángulo con vértices $(a, 0)$, $(0, b)$, $(0, c)$ en la forma $|\sqrt{a^2 + b^2} - \sqrt{a^2 + c^2}| \leq |b - c|$, se encuentra que

$$|\sqrt{\phi(\xi_\nu')^2 + \psi(\xi_\nu'')^2} - \sqrt{\phi(\xi_\nu')^2 + \psi(\xi_\nu')^2}| \leq |\psi(\xi_\nu'') - \psi(\xi_\nu')|$$

de la que se sigue inmediatamente que D tiende a cero.

PROBLEMAS

SECCION 2.1, página 142

1. Sea f una función monótona positiva definida en $[a, b]$, donde $0 < a < b$. Sea ϕ la inversa de f y sean $\alpha = f(a)$, $\beta = f(b)$. Utilizando la interpretación de la integral como un área, mostrar que

$$\int_\alpha^\beta \phi(y)\,dy = b\beta - a\alpha - \int_a^b f(x)\,dx.$$

SECCION 2.2, página 150

1. Probar para cualquier número natural p que

$$\int_a^b x^p\,dx = \frac{1}{p+1}(b^{p+1} - a^{p+1}),$$

utilizando una subdivisión de $[a, b]$ en celdas de igual longitud. Emplear las técnicas del capítulo 1, problemas diversos 5 a 12, para evaluar las sumas de aproximación F_n.

2. Deducir la fórmula para $\int_a^b x^\alpha\,dx$, $a, b > 0$, donde α es racional y negativo; por ejemplo, $\alpha = -r/s$, siendo r y s números naturales. *(Sugerencia:* Hágase $q^{-1/s} = r$, donde $q = \sqrt[s]{b/a}$.)

3. Por el método utilizado para encontrar la integral de $\operatorname{sen} x$, deducir la fórmula

$$\int_a^b \cos x\,dx = \operatorname{sen} b - \operatorname{sen} a.$$

4. Hacer una aseveración general relativa a $\int_{-a}^{a} f(x)\,dx$ cuando $f(x)$ es: (a) una función impar, y (b) una función par.

5. Calcular $\int_0^{\pi/2} \operatorname{sen} x\,dx$ y $\int_0^{\pi/2} \cos x\,dx$. Explicar con argumentos geométricos por qué éstas deben ser iguales. Además, explicar por qué

$$\int_a^{a+2\pi} \operatorname{sen} x\,dx = \int_b^{b+2\pi} \cos x\,dx$$

para todos los valores de a y b.

6. (a) Evaluar $I_n = \int_0^a x^{1/n}\,dx$. ¿Qué es $\lim_{n\to\infty} I_n$? Interpretar geométricamente.

(b) Hacer lo mismo para $I_n = \int_0^a x^n\,dx$.

7. Evaluar

$$\lim_{n\to\infty} \frac{1}{\sqrt{n}}\left(1 + \frac{1}{\sqrt{2}} + \cdots + \frac{1}{\sqrt{n}}\right).$$

SECCION 2.3, página 158

***1.** *Desigualdad de Cauchy para integrales.* Probar que para todas las funciones continuas $f(x)$, $g(x)$, es

$$\int_a^b [f(x)]^2\,dx \int_a^b [g(x)]^2\,dx \geq \left(\int_a^b f(x)g(x)\,dx\right)^2.$$

***2.** Probar que si $f(x)$ es continua y

$$f(x) = \int_0^x f(t)\,dt,$$

entonces $f(x)$ es idénticamente cero.

***3.** Sea $f(x)$ continua en el sentido de Lipschitz en $[0, 1]$; esto es

$$|f(x) - f(y)| < M|x - y|$$

para todo x, y en el intervalo. Probar que

$$\left|\int_0^1 f(x)\,dx - \frac{1}{n}\sum_{k=1}^{n} f\left(\frac{k}{n}\right)\right| < \frac{M}{2n}$$

SECCION 2.5, página 167

1. Probar que

$$\log\frac{p}{q} \leq \frac{p-q}{\sqrt{pq}} \qquad (q \leq p).$$

(Sugerencia: Aplicar la desigualdad de *Cauchy,* Problema 1.)

2. (a) Verificar que $\log(1 + x) = \int_0^x \frac{1}{1+u}\,du$, donde $x > -1$.
(b) Mostrar para $x > 0$ que

$$x - \frac{x^2}{2} < \log(1 + x) < x.$$

*(*c*) Más generalmente, mostrar para $0 < x < 1$ que

$$x - \frac{x^2}{2} + \frac{x^3}{3} - \cdots - \frac{x^{2n}}{2n} < \log(1+x) < x - \frac{x^2}{2} + \frac{x^3}{3} - \cdots + \frac{x^{2n+1}}{2n+1}.$$

(Sugerencia: Comparar $1/(1+u)$ con una progresión geométrica.)

SECCION 2.6, página 171

1. (*a*) Probar que

$$\int_a^b e^x\,dx = e^b - e^a$$

utilizando una subdivisión de $[a, b]$ en celdas iguales. *Sugerencia:* Aplicar también $\log \alpha = \lim_{n\to\infty} n(\sqrt[n]{\alpha} - 1)$.

(*b*) Encontrar $\int_a^b \log x\,dx$. (Véase sección 2.1, problema 1.)

(*c*) Mostrar para $x \geq 0$ que

$$1 + x + \frac{x^2}{2!} + \cdots + \frac{x^n}{n!} \leq e^x \leq 1 + x + \frac{x^2}{2!} + \cdots + \frac{x^n}{n!} + \frac{e^x x^{n+1}}{(n+1)!}.$$

(Sugerencia: Obtener estimaciones superiores e inferiores para $\int_0^x e^u\,du$ e integrar repetidamente.)

Obtener estimaciones del mismo tipo para e^x cuando $x < 0$.

SECCION 2.8c, página 184

Calcular las derivadas de las siguientes funciones dondequiera que estén definidas, directamente como el límite de sus cocientes de incrementos.

1. $\tan x$.

2. $\sec^2 x$.

3. $\operatorname{sen} \sqrt{x}$.

4. $\sqrt{\operatorname{sen} x}$.

5. $\dfrac{1}{\operatorname{sen} x}$

6. $\operatorname{sen} \dfrac{1}{x}$

7. x^α, donde α es racional y negativo.

SECCION 2.8i, página 196

1. Mostrar que $x > \operatorname{sen} x$ para x positivo y $x < \tan x$ para x en $\left(0, \frac{\pi}{2}\right)$.

2. Si $f(x)$ es continua y derivable para $a \leq x \leq b$, mostrar que si $f'(x) \leq 0$ para $a \leq x < \xi$ y $f'(x) \geq 0$ para $\xi < x \leq b$, la función nunca es menor que $f(\xi)$.

***3.** Si la función continua $f(x)$ posee derivada $f'(x)$ en cada punto x de una vecindad de $x = \xi$, y si $f'(x)$ se aproxima a un límite L conforme $x \to \xi$, entonces $f'(\xi)$ existe y es igual a L.

***4.** Sea $f(x)$ definida y derivable en todo el eje x. Mostrar que si $f(0) = 0$ y en todas partes es $|f'(x)| \leq |f(x)|$, entonces $f(x) = 0$ idénticamente.

SECCION 2.9, página 206

***1.** Si una partícula recorre una distancia 1 en un tiempo 1, partiendo del reposo y terminando también en reposo, entonces en algún punto del intervalo deberá haber estado sujeta a una aceleración igual a 4 o más.

SUPLEMENTO, página 213. Existencia de la Integral Indefinida.

1. Sea $f(x)$ definida y acotada en $[a, b]$. Se definen la suma superior Σ y la suma inferior σ para la subdivisión

$$a = x_0 < x_1 < x_2 \cdots < x_n = b$$

como

$$\Sigma = \sum_{i=1}^{n} M_i\,\Delta x_i, \qquad \sigma = \sum_{i=1}^{n} m_i\,\Delta x_i,$$

donde M_i es el extremo superior y m_i es el extremo inferior de $f(x)$ en la celda $[x_{i-1}, x_i]$.

(*a*) Mostrar que en cualquier refinamiento de una subdivisión la suma superior decrece o bien permanece invariable; y que, análogamente, la suma inferior crece o bien permanece invariable.

(*b*) Probar que cada suma superior es mayor que o igual a cada suma inferior.

(*c*) La *integral superior de Darboux* F^+ está definida como el extremo inferior de las sumas superiores, y la *integral inferior de Darboux* F^-, como el extremo superior de las sumas inferiores, sobre todas las subdivisiones. De (*b*), $F^+ \geq F^-$. Si $F^+ = F^-$, al valor común se le denomina *integral de Darboux* de f. Probar que la integral de Darboux de f es en realidad la integral de Riemann ordinaria; además, mostrar que la integral de Riemann existe si y sólo si las integrales superior e inferior de Darboux existen y son iguales.

2. Sea $f(x)$ una función monótona definida en $[a, b]$.

(*a*) Mostrar que la diferencia entre las sumas superior e inferior para una subdivisión en n celdas iguales está dada exactamente por

$$\Sigma - \sigma = |f(b) - f(a)|\,(b-a)/n.$$

y explicar este resultado geométricamente.

(*b*) Usar el resultado de (*a*) para probar que la integral de Darboux existe.

(*c*) Estimar $\Sigma - \sigma$ en términos de $f(a)$, $f(b)$ y la amplitud de la subdivisión si las celdas de la subdivisión pueden ser desiguales.

(*d*) En la mayor parte de los casos $f(x)$, si no es monótona, puede escribirse como la suma de funciones monótonas, $f(x) = \phi(x) + \psi(x)$, donde ϕ es no creciente y ψ no decreciente. Estimar las diferencias entre las sumas superior e inferior en ese caso.

3. Mostrar que si $f(x)$ posee derivada continua en el intervalo cerrado $[a, b]$, entonces $f(x)$ puede escribirse como la suma de funciones monótonas como en el problema 2*d*.

PROBLEMAS DIVERSOS

1. Probar que

(*a*) $\displaystyle\int_1^1 (x^2-1)^2\,dx = \frac{16}{15}$; (*b*) $\displaystyle(-1)^n \int_{-1}^{1} (x^2-1)^n\,dx = \frac{2^{2n+1}(n!)^2}{(2n+1)!}$

2. Probar para el coeficiente binomial $\binom{n}{k}$ que

$$\binom{n}{k} = \left[(n+1) \int_0^1 x^k (1-x)^{n-k}\, dx \right]^{-1}.$$

*3. Si $f(x)$ posee una derivada $f'(x)$ (no necesariamente continua) en cada punto x de $a \leq x \leq b$, y si $f'(x)$ toma los valores m y M, también toma cualquier valor μ entre m y M.

4. Si $f''(x) \geq 0$ para todos los valores de x en $a \leq x \leq b$, la gráfica de $y = f(x)$ está situada sobre o por encima de la línea tangente en cualquier punto $x = \xi$, $y = f(\xi)$ de la gráfica.

5. Si $f''(x) \geq 0$ para todos los valores de x en $u \leq x \leq b$, la gráfica de $y = f(x)$ en el intervalo $x_1 \leq x \leq x_2$ está situada por debajo del segmento de recta que une los dos puntos de la gráfica para los cuales $x = x_1$, $x = x_2$.

6. Si $f''(x) \geq 0$, entonces $f\left(\dfrac{x_1 + x_2}{2}\right) \leq \dfrac{f(x_1) + f(x_2)}{2}$.

*7. Sea $f(x)$ una función tal que $f''(x) \geq 0$ para todos los valores de x y sea $u = u(t)$ una función continua arbitraria. Entonces

$$\frac{1}{a}\int_0^a f[u(t)]\, dt \geq f\left(\frac{1}{a}\int_0^a u(t)\, dt\right).$$

8. (*a*) Derivar directamente y escribir abajo las fórmulas correspondientes de integración: (i) $x^{1/2}$; (ii) $\tan x$.

(*b*) Evaluar

$$\lim_{n\to\infty} \frac{1}{n}\left(1 + \sec^2\frac{\pi}{4n} + \sec^2\frac{2\pi}{4n} + \cdots + \sec^2\frac{n\pi}{4n}\right).$$

9. Sea $f(x)$ una función que posee derivadas primera y segunda para todos los valores reales de x. Probar que si $f(x)$ es positiva en todas partes y cóncava, entonces $f(x)$ es constante.

3

Las técnicas del cálculo

Parte A Derivación e integración de las funciones elementales

3.1 Las reglas más simples para derivar, y sus aplicaciones

Aún cuando los problemas de integración son usualmente de mayor importancia que los de diferenciación, estos últimos ofrecen menor dificultad formal que los primeros. Por ello es natural el procedimiento de conocer a fondo el arte de derivar la más amplia clase posible de funciones; entonces, por el teorema fundamental (sección 2.9), los resultados de la derivación estarán disponibles para la evaluación de integrales. En las siguientes secciones continuaremos tales aplicaciones del teorema fundamental. Hasta cierto punto se hará un nuevo comienzo y se desarrollarán las técnicas de integración sistemáticamente sobre la base de ciertas reglas generales para la diferenciación.

a. Reglas de la derivación

Se supone que en el intervalo bajo consideración las funciones $f(x)$ y $g(x)$ son derivables; entonces las siguientes reglas son básicas.

Regla 1. Multiplicación por una constante. Para cualquier constante c, la función $\phi(x) = cf(x)$ es derivable, y es

$$\phi'(x) = cf'(x) \tag{1}$$

La demostración, trivial, fue dada en el capítulo 2, p. 187.

Regla 2. Derivada de una suma. Si $\phi(x) = f(x) + g(x)$, entonces $\phi(x)$ es derivable y es

$$\phi'(x) = f'(x) + g'(x); \tag{2}$$

esto es, las operaciones de derivación y adición son intercambiables. Lo mismo se cumple para la suma de un número finito n de funciones derivables,

$$\phi(x) = \sum_{\nu=1}^{n} f_\nu(x),$$

para la cual se obtiene

$$\phi'(x) = \sum_{\nu=1}^{n} f_\nu'(x).$$

La demostración es obvia a partir de la definición de derivada.

Regla 3. Derivada de un producto. Si $\phi(x) = f(x)g(x)$, entonces $\phi(x)$ es derivable y es

$$\phi'(x) = f(x)g'(x) + g(x)f'(x). \tag{3}$$

La demostración se sigue a partir de la ecuación

$$\frac{\phi(x+h) - \phi(x)}{h} = \frac{f(x+h)g(x+h) - f(x)g(x)}{h}$$

$$= f(x+h)\frac{g(x+h) - g(x)}{h} + g(x)\frac{f(x+h) - f(x)}{h}.$$

Tomando el límite en esta expresión para $h \to 0$ se obtiene la Ec. (3).

Esta fórmula se vuelve más clara si se divide[1] por $\phi(x) = f(x)g(x)$. Se obtiene entonces

$$\frac{\phi'(x)}{\phi(x)} = \frac{f'(x)}{f(x)} + \frac{g'(x)}{g(x)}.$$

Utilizando la notación de diferenciales (capítulo 2, p. 200), puede también reescribirse la Ec. (3) como

$$d(fg) = f\,dg + g\,df.$$

Por inducción se obtiene para la derivada de un producto de n factores una expresión que consiste de n términos, cada uno de los cuales consta de la derivada de un factor multiplicada por todos los otros factores del producto original:

$$\begin{aligned}\phi'(x) &= \frac{d}{dx}[f_1(x)f_2(x)\cdots f_n(x)] \\ &= f_1'(x)f_2(x)\cdots f_n(x) + f_1(x)f_2'(x)f_3(x)\cdots f_n(x) \\ &\quad + \cdots + f_1(x)f_2(x)\cdots f_n'(x) \\ &= \sum_{\nu=1}^{n} f_\nu'(x)\frac{\phi(x)}{f_\nu(x)},\end{aligned}$$

[1] Debe, por supuesto, suponerse que $\phi(x)$ en ninguna parte es igual a cero.

o bien, dividiendo por $\phi(x) = f_1(x)f_2(x)\cdots f_n(x)$,

$$\frac{\phi'(x)}{\phi(x)} = \frac{f_1'(x)}{f_1(x)} + \frac{f_2'(x)}{f_2(x)} + \cdots + \frac{f_n'(x)}{f_n(x)} = \sum_{\nu=1}^{n} \frac{f_\nu'(x)}{f_\nu(x)},$$

la cual es válida donde se tenga $\phi(x) \neq 0$.

Mediante la aplicación repetida de la regla para la derivada de un producto pueden obtenerse también *fórmulas para la segunda derivada y derivadas de órdenes superiores*. Para la segunda derivada se tiene

$$\begin{aligned}\frac{d^2fg}{dx^2} &= \frac{d}{dx}\left(\frac{dfg}{dx}\right) = \frac{d}{dx}\left(f\frac{dg}{dx} + \frac{df}{dx}g\right)\\ &= \frac{d}{dx}\left(f\frac{dg}{dx}\right) + \frac{d}{dx}\left(\frac{df}{dx}g\right)\\ &= f\frac{d^2g}{dx^2} + 2\frac{df}{dx}\frac{dg}{dx} + \frac{d^2f}{dx^2}g.\end{aligned}$$

Regla de Leibnitz. El lector deberá probar por inducción que la n-ésima derivada de un producto puede encontrarse de acuerdo con la siguiente regla (regla de Leibnitz):

$$\begin{aligned}\frac{d^n}{dx^n}(fg) = f\frac{d^ng}{dx^n} &+ \binom{n}{1}\frac{df}{dx}\frac{d^{n-1}g}{dx^{n-1}}\\ &+ \binom{n}{2}\frac{d^2f}{dx^2}\frac{d^{n-2}g}{dx^{n-2}} + \cdots + \binom{n}{n-1}\frac{d^{n-1}f}{dx^{n-1}}\frac{dg}{dx} + \frac{d^nf}{dx^n}g.\end{aligned}$$

Aquí $\binom{n}{1} = n$, $\binom{n}{2} = [n(n-1)]/2!$, etc., denotan los coeficientes binomiales.

Regla 4. Derivada de un cociente. Para un cociente

$$\phi(x) = \frac{f(x)}{g(x)}$$

se cumple la siguiente regla: La función $\phi(x)$ es derivable en cada punto en el cual $g(x)$ no se hace cero, y

$$\phi'(x) = \frac{g(x)f'(x) - g'(x)f(x)}{[g(x)]^2} \tag{4}$$

Si $\phi(x) \neq 0$, esto puede ser escrito como

$$\frac{\phi'(x)}{\phi(x)} = \frac{f'(x)}{f(x)} - \frac{g'(x)}{g(x)}$$

DEMOSTRACIÓN. Si se supone la derivabilidad de $\phi(x)$, puede aplicarse la regla de un producto a $f(x) = \phi(x)g(x)$ y concluir que

$$f'(x) = \phi(x)g'(x) + g(x)\phi'(x).$$

Mediante la substitución de $f(x)/g(x)$ en vez de $\phi(x)$ en la derecha y despejando $\phi'(x)$, se obtiene la regla 4.

Puede probarse la derivabilidad de $\phi(x)$ así como la regla, si se escribe

$$\frac{\phi(x+h) - \phi(x)}{h} = \frac{\dfrac{f(x+h)}{g(x+h)} - \dfrac{f(x)}{g(x)}}{h}$$

$$= \frac{g(x)\dfrac{f(x+h) - f(x)}{h} - \dfrac{g(x+h) - g(x)}{h}f(x)}{g(x)g(x+h)}.$$

Si ahora se hace h tender a cero se obtiene el resultado establecido, pues, por hipótesis, el denominador no tiende a cero sino al límite $[g(x)]^2$, y los dos términos del numerador poseen límites $g(x)f'(x)$ y $g'(x)f(x)$ respectivamente. Esto prueba tanto la existencia del límite en el miembro izquierdo como la fórmula de derivación.

b. Derivación de las funciones racionales

Primero, se deduce una vez más la fórmula

$$\frac{d}{dx}x^n = nx^{n-1}$$

para todo entero positivo n, invocando la regla para derivar un producto. Considérese x^n como un producto de n factores, $x^n = x\cdots x$, y obténgase así

$$\frac{d}{dx}x^n = 1\cdot x^{n-1} + 1\cdot x^{n-1} + \cdots + 1\cdot x^{n-1} = nx^{n-1}.$$

La segunda derivada de la función x^n se sigue a partir de esta fórmula y la Ec. (1):

$$\frac{d^2}{dx^2}x^n = n(n-1)x^{n-2}.$$

Continuando, se obtienen las derivadas de órdenes superiores

$$\frac{d^3}{dx^3}x^n = n(n-1)(n-2)x^{n-3}$$

. . . . . .

$$\frac{d^n}{dx^n}x^n = 1\cdot 2\cdots n = n!.$$

A partir de la última de estas fórmulas es claro que la n-ésima derivada de x^n es una constante, mientras que la $(n+1)$-ésima derivada se hace cero en todas partes.

Utilizando las dos primeras reglas y la regla para derivar potencias puede derivarse cualquier polinomio $y = a_0 + a_1x + a_2x^2 + \cdots + a_nx^n$, obteniéndose

$$y' = a_1 + 2a_2x + 3a_3x^2 + \cdots + na_nx^{n-1};$$

además,

$$y'' = 2a_2 + 3 \cdot 2a_3x + 4 \cdot 3a_4x^2 + \cdots + n(n-1)a_nx^{n-2},$$

y así sucesivamente.

La derivada de cualquier función racional puede encontrarse ahora con la ayuda de la regla del cociente. En particular, se deduce nuevamente la fórmula de derivación para la función x^n, donde $n = -m$ es un entero negativo. La aplicación de la regla del cociente, junto con el hecho de que la derivada de una constante es igual a cero, da el resultado

$$(d/dx)(1/x^m) = -mx^{m-1}/x^{2m} = -m/x^{m+1};$$

o bien, si se toma $m = -n$,

$$\frac{d}{dx}x^n = nx^{n-1},$$

que coincide formalmente con el resultado para valores positivos de n y con los resultados dados anteriormente (p. 186).

c. Derivación de las funciones trigonométricas

Para las funciones trigonométricas $\operatorname{sen} x$ y $\cos x$ se han obtenida ya (p. 186) las fórmulas de derivación

$$\frac{d}{dx}\operatorname{sen} x = \cos x \qquad \text{y} \qquad \frac{d}{dx}\cos x = -\operatorname{sen} x.$$

La regla del cociente nos permite ahora derivar las funciones

$$y = \tan x = \frac{\operatorname{sen} x}{\cos x} \qquad \text{y} \qquad y = \cot x = \frac{\cos x}{\operatorname{sen} x}.$$

De acuerdo con la regla, la derivada de la primera de estas funciones es

$$y' = \frac{\cos^2 x + \operatorname{sen}^2 x}{\cos^2 x} = \frac{1}{\cos^2 x},$$

de modo que

$$\frac{d}{dx}\tan x = \frac{1}{\cos^2 x} = \sec^2 x = 1 + \tan^2 x.$$

Análogamente, se obtiene

$$\frac{d}{dx}\cot x = -\frac{1}{\operatorname{sen}^2 x} = -\operatorname{cosec}^2 x = -(1 + \cot^2 x).$$

A las fórmulas de derivación para $\operatorname{sen} x$, $\cos x$, $\tan x$ y $\cot x$ corresponden las siguientes fórmulas de integración:

$$\int \cos x\,dx = \operatorname{sen} x, \qquad \int \operatorname{sen} x\,dx = -\cos x,$$

$$\int \frac{1}{\cos^2 x}\,dx = \tan x, \qquad \int \frac{1}{\operatorname{sen}^2 x}\,dx = -\cot x.$$

A partir de estas fórmulas se obtiene, por medio de la regla fundamental de la sección 2.9, p. 211, el valor de la integral definida, entre límites cualesquiera, siendo la única restricción el hecho de que cuando las dos últimas fórmulas son utilizadas el intervalo de integración no debe contener punto alguno de discontinuidad del integrando, tal como un múltiplo impar de $\pi/2$ en el primer caso y un múltiplo par de $\pi/2$ en segundo. Por ejemplo,

$$\int_a^b \cos x\,dx = \operatorname{sen} x\Big|_a^b = \operatorname{sen} b - \operatorname{sen} a.$$

3.2 La derivada de la función inversa

a. Fórmula general

Hemos visto en la p. 69 que una función continua $y = f(x)$ posee una inversa continua en todo intervalo en el cual aquélla es monótona. Precisamente:

Si $a \leq x \leq b$ es un intervalo en el cual la función continua $y = f(x)$ es monótona, y si $f(a) = \alpha$ y $f(b) = \beta$, entonces f posee una función inversa que es continua y monótona en el intervalo entre α y β.

Como se hizo notar en la p. 198, el signo de la derivada proporciona un criterio simple para ver cuándo una función es monótona y posee por consiguiente una inversa. Una función derivable es continua y será monótona creciente en un intervalo donde $f'(x)$ sea mayor que cero, y monótona decreciente en un intervalo en el cual $f'(x)$ sea en todas partes menor que cero.

Se caracterizará ahora la derivada de la función inversa probando el siguiente teorema.

TEOREMA. *Si la función $y = f(x)$ es derivable en el intervalo $a < x < b$ y si $f'(x) > 0$, o bien $f'(x) < 0$, en todo el intervalo, entonces la función*

inversa $x = \phi(y)$ *posee también una derivada en cada punto interior de su intervalo de definición: la derivada de* $y = f(x)$ *y la de su inversa* $x = \phi(y)$ *satisfacen la relación* $f'(x) \cdot \phi'(y) = 1$ *en valores* x, y *correspondientes.*

Esta relación puede también expresarse en la forma

$$\frac{dy}{dx} = \frac{1}{\dfrac{dx}{dy}}. \tag{5}$$

Esta última fórmula nuevamente ilustra lo adecuado de la notación de Leibnitz: el cociente simbólico dy/dx puede ser tratado en las fórmulas como si fuera realmente una fracción.

DEMOSTRACIÓN. La demostración del teorema es sencilla. Escribiendo la derivada como el límite de un cociente de incrementos, se tiene

$$y' = f'(x) = \lim_{\Delta x \to 0} \frac{\Delta y}{\Delta x} = \lim_{x_1 \to x} \frac{y_1 - y}{x_1 - x},$$

donde x e $y = f(x)$, y x_1 e $y_1 = f(x_1)$, respectivamente denotan parejas de valores correspondientes. Por hipótesis el primero de estos valores límite no es igual a cero. Debido a la continuidad de $y = f(x)$ y $x = \phi(y)$, las relaciones $y_1 \to y$ y $x_1 \to x$ son equivalentes. Por tanto, el valor límite

$$\lim_{x_1 \to x} \frac{x_1 - x}{y_1 - y} = \lim_{y_1 \to y} \frac{x_1 - x}{y_1 - y}$$

existe y es igual a $1/f'(x)$. Por otra parte, el valor límite en el miembro derecho es por definición la derivada $\phi'(y)$ de la función inversa $\phi(y)$, y nuestra fórmula queda probada.

El sencillo significado geométrico de la fórmula está mostrado claramente en la Fig. 3.1. La tangente a la curva $y = f(x)$, o bien $x = \phi(y)$, forma un ángulo α con el eje positivo x y un ángulo β con el eje positivo y; y, a partir de la interpretación geométrica de la derivada de una función como la pendiente de la tangente, se tiene

$$f'(x) \tan \alpha, \qquad \phi'(y) = \tan \beta.$$

Puesto que la suma de los ángulos α y β es $\pi/2$, $\tan \alpha \tan \beta = 1$, y esta relación es exactamente equivalente a nuestra fórmula de derivación.

Puntos críticos

Hasta ahora se ha supuesto expresamente que $f'(x) > 0$, o bien $f'(x) < 0$, esto es, que $f'(x)$ nunca es cero. ¿Qué sucede entonces si $f'(x) = 0$? Si $f'(x) = 0$ en todas partes de un intervalo, entonces f es

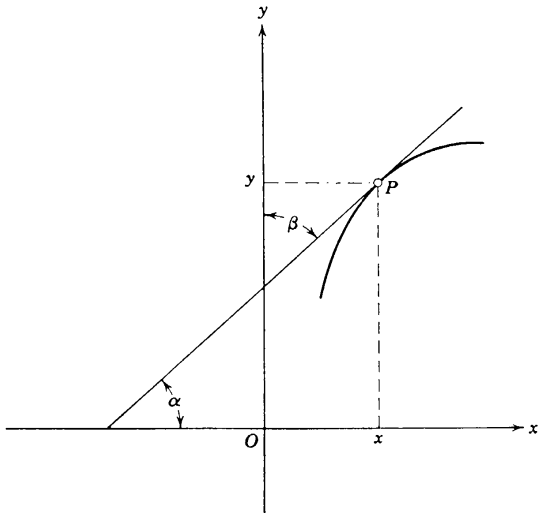


Figura 3.1 Derivación de la función inversa.

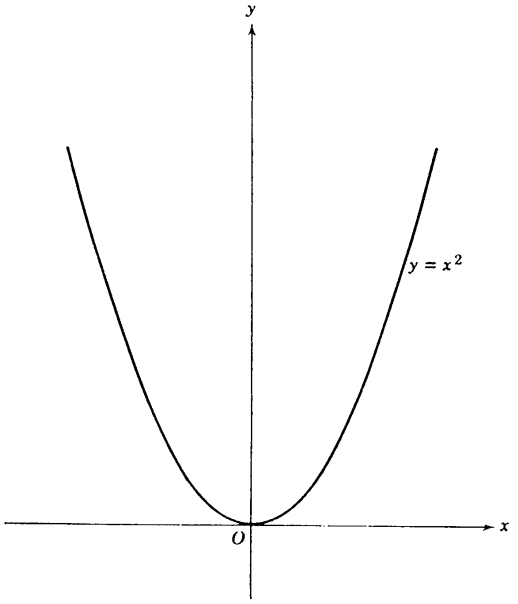


Figura 3.2 Parábola.

constante en él, y consecuentemente no posee inversa debido a que el mismo valor de y corresponde a todos los valores de x en el intervalo. Si $f'(x) = 0$ solamente en puntos "críticos" aislados (y si $f'(x)$ se supone continua), entonces se tienen dos casos, según que al pasar por estos puntos $f'(x)$ cambie o no de signo. En el primer caso este punto separa una región donde la función es monótona creciente de otra donde es monótona decreciente. En la vecindad de uno de tales puntos no puede haber función inversa univaluada. En el segundo caso, el que la derivada se haga cero no contradice el carácter monótono de la función $y = f(x)$, de modo que existe una función inversa univaluada. Sin embargo, la función inversa ya no sigue siendo derivable en el correspondiente punto; de hecho, la derivada es allí infinita. Las funciones $y = x^2$ e $y = x^3$ en el punto $x = 0$ ofrecen ejemplos de los dos tipos. La Fig. 3.2 y la Fig. 3.3 ilustran el comportamiento de las dos funciones al pasar por el origen, y al mismo tiempo muestran que la función $y = x^3$ posee una inversa univaluada, mientras que la otra función, $y = x^2$, no.

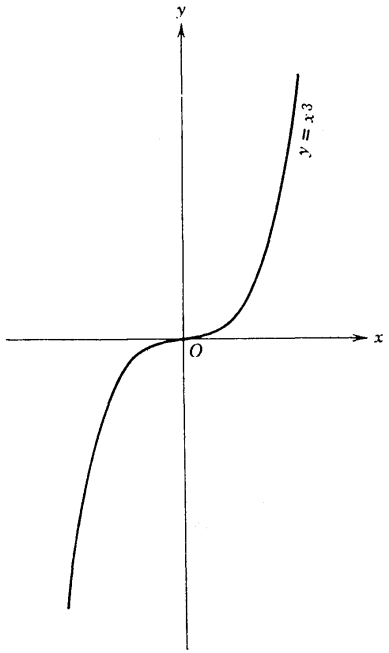


Figura 3.3 Parábola cúbica.

b. La inversa de la n-ésima potencia: la n-ésima raíz

El ejemplo más sencillo lo constituye la inversa de la función $y = x^n$ para enteros positivos n; primero se suponen valores positivos de x, por lo tanto también $y > 0$. Bajo estas condiciones y' es siempre positivo, de modo que para todos los valores positivos de y puede formarse la función inversa única

$$x = \sqrt[n]{y} = y^{1/n}.$$

La derivada de esta función inversa se obtiene inmediatamente por la regla general anterior como sigue:

$$\frac{d(y^{1/n})}{dy} = \frac{dx}{dy} = \frac{1}{dy/dx} = \frac{1}{nx^{n-1}} = \frac{1}{n}\frac{1}{y^{(n-1)/n}} = \frac{1}{n}y^{(1/n)-1}.$$

Si se cambia ahora la notación y se denota nuevamente la variable independiente por x, puede finalmente escribirse

$$\frac{d\sqrt[n]{x}}{dx} = \frac{d}{dx}(x^{1/n}) = \frac{1}{n}x^{(1/n)-1},$$

que coincide con el resultado obtenido en la p. 185.

Para $n > 1$, el punto $x = 0$ requiere una consideración especial. Si x se aproxima a cero a través de valores positivos, $d(x^{1/n})/dx$ crecerá obviamente más allá de toda cota; esto corresponde al hecho de que para $n > 1$ la derivada de la n-ésima potencia $f(x) = x^n$ se hace cero en el origen. Geométricamente esto significa que las curvas $y = x^{1/n}$ para $n > 1$ tocan el eje y en el origen (véase Fig. 1.35, p. 72).

Se deberá notar que para valores impares de n la suposición $x > 0$ puede omitirse y la función $y = x^n$ es monótona y posee una inversa sobre todo el dominio de números reales. La fórmula

$$\frac{d(\sqrt[n]{y})}{dy} = (1/n)y^{(1/n)-1}$$

se sigue cumpliendo para valores negativos de y, pero para $x = 0$, con $n > 1$, se tiene $d(x^n)/dx = 0$, que corresponde a una derivada infinita dx/dy de la función inversa en el punto $y = 0$.

c. Las funciones trigonométricas inversas—Multivalencia

Para formar las inversas de las funciones trigonométricas nuevamente se consideran las gráficas[1] de sen x, cos x, tan x, y cot x. Se ve de inmediato a partir de las Figs. 1.37, p. 74 y 1.38, p. 74, que para cada una de estas

[1] La representación gráfica ayudará al lector a superar las ligeras dificultades inherentes a la discusión de la "multivalencia" de las funciones inversas.

funciones es necesario seleccionar un intervalo definido si va a hablarse de una inversa única, pues las rectas $y = c$ paralelas al eje x cortan a las curvas en un número infinito de puntos.

El seno y el coseno inversos

Para la función $y = \operatorname{sen} x$ (Fig. 3.4), la derivada $y' = \cos x$ es positiva en el intervalo $-\pi/2 < x < \pi/2$. En este intervalo $y = \operatorname{sen} x$ posee una función inversa que denotamos por [1]

$$x = \operatorname{arc\,sen} y$$

(léase arco seno y; esto significa el ángulo cuyo seno tiene el valor y). Esta función crece monótonamente de $-\pi/2$ a $+\pi/2$ conforme y recorre el intervalo de -1 a $+1$. Si se desea destacar que estamos considerando la función inversa del seno en este intervalo particular, se habla del *valor principal* del arco seno. Para algún otro intervalo en el cual $\operatorname{sen} x$ es monótona, por ejemplo, el intervalo $+\pi/2 < x < 3\pi/2$, se obtiene otra inversa o "*rama*" del arco seno. Sin la especificación exacta del intervalo en el cual los valores de la función inversa deban estar situados, el símbolo arco seno no representa una función bien definida sino que, de hecho, denota un número *infinito* de valores.[2]

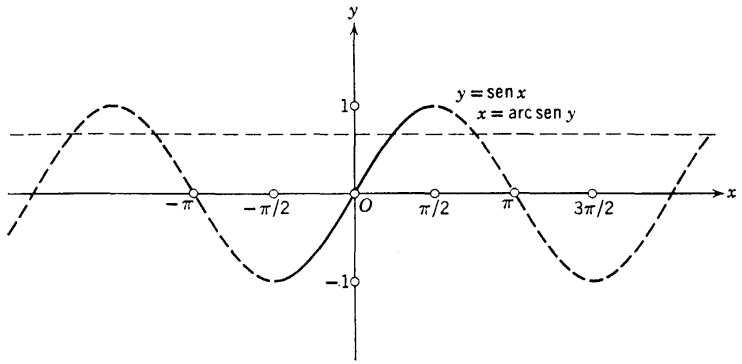


Figura 3.4 Gráfica de $y = \operatorname{sen} x$ (el valor principal está indicado por una curva llena).

La multivalencia de arc sen y está descripta por la propiedad: A cualquier valor y del seno le corresponde no sólo un ángulo x específico sino también cualquier ángulo de la forma $2k\pi + x$, o bien $(2k + 1)\pi - x$, donde k es cualquier entero (véase Fig. 3.4).

[1] La notación simbólica $x = sen^{-1} y$ es utilizada también cuando no existe peligro de confusión con la función *recíproca* $1/\operatorname{sen} x$.

[2] A veces llamada vagamente función *multivaluada*.

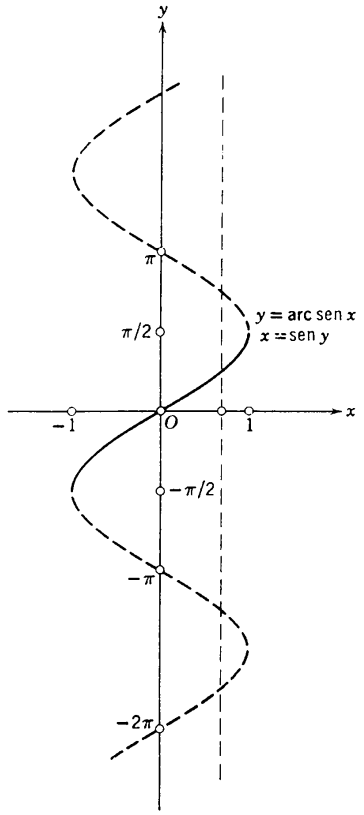


Figura 3.5 Gráfica de $y = \operatorname{arc\,sen} x$ (el valor principal está indicado por una curva llena).

La derivada de la función $x = \operatorname{arc\,sen} y$ se obtiene a partir de la Ec. (5) como sigue:

$$\frac{dx}{dy} = \frac{1}{y'} = \frac{1}{\cos x} = \frac{1}{\pm\sqrt{1 - \operatorname{sen}^2 x}} = \frac{1}{\pm\sqrt{1 - y^2}},$$

donde la raíz cuadrada ha de tomarse como positiva si nos restringimos al primer intervalo mencionado, esto es, $-\pi/2 < x < \pi/2$.[1]

Finalmente, cambiamos el nombre de la variable independiente, de y al comúnmente usado x (Fig. 3.5); entonces la derivada de arc sen x está expresada por

[1] Si, en vez de éste, se hubiese elegido el intervalo $\pi/2 < x < 3\pi/2$, correspondiente a la substitución de x por $x + \pi$, se habría tenido que utilizar la raíz cuadrada negativa, puesto que $\cos x$ es negativo en este intervalo.

$$\frac{d}{dx}\operatorname{arc\,sen} x = \frac{1}{\sqrt{1-x^2}}.$$

Aquí se supone que arc sen es el valor principal que está situado entre $-\pi/2$ y $+\pi/2$, y el signo de la raíz cuadrada es elegido positivo.

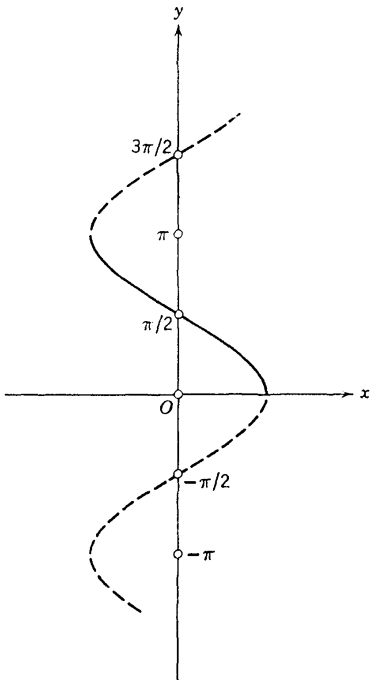


Figura 3.6 Gráfica de $y = \operatorname{arc\,cos} x$ (el valor principal está indicado por una curva llena).

Para la función inversa de $y = \cos x$, denotada (después de intercambiar nuevamente los nombres x e y) por $\operatorname{arc\,cos} x$, se obtiene la fórmula

$$\frac{d}{dx}\operatorname{arc\,cos} x = \mp\frac{1}{\sqrt{1-x^2}}$$

exactamente de la misma manera. Aquí se toma el signo negativo de la raíz si el valor de $\operatorname{arc\,cos} x$ se toma en el intervalo entre 0 y π (no entre $-\pi/2$ y $+\pi/2$, como en el caso de $\operatorname{arc\,sen} x$) (véase Fig. 3.6).

Las derivadas se vuelven infinitas al aproximarse a los puntos extremos $x = -1$ y $x = +1$, lo cual corresponde al hecho de que las gráficas del seno inverso y del coseno inverso tienen tangentes verticales en estos puntos.

Tangente y cotangente inversas

Las funciones inversas de la tangente y de la cotangente se tratan en forma análoga. La función $y = \tan x$, que posee en todas partes una derivada positiva $1/\cos^2 x$ para $x \neq \pi/2 + k\pi$, posee una inversa única en el intervalo $-\pi/2 < x < \pi/2$. Denominamos a esta función inversa (la Rama Principal de) $x = \operatorname{arc\,tan} y$. Se ve de inmediato a partir de la Fig. 3.7 que para cada x pudo haberse elegido en vez de y cualesquiera de los

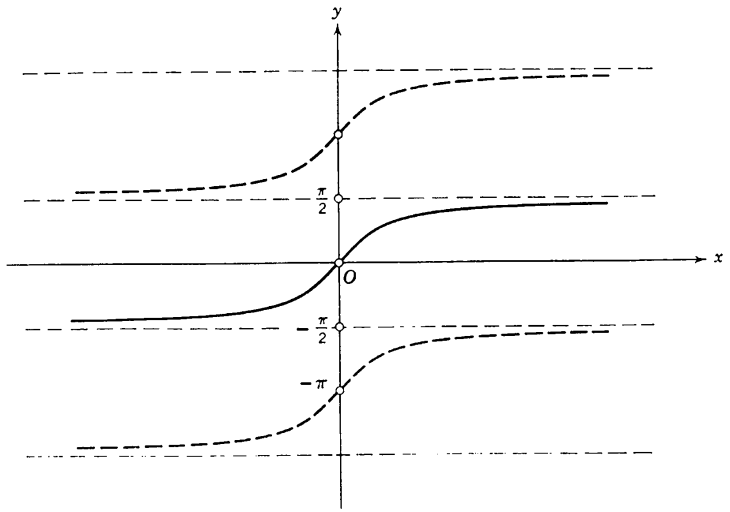


Figura 3.7 Gráfica de $y = \operatorname{arc\,tan} x$ (la curva llena para el valor principal).

valores $y + k\pi$ (donde k es un entero). Análogamente, la función $y = \cot x$ posee una inversa $x = \operatorname{arc\,cot} y$ que está determinada de manera única si se exige que su valor esté situado en el intervalo de 0 a π; si no, la multivalencia de $\operatorname{arc\,cot} x$ es la misma que para $\operatorname{arc\,tan} x$.

Las fórmulas de derivación son como sigue:

$$x = \operatorname{arc\,tan} y, \qquad \frac{dx}{dy} = \frac{1}{dy/dx} = \cos^2 x = \frac{1}{1+\tan^2 x} = \frac{1}{1+y^2},$$

$$x = \operatorname{arc\,cot} y, \qquad \frac{dx}{dy} = -\operatorname{sen}^2 x = -\frac{1}{1+\cot^2 x} = -\frac{1}{1+y^2},$$

o, finalmente, si denotamos nuevamente la variable independiente por x,

$$\frac{d}{dx}\operatorname{arc\,tan} x = \frac{1}{1+x^2},$$

$$\frac{d}{dx}\operatorname{arc\,cot} x = -\frac{1}{1+x^2}.$$

d. Las fórmulas integrales correspondientes

Expresadas en términos de integrales indefinidas, las fórmulas que se acaban de deducir se escriben como sigue:

$$\int \frac{1}{\sqrt{1-x^2}}\,dx = \operatorname{arc\,sen} x, \qquad \int \frac{1}{\sqrt{1-x^2}}\,dx = -\operatorname{arc\,cos} x,$$

$$\int \frac{1}{1+x^2}\,dx = \operatorname{arc\,tan} x, \qquad \int \frac{1}{1+x^2}\,dx = -\operatorname{arc\,cot} x.$$

Aún cuando las dos fórmulas de cada fila expresan diferentes funciones mediante integrales indefinidas idénticas, no hay contradicción entre ellas. En efecto, ilustran lo que se aprendió anteriormente (ver sección 2.9), a saber: que todas las integrales indefinidas de la misma función difieren solamente en constantes. Aquí las constantes son $\pi/2$ puesto que $\operatorname{arc\,cos} x + \operatorname{arc\,sen} x = \pi/2$, $\operatorname{arc\,tan} x + \operatorname{arc\,cot} x = \pi/2$.

Las fórmulas para integrales indefinidas pueden aplicarse de inmediato para encontrar integrales definidas, como en la p. 165. En particular,

$$\int_a^b \frac{dx}{1+x^2} = \operatorname{arc\,sen} x \Big|_a^b = \operatorname{arc\,tan} b - \operatorname{arc\,tan} a.$$

Si se hacen $a = 0$, $b = 1$ y se recuerda que $\tan 0 = 0$ y $\tan \pi/4 = 1$, se obtiene la importante fórmula

$$\frac{\pi}{4} = \int_0^1 \frac{1}{1+x^2}\,dx. \tag{6}$$

El número π, que surgió originalmente de la consideración del círculo, es puesto por esta fórmula en una relación muy simple con la función racional $1/(1+x^2)$ y representa el área indicada en la Fig. 3.8. Esta fórmula para π, a la cual regresaremos posteriormente (Sec. 5.2*a*), constituye uno de los primeros triunfos de la potencia del cálculo.

Más generalmente, las fórmulas integrales de esta sección nos permiten definir las funciones trigonométricas de manera puramente analítica, sin referencia alguna a objetos geométricos tales como triángulos o círculos. Por ejemplo, la relación entre el ángulo y y su tangente, $x = \tan y$, está descrita completamente por la ecuación

$$y = \int_0^x \frac{du}{1+u^2}$$

(al menos para $-\pi/2 < y < \pi/2$). Con esta relación podemos *definir* ahora, sin recurrir a la intuición, un valor numérico para el ángulo y en un triángulo rectángulo con lados a (adyacente) y b (opuesto) para el cual

$b/a = x$. Tal definición analítica en términos de cantidades numéricas hace lícito el uso de ángulos y funciones trigonométricas en el análisis superior independientemente de una definición por construcción geométrica.

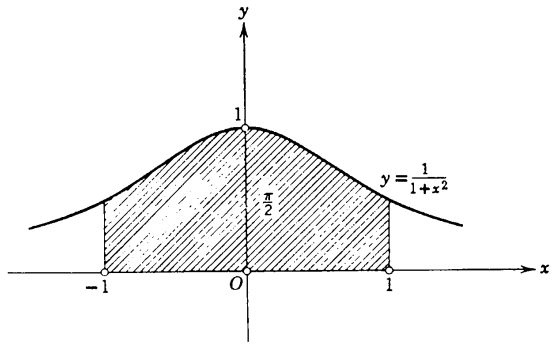


Figura 3.8 $\pi/2$ ilustrado mediante un área.

e. Derivada e integral de la función exponencial

En el capítulo 2, p. 172, se introdujo la función exponencial como la inversa del logaritmo. Más precisamente, las relaciones $y = e^x$ y $x = \log y$ fueron así definidas como equivalentes. En consecuencia, sus derivadas satisfacen la relación [ver (5), p. 229]

$$\frac{de^x}{dx} = \frac{dy}{dx} = \frac{1}{\dfrac{dx}{dy}} = \frac{1}{\dfrac{d\log y}{dy}} = \frac{1}{\dfrac{1}{y}} = y = e^x.$$

Por tanto, *la función exponencial es igual a su derivada:*

$$\frac{de^x}{dx} = e^x.$$

Más generalmente, para cualquier a positivo la función $y = a^x$ posee como inversa

$$x = \log_a y = \frac{\log y}{\log a},$$

y la derivada de a^x es

$$\frac{da^x}{dx} = \frac{1}{\dfrac{d\log_a y}{dy}} = (\log a)y = (\log a)a^x.$$

Así, para cualquier constante positiva a la derivada de la función $y = a^x$ es proporcional a la función misma. El factor de proporcionalidad $\log a$ es

1 cuando a es el número e. En la p. 244 se mostrará, recíprocamente que cualquier función que es proporcional a su derivada deberá ser de la forma $y = ce^x$, donde c denota un factor constante.

Por el teorema fundamental del cálculo, como antes, pueden traducirse las fórmulas para las derivadas de e^x y de a^x en fórmulas para integrales indefinidas:

$$\int e^x\,dx = e^x,$$

$$\int a^x\,dx = \frac{1}{\log a}\,a^x.$$

3.3 Derivación de funciones compuestas

a. Definiciones

Las reglas anteriores nos permiten encontrar las derivadas de funciones que se obtienen como expresiones racionales, en términos de funciones derivadas ya conocidas. Para encontrar expresiones explícitas para las derivadas de otras funciones que ocurren en análisis debemos llegar más lejos deduciendo una regla general para derivar funciones compuestas. Bastante a menudo se encuentran funciones $f(x)$ construidas por el proceso de composición de otras más sencillas (ver capítulo 1, p. 76): $f(x) = g(\phi(x))$, donde $\phi(x)$ está definida en un intervalo cerrado $a \leq x \leq b$ y tiene ahí el rango $\alpha \leq \phi \leq \beta$, y donde $g(\phi)$ está definida en este último intervalo.

En relación con esto es útil recordar la interpretación de funciones como "operadores" o transformaciones. Como en el capítulo 1, se escribe la función compuesta simplemente como

$$f = g\phi$$

y se denomina a $g\phi$ el "producto" (simbólico) de los operadores o transformaciones g y ϕ.

b. La regla de la cadena

Para funciones g y ϕ que son continuas en sus respectivos intervalos de definición, la función compuesta $f(x) = g[\phi(x)]$ es también continua (ver capítulo 1, p. 79).

Las funciones $\phi(x)$ y $g(\phi)$ se suponen ahora no sólo continuas sino también derivables. Se tiene entonces el siguiente teorema fundamental, de la *"regla de la cadena" para la derivación:*

La función $f(x) = g[\phi(x)]$ es derivable y su derivada está dada por la ecuación

$$f'(x) = g'(\phi)\cdot\phi'(x), \tag{7}$$

o bien, en la notación de Leibnitz,

$$\frac{df}{dx} = \frac{df}{d\phi}\frac{d\phi}{dx}.$$

Por consiguiente, la derivada de la función compuesta es el producto de las derivadas de las funciones que la constituyen. O bien: La derivada del producto simbólico de funciones es el producto real de sus derivadas con respecto a sus correspondientes variables independientes.

Intuitivamente esta regla de la cadena es plausible. La cantidad $\phi'(x) = \lim \Delta\phi/\Delta x$ es el cociente local en el cual pequeños intervalos son amplificados por el "mapeo" ϕ. Análogamente, $g'(\phi)$ es la amplificación dada por el "mapeo" g. Aplicar primero ϕ y después g equivale a amplificar un intervalo x primero ϕ' veces y luego agrandar el intervalo ϕ resultante g' veces, obteniéndose así un cociente total de amplificación igual a $g'\phi'$, que debe ser el cociente de amplificación para el "mapeo" compuesto $f = g\phi$.

El teorema se sigue muy fácilmente a partir de la definición de la derivada. De hecho, se vuelve intuitivamente casi obvio si se supone $\phi'(x) \neq 0$ en el intervalo cerrado x que se considera. Entonces, para $\Delta x = x_2 - x_1 \neq 0$ se tiene por el teorema del valor medio

$$\Delta\phi = \phi_2 - \phi_1 = \phi(x_2) - \phi(x_1) = \phi'(\xi)\Delta x \neq 0 \quad \text{con} \quad x_1 \leq \xi \leq x_2,$$

y, con $\Delta g = g(\phi_2) - g(\phi_1)$ y $\Delta f = f(x_2) - f(x_1)$, puede escribirse

$$\frac{\Delta f}{\Delta x} = \frac{\Delta g}{\Delta\phi}\frac{\Delta\phi}{\Delta x}$$

que es una identidad con sentido, puesto que $\Delta\phi \neq 0$. Ahora $\Delta\phi \to 0$ para $\Delta x \to 0$, esto es, para $x_2 \to x_1$; por tanto, para $\Delta x \to 0$ los cocientes incrementales tienden a sus respectivas derivadas y el teorema está probado.

Para evitar la hipótesis explícita $\phi'(x) \neq 0$ podemos prescindir de la división por $\Delta\phi$ de la siguiente manera un poco más ingeniosa:

Por la hipótesis de la derivabilidad de $g(\phi)$ en el punto ϕ se sabe que la cantidad $\varepsilon = \Delta g/\Delta\phi - g'(\phi)$, como función de $\Delta\phi$ para ϕ fijo y $\Delta\phi \neq 0$, posee el límite cero para $\Delta\phi \to 0$. Si se *define* $\varepsilon = 0$ para $\Delta\phi = 0$, sin restricción sobre $\Delta\phi$ se tiene

$$\Delta g = [g'(\phi) + \varepsilon]\,\Delta\phi.$$

Análogamente, para x fijo,

$$\Delta\phi = \phi(x + \Delta x) - \phi(x) = [\phi'(x) + \eta]\,\Delta x,$$

donde $\lim_{\Delta x \to 0} \eta = 0$. Entonces, para $\Delta x \neq 0$ y $\phi = \phi(x)$, es

$$\frac{\Delta g}{\Delta x} = [g'(\phi) + \varepsilon]\frac{\Delta\phi}{\Delta x} = [g'(\phi) + \varepsilon][\phi'(x) + \eta].$$

Para Δx tendiendo a cero a través de valores distintos de cero se tiene $\lim_{\Delta x \to 0} \Delta\phi = 0$ y por lo tanto $\lim_{\Delta x \to 0} \varepsilon = 0$, de modo que

$$\lim_{\Delta x \to 0} \frac{\Delta g}{\Delta x} = \lim_{\Delta x \to 0} [g'(\phi) + \varepsilon] \lim_{\Delta x \to 0} [\phi'(x) + \eta] = g'(\phi)\phi'(x),$$

lo que demuestra la regla de la cadena.

Por aplicación sucesiva esta regla puede extenderse inmediatamente a funciones que resultan de la *composición de más de dos funciones.* Si, por ejemplo,

$$y = g(u), \qquad u = \phi(v), \qquad v = \psi(x),$$

entonces $y = f(x) = g[\phi(\psi(x))]$ es una función compuesta de x; su derivada está dada por la regla

$$\frac{dy}{dx} = y' = g'(u)\phi'(v)\psi'(x) = \frac{dy}{du} \cdot \frac{du}{dv} \cdot \frac{dv}{dy};$$

relaciones similares son ciertas para funciones que están compuestas de un número arbitrario de funciones.

Derivadas superiores de una función compuesta $y = g[\phi(x)]$. Pueden encontrarse fácilmente mediante la aplicación repetida de la regla de la cadena y las reglas anteriores:

$$y' = \frac{dy}{d\phi}\frac{d\phi}{dx} = g'\phi',$$

$$y'' = g''\phi'^2 + g'\phi'',$$

$$y''' = g'''\phi'^3 + 3g''\phi'\phi'' + g'\phi'''.$$

Fórmulas análogas para y'''', etc., pueden ser deducidas sucesivamente.

Finalmente, examinemos la *composición de dos funciones que son inversas* una con respecto a la otra. La función $g(y)$ es la inversa de $y = \phi(x)$ si $f(x) = g[\phi(x)] = x$. Se sigue que

$$f'(x) = g'(y)\phi'(x) = 1$$

que es exactamente el resultado de la sección 3.2, p. 228.

Ejemplos. Como un ejemplo sencillo, pero importante, relativo a una aplicación de la regla de la cadena derivemos $x^\alpha (x > 0)$ para una potencia real arbitraria α. En el capítulo 2, p. 174, se definió

$$x^\alpha = e^{\alpha \log x};$$

y se probó también para $\phi(x) = \log x, \psi(u) = \alpha u, g(y) = e^y$ que

$$\phi'(x) = \frac{1}{x} \qquad \psi'(u) = \alpha, \qquad g'(y) = e^y.$$

Ahora x^α es la función composición $g\{\psi[\phi(x)]\}$. Aplicando la regla de la cadena se obtiene la fórmula general

$$\frac{d}{dx}(x^\alpha) = g'(y)\psi'(u)\phi'(x)$$

$$= e^y \cdot \alpha \cdot \frac{1}{x}$$

$$= \frac{\alpha e^{\alpha \log x}}{x}$$

$$= \alpha \frac{x^\alpha}{x};$$

por tanto

$$\frac{d}{dx}(x^\alpha) = \alpha x^{\alpha-1},$$

un resultado que solamente se podría probar con alguna dificultad si se intentara proceder directamente a partir de la definición de x^α para α irracional como el límite de potencias con exponentes racionales.

Una consecuencia inmediata de esta derivación es, nuevamente, la fórmula integral

$$\int x^\alpha \, dx = \frac{x^{\alpha+1}}{\alpha + 1} \qquad (\alpha \neq -1).$$

Como un segundo ejemplo, considérese

$$y = \sqrt{1 - x^2}, \qquad \text{o sea} \quad y = \sqrt{\phi},$$

donde $\phi = 1 - x^2$ con $-1 < x < 1$. La regla de la cadena conduce a

$$y' = \frac{1}{2\sqrt{\phi}} \cdot (-2x) = \frac{x}{\sqrt{1 - x^2}}.$$

Los siguientes cálculos breves proporcionan ejemplos adicionales.

1. $y = \operatorname{arc\,sen} \sqrt{1 - x^2}, \qquad (-1 \leq x \leq 1, x \neq 0).$

$$\frac{dy}{dx} = \frac{1}{\sqrt{1 - (1 - x^2)}} \frac{d\sqrt{1 - x^2}}{dx}$$

$$= \frac{1}{|x|} \frac{-x}{\sqrt{1 - x^2}} = \frac{-1}{\sqrt{1 - x^2}} \operatorname{sgn}(x).$$

2. $y = \sqrt{\dfrac{1 + x}{1 - x}}, \qquad (-1 < x < 1).$

$$\frac{dy}{dx} = \frac{1}{2\sqrt{\dfrac{1 + x}{1 - x}}} \cdot \frac{d\left(\dfrac{1 + x}{1 - x}\right)}{dx}$$

$$= \frac{\sqrt{1 - x}}{2\sqrt{1 + x}} \cdot \frac{2}{(1 - x)^2} = \frac{1}{(1 + x)^{1/2}(1 - x)^{3/2}}.$$

3. $y = \log|x|$. Esta función [1] puede expresarse como $\log x$ para $x > 0$ y como $\log(-x)$ para $x < 0$. Para $x > 0$ es

$$\frac{d\log|x|}{dx} = \frac{d\log x}{dx} = \frac{1}{x}.$$

Para $x < 0$, se obtiene a partir de la regla de la cadena que

$$\frac{d\log|x|}{dx} = \frac{d\log(-x)}{dx} = \frac{1}{-x}\frac{d(-x)}{dx} = \frac{1}{x}$$

Por tanto, en forma general para $x \neq 0$ es

$$\frac{d\log|x|}{dx} = \frac{1}{x}$$

4. $y = a^x$. Por definición de a^x (ver p. 174), se tiene

$$a^x = e^{\phi(x)},$$

donde $\phi(x) = (\log a)x$. Entonces,

$$\frac{da^x}{dx} = \frac{de^\phi}{d\phi}\frac{d\phi}{dx} = e^\phi(\log a) = (\log a)a^x.$$

El mismo resultado fue obtenido ya en la p. 238 a partir de la regla para la derivada de la función inversa.

5. $y = [f(x)]^{g(x)}$. Puesto que

$$[f(x)]^{g(x)} = e^{\phi(x)},$$

con $\phi(x) = g(x)\log[f(x)]$, se encuentra que

$$\frac{d}{dx}[f(x)]^{g(x)} = e^\phi\left(g'\log f + g\frac{1}{f}f'\right)$$

$$= [f(x)]^{g(x)}\left(g'(x)\log[f(x)] + \frac{g(x)f'(x)}{f(x)}\right).$$

Por ejemplo, cuando $g(x) = f(x) = x$ se tiene

$$\frac{dx^x}{dx} = x^x(\log x + 1).$$

c. El teorema generalizado del valor medio del cálculo diferencial

Como una aplicación de la regla de la cadena se deduce el *teorema generalizado del valor medio* del cálculo diferencial. Considérense dos funciones $F(x)$ y $G(x)$ continuas en un intervalo cerrado $[a, b]$ del eje x y derivables en el interior de este

[1] La función $\log x$ está definida solamente para $x > 0$, mientras que $\log|x|$ está definida en todas partes excepto para $x = 0$.

intervalo. Se supone $G'(x)$ positiva. El teorema ordinario del valor medio del cálculo diferencial aplicado separadamente a F y a G proporciona una expresión para el cociente de incrementos $\dfrac{F(b) - F(a)}{G(b) - G(a)}$:

$$\frac{F(b) - F(a)}{G(b) - G(a)} = \frac{F'(\xi)(b - a)}{G'(\eta)(b - a)} = \frac{F'(\xi)}{G'(\eta)}$$

donde ξ y η son valores intermedios apropiados en el intervalo abierto (a, b). El teorema generalizado del valor medio establece que el cociente de incrementos puede ser escrito en la forma más sencilla

$$\frac{F(b) - F(a)}{G(b) - G(a)} = \frac{F'(\zeta)}{G'(\zeta)}$$

donde F' y G' están evaluadas en *el mismo valor intermedio* ζ.

Para la demostración introducimos $u = G(x)$ como una variable independiente en F. De la hipótesis $G' > 0$ se concluye que la función $u = G(x)$ es monótona en el intervalo $[a, b]$, y, por tanto, que posee una inversa $x = g(u)$ definida en el intervalo $[\alpha, \beta]$, con $\alpha = G(a)$, $\beta = G(b)$. La función compuesta $F[g(u)] = f(u)$ está, por tanto, definida para u en el intervalo $[\alpha, \beta]$. A partir del teorema ordinario del valor medio se encuentra que

$$F(b) - F(a) = f(\beta) - f(\alpha) = f'(\gamma)(\beta - \alpha) = f'(\gamma)[G(b) - G(a)],$$

donde γ es un valor apropiado entre α y β. Por la regla de la cadena se infiere que

$$f'(u) = \frac{d}{du}F[g(u)] = F'[g(u)]g'(u) = \frac{F'(x)}{G'(x)}.$$

Al valor $u = \gamma$ le corresponde un valor $x = g(\gamma) = \zeta$ en el intervalo (a, b). Entonces $f'(\gamma) = F'(\zeta)/G'(\zeta)$, y se sigue el teorema generalizado del valor medio.

3.4 Algunas aplicaciones de la función exponencial

Algunos problemas diversos que involucran la función exponencial ilustrarán la importancia fundamental de esta función en toda clase de aplicaciones.

a. Definición de la función exponencial por medio de una ecuación diferencial

Podemos definir la función exponencial mediante una propiedad sencilla cuyo uso evita muchos argumentos detallados en casos particulares.

Si una función $y = f(x)$ satisface una ecuación de la forma

$$y' = \alpha y,$$

donde α es una constante, entonces y tiene la expresión

$$y = f(x) = ce^{\alpha x}, \tag{8}$$

donde c es también una constante; recíprocamente, toda función de la forma $ce^{\alpha x}$ satisface la ecuación $y' = \alpha y$.

Puesto que la Ec. (8) expresa una relación entre la función y su derivada, aquélla se denomina la *ecuación diferencial* de la función exponencial.

Es claro que $y = ce^{\alpha x}$ satisface esta ecuación para cualquier constante arbitraria c. Recíprocamente, ninguna otra función satisface la ecuación diferencial $y' - \alpha y = 0$. En efecto, si y es tal función, considérese la función $u = ye^{-\alpha x}$. Se tiene entonces

$$u' = y'e^{-\alpha x} - \alpha y e^{-\alpha x} = e^{-\alpha x}(y' - \alpha y).$$

Sin embargo, el miembro derecho se anula, ya que se ha supuesto que $y' = \alpha y$; por lo tanto $u' = 0$, de modo que (ver p. 199) u es una constante c, y así resulta $y = ce^{\alpha x}$, como se quería probar.

Aplicaremos ahora este teorema a cierto número de ejemplos.

b. Interés compuesto continuamente Desintegración radiactiva

Una suma "capital" o "principal", aumentada por su interés en períodos regulares de tiempo, se incrementa por saltos en estos períodos de la siguiente manera. Si 100α es el porcentaje de interés y, además, si el interés acumulado se suma, al finalizar cada año, al capital invertido (original), después de x años la cantidad acumulada de un capital original igual a 1 será

$$(1 + \alpha)^x.$$

Sin embargo, si al capital le fuera sumado el interés no al finalizar cada año sino al finalizar cada n-ésima parte de un año, después de x años el capital ascendería a

$$\left(1 + \frac{\alpha}{n}\right)^{nx}.$$

Tomando, por simplicidad, $x = 1$ se encuentra que el capital 1 ha aumentado después de un año a

$$\left(1 + \frac{\alpha}{n}\right)^{n}.$$

Si ahora se deja crecer n más allá de toda cota, esto es, si se deja que el interés sea abonado en intervalos cada vez más cortos, el caso límite significará en cierto sentido que el interés compuesto es abonado a cada instante; entonces la cantidad total después de un año será e^{α} veces el capital inicial u original (véase p. 175). Análogamente, si el interés se calcula de esta manera, un capital original igual a 1 habrá crecido después de x años hasta una cantidad $e^{\alpha x}$; aquí x puede ser cualquier número, entero o de otra forma.

La discusión de la sección 3.4a constituye un marco de referencia en el cual ejemplos de este tipo encajan fácilmente. Considérese una cantidad, dada por el número y, que crece (o decrece) con el tiempo de modo que la medida en que la cantidad aumenta o disminuye es proporcional a la cantidad total. Entonces, tomando el tiempo como la variable independiente x se obtiene una ley de la forma $y' = \alpha y$ para el grado de aumento (o decrecimiento), donde α, el factor de proporcionalidad, es positivo o negativo dependiendo de que la cantidad esté creciendo o decreciendo. Entonces, de acuerdo con la sección 3.4a, la cantidad y está representada por la fórmula

$$y = ce^{\alpha x},$$

donde el significado de la constante c es obvio si se considera el instante $x = 0$. En este instante $e^{\alpha x} = 1$, y se encuentra que $c = y_0$ es la cantidad al comienzo del tiempo considerado, de modo que se puede escribir

$$y = y_0 e^{\alpha x}.$$

Un ejemplo característico es el de la *desintegración radiactiva.* La medida en que la cantidad total y de la substancia radiactiva está disminuyendo es proporcional en cualquier instante a la cantidad total presente en ese instante; esto es plausible *a priori* pues cada porción de la substancia decrece tan rápidamente como cualquier otra porción. Por tanto, la cantidad y de la substancia, expresada como función del tiempo, satisface una relación de la forma $y = -ky$, donde k debe tomarse positiva puesto que se está tratando con una cantidad que disminuye. La cantidad de substancia queda así expresada como función del tiempo por $y = y_0 e^{-kx}$, donde y_0 es la cantidad de la substancia al comienzo del tiempo considerado (tiempo $x = 0$).

Después de un cierto tiempo τ la substancia radiactiva habrá disminuido a la mitad de la cantidad original. Este tiempo, llamado *vida media,* está dado por la ecuación

$$\tfrac{1}{2}y_0 = y_0 e^{-k\tau},$$

de la cual se obtiene de inmediato $\tau = (\log 2)/k$.

c. Enfriamiento o calentamiento de un cuerpo por el medio ambiente

Otro ejemplo típico en el que aparece la función exponencial lo constituye el enfriamiento de un cuerpo; por ejemplo, una placa de metal de temperatura uniforme que es sumergida en un gran "baño", o medio circundante o "ambiente", de baja temperatura. Se supone que la extensión del medio ambiente es tan grande que su temperatura no es afectada por el proceso de enfriamiento. Se supone además que en cada instante todas las partes del cuerpo sumergido se encuentran a la misma temperatura, y

que la medida en que la temperatura varía es proporcional a la diferencia entre la temperatura del cuerpo y la del medio circundante (Ley de enfriamiento de Newton).

Si se denota el tiempo por x y la diferencia de temperatura entre el cuerpo y el baño por $y = y(x)$, esta ley de enfriamiento está expresada por la ecuación

$$y' = -ky,$$

donde k es una constante positiva (cuyo valor es una característica física de la substancia del cuerpo). A partir de esta ecuación diferencial, que expresa el efecto del proceso de enfriamiento en un instante dado, se obtiene por medio de la Ec. (8), p. 244, una "ley integral" que nos da la temperatura en un instante arbitrario x en la forma

$$y = ce^{-kx}.$$

Esto muestra que la temperatura decrece "exponencialmente" y tiende a ser igual a la temperatura externa. La rapidez con la que esto sucede está expresada por el número k. Como antes, el significado de la constante c es el de la temperatura inicial en el instante $x = 0$, $y_0 = c$, de modo que nuestra ley de enfriamiento puede ser escrita en la forma

$$y = y_0e^{-kx}.$$

Evidentemente, la misma discusión se aplica al calentamiento de un cuerpo. Lo único que cambia es que la diferencia inicial de temperaturas y_0 es en este caso negativa en vez de positiva.

d. Variación de la presión atmosférica con la altura sobre la superficie de la Tierra

Un ejemplo más en el que aparece la fórmula exponencial es el de la variación de la presión atmosférica con la altura: Se hará uso (1) del hecho físico de que la presión atmosférica es igual al peso de la columna de aire que está verticalmente por encima de una superficie de área uno, y (2) de la Ley de Boyle, de acuerdo con la cual la presión p del aire a una temperatura constante es proporcional a su densidad σ. La Ley de Boyle, expresada en símbolos, es $p = a\sigma$, donde a es una constante que depende de una propiedad física específica del aire. Nuestro problema es el de determinar $p = f(h)$ como una función de la altura h por encima de la superficie de la Tierra.

Si por p_0 se denota la presión atmosférica en la superficie de la Tierra, esto es, el peso total de la columna de aire sostenida por un área unitaria, por g la constante gravitacional y por $\sigma(\lambda)$ la densidad del aire a la altura λ sobre la superficie terrestre, el peso [1] de la columna de altura h está dado por la integral $g\int_0^h \sigma(\lambda)\,d\lambda$. La presión a una altura h es, por lo tanto,

$$p = f(h) = p_0 - g\int_0^h \sigma(\lambda)\,d\lambda.$$

Derivando esta expresión se obtiene la siguiente relación entre la presión $p = f(h)$ y la densidad $\sigma(h)$:

$$g\sigma(h) = -f'(h) = -p'.$$

Utilizando ahora la Ley de Boyle para eliminar la cantidad σ de esta ecuación, resulta una relación $p' = -(g/a)p$ que involucra solamente la función presión desconocida. De la Ec. (8), p. 244, se sigue que

$$p = f(h) = ce^{-gh/a}.$$

Si, como antes, la presión $f(0)$ en la superficie de la Tierra se denota por p_0, se sigue de inmediato que $c = p_0$, y consecuentemente

$$p = f(h) = p_0e^{-gh/a}.$$

Tomando logaritmos esto conduce a

$$h = \frac{a}{g}\log\frac{p_0}{p}.$$

Estas dos fórmulas se aplican frecuentemente. Por ejemplo, la constante a, si es conocida, permite determinar la altura de un lugar a partir de la presión barométrica, o bien determinar la diferencia entre las alturas de dos lugares midiendo la presión atmosférica en cada sitio. Por otra parte, si la presión atmosférica y la altura h son conocidas, puede determinarse la constante a, la cual es de gran importancia en la teoría de los gases.

e. Progreso de una reacción química

Se considerará ahora un ejemplo de la Química, la así llamada *reacción unimolecular*. Se supone que una substancia es disuelta en una gran cantidad de solvente; por ejemplo, una cierta cantidad de azúcar de caña en agua. Si ocurre una reacción química, la ley química de acción de las masas establece en este caso que la medida de reacción es proporcional a la cantidad de substancia reaccionante presente. Se supone que el azúcar de caña está siendo transformada mediante acción catalítica en azúcar invertida, y se denota por $u(x)$ la cantidad de azúcar de caña que al tiempo x permanece invariable aún. La velocidad de reacción es entonces $-du/dx$,

[1] $g\sigma(\lambda)$ es el peso del aire por unidad de volumen a la altura λ.

y, de acuerdo con la ley de acción de masas, se cumple una ecuación de la forma

$$\frac{du}{dx} = -ku$$

donde k es una constante que depende de las substancias reaccionantes. A partir de esta ley instantánea o diferencial se obtiene de inmediato, como en la p. 244, una ley integral que da la cantidad de azúcar de caña como función del tiempo:

$$u(x) = ae^{-kx}.$$

Esta fórmula muestra claramente cómo la reacción química tiende asintóticamente a su estado final $u = 0$, esto es, la transformación completa de la substancia reaccionante. La constante a es obviamente la cantidad de azúcar de caña presente en el tiempo $x = 0$.

f. Apertura y cierre de un circuito eléctrico

Como un ejemplo final, considérese el crecimiento de una corriente eléctrica directa o "continua" cuando se cierra un circuito, o bien su decaimiento cuando el circuito se abre. Si R es la resistencia del circuito y E la fuerza electromotriz (voltaje), la corriente I se incrementa gradualmente desde su valor original cero hasta un valor estacionario final E/R. Tenemos, por tanto, que considerar I como una función del tiempo x. El aumento de la corriente depende de la autoinducción del circuito; y el circuito posee una constante característica L, el *coeficiente de autoinducción,* de naturaleza tal que conforme la corriente crece se desarrolla una fuerza electromotriz de magnitud $L\,dI/dx$ que se opone a la fuerza electromotriz externa E. A partir de la ley de Ohm, que afirma que el producto de la resistencia y de la corriente es en cada instante igual al voltaje efectivo real, se obtiene la relación

$$IR = E - L\frac{dI}{dx}.$$

para

$$f(x) = I(x) - \frac{E}{R}$$

se encuentra de inmediato que $f'(x) = -(R/L)f(x)$, de modo que por la Ec. (8), p. 244, es $f(x) = f(0)e^{-Rx/L}$. Recordando que $I(0) = 0$, se encuentra que $f(0) = -E/R$; y así se obtiene la expresión

$$I = f(x) + \frac{E}{R} = \frac{E}{R}(1 - e^{-Rx/L})$$

para la corriente como una función del tiempo.

Esta expresión muestra cómo la corriente tiende asintóticamente a su valor estacionario E/R cuando se cierra el circuito.

3.5 Las funciones hiperbólicas

a. Definición analítica

En muchas aplicaciones la función exponencial aparece en combinaciones de la forma

$$\tfrac{1}{2}(e^x + e^{-x}) \qquad \text{o bien} \quad \tfrac{1}{2}(e^x - e^{-x}).$$

Es conveniente introducir éstas y combinaciones similares como funciones especiales, que se definen como sigue:

$$\text{(9a)} \qquad \operatorname{senh} x = \frac{e^x - e^{-x}}{2}, \qquad \cosh x = \frac{e^x + e^{-x}}{2},$$

$$\text{(9b)} \qquad \tanh x = \frac{e^x - e^{-x}}{e^x + e^{-x}}, \qquad \coth x = \frac{e^x + e^{-x}}{e^x - e^{-x}},$$

llamadas *seno hiperbólico, coseno hiperbólico, tangente hiperbólica,* y *cotangente hiperbólica* respectivamente. Las funciones $\operatorname{senh} x$, $\cosh x$, y $\tanh x$ están definidas para todos los valores de x, mientras que para $\coth x$ el punto $x = 0$ debe ser excluido. Los nombres son elegidos para expresar una cierta analogía con las funciones trigonométricas; y esta analogía, que ha de estudiarse en detalle, es la que justifica una consideración especial de estas nuevas funciones. En las Figs. 3.9, 3.10 y 3.11 se muestran las gráficas de las funciones hiperbólicas; y las líneas punteadas en la Fig. 3.9 son las gráficas de $y = (\frac{1}{2})e^x$ y de $y = (\frac{1}{2})e^{-x}$, a partir de las cuales pueden construirse fácilmente las gráficas de $\operatorname{senh} x$ y $\cosh x$.

Cosh x es, evidentemente, una *función par,* esto es, una función que permanece invariable cuando x es reemplazado por $-x$, mientras que $\operatorname{senh} x$ es una *función impar,* esto es, una función que cambia de signo cuando x es reemplazado por $-x$ (ver p. 53).

Por su definición, la función

$$\cosh x = \frac{e^x + e^{-x}}{2}$$

es positiva y no menor que uno para todos los valores de x. Posee su valor mínimo cuando $x = 0$: $\cosh 0 = 1$.

La relación fundamental entre $\cosh x$ y $\operatorname{senh} x$,

$$\cosh^2 x - \operatorname{senh}^2 x = 1,$$

se sigue de inmediato a partir de las definiciones. Si se denota ahora la variable independiente por t, en vez de x, y se escribe

$$x = \cosh t, \qquad y = \operatorname{senh} t,$$

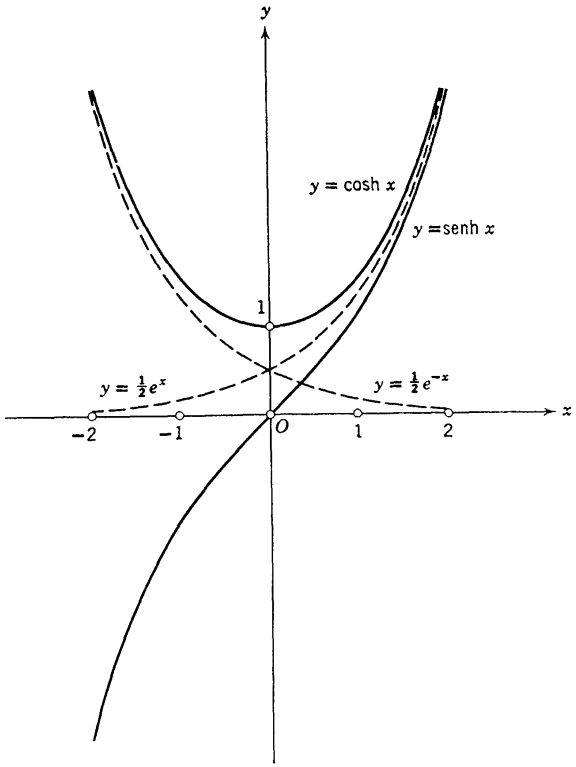


Figura 3.9.

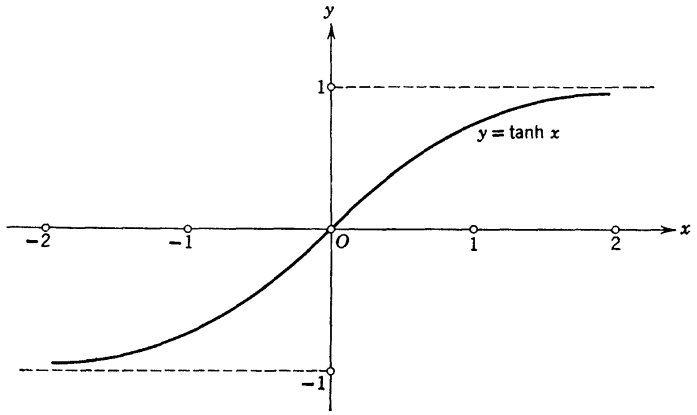


Figura 3.10.

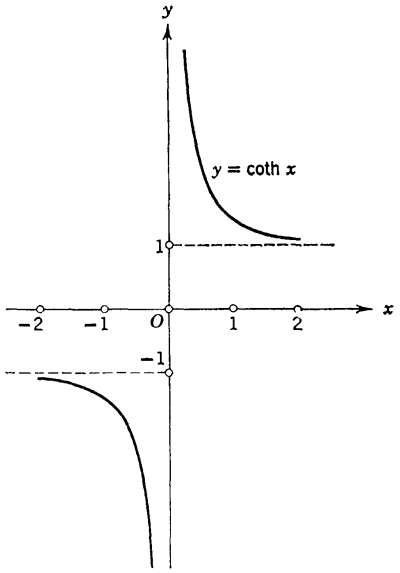


Figura 3.11.

se tiene

$$x^2 - y^2 = 1;$$

esto es, el punto de coordenadas $x = \cosh t$, $y = \operatorname{senh} t$ se mueve a lo largo de la hipérbola rectangular $x^2 - y^2 = 1$ conforme t recorre toda la escala de valores de $-\infty$ a $+\infty$. De acuerdo con la ecuación de definición, $x \geq 1$, y la fórmula hace evidente que y recorre toda la escala de valores de $-\infty$ a $+\infty$ conforme t lo hace; en efecto, si t tiende a infinito, también e^t, mientrsa que e^{-t} tiende a cero. Se puede por tanto establecer más exactamente: Conforme t corre de $-\infty$ a $+\infty$ las ecuaciones $x = \cosh t$, $y = \operatorname{senh} t$ dan una rama de la hipérbola rectangular, a saber, la del lado derecho.

b. Teoremas de adición y fórmulas para derivar

A partir de su definición se obtienen los *teoremas de adición* para las funciones hiperbólicas:

$$\begin{aligned} \cosh(a+b) &= \cosh a \cosh b + \operatorname{senh} a \operatorname{senh} b, \\ \operatorname{senh}(a+b) &= \operatorname{senh} a \cosh b + \cosh a \operatorname{senh} b. \end{aligned} \tag{10}$$

Las demostraciones se obtienen de inmediato si se escribe

$$\cosh(a+b) = \frac{e^a e^b + e^{-a}e^{-b}}{2} \qquad \operatorname{senh}(a+b) = \frac{e^a e^b - e^{-a}e^{-b}}{2},$$

y se insertan en estas ecuaciones

$$e^a = \cosh a + \operatorname{senh} a, \qquad e^{-a} = \cosh a - \operatorname{senh} a,$$

$$e^b = \cosh b + \operatorname{senh} b, \qquad e^{-b} = \cosh b - \operatorname{senh} b.$$

Entre estas fórmulas y las correspondientes fórmulas trigonométricas existe una notable analogía. La única diferencia entre los teoremas de adición es un signo en la primera fórmula.

Una analogía correspondiente se cumple para las fórmulas de diferenciación. Recordando que $d(e^x)/dx = e^x$, se encuentra fácilmente que

$$\begin{aligned} &\frac{d}{dx}\cosh x = \operatorname{senh} x, \qquad \frac{d}{dx}\operatorname{senh} x = \cosh x, \\ &\frac{d}{dx}\tanh x = \frac{1}{\cosh^2 x}, \qquad \frac{d}{dx}\coth x = -\frac{1}{\operatorname{senh}^2 x}. \end{aligned} \tag{11}$$

De las dos primeras ecuaciones se sigue de inmediato que $y = \cosh x$ e $y = \operatorname{senh} x$ son soluciones de la *ecuación diferencial*

$$\frac{d^2y}{dx^2} = y, \tag{12}$$

la cual difiere sólo en signo de la ecuación análoga que es satisfecha por las funciones trigonométricas $\cos x$ y $\operatorname{sen} x$ (véase p. 192).

c. Las funciones hiperbólicas inversas

A las funciones hiperbólicas $x = \cosh t$, $y = \operatorname{senh} t$, les corresponden las funciones inversas que denotamos [1] por

$$t = \operatorname{arc\,cosh} x, \qquad t = \operatorname{arc\,senh} y.$$

Puesto que la función $\operatorname{senh} t$ es monótona creciente [2] en todo el intervalo $-\infty < t < +\infty$, su función inversa está determinada de manera única para todos los valores de y; por otra parte, una ojeada a la gráfica (véase Fig. 3.9, p. 250) muestra que $t = \operatorname{arc\,cosh} x$ no está determinado de manera única sino que posee una ambigüedad de signo debido a que a un valor dado de x le corresponde no solamente el número t sino también el número $-t$. Puesto que $\cosh t \geq 1$ para todos los valores de t, su inversa $\operatorname{arc\,cosh} x$ está definida solamente para $x \geq 1$.

Estas funciones inversas pueden expresarse fácilmente en términos del logaritmo, interpretando la cantidad $e^t = u$ en las definiciones

$$x = \frac{e^t + e^{-t}}{2}, \qquad y = \frac{e^t - e^{-t}}{2}$$

[1] La notación *simbólica* $\cosh^{-1} x$, etc., es también utilizada (ver nota, página, 78).

[2] $(d/dt)\operatorname{senh} t = \cosh t > 0$.

como incógnita y resolviendo estas ecuaciones (cuadráticas) para u:

$$u = x \pm \sqrt{x^2 - 1}, \qquad u = y + \sqrt{y^2 + 1};$$

y puesto que $u = e^t$ sólo puede poseer valores positivos, la raíz cuadrada en la segunda ecuación debe tomarse con el signo positivo, mientras que en la primera cualquier signo es posible (lo que corresponde a la ambigüedad mencionada arriba). En la forma logarítmica, $t = \log u$, y por tanto

$$\begin{aligned} t &= \log\left(x \pm \sqrt{x^2 - 1}\right) = \operatorname{arc\,cosh} x, \\ t &= \log\left(y + \sqrt{y^2 + 1}\right) = \operatorname{arc\,senh} y. \end{aligned} \tag{13}$$

En el caso de $\operatorname{arc\,cosh} x$ la variable x está restringida al intervalo $x \geq 1$, mientras que $\operatorname{arc\,senh} y$ está definido para todos los valores de y.

La ecuación (13) da los valores,

$$\log\left(x + \sqrt{x^2 - 1}\right) \qquad \text{y} \qquad \log\left(x - \sqrt{x^2 - 1}\right)$$

para $\operatorname{arc\,cosh} x$, que corresponden a las dos ramas de $\operatorname{arc\,cosh} x$. Puesto que

$$\left(x + \sqrt{x^2 - 1}\right)\left(x - \sqrt{x^2 - 1}\right) = 1,$$

la suma de estos dos valores de $\operatorname{arc\,cosh} x$ es cero, lo que concuerda con la ambigüedad en el signo de t mencionada antes.

Las inversas de la tangente hiperbólica y la cotangente hiperbólica pueden ser definidas análogamente, y también pueden ser expresadas en términos de logaritmos. Estas funciones las denotamos por $\operatorname{arc\,tanh} x$ y $\operatorname{arc\,coth} x$. Expresando la variable independiente en todas partes por x, fácilmente se obtiene

$$\begin{aligned} \operatorname{arc\,tanh} x &= \frac{1}{2}\log\frac{1+x}{1-x} \qquad \text{en el intervalo } -1 < x < 1, \\ \operatorname{arc\,coth} x &= \frac{1}{2}\log\frac{x+1}{x-1} \qquad \text{en los intervalos } x < -1,\ x > 1. \end{aligned} \tag{14}$$

La diferenciación de estas funciones inversas puede ser llevada a cabo por el mismo lector, quien puede hacer uso ya sea de la regla para derivar una función inversa, o bien de la regla de la cadena simultáneamente con estas expresiones para las funciones inversas en términos de logaritmos. Si x es la variable independiente, los resultados son

$$\begin{aligned} &\frac{d}{dx}\operatorname{arc\,cosh} x = \pm\frac{1}{\sqrt{x^2 - 1}} \qquad \frac{d}{dx}\operatorname{arc\,senh} x = \frac{1}{\sqrt{x^2 + 1}}, \\ &\frac{d}{dx}\operatorname{arc\,tanh} x = \frac{1}{1 - x^2}, \qquad \frac{d}{dx}\operatorname{arc\,coth} x = \frac{1}{1 - x^2}. \end{aligned} \tag{15}$$

Las últimas dos fórmulas no se contradicen una con otra puesto que la primera se cumple solamente para $-1 < x < 1$ y la segunda solamente para $x < -1$ y $1 < x$. Los dos valores de la derivada $d(\operatorname{arc\,cosh} x)/dx$, expresados por el signo $\pm$ en esta primera fórmula, corresponden a las dos ramas diferentes de la curva $y = \operatorname{arc\,cosh} x = \log(x \pm \sqrt{x^2 - 1})$.

d. Otras analogías

Las semejanzas entre las funciones hiperbólicas y las trigonométricas no son accidentales. El origen profundo de estas analogías se manifiesta cuando se consideran estas funciones con argumentos imaginarios, como se hará posteriormente en la sección 7.7a. Se estará entonces en condiciones de identificar $\cosh x$ con $\cos(ix)$ y $\operatorname{senh} x$ con $(1/i)\operatorname{sen}(ix)$, donde $i = \sqrt{-1}$. Este hecho hace evidente que toda relación que involucra funciones trigonométricas tiene su contraparte para funciones hiperbólicas. Muchas de esas analogías poseen interesantes interpretaciones geométricas o físicas. (Véase también el capítulo 4, Sec. 4.1*j*.)

En la representación anterior de la hipérbola rectangular mediante la cantidad t no se atribuyó ningún significado geométrico a este "parámetro". Volviendo ahora a este tema se encuentra otra analogía entre las funciones trigonométricas y las hiperbólicas. Si la circunferencia cuya ecuación es $x^2 + y^2 = 1$ se representa por medio de un parámetro t en la forma $x = \cos t$, $y = \operatorname{sen} t$, la cantidad t puede interpretarse como un ángulo o bien como una longitud de arco medida a lo largo de la circunferencia. Podemos, sin embargo, interpretar t también como el doble del área del sector circular que corresponde a este ángulo, siendo considerada el área positiva o negativa dependiendo de si el ángulo es positivo o negativo.

Se establece ahora análogamente que para las funciones hiperbólicas la cantidad t es el doble del área del sector hiperbólico para $x^2 - y^2 = 1$, mostrado sombreado en la Fig. 3.12.[1] Es esta interpretación de t en términos de áreas la que explica y justifica los nombres $t = \operatorname{arc\,cosh} x$ y $t = \operatorname{arc\,senh} y$ dados a las funciones hiperbólicas inversas.[2] La demostración se obtiene sin ninguna dificultad si se refiere la hipérbola a sus asíntotas tomadas como ejes, por medio de la transformación de coordenadas

$$x - y = \sqrt{2}\xi, \qquad x + y = \sqrt{2}\eta,$$

o bien

$$x = \frac{1}{\sqrt{2}}(\xi + \eta), \qquad y = \frac{1}{\sqrt{2}}(\eta - \xi);$$

[1] Para una demostración diferente, véase Sec. 4.1*j*.

[2] Tal como la notación $t = \operatorname{arc\,cos} x$ se refiere a un *arco* del círculo unitario, la notación $t = \operatorname{arc\,cosh} x$ se refiere a un *área* conectada con una hipérbola rectangular $x^2 - y^2 = 1$. Debe notarse que t *no es* la longitud del arco hiperbólico.

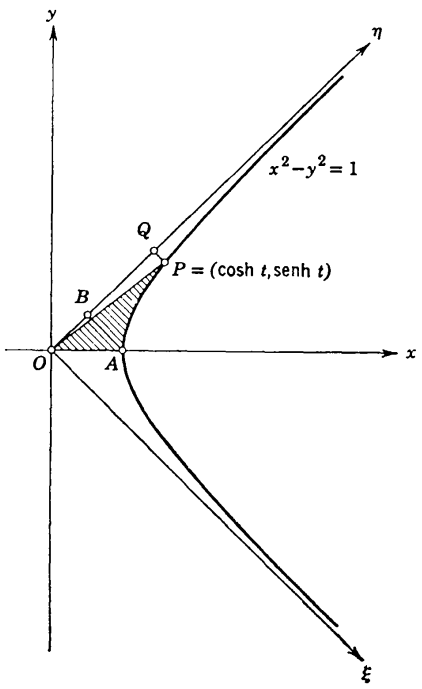


Figura 3.12.

con estas nuevas coordenadas la ecuación de la hipérbola es $\xi\eta = \frac{1}{2}$. Por tanto, los dos triángulos rectángulos OPQ y OAB poseen ambos un área igual a $\frac{1}{4}$, pues las longitudes de OQ y QP son, respectivamente, η y $1/2\eta$, y el área en cuestión es igual a la de la figura $ABQP$. Obviamente las coordenadas de los puntos A y B son

$$\xi = \frac{1}{\sqrt{2}}, \eta = \frac{1}{\sqrt{2}} \qquad \text{y} \qquad \xi = \frac{x - y}{\sqrt{2}}, \eta = \frac{x + y}{\sqrt{2}}$$

respectivamente, y para el doble del área de la figura en cuestión se obtiene así

$$2\int_{1/\sqrt{2}}^{(x+y)/\sqrt{2}} (1/2\eta)\, d\eta = \log(x + y) = \log(x \pm \sqrt{x^2 - 1}),$$

pero, por la Ec. (13), p. 254, esto es igual a t, lo que prueba la aseveración hecha.

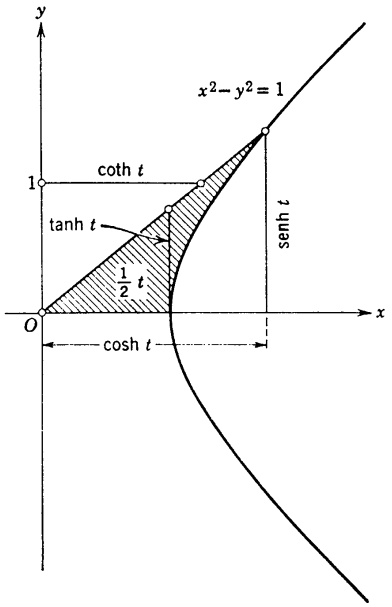


Figura 3.13 Para ilustrar las funciones hiperbólicas.

En conclusión, puede señalarse que, como se muestra en la Fig. 3.13, las funciones hiperbólicas pueden ser representadas gráficamente sobre la hipérbola, de la misma manera que las funciones trigonométricas pueden ser representadas sobre el círculo.

3.6 Máximos y mínimos

Como la primera de una gran variedad de aplicaciones se considerará ahora la teoría de máximos y mínimos de una función, junto con una discusión geométrica de la segunda derivada.

a. Convexidad y concavidad de curvas

Por definición, la derivada $f'(x) = df(x)/dx$ representa la pendiente de la curva $y = f(x)$. La derivada de la función $f'(x)$, es decir, de la pendiente de la curva $y = f(x)$, está dada por la derivada $df'(x)/dx = f''(x)$, la *segunda derivada* de $f(x)$; y así sucesivamente. Si la segunda derivada $f''(x)$ es positiva en un punto x—de modo que debido a la continuidad

(que se supone) es positiva en alguna vecindad [1] de este punto x—entonces a lo largo de esta vecindad la derivada $f'(x)$ aumenta con valores crecientes de x. Por lo tanto, la curva $y = f(x)$ tiene su lado convexo hacia abajo y es "abierta" hacia arriba. A la función $f(x)$, o a la curva $y = f(x)$, la denominamos *convexa*. Si $f''(x)$ es negativa, la curva y la función son *cóncavas*. Por lo tanto, cuando $f''(x) > 0$ la curva en la vecindad del punto

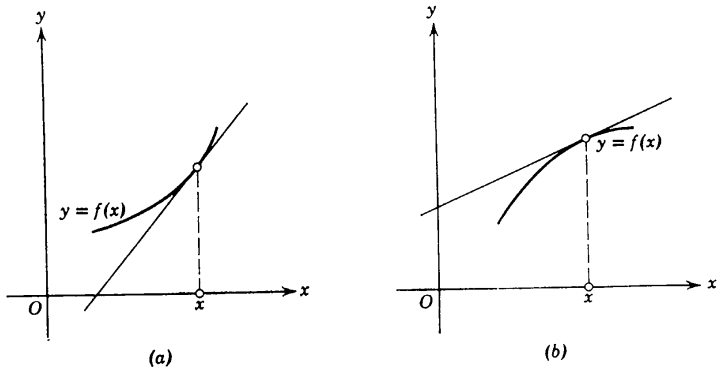


Figura 3.14 $(a) f''(x) > 0$. $(b) f''(x) < 0$.

está situada por encima de la tangente, mientras que cuando $f''(x) < 0$ está situada por debajo de la tangente (véanse las Figs. 3.14a y 3.14b) (ver problema 4, p. 221 y sección 5.6).

Punto de inflexión

Se requiere una consideración especial solamente en los puntos donde $f''(x) = 0$. Al pasar a través de uno de tales puntos la segunda derivada $f''(x)$ cambiará generalmente de signo. Tal punto será entonces un punto de transición entre los dos casos que se acaban de indicar; esto es, en un lado la tangente está por encima de la curva y por el otro lado está por debajo, mientras que en este punto cruza la curva (véase Fig. 3.15). Tal punto es denominado un *punto de inflexión* de la curva, y la tangente correspondiente se denomina *tangente de inflexión*.

[1] Se hace uso aquí de la observación intuitivamente obvia: una función *continua* $g(x)$ que es positiva en un punto x_0 es también positiva para todos los puntos de una vecindad suficientemente pequeña de x_0 (siempre que pertenezcan al dominio de g). La demostración rigurosa es sencilla. A partir de la continuidad de g en x_0, se sabe que para *todo* ε positivo se cumple la desigualdad $|g(x) - g(x_0)| < \varepsilon$ para todo x en una vecindad suficientemente pequeña $|x - x_0| < \delta$ del punto x_0. Como $g(x_0) > 0$, se tiene libertad de elegir para ε el valor $\frac{1}{2}g(x_0)$, de modo que $|g(x) - g(x_0)| < \frac{1}{2}g(x_0)$ en alguna vecindad. Puesto que entonces es $g(x_0) - g(x) \leq |g(x) - g(x_0)| < \frac{1}{2}g(x_0)$, se sigue que $g(x) > \frac{1}{2}g(x_0) > 0$.

El ejemplo más sencillo está dado por la función $y = x^3$, la parábola cúbica, para la cual el eje x mismo es una tangente de inflexión en el punto de inflexión $x = 0$ (véase Fig. 3.3, p. 231). Otro ejemplo está dado por la función $f(x) = \operatorname{sen} x$, para la cual

$$f'(x) = d(\operatorname{sen} x)/dx = \cos x \qquad \text{y} \qquad f''(x) = d^2(\operatorname{sen} x)/dx^2 = -\operatorname{sen} x.$$

Consecuentemente, $f'(0) = 1$ y $f''(0) = 0$; y, puesto que el signo de $f''(x)$ cambia en $x = 0$, la curva seno posee en el origen una tangente de inflexión inclinada a un ángulo de 45 grados con respecto al eje x.

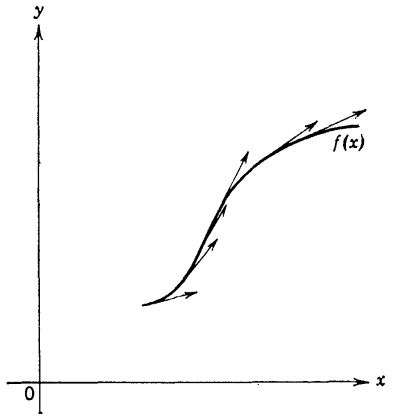


Figura 3.15 Punto de inflexión.

Debe notarse, sin embargo, que pueden existir puntos donde $f''(x) = 0$ y el signo de $f''(x)$ no varíe al crecer x, en tanto que la tangente no corta a la curva sino que permanece por completo a un lado de ésta. Por ejemplo, la curva $y = x^4$ está situada enteramente por encima del eje x, aun cuando la segunda derivada $f''(x) = 12x^2$ se anula para $x = 0$.

b. Máximos y mínimos—Extremos relativos. Puntos estacionarios

Una función $f(x)$ posee un *máximo* en el punto ξ si el valor de f en el punto ξ no es excedido por el valor de f en cualquier otro punto x del dominio de f; esto es, $f(\xi) \geq f(x)$ para todo x donde f esté definida.[1] Análogamente, f posee un *mínimo* en ξ si $f(\xi) \leq f(x)$ para todo x en el dominio. La palabra *extremos* es utilizada para designar tanto a máximos como a mínimos.

[1] Se habla de un *punto máximo estricto* ξ si $f(\xi) > f(x)$ para *todo* x *en el dominio* de f que sea diferente de ξ.

La función $f(x) = \sqrt{1 - x^2}$, por ejemplo, que está definida para $-1 \leq x \leq 1$, posee mínimos en $x = \pm 1$ y un máximo en $x = 0$. Es fácil dar ejemplos de funciones continuas que no poseen máximos o no poseen mínimos. Así, la función $f(x) = 1/(1 + x^2)$ (Fig. 3.8, p. 238) en el dominio $-\infty < x < +\infty$ no posee mínimos; la función $f(x) = 1/x$ definida para $0 < x < \infty$ no posee ningún punto extremo. Sin embargo, recuérdese del capítulo 1, p. 124, el teorema de Weierstrass, de acuerdo con el cual una función continua definida en un intervalo finito cerrado posee siempre un máximo (y análogamente un mínimo).

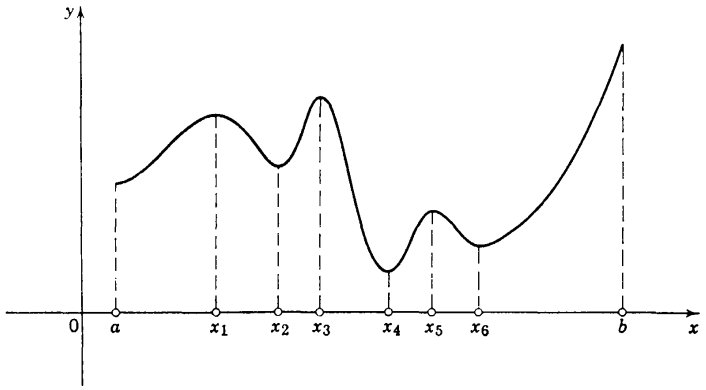


Figura 3.16 Gráfica de una función definida en el intervalo $[a, b]$ con mínimos relativos en $x = a$, x_2, x_4, x_6, máximos relativos en x_1, x_3, x_5, b, máximo absoluto en b, y mínimo absoluto en x_4.

El objetivo aquí es encontrar un medio de localizar los extremos de una función o curva. Este problema, que se encuentra muy frecuentemente en geometría, mecánica, física y otros campos, fue uno de los principales estímulos para el desarrollo del cálculo en el siglo XVII.

El cálculo no proporciona un método directo para determinar los extremos de una función $f(x)$, sino que permite localizar los así llamados puntos *extremos relativos,* entre los cuales los máximos y mínimos reales deben ocurrir. El punto ξ es un máximo *relativo* (mínimo de f si f posee su valor máximo (mínimo) en ξ cuando es comparado no con todos los valores posibles de $f(x)$ sino con sólo aquellos valores de $f(x)$ para x en alguna *vecindad* de ξ. Por una vecindad o entorno del punto ξ se entiende aquí cualquier intervalo abierto $\alpha < x < \beta$ que contiene al punto ξ pero que puede ser arbitrariamente pequeño. Un punto extremo relativo ξ de f es entonces un punto que es un extremo cuando f está restringida

a todos aquellos puntos de su dominio situados suficientemente cerca de ξ.[1] Obviamente, los extremos de una función están incluidos entre los extremos relativos. Para evitar confusión se emplearán los términos *máximos absolutos (mínimos)* para los máximos y mínimos de f en todo su dominio (véase Fig. 3.16).

Geométricamente hablando, los máximos y mínimos relativos, si no están localizados en los puntos extremos del intervalo de definición, son respectivamente las crestas y depresiones de la curva. Una ojeada a la Fig. 3.16 muestra que el valor del máximo relativo en un punto x_5 puede muy bien ser menor que el valor de un mínimo relativo en otro punto x_2. El diagrama sugiere también el hecho de que máximos y mínimos relativos de funciones continuas se alternan: Entre dos máximos relativos sucesivos está siempre localizado un mínimo relativo.

Sea $f(x)$ una función derivable definida en el intervalo cerrado $a \leq x \leq b$. Se ve de inmediato que en un punto extremo relativo que está localizado en el interior del intervalo la tangente a la curva debe ser horizontal. (La demostración rigurosa se da a continuación.) Por lo tanto, la condición

$$f'(\xi) = 0$$

es necesaria para un extremo relativo en un punto ξ tal que $a < \xi < b$. Si, no obstante, $f(\xi)$ es un extremo relativo y ξ coincide con uno de los puntos extremos del intervalo de definición, la ecuación $f'(\xi) = 0$ no necesariamente se cumple. Sólo puede decirse que si el punto extremo izquierdo es un punto máximo relativo (mínimo), la pendiente $f'(a)$ de la curva no puede ser positiva (negativa), mientras que si el punto extremo derecho b es un máximo relativo (mínimo) entonces $f'(b)$ no puede ser negativa (positiva).

Los puntos en los cuales la tangente a la curva $y = f(x)$ es horizontal, que corresponden a las raíces ξ de la ecuación $f'(\xi) = 0$, se denominan los puntos *críticos* o puntos *estacionarios* de f. Todos los extremos relativos de una función derivable f que son puntos *interiores* del dominio de f son puntos estacionarios. Por lo tanto: *un máximo absoluto o un mínimo absoluto de la función coincide ya sea con un punto crítico de la función o bien con un punto frontera de su dominio.* Para localizar los máximos absolutos (mínimos) de la función debemos solamente comparar los valores de f en los puntos críticos y en los dos puntos frontera o extremos del intervalo, y ver cuáles de estos valores son los mayores (menores). Si f no es derivable en un número finito de puntos, se deberá sólo sumar estos puntos a la lista de localizaciones posibles de un extremo y también verificar los valores de f en estos puntos. Así, la labor principal en la determinación de los extremos de una función se reduce a la de encontrar los ceros de la derivada de la función, que usualmente existen en número finito.

[1] La definición rigurosa de un punto de *máximo relativo*, ξ, establecería que existe un intervalo abierto que contiene a ξ tal que $f(\xi) \geq f(x)$ para todo x de este intervalo para el cual f esté definida.

Para tomar un ejemplo sencillo, determinemos los valores máximos y mínimos de la función $f(x) = \frac{1}{10}x^6 - \frac{3}{10}x^2$ en el intervalo $-2 \leq x \leq 2$. Aquí los puntos críticos, las raíces de la ecuación $f'(x) = 6(x^5 - x)/10 = 0$, están localizados en $x = 0, +1, -1$. Calculando los valores de f en estos puntos y también en los puntos extremos del intervalo encontramos

x	-2	-1	0	1	2
$f(x)$	5.2	-0.2	0	-0.2	5.2

Es claro que los puntos $x = \pm 1$ representan mínimos relativos, mientras que los máximos relativos ocurren en $x = 0$ y $x = \pm 2$. El valor máximo de la función, alcanzado en los puntos extremos del intervalo, es 5.2; el valor mínimo, alcanzado en los puntos $x = \pm 1$, es -0.2 (véase Fig. 3.17).

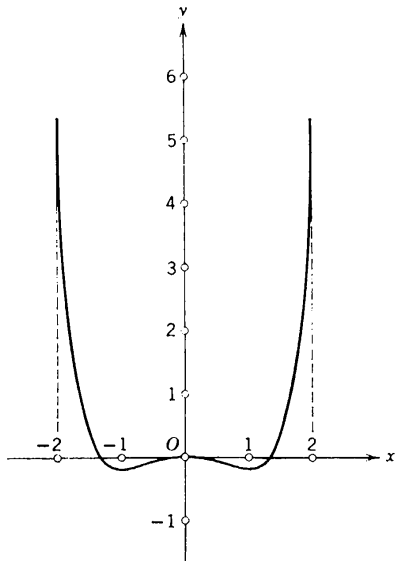


Figura 3.17 $y = (x^6 - 3x^2)/10$.

Sin apelar a la intuición, se puede probar fácilmente mediante métodos puramente analíticos que $f'(\xi) = 0$ siempre y cuando ξ es un punto extremo relativo en el interior del dominio de f, suponiendo f derivable en ξ. (Comparar las consideraciones exactamente análogas para el teorema de Rolle,

p. 196.) Si la función $f(x)$ posee un máximo relativo en el punto ξ, entonces para todos los valores suficientemente pequeños de h diferentes de cero la expresión $f(\xi + h) - f(\xi)$ debe ser negativa o cero. Por lo tanto,

$$\frac{[f(\xi + h) - f(\xi)]}{h} \leq 0$$

para $h > 0$, mientras que

$$\frac{[f(\xi + h) - f(x)]}{h} \geq 0$$

para $h < 0$. Así, si h tiende a cero a través de valores positivos el límite no puede ser positivo, mientras que si h tiende a cero a través de valores negativos el límite no puede ser negativo. Sin embargo, ya que se ha supuesto que la derivada en ξ existe, estos dos límites deben ser iguales entre sí y, precisamente, iguales al valor $f'(\xi)$ que, por consiguiente, solamente puede ser cero; o sea, debe tenerse necesariamente $f'(\xi) = 0$. Una demostración análoga se cumple para un mínimo relativo. La demostración prueba también que si el punto extremo izquierdo $\xi = a$ es un punto máximo relativo (mínimo), entonces por lo menos $f'(a) \leq 0\,[f'(a) \geq 0]$; si el punto extremo derecho b es un punto máximo relativo (mínimo) entonces $f'(b) \geq 0\,[f'(b) \leq 0]$.

La condición $f'(\xi) = 0$ que caracteriza a los puntos críticos no es de ninguna manera suficiente para que existan extremos relativos. Pueden existir puntos en los cuales la derivada se anula, esto es, en los cuales la tangente es horizontal, aun cuando la curva no posee máximo relativo ni mínimo relativo ahí. Esto ocurre si en el punto dado la curva posee una *tangente de inflexión horizontal* que la corta, como en el ejemplo de la función $y = x^3$ en el punto $x = 0$.

El siguiente teorema da las condiciones necesarias y suficientes bajo las cuales un punto crítico es un punto de máximo o mínimo relativo (y, por tanto, un *criterio* para reconocerlo). Se aplica a una función continua, f, que posee una derivada continua f' que se anula en a lo más un número finito de puntos, o, más generalmente, a funciones derivables f para las cuales f' cambia de signo en a lo más un número finito de puntos:

La función $f(x)$ posee un extremo relativo en un punto interior ξ de su dominio si y sólo si la derivada $f'(x)$ cambia de signo cuando x pasa a través de este punto; en particular, la función posee un mínimo relativo si cerca de ξ la derivada es negativa a la izquierda de ξ y positiva a la derecha, mientras que en el caso contrario posee un máximo.

Se demuestra esto rigurosamente utilizando el teorema del valor medio. Primero se observa que a la izquierda y a la derecha de ξ existen intervalos, $\xi_1 < x < \xi$ y $\xi < x < \xi_2$, en cada uno de los cuales $f'(x)$ posee un solo signo, puesto que f' se anula solamente en un número finito de puntos.

(Aquí ξ_1 y ξ_2 pueden ser tomados como los puntos más cercanos a ξ en los cuales f' se anula, si tales puntos existen.) Si los signos de $f'(x)$ en estos dos intervalos son diferentes, entonces $f(\xi + h) - f(\xi) = hf'(\xi + \theta h)$ posee el mismo signo para todos los valores suficientemente pequeños de h, ya sea que h se tome valor positivo o negativo; de modo que ξ es un extremo relativo. Si $f'(x)$ posee el mismo signo en ambos intervalos, entonces $hf'(\xi + \theta h)$ cambia de signo cuando h lo hace, de modo que $f(\xi + h)$ es mayor que $f(\xi)$ en un lado y menor que $f(\xi)$ en el otro lado, y así no existe valor extremo. Nuestro teorema está así probado.

Al mismo tiempo se ve que *el valor $f(\xi)$ es el valor máximo o mínimo de la función en todo intervalo que contiene al punto ξ, en el cual f es derivable y en el cual el único cambio de signo de $f'(x)$ ocurre en ξ mismo.*

El teorema del valor medio, en el que esta demostración se basa, puede seguirse utilizando aún en el caso en que $f(x)$ no es derivable en un punto extremo del intervalo en el cual se aplica, siempre que sea derivable en todos los otros puntos del intervalo; por lo tanto, esta demostración sigue siendo válida si $f'(x)$ no existe en $x = \xi$. Por ejemplo, la función $y = |x|$ posee un mínimo en $x = 0$, puesto que $y' > 0$ para $x > 0$ e $y' < 0$ para $x < 0$ (ver Fig. 2.24, p. 188). La función $y = \sqrt[3]{x^2}$ posee del mismo modo un mínimo en el punto $x = 0$, aun cuando su derivada $\frac{2}{3}x^{-1/3}$ es infinita ahí (ver Fig. 2.27, p. 190).

El método más sencillo para decidir si un punto crítico ξ es un máximo o mínimo relativo involucra la *segunda derivada* en este punto. Es intuitivamente claro que si $f'(\xi) = 0$ entonces f tiene un máximo relativo en ξ si $f''(\xi) < 0$, y tiene un mínimo relativo si $f''(\xi) > 0$. Pues en el primer caso la curva en la vecindad de este punto está situada completamente por debajo de la tangente y en el segundo caso completamente por encima de la tangente. Este resultado se obtiene analíticamente a partir del criterio (teorema) anterior siempre que $f(x)$ y $f'(x)$ sean continuas y que $f''(\xi)$ exista. En efecto, si $f'(\xi) = 0$ y, digamos $f''(\xi) > 0$, se tiene

$$f''(\xi) = \lim_{h \to 0} \frac{f'(\xi + h) - f'(\xi)}{h} = \lim_{h \to 0} \frac{f'(\xi + h)}{h} > 0.$$

Se sigue que $f'(\xi + h)/h > 0$ para todos los $h \neq 0$ que son suficientemente pequeños en valor absoluto; por lo tanto, $f'(\xi + h)$ y h poseen el mismo signo en una vecindad de ξ. Si x está cerca de ξ, la derivada $f'(x)$ debe ser negativa para x a la izquierda de ξ y positiva para x a la derecha de ξ; y esto implica que existe un mínimo relativo en ξ.

La situación es particularmente simple en el caso en que $f''(x)$ es de uno y el mismo signo en todo el intervalo $[a, b]$ en el cual f está definida:

Un punto ξ en el cual f' se anula es un punto de máximo de f si $f''(x) < 0$ en todo el intervalo (o bien si su curva es cóncava), y es un

punto de mínimo de f si en todo el intervalo $f''(x) > 0$ *(esto es, si la curva es convexa).*[1]

En efecto, si $f''(x) < 0$ la función $f'(x)$ es monótona decreciente, y, por lo tanto, ξ es su único cero. Además, $f' > 0$ para $a \leq x < \xi$, mientras que $f' < 0$ para $\xi < x \leq b$. Por el teorema del valor medio esto implica nuevamente que $f(x) < f(\xi)$ para $x \neq \xi$, de modo que ξ resulta ser un estricto punto de máximo. El mínimo de f debe coincidir con uno de los puntos frontera, a, b, puesto que no existe otro punto crítico aparte de ξ. El mismo argumento se aplica cuando $f'' > 0$ en el intervalo.

Ejemplos

Ejemplo 1. De entre todos los triángulos con base y área dadas determinar aquél con mínimo perímetro.

Para resolver este problema se toma al eje x a lo largo de la base dada AB y el punto medio de AB como el origen (Fig. 3.18). Si C es el vértice del triángulo, h su altura (la cual está determinada por el área y la base), y (x, h) son las coordenadas del vértice, entonces la suma de los dos lados AC y BC del triángulo está dada por

$$f(x) = \sqrt{(x+a)^2 + h^2} + \sqrt{(x-a)^2 + h^2},$$

donde $2a$ es la longitud de la base. De aquí se obtiene

$$f'(x) = \frac{x+a}{\sqrt{(x+a)^2 + h^2}} + \frac{x-a}{\sqrt{(x-a)^2 + h^2}},$$

$$f''(x) = \frac{-(x+a)^2}{\sqrt{[(x+a)^2 + h^2]^3}} + \frac{1}{\sqrt{(x+a)^2 + h^2}} + \frac{-(x-a)^2}{\sqrt{[(x-a)^2 + h^2]^3}} + \frac{1}{\sqrt{(x-a)^2 + h^2}}$$

$$= \frac{h^2}{\sqrt{[(x+a)^2 + h^2]^3}} + \frac{h^2}{\sqrt{[(x+a)^2 + h^2]^3}}$$

Vemos de inmediato que: (1) $f'(0)$ se anula, y (2) $f''(x)$ es siempre positiva; por lo tanto, en $x = 0$ existe un valor mínimo (ver pp. 264 y 265). Este valor mínimo está dado según esto por el triángulo isósceles.

Análogamente se encuentra que de entre todos los triángulos con perímetro y base dados el triángulo isósceles posee el área máxima.

Ejemplo 2. Encontrar un punto sobre una recta dada tal que la suma de sus distancias a dos puntos fijos sea un mínimo.

[1] *También se dice, brevemente, que el punto es un "máximo (mínimo) de* $f(x)$*" (N. del R.).*

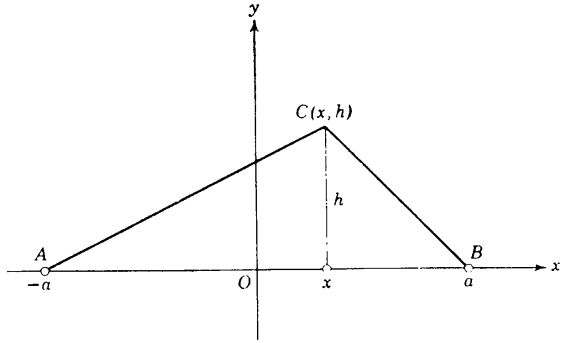


Figura 3.18.

Supónganse dados la recta y dos puntos fijos cualesquiera A y B en el mismo lado de la recta. Se desea encontrar un punto P sobre la línea recta, tal que la distancia $PA + PB$ tenga el mínimo valor posible.[1]

Se toma la recta dada como el eje x y se utiliza la notación de la Fig. 3.19. Entonces, la distancia en cuestión está dada por

$$f(x) = \sqrt{x^2 + h^2} + \sqrt{(x-a)^2 + h_1^2},$$

y se obtiene

$$f'(x) = \frac{x}{\sqrt{x^2 + h^2}} + \frac{x-a}{\sqrt{(x-a)^2 + h_1^2}},$$

$$f''(x) = \frac{h^2}{\sqrt{(x^2 + h^2)^3}} + \frac{h_1^2}{\sqrt{[(x-a)^2 + h_1^2]^3}}.$$

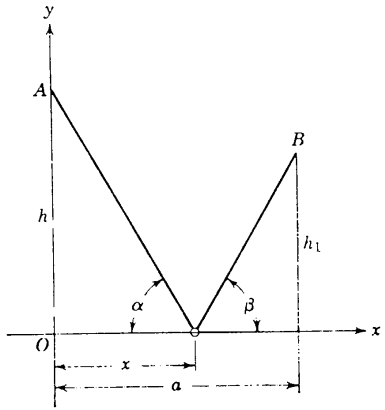


Figura 3.19 Ley de reflexión.

[1] Si A y B están situados en lados opuestos de la recta, obviamente P es, precisamente, la intersección de la recta con el segmento AB.

La ecuación $f'(\xi) = 0$ significa

$$\frac{\xi}{\sqrt{\xi^2 + h^2}} = \frac{a - \xi}{\sqrt{(\xi - a)^2 + h_1^2}},$$

o bien

$$\cos\alpha = \cos\beta;$$

y, por lo tanto, las dos rectas PA y PB deben formar ángulos iguales con la recta dada. El signo positivo de $f''(x)$ muestra que realmente tenemos un valor mínimo.

La solución de este problema está conectada muy de cerca con la ley de reflexión óptica. Por un principio importante de la óptica, conocido

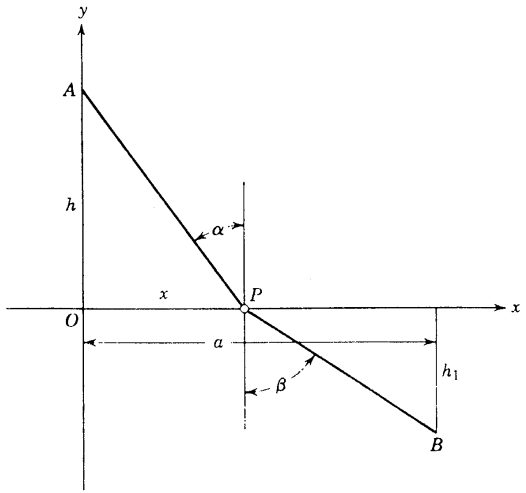


Figura 3.20 Ley de refracción.

como *"principio del tiempo mínimo"* de Fermat, la trayectoria de un rayo de luz está determinada por la propiedad de que el tiempo que tarda la luz en ir de un punto A a un punto B bajo las condiciones dadas debe ser el menor posible. Si se impone la condición de que un rayo de luz en su recorrido de A a B pase por algún punto de una recta dada (digamos, sobre un espejo), se ve que el tiempo mínimo se logrará a lo largo del rayo para el cual el "ángulo de incidencia" sea igual al "ángulo de reflexión".

Ejemplo 3. La Ley de Refracción.[1] Considérense dados dos puntos A y B en lados opuestos del eje x. ¿Cuál es la trayectoria de A a B que

[1] En tanto que los ejemplos anteriores pueden ser tratados también mediante la geometría elemental, el presente no puede ser fácilmente resuelto sin el cálculo.

requiere el menor tiempo posible si la velocidad en un lado del eje x es c_1 y en el otro lado es c_2?

Claramente, esta "trayectoria mínima" debe consistir de dos porciones de líneas rectas que se encuentran una a la otra en un punto P sobre el eje x. Utilizando la notación de la Fig. 3.20 se obtienen las dos expresiones $\sqrt{h^2 + x^2}$ y $\sqrt{h_1^2 + (a - x)^2}$ para las longitudes PA, PB, respectivamente; y el tiempo, T, de viaje a lo largo de esta trayectoria se encuentra dividiendo las longitudes de los dos segmentos por las correspondientes velocidades y después sumando:

$$T = f(x) = \frac{1}{c_1}\sqrt{h^2 + x^2} + \frac{1}{c_2}\sqrt{h_1^2 + (a - x)^2}.$$

Diferenciando, se obtiene

$$f'(x) = \frac{1}{c_1}\frac{x}{\sqrt{h^2 + x^2}} - \frac{1}{c_2}\frac{a - x}{\sqrt{h_1^2 + (a - x)^2}},$$

$$f''(x) = \frac{1}{c_1}\frac{h^2}{\sqrt{(h^2 + x^2)^3}} + \frac{1}{c_2}\frac{h_1^2}{\sqrt{[h_1^2 + (a - x)^2]^3}}.$$

Como fácilmente se ve por la Fig. 3.20, la ecuación $f'(x) = 0$, esto es, la ecuación

$$\frac{1}{c_1}\frac{x}{\sqrt{h^2 + x^2}} = \frac{1}{c_2}\frac{a - x}{\sqrt{h_1^2 + (a - x)^2}},$$

es equivalente a la condición $(1/c_1)\ \text{sen}\,\alpha = (1/c_2)\ \text{sen}\,\beta$, o bien,

$$\frac{\text{sen}\,\alpha}{\text{sen}\,\beta} = \frac{c_1}{c_2}.$$

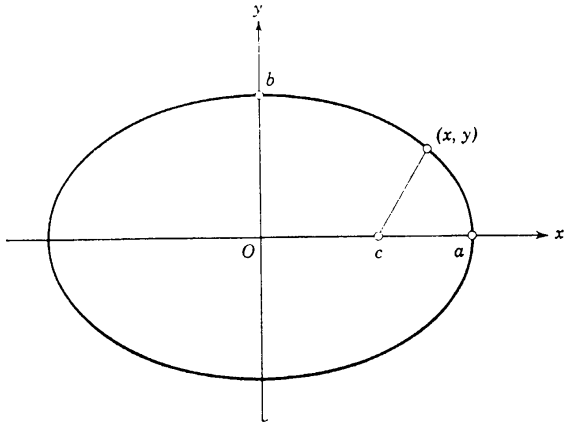


Figura 3.21 Punto de una elipse que está a la mínima distancia de otro sobre el eje mayor.

El lector debe verificar el hecho de que existe sólo *un* punto que satisface la condición, y que este punto proporciona en realidad el valor mínimo requerido.

El significado físico de nuestro ejemplo está precisamente dado por el principio óptico del tiempo mínimo. Un rayo de luz que viaja entre dos puntos describe una trayectoria de tiempo mínimo. Si c_1 y c_2 son las velocidades de la luz en dos medios ópticos distintos, la trayectoria luminosa estará dada por nuestro resultado, el cual es una forma de la *Ley de Snell de la refracción.*

Ejemplo 4. Encontrar el punto de una elipse que está más cerca que cualquier otro de un punto dado sobre su eje mayor (Fig. 3.21).

Tomando la ecuación de la elipse en la forma

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1 \qquad (b < a),$$

y el punto dado sobre el eje mayor como $(c, 0)$, se encuentra para la distancia de cualquier punto (x, y) sobre la elipse al punto $(c, 0)$ la expresión

$$d = \sqrt{(x - c)^2 + b^2(1 - x^2/a^2)},$$

donde $-a \leq x \leq a$. La función $f(x) = d^2$ es convexa $(f'' > 0)$. Posee un mínimo para el mismo x que d. El único punto crítico de f está en $x = c/(1 - b^2/a^2)$. Si este punto está situado en el dominio de d, representa un punto de mínimo; si no, el mínimo de d corresponde al punto extremo del eje mayor que esté más cercano a c. De acuerdo con esto, se encuentran para la mínima distancia los valores

$$d = b\sqrt{1 - \frac{c_2}{a^2 - b^2}} \qquad \text{si} \quad |c| \leq a\left(1 - \frac{b^2}{a^2}\right),$$

$$d = a - |c| \qquad \text{si} \quad |c| \geq a\left(1 - \frac{b^2}{a^2}\right).$$

*3.7 El orden de magnitud de las funciones

Las diferencias entre los comportamientos de las funciones para valores grandes del argumento conducen a la noción de *orden de magnitud.* Debido a su gran importancia este asunto merece una breve discusión aquí, aunque no esté directamente conectado con la idea de la integral o de la derivada.

a. El concepto de orden de magnitud. Los casos más simples

Si la variable x crece más allá de toda cota, entonces, para $\alpha > 0$, las funciones x^α, $\log x$, e^x, $e^{\alpha x}$ crecen también más allá de toda cota. Estas, sin

embargo, se incrementan en formas esencialmente diferentes. Por ejemplo, la función x^3 se torna "infinita en un orden superior" [1] al de x^2; y por esto se entiende que: conforme x crece, el cociente x^3/x^2 se incrementa más allá de toda cota. Análogamente, la función x^α se torna infinita en un orden superior al de x^β si $\alpha > \beta > 0$, etc.

En forma general se dice de dos funciones, $f(x)$ y $g(x)$, cuyos valores absolutos se incrementan con x más allá de toda cota, que $f(x)$ se *torna infinita de un orden superior al de* $g(x)$ si para $x \to \infty$ el cociente $|f(x)/g(x)|$ se incrementa más allá de toda cota; y diremos que $f(x)$ se torna infinita en un orden menor que el de $g(x)$ si el cociente $|f(x)/g(x)|$ tiende a cero conforme x crece; y se dirá que las dos funciones se tornan *infinitas del mismo orden de magnitud* si, al crecer x, el cociente $|f(x)/g(x)|$ posee un límite diferente de cero o, al menos, permanece entre dos cotas fijas positivas. Por ejemplo, la función $ax^3 + bx^2 + c = f(x)$, donde $a \neq 0$, será del mismo orden de magnitud que la función $x^3 = g(x)$, pues el cociente $|f(x)/g(x)| = |(ax^3 + bx^2 + c)/x^3|$ posee el límite $|a|$ conforme $x \to \infty$. Por otra parte, la función $x^3 + x + 1$ se torna infinita de un orden de magnitud superior al de la función $x^2 + x + 1$.

Una suma de dos funciones $f(x)$ y $\phi(x)$, donde $f(x)$ es de orden de magnitud superior al de $\phi(x)$, posee el mismo orden de magnitud que $f(x)$.

Pues $|(f(x) + \phi(x))/f(x)| = |1 + \phi(x)/f(x)|$, y, por hipótesis, esta expresión tiende a uno cuando x crece más allá de toda cota.

b. Los órdenes de magnitud de la función exponencial y del logaritmo

Se podría estar tentado a medir el orden de magnitud de las funciones mediante una escala, asignando a la cantidad x el orden de magnitud uno y a la potencia x^α $(\alpha > 0)$ el orden de magnitud α. Un polinomio de grado n tendría entonces obviamente el orden de magnitud n; una función racional, el grado de cuyo numerador es superior en h al del denominador, tendría el orden de magnitud h.

Salta a la vista, sin embargo, que cualquier intento para describir el orden de magnitud de funciones arbitrarias mediante tal escala debe fallar. Esto es así porque existen funciones que se tornan infinitas en un orden superior al de la potencia x^α de x, no importa cuán grande se elija α; además, existen funciones que se tornan infinitas en un orden menor que el de la potencia x^α, no importa cuán pequeño se elija el número positivo α. Estas funciones, por lo tanto, no encajan en nuestra escala.

Sin entrar en una teoría detallada, se establece el siguiente teorema.

[1] O "infinita de orden superior" al de x^2.

TEOREMA. *Si a es un número arbitrario mayor que uno, entonces el cociente a^x/x tiende a infinito cuando x crece.*

DEMOSTRACIÓN. Para probar esto se construye la función

$$\phi(x) = \log \frac{a^x}{x} = x \log a - \log x;$$

y es obvio que será suficiente mostrar que $\phi(x)$ crece más allá de toda cota si x tiende a $+\infty$. Con este objeto, se considera la derivada

$$\phi'(x) = \log a - \frac{1}{x}$$

y se observa que para $x \geq c = 2/\log a$ no es menor que el número positivo $\frac{1}{2} \log a$. Por lo tanto, se sigue que para $x \geq c$ es

$$\phi(x) - \phi(c) = \int_c^x \phi'(t)\, dt \geq \int_c^x \tfrac{1}{2} \log a\, dt \geq \tfrac{1}{2}(x - c) \log a,$$

$$\phi(x) \geq \phi(c) + \tfrac{1}{2}(x - c) \log a,$$

y el miembro derecho se torna infinito para $x \to \infty$.

Se da en seguida una segunda demostración de este importante teorema: con $\sqrt{a} = b = 1 + h$, se tiene $b > 1$ y $h > 0$. Sea n el entero tal que $n \leq x \leq n + 1$; y puede tomarse $x > 1$, de modo que $n \geq 1$. Aplicando el lema de la p. 87 se tiene

$$\sqrt{\frac{a^x}{x}} = \frac{b^x}{\sqrt{x}} = \frac{(1+h)^x}{\sqrt{x}} > \frac{(1+h)^n}{\sqrt{n+1}} > \frac{1 + nh}{\sqrt{n+1}} > \frac{nh}{\sqrt{2n}} = \frac{h}{\sqrt{2}}\sqrt{n},$$

de modo que

$$\frac{a^x}{x} > \frac{h^2}{2} \cdot n,$$

y, por lo tanto, el cociente tiende a infinito con x.

De este hecho que se acaba de probar se siguen muchos otros. Por ejemplo: para todo índice positivo α y todo número $a > 1$, el cociente a^x/x^α tiende a infinito cuando x crece. Esto es, vale el siguiente

TEOREMA. *La función exponencial se torna infinita de un orden de magnitud superior al de cualquier potencia de x.*[1]

Para demostrarlo sólo se necesita mostrar que la α-ésima raíz de la expresión, esto es,

$$\frac{a^{x/\alpha}}{\alpha} = \frac{1}{\alpha} \frac{a^{x/\alpha}}{x/\alpha} = \frac{1}{\alpha} \cdot \frac{a^y}{y} \qquad \left(y = \frac{x}{\alpha}\right),$$

[1] También se dice frecuentemente; por brevedad: "de orden superior a cualquier potencia de x". Otra expresión equivalente es: "a^x crece a infinito *más rápidamente* que x^α cuando $x \to \infty$".

tiende a infinito. Esto, sin embargo, se sigue inmediatamente a partir del teorema anterior cuando x es reemplazado por $y = x/\alpha$.

De un modo análogo se demuestra el siguiente teorema. Para todo valor positivo de α el cociente $(\log x)/x^\alpha$ tiende a cero cuando $x \to \infty$; esto es

TEOREMA. *El logaritmo se torna infinito de un orden menor de magnitud que el de cualquier potencia positiva arbitrariamente pequeña de x.*[1]

DEMOSTRACIÓN. La demostración se obtiene de inmediato si se expresa $\log x = y$ y así el cociente es transformado en $y/e^{\alpha y}$. Se hace entonces $e^\alpha = a$; entonces $a > 1$, y el cociente y/a^y se aproxima a cero cuando y tiende a infinito. Puesto que y se aproxima a infinito conforme x lo hace, el teorema está demostrado.[2]

Sobre la base de estos resultados pueden construirse funciones de un orden de magnitud muy superior al de la función exponencial, y otras funciones de un orden de magnitud bastante más pequeño que el del logaritmo. Por ejemplo, la función $e^{(e^x)}$ es de un orden superior al de la función exponencial, y la función $\log \log x$ es de un orden menor que el del logaritmo; además puede iterarse este proceso tantas veces como se quiera, reiterando los símbolos u operaciones $e^{(\ )}$ o log tantas veces como se desee.

Todas las funciones x, $\log x$, $\log (\log x)$, $\log [\log (\log x)]$, etc., eventualmente se tornan arbitrariamente grandes para x suficientemente grande, pero con creciente lentitud. Tomando, por ejemplo, para x el enorme número $x = 10^{100}$ se encuentra que $\log x$ es más o menos 230, mientras que $\log (\log x)$ es sólo 5.4, aproximadamente.

c. Comentarios generales

Estas consideraciones muestran que no es posible asignar a todas las funciones números definidos, como órdenes de magnitud, de modo que de dos funciones, aquélla con el orden de magnitud más alto tenga un número superior. Si, por ejemplo, la función x es de orden de magnitud uno y la función $x^{1+\varepsilon}$ de orden de magnitud $1 + \varepsilon$, entonces la función $x \log x$ debe ser de un orden de magnitud mayor que uno y menor que $1 + \varepsilon$, no importa cuán pequeño sea elegido ε. Pero no existe tal número.

Además, es fácil ver que las funciones no tienen que poseer un orden relativo de magnitud claramente definido. Por ejemplo, la función

[1] También: "crece *menos rápidamente que* x^α, cuando $x \to \infty$, por pequeño que sea α". (Cf. Nota al pie de la página anterior.) *(N. del R.)*

[2] Puede sugerirse otra demostración sencilla: Para $x > 1$ y $\varepsilon > 0$

$$\log x = \int_1^x \frac{d\xi}{\xi} < \int_1^x \xi^{\varepsilon - 1}\, d\xi = \frac{1}{\varepsilon}(x^\varepsilon - 1) \cdot$$

y, si se elige ε igual a α y se dividen ambos miembros de esta desigualdad por x^α, entonces se sigue que $(\log x)/x^\alpha \to 0$ cuando $x \to \infty$.

$[x^2(\operatorname{sen} x)^2 + x + 1]/[x^2(\cos x)^2 + x]$ no se aproxima a un límite definido cuando x crece; por el contrario, para $x = n\pi$ (donde n es un entero) el valor es $1/n\pi$, mientras que para $x = (n + \frac{1}{2})\pi$ es $(n + \frac{1}{2})\pi + 1 + 1/(n + \frac{1}{2})\pi$. Aunque tanto el numerador como el denominador se tornan infinitos, el cociente no permanece entre cotas positivas, ni tiende a cero ni tiende a infinito. El numerador, por lo tanto, no es del mismo orden que el denominador, ni de orden mayor o menor. Esta situación aparentemente alarmante significa solamente que las definiciones anteriores no están diseñadas de tal forma que se puedan comparar dos funciones cualesquiera. Este no es un defecto. No se desea aquí comparar los órdenes de tales funciones como el numerador y el denominador mencionados; pues el conocimiento del valor de uno de ellos no da ninguna información útil acerca del otro.

d. El orden de magnitud de una función en la vecindad de un punto arbitrario

Así como se pueden comparar los comportamientos de funciones para $x \to \infty$, pueden compararse también funciones que se tornan infinitas en el punto finito $x = \xi$.

Se dice que la función $f(x) = 1/|x - \xi|$ se torna *infinita del primer orden* en el punto $x = \xi$, y, correspondientemente, que la función $1/|x - \xi|^\alpha$ se torna *infinita de orden* α, siempre que α sea positiva.

Se reconoce entonces que la función $e^{1/|x-\xi|}$ se torna infinita de orden superior para $x \to \xi$, y la función $\log|x - \xi|$ infinita de menor orden que el de todas estas potencias; esto es, que las relaciones límite

$$\lim_{x\to\xi} (|x - \xi|^\alpha \cdot e^{1/|x-\xi|}) = \infty \qquad \text{y} \qquad \lim_{x\to\xi} (|x - \xi|^\alpha \cdot \log|x - \xi|) = 0$$

se cumplen.

Para confirmar esto, solamente se hace $1/|x - \xi| = y$; y nuestros asertos se reducen entonces al conocido teorema de la p. 271, puesto que

$$|x - \xi|^\alpha \cdot e^{1/|x-\xi|} = \frac{e^y}{y^\alpha} \qquad \text{y} \qquad |x - \xi|^\alpha \cdot \log|x - \xi| = -\frac{\log y}{y^\alpha},$$

y que y crece más allá de toda cota conforme x tiende a ξ. (El método de reducir el comportamiento en un punto ξ al comportamiento en el infinito mediante la substitución $1/|x - \xi| = y$ resulta ser frecuentemente útil.)

e. El orden de magnitud (pequeñez) de una función que tiende a cero

Tal como se buscó describir la aproximación de una función al infinito por medio del concepto de orden de magnitud, puede especificarse también

la forma en la cual una función se aproxima a cero. Se dice que cuando $x \to \infty$ la cantidad $1/x$ *se anula de (con) primer orden;* y que la cantidad $x^{-\alpha}$, donde α es positiva, se *anula de orden* α.[1] Se encuentra nuevamente que la *función* $1/\log x$ *se anula con orden menor que el de una potencia arbitraria* $x^{-\alpha}$; esto es, para todo α positivo la relación

$$\lim_{x\to\infty} (x^{-\alpha} \cdot \log x) = 0$$

se cumple.

De la misma manera se dice que para $x = \xi$ la cantidad $x - \xi$ se anula en primer orden y la cantidad $|x - \xi|^\alpha$ en orden α. Con nuestros resultados es fácil probar las relaciones

$$\lim_{x\to 0} (|x|^\alpha \cdot \log|x|) = 0, \qquad \lim_{x\to 0} (|x|^{-\alpha} \cdot e^{-1/|x|}) = 0,$$

las cuales son expresadas usualmente como sigue:

La función $1/\log|x|$ *se anula cuando* $x \to 0$ *con un orden menor que el de cualquier potencia positiva de* x*; y la función exponencial* $e^{-1/|x|}$ *se anula con un orden superior al de cualquier potencia positiva de* x.[2]

f. Las notaciones "O" y "o" para órdenes de magnitud

Una forma conveniente de indicar que una función $f(x)$ es de menor orden de magnitud que una función $g(x)$ es la de escribir $f = o(g)$. Esta ecuación simbólica significa solamente que el cociente f/g posee el límite cero, y puede ser utilizada con igual ventaja para funciones que se anulan o se tornan infinitas y para argumentos x que tienden a infinito o se aproximan a un valor ξ.[3]

Muchos de los resultados de la seccción anterior pueden reescribirse en esta notación; por ejemplo,

$x^\alpha = o(x^\beta)$	para $\alpha < \beta$	conforme $x \to \infty$
$\log x = o(x^\alpha)$	para $\alpha > 0$	conforme $x \to \infty$
$e^{-x} = o(x^{-\alpha})$	conforme $x \to \infty$	
$e^{-1/x} = o(x^\alpha)$	conforme $x \to 0$	con valores positivos
$\log\|x\| = o(1/x)$	conforme $x \to 0$	
$1 - \cos x = o(x)$	conforme $x \to 0$.	

[1] También se dice que "$1/x$ *es infinitésimo de primer orden* y $x^{-\alpha}$ $(\alpha > 0)$ *es infinitésimo de orden* α". *(N. del R.)*

[2] O sea: "La función $1/\log|x|$, cuando $x \to 0$, se torna un *infinitésimo de orden inferior* al de cualquier potencia $x^\alpha(\alpha > 0)$; y la función $e^{-1/|x|}$ se torna un *infinitésimo de orden superior* al de cualquier potencia $x^\alpha(\alpha > 0)$". *(N. del R.)*

[3] La letra o es elegida para sugerir la palabra "orden". Obsérvese que la relación $f = o(g)$ para una g que se anula significa que f se anula con un orden superior.

Esta notación, introducida por E. Landau, es útil para indicar el orden de magnitud del error en una fórmula de aproximación. Por ejemplo,

$$\frac{1}{\sqrt{1+4x^2}} = \frac{1}{2x} + o\left(\frac{1}{x}\right) \qquad \text{para } x \to \infty$$

es la notación para la relación

$$\lim_{x\to\infty} \frac{\dfrac{1}{\sqrt{1+4x^2}} - \dfrac{1}{2x}}{1/x} = 0.$$

Análogamente, la relación entre incremento y diferencial de una función f que posee derivada en un punto x puede ser escrita en la forma

$$f(x+h) - f(x) = hf'(x) + o(h) \qquad \text{para } h \to 0.$$

Igualmente útil es la notación simbólica $f = O(g)$ usada para indicar que $f(x)$ es *a lo más* del orden de magnitud de $g(x)$, esto es, que el cociente $f(x)/g(x)$ *está acotado* para los valores de x en cuestión.[1] El uso del símbolo O es, además, muy flexible. Así, la frase "$f = O(g)$ para $x \to \infty$" significa que el cociente f/g está acotado para todo x suficientemente grande, como en

$$\sqrt{10x-1} = O(\sqrt{x}) \qquad \text{para } x \to \infty.$$

Análogamente, "$f = O(g)$ para $x \to \xi$" significa que f/g está acotada en una vecindad suficientemente pequeña del punto $x = \xi$, como en

$$e^x - 1 = O(x) \qquad \text{para } x \to 0.$$

Más generalmente, puede utilizarse la ecuación $f = O(g)$ para indicar la acotación de f/g en cualquier dominio del eje x sin que se requiera que x se aproxime a un límite. Así

$$\log x = O(x) \qquad \text{para } x > 1,$$

$$x = O(\operatorname{sen} x) \qquad \text{para } |x| < \frac{\pi}{2}.$$

Algunos de los primeros ejemplos que involucran el símbolo o pueden ser ahora refinados indicándose una mejor estimación del error con la ayuda del símbolo O. Así se tiene para una función f para la cual f'' está definida y es continua:

$$f(x+h) - f(x) = hf'(x) + O(h^2) \qquad \text{para } h \to 0.$$

[1] Nótese que $f = O(g)$ *no* significa que f/g tenga el límite uno o que el cociente tenga necesariamente un límite, cualquiera que sea.

Otros ejemplos son

$$\frac{1}{\sqrt{1+4x^2}} = \frac{1}{2x} + O\left(\frac{1}{x^2}\right),$$

$$\cos x = 1 + O(x^2) \qquad \text{para todo } x.$$

Las mismas notaciones pueden utilizarse para *sucesiones* a_n, con el índice n que tienda a infinito. En lo siguiente se encontrarán algunos ejemplos interesantes de tales fórmulas "asintóticas" con un término de error de orden superior (ver la fórmula de Stirling para $n!$ Cap. 6, A.1). Una famosa ley asintótica,[1] ya mencionada en el capítulo 1, p. 80, afirma que el número $\pi(n)$ de primos menores que n está dado aproximadamente por $n/(\log n)$. Aquí el orden de magnitud del error ha sido encontrado también y se tiene, más precisamente, el resultado

$$\pi(n) = \frac{n}{\log n} + O\left(\frac{n}{\log^2 n}\right).$$

Apéndice

La dificultad para apreciar un desarrollo riguroso del cálculo surge de un dilema básico: Aunque los procedimientos y conceptos fundamentales como continuidad, suavidad, etc., son motivados por apremiantes necesidades intuitivas, deben hacerse precisos para que tengan significado lógico y las resultantes definiciones rigurosas puedan cubrir fenómenos que están más allá de aquéllos de carácter intuitivo. Así, el concepto riguroso de continuidad requiere inevitablemente un grado de abstracción que no está totalmente reflejado en la noción intuitiva de una curva conexa; y el concepto de derivabilidad es más restrictivo y más abstracto que la vaga idea que sugiere la *suavidad* de una curva. Discrepancias de este tipo son inevitables y pueden arruinar la paciencia y el entendimiento de un principiante o de alguien para quien la sutileza lógica no es de principal interés. Sin embargo, deseamos hacer más clara para el lector la necesidad de dicha precisión, mostrando que, quizá inesperadamente, la precisión y el refinamiento son requeridos aún para los ejemplos más sencillos e intuitivamente comprensibles.

A.1 Algunas funciones especiales

Como regla, tales ejemplos pueden no estar dados en términos de expresiones analíticas sencillas (véanse Figs. 1.28, p. 62 y 1.30, p. 63). Aquí,

[1] La demostración no puede darse en este libro. Véase: A. E. Ingham, *The Distribution of Primes*, Cambridge University Press, 1932.

sin embargo, se desea representar varios tipos de discontinuidades y fenómenos inesperados o "anormales" mediante expresiones muy sencillas construidas a partir de las funciones elementales. Se comienza con un ejemplo en el cual no se presenta ninguna discontinuidad.

a. La función $y = e^{-1/x^2}$

Esta función (ver Fig. 3.22) está definida en primera instancia solamente para valores de x diferentes de cero, y obviamente posee el límite cero cuando $x \to 0$. En efecto, mediante la transformación $1/x^2 = \xi$ la función se vuelve $y = e^{-\xi}$, y $\lim e^{-\xi} = 0$. Por lo tanto, es natural extender la función de modo que sea continua para $x = 0$ definiendo el valor de la función en el punto $x = 0$ como $y(0) = 0$.

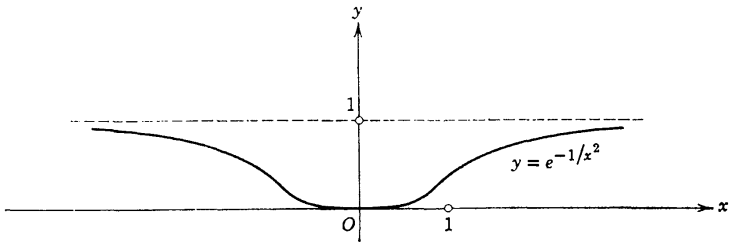


Figura 3.22.

Por la regla de la cadena, la derivada de la función para $x \neq 0$ es $y' = -(2/x^3)e^{-1/x^2} = 2\xi^{3/2}e^{-\xi}$. Si x tiende a cero esta derivada posee también el límite cero, como se deduce inmediatamente de la p. 271. En el punto $x = 0$ la derivada

$$y'(0) = \lim_{h\to 0}\frac{y(h) - y(0)}{h} = \lim_{h\to 0}\frac{e^{-1/h^2}}{h}$$

puede ser también definida en forma continua como cero.

Para las derivadas superiores en $x \neq 0$, obviamente se obtiene siempre el producto de la función e^{-1/x^2} y un polinomio en $1/x$, y el paso al límite $x \to 0$ conduce siempre al límite cero. Por lo tanto, todas las derivadas superiores se anulan, como y', en el punto $x = 0$.

Así, la función es continua en todas partes y derivable tantas veces como se desee; y, aún más, en el punto $x = 0$ se anula con todas sus derivadas, pero no se anula idénticamente. Posteriormente se podrá observar (Apéndice I.1 del capítulo 5) cuán notable o "anormal" es este comportamiento.

b. La función $y = e^{-1/x}$

Como se ve fácilmente, para valores positivos de x esta función se comporta de la misma manera que la función que acabamos de tratar: si x tiende a cero por un lado, con valores positivos, la función tiende a cero, y lo mismo es cierto para todas sus derivadas. Si se define el valor de la función en $x = 0$ como $y(0) = 0$, todas las derivadas laterales derechas en el punto $x = 0$ poseen el valor cero. Otra situación muy diferente se tiene cuando x tiende a cero a través de valores negativos; pues entonces la función y todas sus derivadas se vuelven infinitas, y las derivadas laterales izquierdas no existen en el punto $x = 0$. En el punto $x = 0$, por lo tanto, la función posee una clase notable de discontinuidad bastante diferente de las discontinuidades infinitas de las funciones racionales consideradas en las pp. 60, 61 (ver Fig. 3.23).

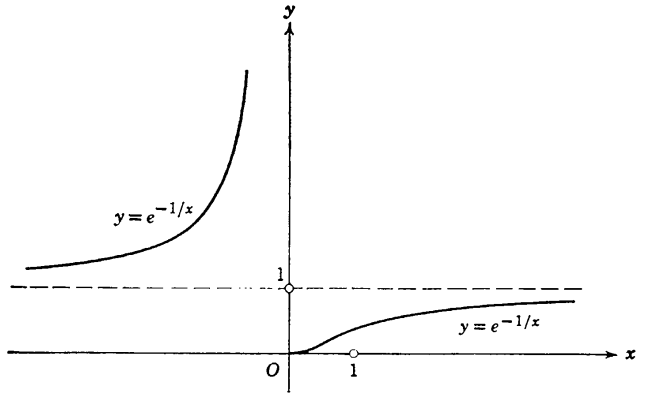


Figura 3.23.

c. La función $y = \tanh\dfrac{1}{x}$

Como se ha visto ya en la p. 88, las funciones con discontinuidades de "salto" pueden obtenerse a partir de funciones elementales mediante el paso al límite. La función exponencial, definida en la p. 172, junto con el principio de la composición de funciones, nos da otro método para construir funciones con tales discontinuidades a partir de funciones elementales, sin ningún otro proceso límite adicional. Un ejemplo de esto es la función

$$y = \tanh\frac{1}{x} = \frac{e^{1/x} - e^{-1/x}}{e^{1/x} + e^{-1/x}}$$

y su comportamiento en el punto $x = 0$. La función en primera instancia no está definida en este punto. Si x se acerca al punto $x = 0$ a través de valores positivos, obviamente se obtiene el límite 1; si, por otra parte, se aproxima al punto $x = 0$ a través de valores negativos, se obtiene el límite -1. Este punto $x = 0$ es así un punto de discontinuidad de salto; conforme x crece pasando por 0 el valor de la función salta dos unidades (ver Fig. 3.24). Por otra parte, la derivada

$$y' = -\frac{1}{\cosh^2(1/x)}\frac{1}{x^2}$$

$$= -\frac{1}{x^2}\frac{4}{(e^{1/x} + e^{-1/x})^2}$$

se aproxima al límite cero desde ambos lados, como surge fácilmente [1] a partir de la sección 3.7, p. 270.

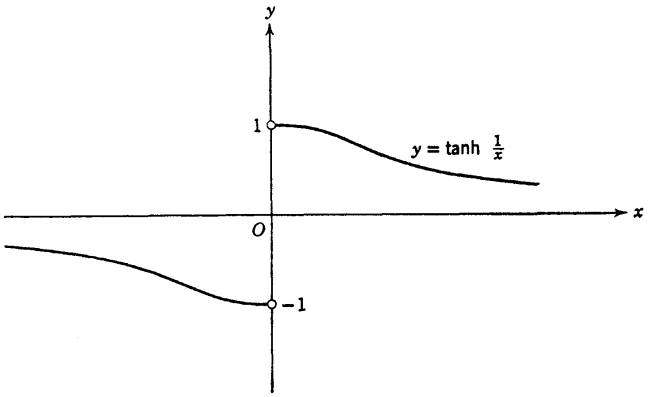


Figura 3.24.

d. La función $y = x \tanh 1/x$

En el caso de la función

$$y = x \tanh\frac{1}{x} = x\frac{e^{1/x} - e^{-1/x}}{e^{1/x} + e^{-1/x}}$$

la discontinuidad anterior es eliminada por el factor x. Esta función tiene el límite cero cuando $x \to 0$ desde cualquier lado, de modo que se puede nuevamente definir en forma apropiada $y(0)$ como igual a cero. Esta función es entonces continua en $x = 0$, pero su primera derivada

[1] Otro ejemplo donde aparece una discontinuidad "de salto" está dado por la función $y = \arctan 1/x$ cuando $x \to 0$.

$$y' = \tanh\frac{1}{x} - \frac{1}{x}\frac{1}{\cosh^2(1/x)}$$

posee justamente la misma clase de discontinuidad del ejemplo anterior. La gráfica de la función es una curva con una esquina (ver Fig. 3.25); y en el punto $x = 0$ la función no posee una derivada, sino una derivada lateral derecha con el valor $+1$ y una derivada lateral izquierda con el valor -1.

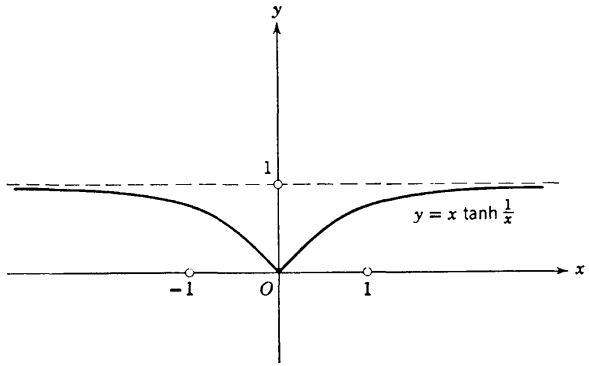


Figura 3.25.

e. La función $y = x \operatorname{sen} 1/x$, $y(0) = 0$

Se ha visto ya que esta función no está compuesta por un número finito de partes monótonas—o, lo que es lo mismo, no es "seccionalmente" monótona o monótona "por pedazos"—pero que, sin embargo, es continua (p. 63 y Fig. 1.31). Su primera derivada

$$y' = \operatorname{sen}\frac{1}{x} - \frac{1}{x}\cos\frac{1}{x} \qquad (x \neq 0),$$

por el contrario, posee una discontinuidad en $x = 0$; pues conforme x tiende a cero esta derivada oscila continuamente entre curvas acotadas, una positiva y una negativa, las cuales a su vez tienden a $+\infty$ y $-\infty$ respectivamente. En el punto $x = 0$ el cociente incremental es $[y(h) - y(0)]/h = \operatorname{sen}(1/h)$; y puesto que este valor oscila hacia arriba y hacia abajo entre 1 y -1 un número infinito de veces conforme $h \to 0$, la función no posee una derivada lateral derecha ni tampoco una derivada lateral izquierda en $x = 0$.

A.2 Comentarios sobre la derivabilidad de funciones

La derivada de una función que es continua y posee derivada en cada punto de un intervalo puede no ser continua.

Como un ejemplo sencillo considérese la función dada por

$$y = f(x) = x^2 \operatorname{sen} \frac{1}{x} \qquad \text{para } x \neq 0$$

y

$$f(0) = 0.$$

Esta función está definida y es continua en todas partes. Para todos los valores de x diferentes de cero la derivada está dada por la expresión

$$f'(x) = -x^2 \left(\cos \frac{1}{x} \right) \frac{1}{x^2} + 2x \operatorname{sen} \frac{1}{x} = -\cos \frac{1}{x} + 2x \operatorname{sen} \frac{1}{x}.$$

Cuando x tiende a cero $f'(x)$ no posee límite. Si, por otra parte, se forma el cociente incremental $[f(h) - f(0)]/h = (h^2 \operatorname{sen} 1/h)/h = h \operatorname{sen} 1/h$, se ve de inmediato que éste tiende a cero cuando h lo hace. La derivada, por lo tanto, existe para $x = 0$ y posee el valor 0.

Para entender intuitivamente la razón de este comportamiento paradójico se representa la función gráficamente (ver Fig. 3.26). Esta oscila entre las curvas $y = x^2$ e $y = -x^2$ a las cuales toca alternadamente. Así, las razones de las alturas de las crestas de onda de la curva y sus distancias al

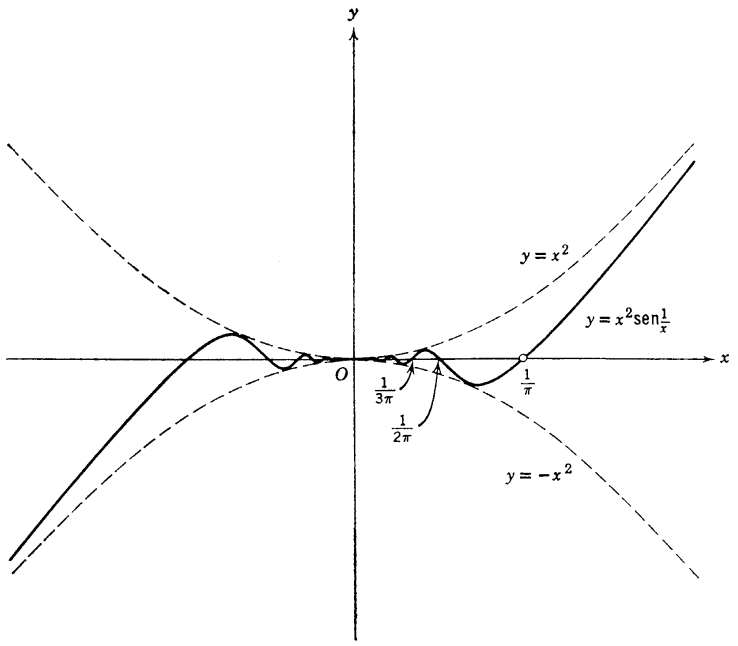


Figura 3.26.

origen se tornan gradualmente más pequeñas. Sin embargo, estas ondas no se tornan más planas pues su pendiente, dada por la derivada

$$f'(x) = 2x \operatorname{sen} 1/x - \cos 1/x,$$

es igual a -1 en los puntos $x = 1/2\ n\pi$, donde $\cos 1/x = 1$, e igual a $+1$ en los puntos $x = 1/(2n + 1)\pi$, donde $\cos 1/x = -1$.

En contraste con la posibilidad ilustrada aquí (de que una derivada existe en todas partes y sin embargo no es continua), se establece el siguiente sencillo teorema, el cual ilumina toda una serie de ejemplos y discusiones anteriores.

TEOREMA. *Si se sabe que en una vecindad de un punto $x = a$ la función $f(x)$ es continua, que para $x \neq a$ también posee derivada $f'(x)$ y si, además, la ecuación $\lim_{x \to a} f'(x) = b$ se cumple, entonces la derivada $f'(x)$ existe también en el punto a y $f'(a) = b$.*

DEMOSTRACIÓN. La demostración se obtiene inmediatamente a partir del teorema del valor medio. Pues se tiene $[f(a + h) - f(a)]/h = f'(\xi)$, donde ξ es un valor intermedio entre a y $a + h$. Si h tiende ahora a cero, por hipótesis $f'(\xi)$ tiende a b, y la tesis surge de inmediato.

Un teorema análogo puede ser probado de un modo similar: Si la función $f(x)$ es continua en $a \leq x \leq b$ y para $a < x < b$ posee una derivada que crece más allá de toda cota conforme x tiende a a, el cociente incremental derecho $[f(a + h) - f(a)]/h$ crece también más allá de toda cota conforme h tiende a cero, de modo que no existe derivada lateral derecha para $x = a$. El significado geométrico de este aserto es que en el punto con coordenadas (finitas) $[a, f(a)]$ la curva posee una tangente vertical.

Parte B Técnicas de integración

Funciones explícitas

Una amplia clase de funciones puede ser construida a partir de las funciones elementales [1] aplicando repetidamente operaciones racionales, esto es, la adición, multiplicación, división, y, además, mediante las operaciones de formar funciones inversas y funciones compuestas. Las funciones así descriptas constituyen la clase de *funciones "explícitas"* o "expresiones exactas" finitas.[2] Como un resultado de la **Parte A** de este capítulo se establece la propiedad bastante general:

Toda función explícita puede derivarse y su derivada es también una función explícita.

Se ha logrado así un dominio bastante completo de la operación o "algoritmo" de la diferenciación. No obstante, el proceso inverso, el de integra-

[1] Se deberá destacar que la distinción entre funciones "elementales" y "explícitas" y otras es, de por sí, algo arbitraria. Para nosotros el término función "elemental" incluye justamente las funciones racionales, las funciones trigonométricas y las exponenciales, y sus inversas.

[2] Este nombre indica que se encontrarán muchas otras funciones que no pueden ser representadas de esta forma pero que pueden ser construidas por medio de procesos de límite tales como las series infinitas.

ción, es, hablando generalmente, más importante y representa el mayor problema. Hasta cierto grado el problema es abordado y resuelto mediante el teorema fundamental del cálculo: A cada fórmula de diferenciación $F'(x) = f(x)$ le corresponde una fórmula equivalente para las funciones primitivas $F(x)$ de $f(x)$, o sea, la integral:

$$\int f(x)\,dx = F(x).$$

(Más precisamente, se tiene $F(x) = \int_a^x f(u)\,du$ + constante). Así, mientras más fórmulas explícitas de diferenciación sean deducidas, más funciones explícitas podrán ser integradas en términos de funciones explícitas. Una primera tabla de integrales se da en la p. 284. En principio, no sería difícil, aunque sí impráctico y confuso, extender mucho más dicha tabla.

En las primeras fases del desarrollo del cálculo muchos matemáticos trataron de encontrar, en forma explícita o exacta, la función integral o primitiva para toda función dada explícitamente.

Pasó algún tiempo hasta que se comprendió que, en principio, este problema no puede ser resuelto. Por el contrario, para algunos integrandos bastante elementales la integral no puede ser expresada en términos de funciones elementales (véase Sec. 3.14*a*). Así, la necesidad de estudiar nuevos tipos de funciones generadas mediante procesos de integración a partir de funciones elementales se convirtió en un estímulo importante para el desarrollo del análisis. Sin embargo, el deseo de integrar—cuando es factible—funciones explícitas dadas explícitamente, sin meterse en los enredos desesperados de la consulta tediosa de tablas o de los cálculos numéricos, ha conducido a algunos recursos sencillos que proporcionan una cierta flexibilidad para transformar integrales dadas. En efecto, estos recursos permiten efectuar la integración mediante la reducción a alguna de las integrales elementales en la Tabla de Integrales.

La sección 3.9 será dedicada al desarrollo de tales útiles recursos. En relación con esto, el principiante deberá estar precavido en contra de memorizar simplemente las muchas fórmulas obtenidas usando estos recursos técnicos. En vez de ello, el estudiante deberá dirigir su esfuerzo hacia la obtención de una comprensión clara de los *métodos* de integración y el aprendizaje de cómo aplicarlos. Más aún, deberá recordar que aun cuando la integración mediante estos recursos sea imposible, la *integral sí existe* (al menos para todas las funciones continuas), y puede ser realmente calculada con un grado tan alto de precisión como se desee por medio de los métodos numéricos que se desarrollarán posteriormente (sección 6.1).

La Parte C de este capítulo se ocupará de extender los conceptos de integración e integral, en forma separada del problema de la técnica de la integración.

Tabla de integrales elementales

	$F'(x) = f(x)$	$F(x) = \int f(x)\,dx$
1.	$x^a \quad (a \neq -1)$	$\dfrac{x^{a+1}}{a+1}$
2.	$\dfrac{1}{x}$	$\log \lvert x \rvert$
3.	e^x	e^x
4.	$a^x \quad (a \neq 1)$	$\dfrac{a^x}{\log a}$
5.	$\operatorname{sen} x$	$-\cos x$
6.	$\cos x$	$\operatorname{sen} x$
7.	$\dfrac{1}{\operatorname{sen}^2 x} \quad (= \operatorname{cosec}^2 x)$	$-\cot x$
8.	$\dfrac{1}{\cos^2 x} \quad (= \sec^2 x)$	$\tan x$
9.	$\operatorname{senh} x$	$\cosh x$
10.	$\cosh x$	$\operatorname{senh} x$
11.	$\dfrac{1}{\operatorname{senh}^2 x} \quad (= \operatorname{cosech}^2 x)$	$-\coth x$
12.	$\dfrac{1}{\cosh^2 x} \quad (= \operatorname{sech}^2 x)$	$\tanh x$
13.	$\dfrac{1}{\sqrt{1-x^2}} \quad (\lvert x \rvert < 1)$	$\begin{cases} \operatorname{arc\,sen} x \\ -\operatorname{arc\,cos} x \end{cases}$
14.	$\dfrac{1}{1+x^2}$	$\begin{cases} \operatorname{arc\,tan} x \\ -\operatorname{arc\,cot} x \end{cases}$
15.	$\dfrac{1}{\sqrt{1+x^2}}$	$\operatorname{arc\,senh} x = \log\,(x + \sqrt{1+x^2})$
16.	$\dfrac{1}{\pm\sqrt{x^2-1}} \quad (\lvert x \rvert > 1)$	$\operatorname{arc\,cosh} x = \log\,(x \pm \sqrt{x^2-1})$
17.	$\dfrac{1}{1-x^2}$ $\begin{cases} \lvert x \rvert < 1 \\ \lvert x \rvert > 1 \end{cases}$	$\operatorname{arc\,tanh} x = \dfrac{1}{2}\log\dfrac{1+x}{1-x}$ $\operatorname{arc\,coth} x = \dfrac{1}{2}\log\dfrac{x+1}{x-1}$

3.8 Tabla de integrales elementales

A cada una de las fórmulas de derivación demostradas anteriormente le corresponde una fórmula equivalente de integración. Puesto que estas integrales elementales son utilizadas una y otra vez como materiales para el arte de la integración, se han reunido en una tabla. La columna derecha contiene un número de funciones elementales y la columna izquierda las correspondientes derivadas. Si se lee la tabla de izquierda a derecha, se obtiene en la columna derecha una integral indefinida de la función que aparece en la columna izquierda.

Se le recuerdan también al lector los teoremas fundamentales del cálculo diferencial e integral, demostrados en la sección 2.9, en particular, el hecho de que cualquier integral definida es obtenida a partir de la integral indefinida $F(x)$ mediante la fórmula [1]

$$\int_a^b f(x)\,dx = F(x)\Big|_a^b = F(b) - F(a).$$

En las secciones siguientes se intentará reducir el cálculo de integrales de funciones dadas de una u otra manera al de las integrales elementales coleccionadas en nuestra tabla. Aparte de algunos artificios especiales que son aprendidos solamente por experiencia, esta reducción está basada esencialmente en dos métodos útiles: "substitución" e "integración por partes". Cada uno de estos métodos permite transformar de muchas maneras una integral dada; y el objeto de tales transformaciones en su mayor parte es el de reducir la integral dada, en un paso o en una sucesión de pasos, a una o más de las fórmulas elementales de integración dadas arriba.

3.9 El método de substitución

Integración de funciones compuestas

El primero de estos métodos consiste en la *introducción de una nueva variable* (esto es, el método de *substitución* o *transformación*). Su objeto es el de reducir la integración de funciones compuestas (tales como funciones de $x - c$ o de $ax + b$) a la de funciones más sencillas.

a. La fórmula de substitución. Integral de una función compuesta

La regla para integrar funciones compuestas se obtiene a partir de la correspondiente regla de la cadena para derivar. Para una función compuesta $G(u) = F[\phi(u)]$ se tiene (véase p. 239)

[1] En este capítulo no se discutirá el problema algo diferente de calcular integrales definidas especiales sin encontrar primero una función primitiva general.

$$\frac{dG(u)}{du} = \frac{dF[\phi(u)]}{du} = F'[\phi(u)]\phi'(u). \tag{16}$$

Para la validez de esta fórmula es suficiente que las funciones $x = \phi(u)$ y $F(x)$ sean continuamente derivables respecto a sus argumentos u, x, respectivamente, y que $F(x)$ esté definida para los valores de x tomados por la función $x = \phi(u)$ (esto es, el rango de la función ϕ debe pertenecer al dominio de F). Integrando la fórmula entre los límites $u = \alpha$ y $u = \beta$ se encuentra

$$G(\beta) - G(\alpha) = F[\phi(\beta)] - F[\phi(\alpha)] = \int_\alpha^\beta F'[\phi(u)]\phi'(u)\,du. \tag{17}$$

Si aquí se hace

$$\phi(\beta) = b, \qquad \phi(\alpha) = a,$$

se tiene

$$F[\phi(\beta)] - F[\phi(\alpha)] = F(b) - F(a) = \int_a^b F'(x)\,dx.$$

Haciendo $F'(x) = f(x)$ se obtiene la *fórmula fundamental de substitución*

$$\int_a^b f(x)\,dx = \int_\alpha^\beta f[\phi(u)]\phi'(u)\,du, \qquad x = \phi(u), \tag{18}$$

o sea, escrita sugestivamente en la notación de Leibnitz con la diferencial $d\phi = \phi'(x)\,dx$,

$$\int f(x)\,dx = \int f(\phi)\,d\phi. \tag{18a}$$

Aquí $x = \phi(u)$ puede ser cualquier función que esté definida y posea una derivada continua en el intervalo J con puntos extremos α y β; y dicha función transforma estos puntos extremos en $x = a$ y $x = b$ respectivamente. Se supone que la función $f(x)$ es continua en un intervalo I que contiene las imágenes de todos los puntos de J bajo la transformación ϕ. Por $F(x)$ puede tomarse *cualquier función primitiva de* $f(x)$.

Como se habrá notado, la regla de la substitución (18) no requiere que la transformación $x = \phi(u)$ transforme puntos entre α y β solamente en puntos entre a y b, o sea, que valores diferentes u sean transformados en diferentes x; todo lo que importa es que α y β sean transformados en a y b y que $f(x)$ esté definida para los valores x tomados por $\phi(u)$ para u entre α y β.

En términos de integrales indefinidas la regla de la substitución adquiere la forma

$$G(u) = \int f[\phi(u)]\phi'(u)\,du = \int f(x)\,dx = F(x) = F[\phi(u)]. \tag{19}$$

Los símbolos diferenciales

$$\phi'(u)\,du = \frac{dx}{du}du \qquad \text{y} \qquad dx$$

se vuelven idénticos si formalmente se cancelan los símbolos du en el numerador y el denominador.

Ejemplos. Se aplica la fórmula (18) al integrando $f(x) = 1/x$ y se hace la substitución $x = \phi(u)$, suponiendo $\phi(u) \neq 0$ en el intervalo considerado; entonces

$$\int \frac{\phi'(u)}{\phi(u)}\,du = \int \frac{dx}{x} = \log|x| = \log|\phi(u)|,$$

o bien, cambiando el nombre de la variable u nuevamente por x,

$$\int \frac{\phi'(x)}{\phi(x)}\,dx = \log|\phi(x)|. \tag{20}$$

Si en esta importante fórmula se substituyen funciones particulares, tales como $\phi(x) = \log x$, $\phi(x) = \operatorname{sen} x$, o $\phi(x) = \cos x$, se obtiene [1]

$$\int \frac{dx}{x \log x} = \log|\log x|,$$

(21)

$$\int \cot x\,dx = \log|\operatorname{sen} x|, \qquad \int \tan x\,dx = -\log|\cos x|.$$

Otros ejemplos

1) $$\int \phi(u)\phi'(u)\,du = \int x\,dx = \tfrac{1}{2}x^2 = \tfrac{1}{2}[\phi(u)]^2,$$

donde $f(x) = x$. De aquí se obtiene para $\phi(u) = \log u$:

$$\int \frac{\log u}{u}\,du = \tfrac{1}{2}(\log u)^2. \tag{22}$$

2) Finalmente se considera

$$\int \operatorname{sen}^n u \cos u\,du.$$

Aquí $x = \operatorname{sen} u = \phi(u)$, y, por lo tanto,

$$\int \operatorname{sen}^n u \cos u\,du = \int x^n\,dx = \frac{x^{n+1}}{n+1} = \frac{\operatorname{sen}^{n+1} u}{n+1}.$$

[1] Estas y las siguientes fórmulas son verificadas fácilmente mostrando que la derivada del resultado da otra vez el integrando.

3) La misma substitución $x = \operatorname{sen} u$ da, para cualquier función $f(x)$ continua en el intervalo $-1 \leq x \leq 1$,

$$\int_\alpha^\beta f(\operatorname{sen} u)\cos u\,du = \int_{\operatorname{sen}\alpha}^{\operatorname{sen}\beta} f(x)\,dx.$$

Tomando aquí $\alpha = 0$ y $\beta = 2\pi$, resulta un ejemplo para aplicar la fórmula de substitución a un caso en que la función transformación $x = \phi(u) = \operatorname{sen} u = x$ no es monótona en el intervalo $\alpha \leq u \leq \beta$. Se encuentra

$$\int_0^{2\pi} f(\operatorname{sen} u)\cos u\,du = \int_0^0 f(x)\,dx = 0.$$

Otras formas de la regla

En muchas aplicaciones la integral que ha de evaluarse está dada en la forma

$$F(u) = \int h[\phi(u)]\,du$$

en la cual el integrando aparece como la función compuesta $h[\phi(u)]$ sin el factor adicional $\phi'(u)$. Puede aplicarse la regla de substitución (18) si se logra escribir el integrando $h[\phi(u)]$ en la forma $f[\phi(u)]\phi'(u)$. Esto puede ser logrado siempre bajo la hipótesis de que la función $x = \phi(u)$ posee una derivada continua $\phi'(u)$ *que no se anula.* Porque entonces existe una función inversa $u = \psi(x)$ con derivada continua $du/dx = \psi'(x) = 1/\phi'(u)$. Tomando por $f(x)$ la función $h(x)\psi'(x)$, se tiene así $h[\phi(u)] = f[\phi(u)]/\psi'(x) = f[\phi(u)]\phi'(u)$ y se obtiene, a partir de la regla de substitución,

$$\begin{aligned}\int h[\phi(u)]\,du &= \int f[\phi(u)]\phi'(u)\,du = \int f(x)\,dx \\ &= \int h(x)\psi'(x)\,dx = \int h(x)\frac{du}{dx}dx.\end{aligned} \tag{23}$$

La hipótesis $\phi'(u) \neq 0$ ha sido introducida para evitar que la expresión dx/du en la fórmula (23) se torne infinita.

El principiante jamás deberá olvidar que al substituir $\psi(x)$ por u en una integral no se debe expresar simplemente la antigua variable x en términos de la nueva, u, y después integrar con respecto a esta nueva variable; en vez de esto, *antes* de integrar se debe multiplicar el integrando por la derivada de la variable original x con respecto a la nueva variable u. Esto, por supuesto, es sugerido por la notación de Leibnitz $h\,dx = h\frac{dx}{du}du$. En la integral definida

$$\int_a^b h[\psi(x)]\,dx = \int_\alpha^\beta h(u)\phi'(u)\,du$$

no debe olvidarse cambiar los límites a, b, para x, por los límites correspondientes $\alpha = \psi(a)$ y $\beta = \psi(b)$ para la variable u.

Ejemplos. Para calcular $\int \operatorname{sen} 2x\, dx$ se toma $u = \psi(x) = 2x$ y $h(u) = \operatorname{sen} u$. Se tiene entonces

$$\frac{du}{dx} = \psi'(x) = 2, \quad \frac{dx}{du} = \frac{1}{2}.$$

Si se introduce ahora $u = 2x$ en la integral como la nueva variable, entonces aquélla se transforma, *no* en $\int \operatorname{sen} u\, du$, sino en

$$\frac{1}{2}\int \operatorname{sen} u\, du = -\frac{1}{2}\cos u = -\frac{1}{2}\cos 2x;$$

esto puede, por supuesto, verificarse de inmediato derivando el lado derecho de la ecuación.

Si se integra con respecto a x entre los límites cero y $\pi/4$, los correspondientes límites para $u = 2x$ son cero y $\pi/2$, y se obtiene

$$\int_0^{\pi/4} \operatorname{sen} 2x\, dx = \frac{1}{2}\int_0^{\pi/2} \operatorname{sen} u\, du = -\frac{1}{2}\cos u \,\Big|_0^{\pi/2} = \frac{1}{2}$$

Otro ejemplo sencillo es la integral $\int_1^4 \frac{dx}{\sqrt{x}}$. Aquí se toma $u = \psi(x) = \sqrt{x}$, de donde $x = \phi(u) = u^2$. Puesto que $\phi'(u) = 2u$, se tiene

$$\int_1^4 \frac{dx}{\sqrt{x}} = \int_1^2 \frac{2u\, du}{u} = 2\int_1^2 du = 2.$$

Como otro ejemplo, se considera la integral de $\operatorname{sen} 1/x$ para el intervalo $\frac{1}{2} \leq x \leq 1$. Se tiene para $u = 1/x$ o bien $x = 1/u$, $dx = -du/u^2$; y, por lo tanto,

$$\int_{1/2}^1 \operatorname{sen}\frac{1}{x}\, dx = -\int_2^1 \frac{\operatorname{sen} u}{u^2}\, du = \int_1^2 \frac{\operatorname{sen} u}{u^2}\, du.$$

*b. Una deducción alternativa de la fórmula de substitución

La fórmula de integración (17), con una ligera variante en la notación, puede ser intepretada también de un modo directo basado en el significado de la integral definida como el límite de una suma, en vez de ser deducida a partir de la regla de la cadena de la derivación.[1] Para calcular la integral

$$\int_a^b h[\psi(x)]\, dx$$

(para el caso $a < b$), se parte con una subdivisión arbitraria del intervalo $a \leq x \leq b$ y se hace entonces la subdivisión cada vez más fina. Estas subdivisiones se inician de la siguiente manera. Si se supone que la función $u = \psi(x)$ es monótona creciente, existe una correspondencia biunívoca entre el intervalo $a \leq x \leq b$ sobre el eje x y un intervalo $\alpha \leq u \leq \beta$ de los

[1] El resultado obtenido de esta manera está restringido nuevamente a substituciones monótonas, y así es menos general que la fórmula (18) proporcionada por la regla de la cadena (pág. 286).

valores de $u = \psi(x)$, donde $\alpha = \psi(a)$ y $\beta = \psi(b)$. Se divide este intervalo x en n partes de longitud[1] Δx; existe *una correspondiente* subdivisión del intervalo u en subintervalos que, en general, no son todos de la misma longitud. Los puntos de división del intervalo x se denotan por

$$x_0 = a, x_1, x_2, \ldots, x_n = b,$$

y las longitudes de las correspondientes celdas u por

$$\Delta u_1, \Delta u_2, \ldots, \Delta u_n.$$

La integral que se está considerando es el límite de la suma

$$\sum_{\nu=1}^{n} h\{\psi(\xi_\nu)\}\, \Delta x,$$

donde el valor ξ_ν es seleccionado arbitrariamente en el ν-ésimo subintervalo de la subdivisión x. Esta suma se escribe ahora en la forma $\sum_{\nu=1}^{n} h(v_\nu)\frac{\Delta x}{\Delta u_\nu}\Delta u_\nu$, donde $v_\nu = \psi(\xi_\nu)$. Por el teorema del valor medio del cálculo diferencial, es $\Delta x/\Delta u_\nu = \phi'(\eta_\nu)$, donde η_ν es un valor intermedio elegido adecuadamente para la variable u en el ν-ésimo subintervalo de la subdivisión u, y $x = \phi(u)$ denota la función inversa de $u = \psi(x)$. Si se selecciona ahora el valor ξ_ν de tal manera que v_ν y η_ν coincidan, esto es, $\eta_\nu = \psi(\xi_\nu)$, $\xi_\nu = \phi(\eta_\nu)$, entonces la suma toma la forma

$$\sum_{\nu=1}^{n} h(\eta_\nu)\phi'(\eta_\nu)\, \Delta u_\nu.$$

Si ahora aquí se pasa al límite haciendo $n \to \infty$,[2] se obtiene la expresión

$$\int_\alpha^\beta h(u)\frac{dx}{du}\, du$$

como el valor límite, esto es, como el valor de la integral que se está considerando, de acuerdo con la fórmula (23) dada anteriormente.

Así se llega al siguiente resultado.

TEOREMA. *Sea $h(u)$ una función continua de u en el intervalo $\alpha \leq u \leq \beta$. Si la función $u = \psi(x)$ es continua y monótona y posee una derivada continua du/dx que no se anula en $a \leq x \leq b$, y si, además, $\psi(a) = \alpha$ y $\psi(b) = \beta$,* entonces

$$\int_a^b h\{\psi(x)\}\, dx = \int_a^b h(u)\, dx = \int_\alpha^\beta h(u)\frac{dx}{du}\, du.$$

[1] La hipótesis de que las longitudes de estos subintervalos son iguales no es de ninguna manera esencial para la demostración.

[2] Este límite existe (para $\Delta x \to 0$) y es la integral, puesto que, de acuerdo con la continuidad uniforme de $u = \psi(x)$, la mayor de las longitudes Δu_ν tiende a cero con Δx.

Esta deducción exhibe el mérito sugestivo de la notación de Leibnitz. Para llevar a cabo la substitución $u = \psi(x)$ se requiere escribir solamente $(dx/du)\,du$ en vez de dx, cambiando los límites de los valores originales de x a los correspondientes valores de u.

c. Ejemplos. Fórmulas de integración

Con la ayuda de la regla de substitución puede evaluarse en muchos casos una integral dada, $\int f(x)\,dx$, si se le reduce por medio de una substitución adecuada, $x = \phi(u)$, a una de las integrales fundamentales de nuestra tabla. Si tales substituciones existen y cómo encontrarlas son problemas a los cuales no se les puede dar una respuesta general; y ésta es, más bien, una cuestión en la cual la práctica y el ingenio, en contraste con los métodos sistemáticos, desempeñan el papel más importante.

Como un ejemplo se evalúa la integral $\displaystyle\int \frac{dx}{\sqrt{a^2 - x^2}}$ por medio de la substitución [1] $x = \phi(u) = au$, $u = \psi(x) = x/a$, $dx = a\,du$, mediante la cual, usando el Nº 13 de nuestra tabla, se obtiene

(24)
$$\int \frac{dx}{\sqrt{a^2 - x^2}} = \int \frac{a\,du}{a\sqrt{1 - u^2}} = \operatorname{arc\,sen} u = \operatorname{arc\,sen} \frac{x}{a}, \qquad \text{para } |x| < |a|.$$

Mediante la misma substitución se obtienen

(25) $$\int \frac{dx}{a^2 + x^2} = \int \frac{a\,du}{a^2(1 + u^2)} = \frac{1}{a} \operatorname{arc\,tan} u = \frac{1}{a} \operatorname{arc\,tan} \frac{x}{a},$$

(26) $$\int \frac{dx}{\sqrt{a^2 + x^2}} = \operatorname{ar\,senh} \frac{x}{a},$$

(27) $$\int \frac{dx}{\sqrt{x^2 - a^2}} = \operatorname{ar\,cosh} \frac{x}{a}, \qquad \text{para } |x| > |a|,$$

(28) $$\int \frac{dx}{a^2 - x^2} = \begin{cases} \dfrac{1}{a} \operatorname{ar\,tanh} \dfrac{x}{a} & \text{para } |x| < |a|, \\ \dfrac{1}{a} \operatorname{ar\,coth} \dfrac{x}{a} & \text{para } |x| > |a|, \end{cases}$$

fórmulas que aparecen muy frecuentemente y que pueden ser fácilmente verificadas derivando el lado derecho de cada ecuación.

[1] Por brevedad nos tomamos nuevamente la libertad de escribir los símbolos dx y du separadamente, esto es $dx = \phi'(u)\,du$ en vez de $dx/du = \phi'(u)$ (Ver. p. 201).

3.10 Otros ejemplos del método de substitución

En esta sección se recopila un número de ejemplos que el lector puede considerar cuidadosamente para ejercitación.

Mediante la substitución $u = 1 \pm x^2$, $du = \pm 2x\,dx$, se deduce que

(29) $$\int \frac{x\,dx}{\sqrt{1 \pm x^2}} = \pm\sqrt{1 \pm x^2},$$

(30) $$\int \frac{x\,dx}{1 \pm x^2} = \pm\tfrac{1}{2} \log |1 \pm x^2|.$$

En estas fórmulas debe tomarse en los tres lugares el mismo signo, ya sea el positivo o bien el negativo.

Mediante la substitución $u = ax + b$, $du = a\,dx$ $(a \neq 0)$, se obtiene

(31) $$\int \frac{dx}{ax + b} = \frac{1}{a} \log |ax + b|,$$

(32) $$\int (ax + b)^\alpha\,dx = \frac{1}{a(\alpha + 1)} (ax + b)^{\alpha+1} \qquad (\alpha \neq -1),$$

(33) $$\int \operatorname{sen} (ax + b)\,dx = -\frac{1}{a} \cos (ax + b);$$

y análogamente, por medio de la substitución $u = \cos x$, $du = -\operatorname{sen} x\,dx$, se obtiene

(34) $$\int \tan x\,dx = -\log |\cos x|;$$

y, por medio de la substitución $u = \operatorname{sen} x$, $du = \cos x\,dx$, se tiene

(35) $$\int \cot x\,dx = \log |\operatorname{sen} x|$$

[Ver (21) p. 287]. Utilizando las substituciones análogas $u = \cosh x$ (que da $du = \operatorname{senh} x\,dx$) y $u = \operatorname{senh} x$ $(du = \cosh x\,dx)$, se obtienen las fórmulas

(36) $$\int \tanh x\,dx = \log \cosh x,$$

(37) $$\int \coth x\,dx = \log |\operatorname{senh} x|.$$

Por medio de la substitución $u = (a/b) \tan x$, $du = (a/b) \sec^2 x\,dx$, se llega a las dos fórmulas

(38) $$\int \frac{dx}{a^2 \operatorname{sen}^2 x + b^2 \cos^2 x} = \frac{1}{b^2} \int \frac{1}{(a^2/b^2) \tan^2 x + 1} \frac{dx}{\cos^2 x}$$

$$= \begin{cases} \dfrac{1}{ab} \operatorname{arc\,tan} \left(\dfrac{a}{b} \tan x \right) \\ \dfrac{1}{ab} \operatorname{arc\,cot} \left(\dfrac{a}{b} \tan x \right) \end{cases},$$

y

$$\int \frac{dx}{a^2\operatorname{sen}^2 x - b^2\cos^2 x} = \begin{cases} -\dfrac{1}{ab}\operatorname{ar\,tanh}\left(\dfrac{a}{b}\tan x\right) \\ -\dfrac{1}{ab}\operatorname{ar\,coth}\left(\dfrac{a}{b}\tan x\right) \end{cases} \tag{39}$$

Se evalúa la integral

$$\int \frac{dx}{\operatorname{sen} x}$$

escribiendo $\operatorname{sen} x = 2 \operatorname{sen}(x/2)\cos(x/2) = 2\tan(x/2)\cos^2(x/2)$ y haciendo $u = \tan(x/2)$, de modo que $du = \frac{1}{2}\sec^2(x/2)\,dx$; la integral se vuelve entonces

$$\int \frac{dx}{\operatorname{sen} x} = \int \frac{du}{u} = \log\left|\tan\frac{x}{2}\right|. \tag{40}$$

Si se reemplaza x por $x + \pi/2$, esta fórmula se convierte en

$$\int \frac{dx}{\cos x} = \log\left|\tan\left(\frac{x}{2} + \frac{\pi}{4}\right)\right|. \tag{41}$$

La substitución $u = 2x$ conduce, si se aplican también las conocidas fórmulas trigonométricas $2\cos^2 x = 1 + \cos 2x$ y $2\operatorname{sen}^2 x = 1 - \cos 2x$, a las fórmulas frecuentemente utilizadas

$$\int \cos^2 x\,dx = \tfrac{1}{2}(x + \operatorname{sen} x\cos x) \tag{42}$$

y

$$\int \operatorname{sen}^2 x\,dx = \tfrac{1}{2}(x - \operatorname{sen} x\cos x). \tag{43}$$

Mediante la substitución $x = \cos u$, equivalente a $u = \arccos x$, o, más generalmente, $x = a\cos u$ $(a \neq 0)$, podemos reducir

$$\int \sqrt{1 - x^2}\,dx \quad \text{y} \quad \int \sqrt{a^2 - x^2}\,dx$$

respectivamente a estas fórmulas. Se obtiene así

$$\int \sqrt{a^2 - x^2}\,dx = -\frac{a^2}{2}\arccos\frac{x}{a} + \frac{x}{2}\sqrt{a^2 - x^2}. \tag{44}$$

Análogamente, mediante la substitución $x = a\cosh u$ se obtiene la fórmula

$$\int \sqrt{x^2 - a^2}\,dx = -\frac{a^2}{2}\operatorname{ar\,cosh}\frac{x}{a} + \frac{x}{2}\sqrt{x^2 - a^2}, \tag{45}$$

y mediante la substitución $x = a\operatorname{senh} u$ se tiene

$$\int \sqrt{a^2 + x^2}\,dx = \frac{a^2}{2}\operatorname{ar\,senh}\frac{x}{a} + \frac{x}{2}\sqrt{a^2 + x^2}. \tag{46}$$

La substitución $u = a/x$, $dx = -(a/u^2)\,du$, conduce a las fórmulas

$$\int \frac{dx}{x\sqrt{x^2 - a^2}} = -\frac{1}{a}\arccos\frac{a}{x}, \tag{47}$$

$$\int \frac{dx}{x\sqrt{x^2 + a^2}} = -\frac{1}{a}\operatorname{ar\,senh}\frac{a}{x}; \tag{48}$$

$$\int \frac{dx}{x\sqrt{a^2 - x^2}} = -\frac{1}{a}\operatorname{ar\,cosh}\frac{a}{x}. \tag{49}$$

Finalmente, se consideran las tres integrales

$$\int \operatorname{sen} mx \operatorname{sen} nx\,dx, \qquad \int \operatorname{sen} mx\cos nx\,dx, \qquad \int \cos mx\cos nx\,dx,$$

donde m y n son enteros positivos. Mediante bien conocidas fórmulas trigonométricas puede dividirse cada una de estas integrales en dos partes, escribiendo

$$\operatorname{sen} mx \operatorname{sen} nx = \tfrac{1}{2}[\cos(m - n)x - \cos(m + n)x],$$
$$\operatorname{sen} mx\cos nx = \tfrac{1}{2}[\operatorname{sen}(m + n)x + \operatorname{sen}(m - n)x],$$
$$\cos mx\cos nx = \tfrac{1}{2}[\cos(m + n)x + \cos(m - n)x].$$

Si se hace ahora uso de las substituciones $u = (m + n)x$ y $u = (m - n)x$ respectivamente, se obtiene directamente el siguiente sistema de fórmulas:

$$\int \operatorname{sen} mx \operatorname{sen} nx\,dx = \begin{cases} \dfrac{1}{2}\left\{\dfrac{\operatorname{sen}(m - n)x}{m - n} - \dfrac{\operatorname{sen}(m + n)x}{m + n}\right\} & \text{si } m \neq n, \\ \dfrac{1}{2}\left(x - \dfrac{\operatorname{sen} 2mx}{2m}\right) & \text{si } m = n; \end{cases} \tag{50}$$

$$\int \operatorname{sen} mx\cos nx\,dx = \begin{cases} -\dfrac{1}{2}\left\{\dfrac{\cos(m + n)x}{m + n} + \dfrac{\cos(m - n)x}{m - n}\right\} & \text{si } m \neq n \\ -\dfrac{1}{2}\left(\dfrac{\cos 2mx}{2m}\right) & \text{si } m = n; \end{cases} \tag{51}$$

$$\int \cos mx\cos nx\,dx = \begin{cases} \dfrac{1}{2}\left\{\dfrac{\operatorname{sen}(m + n)x}{m + n} + \dfrac{\operatorname{sen}(m - n)x}{m - n}\right\} & \text{si } m \neq n, \\ \dfrac{1}{2}\left(\dfrac{\operatorname{sen} 2mx}{2m} + x\right) & \text{si } m = n. \end{cases} \tag{52}$$

Si, en particular, se integra de $-\pi$ a $+\pi$, se obtienen de estas fórmulas las relaciones sumamente importantes

$$\int_{-\pi}^{+\pi} \operatorname{sen} mx \operatorname{sen} nx\, dx = \begin{cases} 0 & \text{si } m \neq n, \\ \pi & \text{si } m = n, \end{cases}$$

$$\int_{-\pi}^{+\pi} \operatorname{sen} mx \cos nx\, dx = 0, \tag{53}$$

$$\int_{-\pi}^{+\pi} \cos mx \cos nx\, dx = \begin{cases} 0 & \text{si } m \neq n, \\ \pi & \text{si } m = n. \end{cases}$$

Estas son las *relaciones de ortogonalidad* de las funciones trigonométricas, las cuales se encontrarán nuevamente en la sección 8.4e.

3.11 Integración por partes

a. Fórmula general

El segundo método ampliamente utilizado para tratar problemas de integración expresa en forma integral la regla para derivar un producto,

$$(fg)' = f'g + fg'.$$

La correspondiente fórmula integral es (Ver. p. 210)

$$f(x)g(x) = \int g(x)f'(x)\, dx + \int f(x)g'(x)\, dx,$$

o bien,

$$f(x)g'(x)\, dx = f(x)g(x) - \int g(x)f'(x)\, dx. \tag{54}$$

Utilizando la notación diferencial de Leibnitz, esta expresión se convierte en

$$\int f\, dg = fg - \int g\, df. \tag{54a}$$

Esta fórmula será mencionada como la *fórmula de integración por partes.* Ella reduce el cálculo de una integral al cálculo de otra integral. Puesto que un integrando dado puede ser interpretado como un producto $f(x)g'(x)$ de muchas maneras diferentes, esta fórmula proporciona una herramienta efectiva para la transformación de integrales.

Escrita como fórmula de *integración definida,* la fórmula para integración por partes es

$$\int_a^b f(x)g'(x)\, dx = f(x)g(x)\Big|_a^b - \int_a^b g(x)f'(x)\, dx \tag{54b}$$

$$= f(b)g(b) - f(a)g(a) - \int_a^b g(x)f'(x)\, dx.$$

Esta se obtiene, ya sea directamente integrando la fórmula de la derivada de un producto entre los límites a y b, o bien formando la diferencia de la fórmula (54) en los puntos b y a.

Puede darse una interpretación geométrica sencilla de la fórmula (54b): Supóngase que $y = f(x)$ y $z = g(x)$ son funciones monótonas y que $f(a) = A$, $f(b) = B$, $g(a) = \alpha$, $g(b) = \beta$; puede entonces formarse la inversa de la primera función y substituir en la segunda ecuación, obteniéndose así z como función de y. Se supone que esta función es monótona creciente. Puesto que $dy = f'(x)\, dx$ y $dz = g'(x)\, dx$, la fórmula para integración por partes puede ser escrita [Ver la regla de substitución (18), p. 286] como

$$\int_\alpha^\beta y\, dz + \int_A^B z\, dy = B\beta - A\alpha,$$

en concordancia con la relación que se pone de manifiesto mediante la Fig. 3.27:

área $NQLK$ + área $PMLQ$ = área $OMLK$ − área $OPQN$.

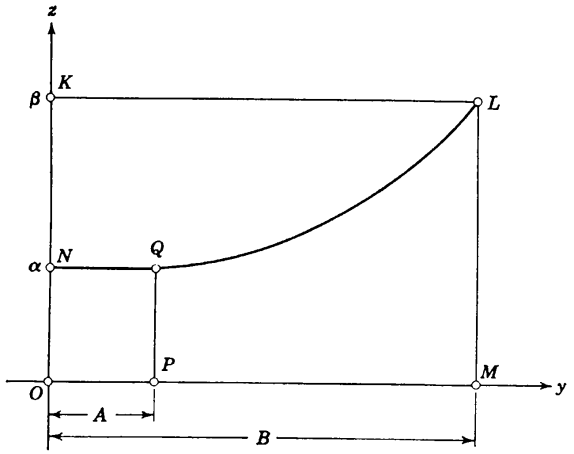


Figura 3.27.

El siguiente ejemplo puede servir como una primera ilustración:

$$\int \log x\, dx = \int \log x \cdot 1\, dx.$$

El integrando se escribe de esta manera para indicar que se hace $f(x) = \log x$ y $g'(x) = 1$, de modo que se tiene $f'(x) = 1/x$ y $g(x) = x$. La fórmula de integración da entonces

$$\int \log x\, dx = x \log x - \int \frac{x}{x}\, dx = x \log x - x. \tag{55}$$

Esta última expresión es, por lo tanto, la integral indefinida del logaritmo, como puede verificarse de inmediato derivando.

b. Otros ejemplos de integración por partes

Con $f(x) = x$, $g'(x) = e^x$, se tiene $f'(x) = 1$, $g(x) = e^x$ y la integral es entonces

$$\int x\, e^x\, dx = e^x(x - 1). \tag{56}$$

De modo análogo se obtiene

$$\int x \operatorname{sen} x\, dx = -x \cos x + \operatorname{sen} x \tag{57}$$

y

$$\int x \cos x\, dx = x \operatorname{sen} x + \cos x. \tag{58}$$

Para $f(x) = \log x$, $g'(x) = x^a$, se tiene la relación

$$\int x^a \log x\, dx = \frac{x^{a+1}}{a+1}\left(\log x - \frac{1}{a+1}\right). \tag{59}$$

Aquí debe suponerse $a \neq -1$. Para $a = -1$ se obtiene

$$\int \frac{1}{x} \log x\, dx = (\log x)^2 - \int \log x \cdot \frac{dx}{x};$$

y pasando la integral del miembro derecho al izquierdo se tiene [Ver (22), p. 287]

$$\int \frac{1}{x} \log x\, dx = \tfrac{1}{2}(\log x)^2. \tag{60}$$

La integral $\int \operatorname{arc\,sen} x\, dx$ se calcula tomando $f(x) = \operatorname{arc\,sen} x$, $g'(x) = 1$. Por tanto,

$$\int \operatorname{arc\,sen} x\, dx = x \operatorname{arc\,sen} x - \int \frac{x\, dx}{\sqrt{1 - x^2}}.$$

La integración en el miembro derecho puede ser efectuada como en (29), p. 292; se encuentra así

$$\int \operatorname{arc\,sen} x\, dx = x \operatorname{arc\,sen} x + \sqrt{1 - x^2}. \tag{61}$$

De la misma manera se encuentra la integral

$$\int \operatorname{arc\,tan} x\, dx = x \operatorname{arc\,tan} x - \tfrac{1}{2} \log (1 + x^2), \tag{62}$$

y muchas otras de un tipo similar.

Los siguientes ejemplos son de una naturaleza algo diferente. Aquí la repetición de la integración por partes lleva nuevamente a la integral original, para la cual se obtiene así una ecuación.

Procediendo en esta forma se obtiene

$$\int e^{ax} \operatorname{sen} bx\, dx = -\frac{1}{b} e^{ax} \cos bx + \frac{a}{b}\int e^{ax} \cos bx\, dx$$

$$= -\frac{1}{b} e^{ax} \cos bx + \frac{a}{b^2} e^{ax} \operatorname{sen} bx - \frac{a^2}{b^2}\int e^{ax} \operatorname{sen} bx\, dx;$$

Resolviendo esta ecuación para la integral $\int e^{ax} \operatorname{sen} bx\, dx$:

$$\int e^{ax} \operatorname{sen} bx\, dx = \frac{1}{a^2 + b^2} e^{ax}(a \operatorname{sen} bx - b \cos bx). \tag{64}$$

De manera análoga se sigue que

$$\int e^{ax} \cos bx\, dx = \frac{1}{a^2 + b^2} e^{ax}(a \cos bx + b \operatorname{sen} bx). \tag{63}$$

c. Fórmula integral para $f(b) + f(a)$

Como un último ejemplo se deduce una fórmula notable que expresa la suma $f(b) + f(a)$ como una integral definida (en vez de la diferencia $f(b) - f(a)$ dada por la fórmula fundamental). La integración por partes será aplicada introduciendo $1 = g'(x)$, donde $g(x) = x - m$ con una constante m a nuestra disposición. Se tiene entonces para las integrales indefinidas

$$\int f(x)\, dx + \int f'(x)(x - m)\, dx = f(x)(x - m),$$

y para la integral entre a y b

$$\int_a^b f(x)\, dx + \int_a^b f'(x)(x - m)\, dx = f(b)(b - m) - f(a)(a - m).$$

Si con a y b arbitrarios se elige para m el valor medio $m = (a + b)/2$ entre a y b, se obtiene, como el lector puede verificar fácilmente,

$$\frac{b - a}{2}[f(a) + f(b)] = \int_a^b f(x)\, dx + \int_a^b (x - m) f'(x)\, dx.$$

d. Fórmulas recursivas

En muchos casos el integrando no sólo es función de la variable independiente sino que depende también de un índice entero n. Al integrar por partes se obtiene en ocasiones, en vez del valor de la integral, otra expresión similar en la cual el índice n posee un valor menor. Puede llegarse así, después de cierto número de pasos, a una integral que puede ser tratada por medio de la tabla de integrales de la p. 284. Tal proceso es llamado *recursivo*.

Los siguientes ejemplos son ilustraciones: Mediante la repetida integración por partes pueden calcularse las integrales trigonométricas

$$\int \cos^n x\,dx, \qquad \int \operatorname{sen}^n x\,dx, \qquad \int \operatorname{sen}^m x \cos^n x\,dx,$$

siempre que m y n sean enteros positivos. Pues utilizando $f(x) = \cos^{n-1} x$, $g(x) = \operatorname{sen} x$, se encuentra para la primera integral que

$$\int \cos^n x\,dx = \cos^{n-1} x \operatorname{sen} x + (n-1)\int \cos^{n-2} x \operatorname{sen}^2 x\,dx;$$

y el miembro derecho puede ser escrito en la forma

$$\cos^{n-1} x \operatorname{sen} x + (n-1)\int \cos^{n-2} x\,dx - (n-1)\int \cos^n x\,dx.$$

Así se ha obtenido una relación recursiva:

$$\int \cos^n x\,dx = \frac{1}{n}\cos^{n-1} x \operatorname{sen} x + \frac{n-1}{n}\int \cos^{n-2} x\,dx. \tag{65}$$

Esta fórmula permite disminuir el índice en el integrando paso por paso hasta que finalmente se llega a la integral

$$\int \cos x\,dx = \operatorname{sen} x \qquad \text{o} \qquad \int dx = x,$$

dependiendo de que n sea impar o par. De un modo análogo se obtienen las fórmulas recursivas similares

$$\int \operatorname{sen}^n x\,dx = -\frac{1}{n}\operatorname{sen}^{n-1} x \cos x + \frac{n-1}{n}\int \operatorname{sen}^{n-2} x\,dx \tag{66}$$

y

$$\int \operatorname{sen}^m x \cos^n x\,dx = \frac{\operatorname{sen}^{m+1} x \cos^{n-1} x}{m+n} + \frac{n-1}{m+n}\int \operatorname{sen}^m x \cos^{n-2} x\,dx. \tag{67}$$

En particular, se calculan las integrales

$$\int \operatorname{sen}^2 x\,dx = \frac{1}{2}(x - \operatorname{sen} x \cos x)$$

y

$$\int \cos^2 x\,dx = \frac{1}{2}(x + \operatorname{sen} x \cos x),$$

como se ha hecho ya por el método de substitución [Ecs. (42), (43), p. 293].

Casi no es necesario mencionar que las integrales correspondientes para las funciones hiperbólicas pueden ser calculadas de exactamente el mismo modo:

$$\int \operatorname{senh}^2 x\,dx = \frac{1}{2}(-x + \operatorname{senh} x \cosh x), \tag{68}$$

$$\int \cosh^2 x\,dx = \frac{1}{2}(x + \operatorname{senh} x \cosh x). \tag{69}$$

Otras fórmulas recursivas están dadas por las siguientes transformaciones:

$$\int (\log x)^m\,dx = x(\log x)^m - m\int (\log x)^{m-1}\,dx, \tag{70}$$

$$\int x^m e^x\,dx = x^m e^x - m\int x^{m-1} e^x\,dx, \tag{71}$$

$$\int x^m \operatorname{sen} x\,dx = -x^m \cos x + m\int x^{m-1} \cos x\,dx, \tag{72}$$

$$\int x^m \cos x\,dx = x^m \operatorname{sen} x - m\int x^{m-1} \operatorname{sen} x\,dx, \tag{73}$$

$$\int x^a (\log x)^m\,dx - \frac{x^{a+1}(\log x)^m}{a+1} - \frac{m}{a+1}\int x^a (\log x)^{m-1}\,dx \quad (a \neq -1). \tag{74}$$

e. El producto infinito de Wallis para π

La fórmula recursiva para la integral $\int \operatorname{sen}^n x\,dx$ con $n > 1$ conduce a una expresión fascinante para el número π como "producto infinito". En la fórmula

$$\int \operatorname{sen}^n x\,dx = -\frac{1}{n}\operatorname{sen}^{n-1} x \cos x + \frac{n-1}{n}\int \operatorname{sen}^{n-2} x\,dx$$

se insertan los límites 0 y $\pi/2$, obteniéndose así

$$\int_0^{\pi/2} \operatorname{sen}^n x\,dx = \frac{n-1}{n}\int_0^{\pi/2} \operatorname{sen}^{n-2} x\,dx \qquad \text{para } n > 1. \tag{75}$$

Si repetidamente se aplica la fórmula recursiva, se obtiene, distinguiendo entre los casos $n = 2m$ y $n = 2m+1$, las integrales

$$\int_0^{\pi/2} \operatorname{sen}^{2m} x\,dx = \frac{2m-1}{2m}\cdot\frac{2m-3}{2m-2}\cdots\frac{1}{2}\cdot\int_0^{\pi/2} dx, \tag{76}$$

$$\int_0^{\pi/2} \operatorname{sen}^{2m+1} x\,dx = \frac{2m}{2m+1}\cdot\frac{2m-2}{2m-1}\cdots\frac{2}{3}\cdot\int_0^{\pi/2} \operatorname{sen} x\,dx, \tag{76a}$$

por lo que

$$(77)\qquad \int_0^{\pi/2} \operatorname{sen}^{2m} x\,dx = \frac{2m-1}{2m}\cdot\frac{2m-3}{2m-2}\cdots\frac{1}{2}\cdot\frac{\pi}{2},$$

$$(77a)\qquad \int_0^{\pi/2} \operatorname{sen}^{2m+1} x\,dx = \frac{2m}{2m+1}\cdot\frac{2m-2}{2m-1}\cdots\frac{2}{3}.$$

Por división resulta

$$(78)\qquad \frac{\pi}{2} = \frac{2\cdot 2}{1\cdot 3}\cdot\frac{4\cdot 4}{3\cdot 5}\cdot\frac{6\cdot 6}{5\cdot 7}\cdots\frac{2m\cdot 2m}{(2m-1)\cdot(2m+1)}\,\frac{\displaystyle\int_0^{\pi/2}\operatorname{sen}^{2m} x\,dx}{\displaystyle\int_0^{\pi/2}\operatorname{sen}^{2m+1} x\,dx}$$

El cociente de las dos integrales en el miembro derecho converge a 1 cuando m crece más allá de toda cota, como se deduce a partir de las siguientes consideraciones. En el intervalo $0 < x < \pi/2$, donde $0 < \operatorname{sen} x < 1$, se tiene

$$0 < \operatorname{sen}^{2m+1} x \le \operatorname{sen}^{2m} x \le \operatorname{sen}^{2m-1} x;$$

consecuentemente

$$0 < \int_0^{\pi/2} \operatorname{sen}^{2m+1} x\,dx \le \int_0^{\pi/2} \operatorname{sen}^{2m} x\,dx \le \int_0^{\pi/2} \operatorname{sen}^{2m-1} x\,dx.$$

Si se divide aquí cada término por $\displaystyle\int_0^{\pi/2} \operatorname{sen}^{2m+1} x\,dx$ y se observa que, por la fórmula (75), es

$$\frac{\displaystyle\int_0^{\pi/2}\operatorname{sen}^{2m-1} x\,dx}{\displaystyle\int_0^{\pi/2}\operatorname{sen}^{2m+1} x\,dx} = \frac{2m+1}{2m} = 1 + \frac{1}{2m},$$

se tiene que

$$1 \le \frac{\displaystyle\int_0^{\pi/2}\operatorname{sen}^{2m} x\,dx}{\displaystyle\int_0^{\pi/2}\operatorname{sen}^{2m+1} x\,dx} \le 1 + \frac{1}{2m},$$

de la cual surge la propiedad antedicha.

Consecuentemente, la relación

$$(79)\qquad \frac{\pi}{2} = \lim_{m\to\infty} \frac{2}{1}\frac{2}{3}\frac{4}{3}\frac{4}{5}\frac{6}{5}\frac{6}{7}\cdots\frac{2m}{2m-1}\frac{2m}{2m+1}$$

se cumple.

Esta fórmula producto (debida a Wallis), con su sencilla ley de formación da una relación notable entre el número π y los enteros.

Producto para $\sqrt{\pi}$

Como una consecuencia sencilla, puede deducirse una expresión igualmente notable para $\sqrt{\pi}$. Si se observa que

$$\lim_{m\to\infty} \frac{2m}{2m+1} = 1,$$

entonces puede escribirse

$$\lim_{m\to\infty} \frac{2^2\cdot 4^2\cdots(2m-2)^2}{3^2\cdot 5^2\cdots(2m-1)^2}\,2m = \frac{\pi}{2},$$

y tomando la raíz cuadrada y multiplicando después el numerador y el denominador por $2\cdot 4\cdots(2m-2)$, se encuentra

$$\sqrt{\frac{\pi}{2}} = \lim_{m\to\infty} \frac{2\cdot 4\cdots(2m-2)}{3\cdot 5\cdots(2m-1)}\sqrt{2m} = \lim_{m\to\infty} \frac{2^2\cdot 4^2\cdots(2m-2)^2}{(2m-1)!}\sqrt{2m}$$

$$= \lim_{m\to\infty} \frac{2^2\cdot 4^2\cdots(2m)^2}{(2m)!}\,\frac{\sqrt{2m}}{2m}.$$

$$= \lim \frac{(2^2\cdot 1^2)(2^2\cdot 2^2)(2^2\cdot 3^2)\cdots(2^2\cdot m^2)}{(2m)!\sqrt{2m}}$$

De aquí se obtiene finalmente

$$(80)\qquad \lim_{m\to\infty} \frac{(m!)^2 2^{2m}}{(2m)!\sqrt{m}} = \sqrt{\pi},$$

una forma del producto de Wallis que será utilizada posteriormente (Ver apéndice, capítulo 6).

*3.12 Integración de funciones racionales

Durante los siglos XVII y XVIII los matemáticos estuvieron preocupados por descubrir clases de funciones explícitas elementales que pudieran ser integradas explícitamente. Fueron inventados gran número de medios ingeniosos y al mismo tiempo se creó la base para una comprensión más profunda. Cuando posteriormente se entendió que el lograr la integración de *todas* las funciones explícitas en forma exacta no era una meta alcanzable ni tampoco una meta realmente importante, las tediosas técnicas que se habían desarrollado en conexión con tales problemas fueron gradualmente pasando a segundo plano. Sin embargo, prevaleció un significativo resultado general:

Todas las funciones racionales $R(x)$ de una variable x pueden ser integradas explícitamente en términos de las integrales elementales registradas en la Tabla 3.1.

Este resultado general puede ser obtenido mucho más fácilmente en el contexto de la teoría, más avanzada, de funciones de una variable compleja. No obstante, vale la pena bosquejar una deducción elemental empleando solamente variables reales.

Las funciones racionales son aquéllas de la forma

$$R(x) = \frac{f(x)}{g(x)}, \tag{81}$$

donde $f(x)$ y $g(x)$ son polinomios:

$$f(x) = a_m x^m + a_{m-1}x^{m-1} + \cdots + a_0,$$
$$g(x) = b_n x^n + b_{n-1}x^{n-1} + \cdots + b_0 \quad (b_n \neq 0).$$

Como se recuerda, todo polinomio puede ser integrado de inmediato y su integral es también un polinomio. Basta por ello considerar solamente aquellas funciones racionales para las cuales el denominador $g(x)$ no es una constante. Además, puede suponerse siempre que el grado del numerador es menor que el grado n del denominador. Pues, de lo contrario, dividiendo el polinomio $f(x)$ por el polinomio $g(x)$ se obtiene un residuo de grado menor que n; en otras palabras, puede escribirse $f(x) = q(x)g(x) + r(x)$, donde $q(x)$ y $r(x)$ son también polinomios y $r(x)$ es de grado menor que n. La integración de $f(x)/g(x)$ es reducida entonces a la integración del polinomio $q(x)$ y a la de la fracción "propia" $r(x)/g(x)$. Se observa además que la función $f(x)/g(x)$ puede ser representada como la suma de las funciones $a_\nu x^\nu/g(x)$, de modo que se requiere considerar solamente integrandos de la forma $x^\nu/g(x)$.

a. Los tipos fundamentales

Se procederá por pasos en la integración de las funciones racionales más generales del tipo (81), estudiando primero sólo aquellas funciones con denominador $g(x)$ del tipo particularmente sencillo

$$g(x) = x^n,$$

o bien

$$g(x) = (1 + x^2)^n,$$

donde n es cualquier entero positivo.

A este caso puede reducirse el caso algo más general en el que $g(x) = (\alpha x + \beta)^n$, una potencia entera $(n > 1)$ de una expresión lineal $\alpha x + \beta$ $(\alpha \neq 0)$, o bien $g(x) = (ax^2 + 2bx + c)^n$, una potencia de una expresión cuadrática definida.[1] Si $g(x) = (\alpha x + \beta)^n$, se introduce $\xi = \alpha x + \beta$ como

[1] Una expresión cuadrática $Q(x) = ax^2 + 2bx + c$ se dice que es *definida* si para todos los valores reales de x toma valores que poseen uno y el mismo signo, esto es, si la ecuación $Q(x) = 0$ no posee raíces reales. Para esto es necesario y suficiente que el "discriminante" $ac - b^2$ sea positivo. Esto surge, por supuesto, a partir de la fórmula explícita $(-b \pm \sqrt{b^2 - ac})/a$ para las raíces. Equivalentemente, una expresión cuadrática definida es aquélla que no puede ser factorizada en dos factores lineales *reales*.

una nueva variable. Entonces $d\xi/dx = \alpha$ y $x = (\xi - \beta)/\alpha$ es también una función lineal de ξ. Cada numerador $f(x)$ se vuelve un polinomio $\phi(\xi)$ del mismo grado, y, consecuentemente,

$$\int \frac{f(x)}{(\alpha x + \beta)^n}\,dx = \frac{1}{\alpha}\int \frac{\phi(\xi)}{\xi^n}\,d\xi.$$

En el segundo caso se escribe

$$ax^2 + 2bx + c = \frac{1}{a}(ax + b)^2 + \frac{d^2}{a} \qquad (d^2 = ac - b^2,\, d > 0);$$

y como se ha supuesto que la expresión es cuadrática y definida, $ac - b^2$ debe ser positivo y $a \neq 0$. Introduciendo la nueva variable

$$\xi = \frac{ax + b}{d}$$

se obtiene una integral con denominador $[(d^2/a)(1 + \xi^2)]^n$, en el integrando.

Por lo tanto, para integrar funciones racionales cuyos denominadores son potencias de una expresión lineal o de expresiones cuadráticas definidas es suficiente con poder integrar los siguientes tipos de funciones:

$$\frac{1}{x^n}, \qquad \frac{x^{2\nu}}{(x^2 + 1)^n}, \qquad \frac{x^{2\nu+1}}{(x^2 + 1)^n}.$$

Se verá, en efecto, que aún estos tipos no requieren ser tratados en general, pues se puede reducir la integración de cada función racional a la integración de las formas muy especiales obtenidas de estas tres funciones tomando $\nu = 0$. De acuerdo con esto se considerará ahora la integración de las tres expresiones

$$\frac{1}{x^n}, \qquad \frac{1}{(x^2 + 1)^n}, \qquad \frac{x}{(x^2 + 1)^n}.$$

b. Integración de los tipos fundamentales

La integración del primer tipo de función, $1/x^n$, conduce inmediatamente a la expresión $\log|x|$ si $n = 1$, y a la expresión $-1/(n-1)x^{n-1}$ si $n > 1$, de modo que en ambos casos la integral es también una función elemental. Las funciones del tercer tipo pueden ser integradas rápidamente introduciendo la nueva variable $\xi = x^2 + 1$, de lo cual se obtiene $2x\,dx = d\xi$ y

$$\int \frac{x}{(x^2 + 1)^n}\,dx = \frac{1}{2}\int \frac{d\xi}{\xi^n} = \begin{cases} \tfrac{1}{2}\log(x^2 + 1) & \text{si } n = 1, \\ -\dfrac{1}{2(n-1)(x^2 + 1)^{n-1}} & \text{si } n > 1. \end{cases}$$

Finalmente, para calcular la integral

$$I_n = \int \frac{dx}{(x^2+1)^n},$$

donde n tiene cualquier valor entero mayor que uno, se hace uso del método recursivo: Si se pone

$$\frac{1}{(x^2+1)^n} = \frac{1}{(x^2+1)^{n-1}} - \frac{x^2}{(x^2+1)^n},$$

de modo que

$$\int \frac{dx}{(x^2+1)^n} = \int \frac{dx}{(x^2+1)^{n-1}} - \int \frac{x^2\,dx}{(x^2+1)^n},$$

puede transformarse el miembro derecho mediante integración por partes, utilizando la fórmula (54) de la p. 295 con

$$f(x) = x, \quad g'(x) = \frac{x}{(x^2+1)^n}.$$

Entonces, tal como se ha encontrado,

$$g(x) = -\frac{1}{2}\,\frac{1}{(n-1)(x^2+1)^{n-1}},$$

y, consecuentemente, se obtiene

$$I_n = \int \frac{dx}{(x^2+1)^n} = \frac{x}{2(n-1)(x^2+1)^{n-1}} + \frac{2n-3}{2(n-1)} \int \frac{dx}{(x^2+1)^{n-1}}.$$

El cálculo de la integral I_n es reducido así al de la integral I_{n-1}. Si $n - 1 > 1$, se aplica el mismo proceso a esta última integral y se continúa hasta que finalmente se llega a la expresión

$$\int \frac{dx}{x^2+1} = \text{arc tan } x.$$

Se ve así que la integral [1] I_n puede ser expresada explícitamente en términos de funciones racionales y de la función arc tan x.

Por otra parte, nótese que pudo haberse integrado la función $1/(x^2+1)^n$ también en forma directa, utilizando la substitución $x = \tan t$; se habría obtenido entonces $dx = \sec^2 t\,dt$ y $1/(1+x^2) = \cos^2 t$, de modo que

$$\int \frac{dx}{(x^2+1)^n} = \int \cos^{2n-2} t\,dt,$$

y se ha aprendido ya [Ec. (65), p. 299] cómo evaluar esta integral.

[1] La integral de la función $1/(x^2-1)^n$ puede ser calculada de la misma manera; por el correspondiente método de recurrencia se reduce ésta a la integral

$$\int \frac{dx}{1-x^2} = \text{ar tanh } x \qquad (\text{o bien ar coth } x).$$

c. Fracciones parciales

Se está ahora en condiciones de integrar las funciones racionales más generales. Se hace uso del hecho de que cualquiera de tales funciones puede ser representada como la suma de las así llamadas *fracciones parciales,* esto es, como la suma de un polinomio y un número finito de funciones racionales cada una de las cuales tiene, ya sea una potencia de una expresión lineal como denominador y una constante definida como numerador, o bien una potencia de una expresión cuadrática definida como denominador y una función lineal como numerador. Si el grado del numerador $f(x)$ es menor que el del denominador $g(x)$, el polinomio no aparece. Se sabe ya cómo integrar cada fracción parcial. Pues, de acuerdo con lo visto en la p. 304, el denominador puede ser reducido a una de las formas especiales x^n y $(x^2+1)^n$, y la fracción es entonces una combinación de los tipos fundamentales integrados en la p. 304.

No se dará la demostración general de la posibilidad de esta resolución en fracciones parciales. Simplemente se tratará de hacer inteligible para el lector el contenido del teorema y mostrar mediante ejemplos cómo la resolución en fracciones parciales puede ser llevada a cabo en casos típicos. En la práctica se tratan solamente funciones comparativamente sencillas, pues de otra forma los cálculos se tornan demasiado complicados.

Como se sabe del álgebra elemental, todo polinomio real $g(x)$ puede ser escrito en la forma [1]

$$g(x) = a(x-\alpha_1)^{l_1}(x-\alpha_2)^{l_2}\cdots \cdots (x^2+2b_1x+c_1)^{r_1}(x^2+2b_2x+c_2)^{r_2}\cdots.$$

Aquí los números distintos $\alpha_1, \alpha_2, \ldots$ son las raíces reales de la ecuación $g(x) = 0$, y los enteros positivos $l_1, l_2, \ldots$ indican la multiplicidad de estas raíces; los factores $x^2 + 2b_\nu x + c_\nu$ indican expresiones cuadráticas definidas, de las cuales no hay dos iguales, con raíces complejas conjugadas, y los enteros postitivos $r_1, r_2, \ldots$ dan la multiplicidad de estas raíces.

Se supone que el denominador está dado en esta forma, o bien que ha sido llevado a esta forma calculando las raíces reales e imaginarias. Se supone además que el numerador $f(x)$ es de menor grado que el denominador (Ver p. 303). Entonces el teorema sobre resolución en fracciones parciales puede establecerse como sigue: *Para cada factor* $(x-\alpha)^l$, *donde* α *es cualquiera de las raíces reales de multiplicidad* l, *se puede determinar una expresión de la forma*

[1] La *demostración* real de este teorema, llamado *teorema fundamental del álgebra,* no pertenece al álgebra. Se la logra más fácilmente por métodos que pertenecen a la teoría de funciones de una *variable compleja.*

$$\frac{A_1}{x-\alpha}+\frac{A_2}{(x-\alpha)^2}+\cdots+\frac{A_l}{(x-\alpha)^l},$$

y para cada factor cuadrático $Q(x)=x^2+2bx+c$ *en el producto, que está elevado a la potencia r, se puede determinar una expresión de la forma*

$$\frac{B_1+C_1x}{Q}+\frac{B_2+C_2x}{Q^2}+\cdots+\frac{B_r+C_rx}{Q^r},$$

de tal manera que la función $f(x)/g(x)$ *es la suma de todas estas expresiones* (A_v, B_v, C_v, *son constantes*). En otras palabras, el cociente $f(x)/g(x)$ puede ser representado como una suma de fracciones cada una de las cuales pertenece a uno de los tipos integrados arriba.[1]

En casos particulares la descomposición en fracciones parciales puede efectuarse fácilmente por inspección. Si, por ejemplo, $g(x)=x^2-1$, se ve de inmediato que

$$\frac{1}{x^2-1}=\frac{1}{2}\,\frac{1}{(x-1)}-\frac{1}{2}\,\frac{1}{(x+1)},$$

de modo que

$$\int\frac{dx}{x^2-1}=\tfrac{1}{2}\log\left|\frac{x-1}{x+1}\right|.$$

Más generalmente, si $g(x)=(x-\alpha)(x-\beta)$, esto es, si $g(x)$ es una expresión cuadrática no definida con dos ceros reales α y β, se tiene

$$\frac{1}{(x-\alpha)(x-\beta)}=\frac{1}{(\alpha-\beta)}\,\frac{1}{(x-\alpha)}-\frac{1}{(\alpha-\beta)}\,\frac{1}{(x-\beta)},$$

[1] Bosquejamos brevemente un método mediante el cual puede probarse la posibilidad de esta descomposición en fracciones parciales sin usar la teoría de funciones de variables complejas, suponiendo que $g(x)$ puede factorizarse completamente en factores lineales. Si $g(x)=(x-\alpha)^k h(x)$ y $h(\alpha)\neq 0$, entonces el numerador en el lado derecho de la ecuación

$$\frac{f(x)}{g(x)}-\frac{f(\alpha)}{h(\alpha)(x-\alpha)^k}=\frac{1}{h(\alpha)}\,\frac{f(x)h(\alpha)-f(\alpha)h(x)}{(x-\alpha)^k h(x)}$$

obviamente se anula para $x=\alpha$; por consiguiente, es de la forma $h(\alpha)(x-\alpha)^m f_1(x)$, donde $f_1(x)$ es también un polinomio, el entero m es ≥ 1 y $f_1(\alpha)\neq 0$. Escribiendo $f(\alpha)/h(\alpha)=\beta$, la anterior da

$$\frac{f(x)}{g(x)}-\frac{\beta}{(x-\alpha)^k}=\frac{f_1(x)}{(x-\alpha)^{k-m}h(x)}.$$

Continuando este proceso, podemos hacer que disminuya paso a paso el grado de la potencia de $(x-\alpha)$ que aparece en el denominador, hasta que finalmente no aparezca tal factor. Repetimos el proceso con la fracción restante considerando alguna otra raíz de $g(x)$, y hacemos lo mismo tantas veces como factores distintos tenga $g(x)$. Haciendo esto no sólo para las raíces reales sino también para las complejas, y combinando fracciones complejas conjugadas, finalmente se llegará a la descomposición completa en fracciones parciales.

de modo que

$$\int\frac{dx}{(x-\alpha)(x-\beta)}=\frac{1}{\alpha-\beta}\log\left|\frac{x-\alpha}{x-\beta}\right|$$

d. Ejemplos de resolución en fracciones parciales. El método de los coeficientes indeterminados

Si $g(x)=(x-\alpha_1)(x-\alpha_2)\cdots(x-\alpha_n)$, donde $\alpha_i\neq\alpha_k$ si $i\neq k$, esto es, si la ecuación $g(x)=0$ posee solamente raíces reales simples, y si $f(x)$ es cualquier polinomio de grado $<n$, la expresión en términos de fracciones parciales tiene la sencilla forma

$$\frac{f(x)}{g(x)}=\frac{a_1}{x-\alpha_1}+\frac{a_2}{x-\alpha_2}+\cdots+\frac{a_n}{x-\alpha_n}.$$

Se obtienen expresiones explícitas para los coeficientes $a_1, a_2, \ldots$ si se multiplican ambos lados de esta ecuación por $(x-\alpha_1)$, se cancela el factor común $(x-\alpha_1)$ en el numerador y el denominador de la izquierda y en el primer término de la derecha, y luego se hace $x=\alpha_1$. Esto da

$$a_1=\frac{f(\alpha_1)}{(\alpha_1-\alpha_2)(\alpha_1-\alpha_3)\cdots(\alpha_1-\alpha_n)}.$$

El lector observará a partir de la regla para la derivada de un producto que el denominador a la derecha es $g'(\alpha_1)$, esto es, la derivada de la función $g(x)$ en el punto $x=\alpha_1$. Fórmulas análogas para $a_2, a_3, \ldots$, obtenidas de esta manera, conducen a la expresión explícita en fracciones parciales

$$\frac{f(x)}{g(x)}=\frac{f(\alpha_1)}{g'(\alpha_1)(x-\alpha_1)}+\frac{f(\alpha_2)}{g'(\alpha_2)(x-\alpha_2)}+\cdots+\frac{f(\alpha_n)}{g'(\alpha_n)(x-\alpha_n)}.$$

Como un ejemplo típico de un denominador $g(x)$ con raíces múltiples se considera la función $1/[x^2(x-1)]$. Esta tiene por representación

$$\frac{1}{x^2(x-1)}=\frac{a}{x-1}+\frac{b}{x}+\frac{c}{x^2},$$

en concordancia con la p. 307. Si se multiplican ambos miembros de esta ecuación por $x^2(x-1)$, se obtiene la ecuación

$$1=(a+b)x^2-(b-c)x-c,$$

válida para todos los valores de x, a partir de la cual han de determinarse los coeficientes a, b, c. Esta condición no puede cumplirse a menos que todos los coeficientes del polinomio $(a+b)x^2-(b-c)x-c-1$ sean cero; esto es, se debe tener $a+b=b-c=c+1=0$, o bien $c=-1$, $b=-1$, $a=1$. Se obtiene así la resolución

$$\frac{1}{x^2(x-1)}=\frac{1}{x-1}-\frac{1}{x}-\frac{1}{x^2},$$

y, consecuentemente,

$$\int \frac{dx}{x^2(x-1)} = \log|x-1| - \log|x| + \frac{1}{x}.$$

A continuación se descompone la función $1/[x(x^2+1)]$, cuyo denominador posee ceros complejos, de acuerdo con la ecuación

$$\int \frac{1}{x(x^2+1)} = \frac{a}{x} + \frac{bx+c}{x^2+1}.$$

Para los coeficientes se obtiene $a + b = c = a - 1 = 0$, de modo que

$$\frac{1}{x(x^2+1)} = \frac{1}{x} - \frac{x}{x^2+1},$$

y, por lo tanto,

$$\int \frac{dx}{x(x^2+1)} = \log|x| - \tfrac{1}{2}\log(x^2+1).$$

Como un tercer ejemplo se considera la función $1/(x^4+1)$, cuya integración era un serio problema aun en los tiempos de Leibnitz. Se puede representar el denominador como el producto de dos factores cuadráticos:[1]

$$x^4 + 1 = (x^2+1)^2 - 2x^2 = (x^2 + 1 + \sqrt{2}x)(x^2 + 1 - \sqrt{2}x).$$

Se sabe, por lo tanto, que la resolución en fracciones parciales tendrá la forma

$$\frac{1}{x^4+1} = \frac{ax+b}{x^2+\sqrt{2}x+1} + \frac{cx+d}{x^2-\sqrt{2}x+1}.$$

Para determinar los coeficientes a, b, c, d, se utiliza la ecuación

$$(a+c)x^3 + (b+d-a\sqrt{2}+c\sqrt{2})x^2 + (a+c-b\sqrt{2}+d\sqrt{2})x + (b+d-1) = 0,$$

la cual se satisface para los valores

$$a = \frac{1}{2\sqrt{2}}, \qquad b = \tfrac{1}{2}, \qquad c = -\frac{1}{2\sqrt{2}}, \qquad d = \tfrac{1}{2}.$$

Se tiene entonces

$$\frac{1}{x^4+1} = \frac{1}{2\sqrt{2}}\cdot\frac{x+\sqrt{2}}{x^2+\sqrt{2}x+1} - \frac{1}{2\sqrt{2}}\cdot\frac{x-\sqrt{2}}{x^2-\sqrt{2}x+1},$$

[1] La factorización de x^4+1 en factores cuadráticos reales corresponde a la factorización en factores lineales complejos conjugados

$$x^4 + 1 = [(x-\varepsilon)(x-\varepsilon^{-1})][(x-\varepsilon^3)(x-\varepsilon^{-3})],$$

donde

$$\varepsilon = \cos\frac{\pi}{4} + i\,\mathrm{sen}\,\frac{\pi}{4} = \frac{1}{2}\sqrt{2}(1+i)$$

es una de las ocho raíces de $+1$, y una cuarta raíz de -1 (véase pp. 128-129).

y, aplicando el método dado en la p. 304, se obtiene

$$\int \frac{dx}{x^4+1} = \frac{1}{4\sqrt{2}}\log|x^2+\sqrt{2}x+1| - \frac{1}{4\sqrt{2}}\log|x^2-\sqrt{2}x+1| + \frac{1}{2\sqrt{2}}\arctan(\sqrt{2}x+1) + \frac{1}{2\sqrt{2}}\arctan(\sqrt{2}x-1),$$

lo cual puede ser verificado fácilmente por derivación.

Los ejemplos anteriores ilustran un método general de integrar una función racional $f(x)/g(x)$. Se divide primero y se reduce el problema al caso en que el grado de f es menor que el de g. Se descompone $g(x)$ en factores lineales cuadráticos definidos, agrupando el producto en potencias de tales factores. Se escribe la representación en fracciones parciales *apropiadas* de f/g, con coeficientes *indeterminados* $a, b, c, \ldots$. Multiplicando todo por $g(x)$ y *comparando coeficientes* de potencias iguales en la identidad polinomial resultante se obtiene un sistema de ecuaciones lineales para los coeficientes desconocidos, que justamente debe ser adecuado para determinar estos coeficientes si realmente se tiene la forma correcta para la descomposición en fracciones parciales. Se está entonces en posibilidad de integrar cualquiera de las fracciones parciales resultantes mediante las reglas discutidas anteriormente.

3.13 Integración de algunas otras clases de funciones

a. Comentarios preliminares sobre la representación[1] racional de la circunferencia y de la hipérbola

La integración de algunas otras clases generales de funciones puede ser reducida a la integración de funciones racionales. Se estará en condiciones de entender mejor esta reducción estableciendo primero ciertas propiedades elementales de las funciones trigonométricas e hiperbólicas. Si se hace $t = \tan(x/2)$, la trigonometría elemental conduce a las fórmulas sencillas

$$\mathrm{sen}\,x = \frac{2t}{1+t^2}, \qquad \cos x = \frac{1-t^2}{1+t^2};$$

en efecto, a partir de

$$\frac{1}{1+t^2} = \cos^2\frac{x}{2} \qquad \text{y} \qquad \frac{t^2}{1+t^2} = \mathrm{sen}^2\frac{x}{2},$$

y de las fórmulas elementales

$$\mathrm{sen}\,x = 2\cos^2\frac{x}{2}\tan\frac{x}{2} \qquad \text{y} \qquad \cos x = \cos^2\frac{x}{2} - \mathrm{sen}^2\frac{x}{2},$$

se obtienen esas ecuaciones. Ellas muestran que $\mathrm{sen}\,x$ y $\cos x$ *pueden ser ambas expresadas racionalmente en términos de la cantidad* $t = \tan x/2$. Derivando se tiene

$$\frac{dt}{dx} = \frac{1}{2\cos^2 x/2} = \frac{1+t^2}{2},$$

[1] Llamada a veces "uniformización".

de modo que

$$\frac{dx}{dt} = \frac{2}{1+t^2}; \tag{82}$$

por lo que la derivada dx/dt es también una expresión racional en t.

*La representación geométrica de nuestras fórmulas y su significado geométrico se da en la Fig. 3.28. Aquí se muestra el círculo $u^2 + v^2 = 1$ en un plano u, v. Si x denota el ángulo TOP en la figura, entonces $u = \cos x$ y $v = \operatorname{sen} x$. El ángulo OSP, con su vértice en el punto $u = -1$, $v = 0$, es igual a $x/2$, por un teorema de la geometría elemental, y puede extraerse el significado geométrico del parámetro t de la figura; $t = \tan \frac{1}{2}x = OR$ donde R es la "proyección" sobre el eje v del punto

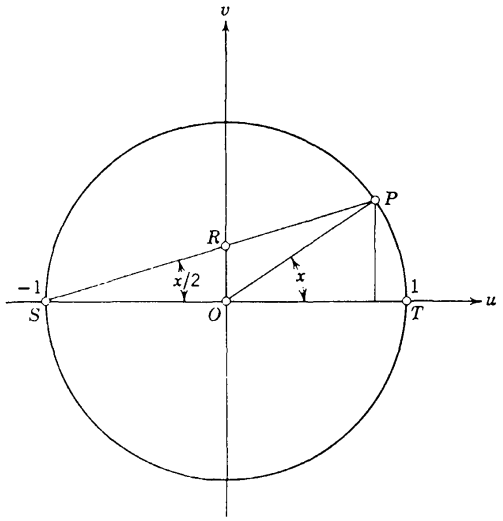


Figura 3.28 Representación paramétrica de las funciones trigonométricas.

P del círculo desde S. Si el punto P parte desde S y describe una vez el círculo en la dirección positiva, esto es, si x recorre el intervalo desde $-\pi$ hasta $+\pi$, la cantidad t recorrerá todo el rango de valores desde $-\infty$ hasta $+\infty$ una vez exactamente. (Nótese que el punto S mismo corresponde a $t = \pm\infty$.) Se tiene aquí una representación del punto general (u, v) del círculo $u^2 + v^2 = 1$ en términos de las funciones racionales $u = (1 - t^2)/(1 + t^2)$, y $v = 2t/(1 + t^2)$ del parámetro t. Estas fórmulas definen entonces una transformación racional de la recta t sobre un círculo del plano u, v (la cual es la análoga bidimensional de la proyección estereográfica sobre una esfera, mencionada en la p. 45). En la base de esta representación racional del círculo se encuentra, evidentemente, la identidad

$$(t^2 - 1)^2 + (2t)^2 = (t^2 + 1)^2.$$

Es por demás curioso que esta fórmula sea de interés también en la teoría de números, puesto que genera para cada entero t *enteros pitagóricos* $a = t^2 - 1$, $b = 2t$, y $c = t^2 + 1$ que satisfacen la identidad $a^2 + b^2 = c^2$; esto es, determina un triángulo rectángulo con lados conmensurables. Así $t = 2$ da lugar a la terna bien conocida $a = 3$, $b = 4$, $c = 5$; para $t = 4$ se obtiene $a = 15$, $b = 8$, $c = 17$, etc. Es notable, y, por supuesto, no accidental, que la misma identidad algebraica es de significación en contextos tan diversos como la integración en forma exacta, la geometría y la teoría de números. Ligar campos diferentes de tal manera es una tendencia típica en las matemáticas modernas; aunque cuando el ejemplo en cuestión se remonta a la antigüedad.

Análogamente pueden expresarse las funciones hiperbólicas

$$\cosh x = \tfrac{1}{2}(e^x + e^{-x})$$

y $\operatorname{senh} x = \frac{1}{2}(e^x - e^{-x})$ como funciones racionales de una tercera cantidad. El camino más obvio es hacer $e^x = \tau$, de modo que se tiene

$$\cosh x = \frac{1}{2}\left(\tau + \frac{1}{\tau}\right), \qquad \operatorname{senh} x = \frac{1}{2}\left(\tau - \frac{1}{\tau}\right),$$

que son expresiones racionales para $\operatorname{senh} x$ y $\cosh x$. Aquí, nuevamente, $dx/d\tau = 1/\tau$ es racional en τ. Sin embargo, se obtiene una analogía más cercana con las funciones trigonométricas introduciendo la cantidad $t = \tanh(x/2) = (\tau - 1)/(\tau + 1)$; y se llega entonces a las fórmulas

$$\cosh x = \frac{1 + t^2}{1 - t^2}, \qquad \operatorname{senh} x = \frac{2t}{1 - t^2}.$$

Derivando $t = \tanh(x/2)$ se obtiene, como en la Ec. (82) de la p. 311, la expresión racional

$$\frac{dx}{dt} = \frac{2}{1 - t^2} \tag{83}$$

para la derivada dx/dt. Aquí, nuevamente, la cantidad t posee un significado geométrico similar al que tiene para las funciones trigonométricas, como se ve de inmediato de la Fig. 3.29.

Se tiene aquí una representación racional de la hipérbola $u^2 - v^2 = 1$ en el plano u, v, por medio de las ecuaciones $u = (1 + t^2)/(1 - t^2)$, $v = 2t/(1 - t^2)$. Los puntos en la rama derecha de la curva son de la forma $u = \cosh x$, $v = \operatorname{senh} x$, y corresponden a valores de t con $|t| < 1$. La otra rama se obtiene para $|t| > 1$.

Se prosigue ahora con los problemas de integración.

*b. Integración de $R(\cos x, \operatorname{sen} x)$

Denótese por $R(\cos x, \operatorname{sen} x)$ una expresión que es racional en las dos funciones $\operatorname{sen} x$ y $\cos x$, esto es, una expresión que se forma racionalmente con estas dos funciones y algunas constantes, tal como

$$\frac{3\operatorname{sen}^2 x + \cos x}{3\cos^2 x + \operatorname{sen} x}$$

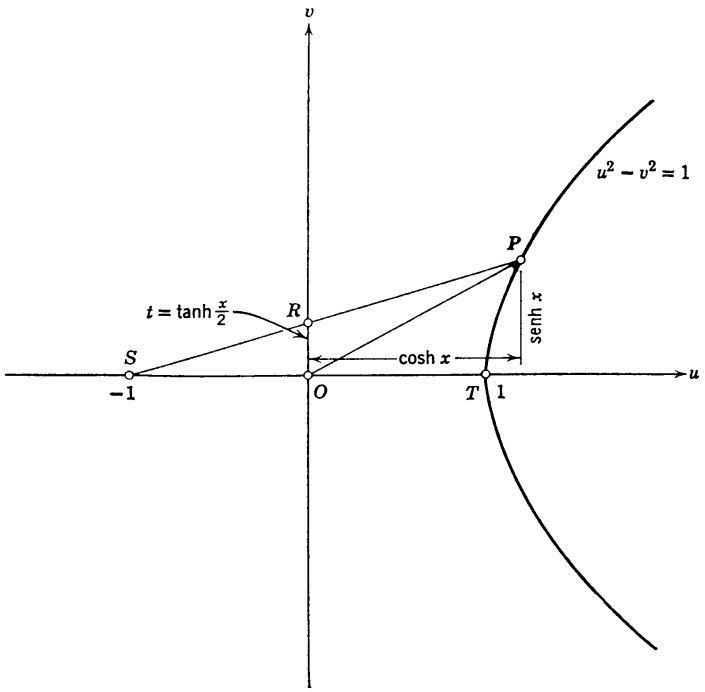


Figura 3.29 Representación paramétrica de las funciones hiperbólicas.

Si se aplica la substitución $t = \tan x/2$, la integral

$$\int R(\cos x, \operatorname{sen} x)\, dx$$

es transformada en la integral

$$\int R\left(\frac{1-t^2}{1+t^2}, \frac{2t}{1+t^2}\right)\frac{2}{1+t^2}\, dt,$$

y bajo el signo de integral se tiene ahora una función racional de t. Así se ha obtenido, en principio, la integral de nuestra expresión, puesto que puede efectuarse ahora la integración por los métodos de la sección anterior.

c. Integración de $R(\cosh x, \operatorname{senh} x)$

Del mismo modo, si $R(\cosh x, \operatorname{senh} x)$ es una expresión racional en términos de las funciones hiperbólicas $\cosh x$ y $\operatorname{senh} x$, puede efectuarse su integración por medio de la substitución $t = \tanh x/2$. Recordando la Ec. (83), se tiene

$$\int R(\cosh x, \operatorname{senh} x)\, dx = \int R\left(\frac{1+t^2}{1-t^2}, \frac{2t}{1-t^2}\right)\frac{2}{1-t^2}\, dt.$$

(De acuerdo con un comentario anterior pudo haberse introducido también $\tau = e^x$ como una nueva variable y expresado $\cosh x$ y $\operatorname{senh} x$ en términos de τ). La integración queda nuevamente reducida a la de una función racional.

*d. Integración de $R(x, \sqrt{1-x^2})$

La integral $\int R(x, \sqrt{1-x^2})\, dx$ puede ser reducida a la del tipo tratado en la sección 3.13*b* utilizando la substitución

$$x = \cos u, \qquad \sqrt{1-x^2} = \operatorname{sen} u, \qquad dx = -\operatorname{sen} u\, du.$$

A partir de este paso la transformación $t = \tan u/2$ conduce a la integración de una función racional. Obsérvese que pudo haberse efectuado la reducción en un paso, en vez de dos, utilizando la substitución

$$t = \sqrt{\frac{1-x}{1+x}}; \qquad x = \frac{1-t^2}{1+t^2};$$

$$\sqrt{1-x^2} = \frac{2t}{1+t^2}; \qquad \frac{dx}{dt} = \frac{-4t}{(1+t^2)^2};$$

esto es, pudo haberse introducido $t = \tan u/2$ directamente como la nueva variable y con ello obtenido un integrando racional.

*e. Integración de $R(x, \sqrt{x^2-1})$

La integral $\int R(x, \sqrt{x^2-1})\, dx$ es transformada mediante la substitución $x = \cosh u$ en una del tipo tratado en la sección 3.13*c*. Aquí también puede llegarse a la meta directamente introduciendo

$$t = \sqrt{\frac{x-1}{x+1}} = \tanh \frac{u}{2}.$$

*f. Integración de $R(x, \sqrt{x^2+1})$

La integral $\int R(x, \sqrt{(x^2+1)})\, dx$ es reducida mediante la transformación $x = \operatorname{senh} u$ a una del tipo considerado en la sección 3.13*c* (pp. 313-314) y puede, por ello, ser integrada en términos de funciones elementales. En vez de la reducción posterior a la integral de una función racional por medio de la substitución $e^u = \tau$, o de $\tanh u/2 = t$, podría llegarse a la integral de una función racional en un solo paso por cualquiera de las substituciones

$$x = \frac{\xi^2 - \beta}{\alpha} \qquad \text{y} \qquad \frac{dx}{d\xi} = \frac{2\xi}{\alpha};$$

*g. Integración de $R(x, \sqrt{ax^2 + 2bx + c})$

La integral $\int R(x, \sqrt{ax^2 + 2bx + c})\, dx$ de una expresión que es racional en términos de x y de la raíz cuadrada de un polinomio arbitrario de segundo grado en x puede ser reducida inmediatamente a uno de los tipos tratados. Se escribe (Ver p. 304)

$$ax^2 + 2bx + c = \frac{1}{a}(ax + b)^2 + \frac{ac - b^2}{a}.$$

Si $ac - b^2 > 0$, se introduce una nueva variable ξ por medio de la transformación $\xi = (ax + b)/\sqrt{ac - b^2}$, después de lo cual el radical adquiere la forma $\sqrt{(ac - b^2)(\xi^2 + 1)/a}$. Por lo tanto, la integral cuando está expresada en términos de ξ es del tipo de la sección 3.13*f*. La constante a debe ser aquí positiva de modo que la raíz cuadrada pueda tener valores reales.

Si $ac - b^2 = 0$ y $a > 0$, entonces por medio de la fórmula

$$\sqrt{ax^2 + 2bx + c} = \sqrt{a}\left(x + \frac{b}{a}\right)$$

se ve que el integrando era racional desde un principio.

Si, finalmente, $ac - b^2 < 0$, se hace $\xi = (ax + b)/\sqrt{b^2 - ac}$ y se obtiene para el radical la expresión $\sqrt{(ac - b^2)(\xi^2 - 1)/a}$. Si a es positiva, la integral es así reducida al tipo de la sección 3.13*e*; si, por otra parte, a es negativa el radical se escribe en la forma

$$\sqrt{(b^2 - ac)(1 - \xi^2)/(-a)},$$

y se ve que la integral es así reducida al tipo de la sección 3.13*d*.

*h. Otros ejemplos de reducción a integrales de funciones racionales

De otros tipos de funciones que pueden ser integradas mediante reducción a funciones racionales, serán mencionados solamente dos: 1) expresiones racionales que involucran dos raíces cuadradas diferentes de expresiones lineales, $R(x, \sqrt{ax + b}, \sqrt{\alpha x + \beta})$; 2) expresiones de la forma $R(x, \sqrt[n]{(ax + b)/(\alpha x + \beta)})$, donde a, b, α, β son constantes. En las del primer tipo se introduce la nueva variable $\xi = \sqrt{\alpha x + \beta}$, de modo que $\alpha x + \beta = \xi^2$, y, consecuentemente,

$$x = \frac{\xi^2 - \beta}{\alpha} \qquad \text{y} \qquad \frac{dx}{d\xi} = \frac{2\xi}{\alpha};$$

entonces

$$\int R(x, \sqrt{ax + b}, \sqrt{\alpha x + \beta})\, dx = \int R\left\{\frac{\xi^2 - \beta}{\alpha}, \sqrt{\frac{1}{\alpha}[a\xi^2 - (a\beta - b\alpha)]}, \xi\right\}\frac{2\xi}{\alpha}\, d\xi,$$

que es del tipo discutido en la sección 3.13*g*.

Si en las del segundo tipo se introduce la nueva variable

$$\xi = \sqrt[n]{\frac{ax + b}{\alpha x + \beta}},$$

se tiene

$$\xi^n = \frac{ax + b}{\alpha x + \beta}, \qquad x = \frac{-\beta\xi^n + b}{\alpha\beta^n - a}, \qquad \frac{dx}{d\xi} = \frac{a\beta - b\alpha}{(\alpha\xi^n - a)^2}\, n\xi^{n-1},$$

e inmediatamente se llega a la fórmula

$$\int R\left(x, \sqrt[n]{\frac{ax + b}{\alpha x + \beta}}\right) dx = \int R\left(\frac{-\beta\xi^n + b}{\alpha\xi^n - a}, \xi\right)\frac{a\beta - b\alpha}{(a\xi^n - a)^2}\, n\xi^{n-1}\, d\xi,$$

que es la integral de una función racional.

i. Comentarios sobre los ejemplos

Las discusiones anteriores son principalmente de interés teórico. En expresiones complicadas los cálculos reales podrían ser demasiado intrincados. Es, por lo tanto, conveniente aprovechar cuando sea posible la forma especial del integrando para simplificar el trabajo. Por ejemplo, para integrar $1/(a^2 \operatorname{sen}^2 x + b^2 \cos^2 x)$ es mejor utilizar la substitución $t = \tan x$ en vez de la dada en las pp. 313-314; pues $\operatorname{sen}^2 x$ y $\cos^2 x$ pueden ser expresados racionalmente en términos de $\tan x$, y es entonces innecesario retroceder a $t = \tan x/2$. Lo mismo es cierto para toda expresión formada racionalmente a partir de [1] $\operatorname{sen}^2 x$, $\cos^2 x$ y $\operatorname{sen} x \cos x$. Además, para el cálculo de muchas integrales es preferible una forma trigonométrica a una racional, siempre que la forma trigonométrica pueda ser evaluada por algún método sencillo de recurrencia. Por ejemplo, aunque el integrando en $\int x^n(\sqrt{1 - x^2})^m\, dx$ puede ser reducido a una forma racional, es mejor escribir $x = \operatorname{sen} u$ y llevarla a la forma $\int \operatorname{sen}^n u \cos^{m+1} u\, du$, puesto que ésta puede ser tratada fácilmente por el método de recurrencia de la p. 299 (o bien, utilizando los teoremas de adición para reducir las potencias del seno y coseno a senos y cosenos de ángulos múltiples).

[1] Pues $\operatorname{sen} x \cos x = \tan x \cos^2 x$ puede, por supuesto, ser expresado racionalmente en términos de $\tan x$.

Para la evaluación de la integral

$$\int \frac{dx}{a\cos x + b \operatorname{sen} x} \qquad (a^2 + b^2 > 0),$$

en vez de recurrir a la teoría general se escribe

$$A = \sqrt{a^2 + b^2}, \qquad \operatorname{sen}\theta = \frac{a}{A}, \qquad \cos\theta = \frac{b}{A}.$$

La integral toma entonces la forma

$$\frac{1}{A}\int \frac{dx}{\operatorname{sen}(x + \theta)},$$

e introduciendo la nueva variable $x + \theta$ se encuentra [(cfr. Ec. (40), p. 293)] que el valor de la integral es

$$\frac{1}{A}\log\left|\tan\frac{x+\theta}{2}\right|.$$

Parte C Otros pasos en la teoría del cálculo integral

3.14 Integrales de funciones elementales

a. Definición de funciones por medio de integrales. Integrales y funciones elípticas

Con los ejemplos que se han dado de los tipos de funciones que pueden ser integradas mediante la reducción a funciones racionales se ha terminado prácticamente la lista de funciones que son integrables en términos de funciones elementales. Intentos de expresar integrales indefinidas tales como (para $n > 2$)

$$\int \frac{dx}{\sqrt{a_0 + a_1x + \cdots + a_nx^n}},$$

$$\int \sqrt{a_0 + a_1x + \cdots + a_nx^n}\,dx,$$

o bien

$$\int \frac{e^x}{x}\,dx$$

en términos de funciones elementales han fallado. En el siglo XIX finalmente se probó que es realmente *imposible* llevar a cabo estas integraciones en términos de funciones elementales.

Si el objeto del cálculo integral fuera integrar funciones explícitamente, se habría llegado a un alto definitivo. Sin embargo, tal restringido objetivo no posee ninguna justificación intrínseca; es de una naturaleza artificial.

Se sabe que la integral de toda función continua *existe* como un límite y es ella misma una función continua del límite superior, bien sea que la integral pueda ser integrada en términos de funciones elementales o no. Los rasgos distintivos de las funciones elementales están basados en el hecho de que sus propiedades son fácilmente reconocidas, que su aplicación a problemas numéricos es facilitada mediante tablas convenientes, o bien que pueden ser calculadas con un grado tan grande de precisión como se desee.

Siempre que la integral de una función no pueda ser expresada por medio de funciones con las cuales ya se tenga familiaridad, no existe objeción para introducir esta integral como una nueva función "superior", lo que no es otra cosa que darle a la integral un nombre. El que la introducción de tal nueva función sea conveniente depende de las propiedades que posee, la frecuencia con la que aparece y la facilidad con la cual puede ser manipulada en la teoría y en la práctica. En este sentido el proceso de integración es un principio general para la generación de nuevas funciones.

Cierta familiaridad con este principio se ha obtenido ya al tratar con las funciones elementales. Así, se tuvo necesidad de introducir (p. 167) la integral de $1/x$ como una nueva función, la cual se denominó logaritmo y cuyas propiedades pudieron fácilmente ser deducidas. Las funciones trigonométricas pudieron haberse introducido del mismo modo, haciendo uso solamente de las funciones racionales, el proceso de integración y la operación de inversión. Con este objeto, se requiere tomar solamente una u otra de las ecuaciones

$$\arctan x = \int_0^x \frac{dt}{1 + t^2},$$

o

$$\operatorname{arc\,sen} x = \int_0^x \frac{dt}{\sqrt{1 - t^2}},$$

como la *definición* de la función arc tan x, o bien arc sen x, respectivamente, y obtener entonces por inversión las funciones trigonométricas. Mediante este proceso la definición de estas funciones está divorciada de la intuición geométrica (en particular, de la noción intuitiva de "ángulo"), pero queda entonces la tarea de desarrollar estas propiedades independientemente de la geometría.[1] (Posteriormente, en la sección 3.16 se dará otra discusión puramente analítica de las funciones trigonométricas.)

**Integrales elípticas*

El primer ejemplo importante que rebasa los límites del conjunto de las funciones elementales está dado por las *integrales elípticas*. Estas son

[1] No continuaremos aquí con el desarrollo de estas ideas. El paso esencial es el de probar los teoremas de adición para las funciones inversas, esto es, para el seno y para la tangente.

integrales en las que el integrando depende racionalmente de la raíz cuadrada de un polinomio de tercer o cuarto grado. Entre estas integrales, la función

$$u(s) = \int_0^s \frac{dx}{\sqrt{(1 - x^2)(1 - k^2x^2)}}$$

ha llegado a ser particularmente importante. Su función inversa $s(u)$ también desempeña un papel importante.[2] Esta función $s(u)$ ha sido examinada detenidamente y tabulada, como se hizo con las funciones elementales.[3]

Es el prototipo de las llamadas *funciones elípticas,* que tienen un lugar principalísimo en la teoría de funciones de variable compleja y aparecen en muchas aplicaciones físicas (por ejemplo, en conexión con el movimiento de un péndulo simple; ver p. 429).

El nombre "integral elíptica" surge del hecho de que tales integrales aparecen en el problema de determinar la longitud de arco de una elipse (Ver capítulo 4, pp. 395-396).

Debe hacerse notar además que integrales que a primera vista tienen una apariencia diferente resultan ser integrales elípticas después de una substitución sencilla. Como un ejemplo, la integral

$$\int \frac{dx}{\sqrt{\cos\alpha - \cos x}}$$

es transformada por medio de la substitución $u = \cos x/2$ en la integral

$$-k\sqrt{2}\int \frac{du}{\sqrt{(1 - u^2)(1 - k^2u^2)}}, \qquad k = \frac{1}{\cos(\alpha/2)};$$

y la integral

$$\int \frac{dx}{\sqrt{\cos 2x}}$$

se convierte por medio de la substitución $u = \operatorname{sen} x$ en

$$\int \frac{du}{\sqrt{(1 - u^2)(1 - 2u^2)}},$$

y, finalmente, la integral

$$\int \frac{dx}{\sqrt{1 - k^2 \operatorname{sen}^2 x}}$$

es transformada por medio de la substitución $u = \operatorname{sen} x$ en

$$\int \frac{du}{\sqrt{(1 - u^2)(1 - k^2u^2)}}.$$

[2] Para el valor especial $k = 0$ se obtiene $u(s) = \operatorname{arc\,sen} x$ y $s(u) = \operatorname{sen} u$ respectivamente.

[3] La función $s(u)$, una de las llamadas *funciones elípticas jacobianas,* es denotada usualmente por el símbolo $\operatorname{sn} u$ para indicar que es una generalización de la función seno ordinaria.

b. Sobre la derivación y la integración

Debe agregarse otro comentario sobre la relación entre derivación e integración. La derivación puede ser considerada como un proceso más elemental que la integración, puesto que ella no conduce fuera del dominio de las funciones "conocidas". Por otra parte, debe recordarse que la derivabilidad de una función continua, arbitraria, no es de ninguna manera una conclusión predeterminada sino una hipótesis estipulada. En efecto, como se ha visto, existen funciones continuas que no son derivables en ciertos puntos aislados; en tanto que desde los tiempos de Weierstrass han sido construidos muchos ejemplos de funciones continuas que no poseen derivada en ninguna parte.[1] En contraste, aun cuando la integración en términos de funciones elementales no es posible en general, es cierto, al menos, que la integral de una función continua existe.

Tomadas en conjunto, la integración y la derivación no pueden ser comparadas entre sí sencillamente como operaciones más elementales o menos elementales una respecto a la otra; desde algunos puntos de vista, la primera puede ser considerada como más elemental; desde otros puntos de vista, la segunda.

En cuanto a lo que concierne al concepto de integral, no se utilizará en la siguiente sección la hipótesis de que el integrando es continuo en todas partes; y se verá que la integral puede ser extendida a una amplia clase de funciones que posee discontinuidades.

3.15 Extensión del concepto de integral

a. Introducción. Definición de integrales "impropias"

En el capítulo 2 (p. 150) se definió $\int_a^b f(x)\,dx$ formando las "sumas de Riemann",

$$F_n = \sum_{i=1}^{n} f(\xi_i)\,\Delta x_i,$$

basadas en la subdivisión del intervalo $[a, b]$ en n subintervalos de longitudes Δx_i y en una elección de puntos intermedios ξ_i de esos subintervalos. Si las sucesiones F_n tienden al mismo límite, F_a^b, para cualquier sucesión de subdivisiones y puntos intermedios, cuando el valor Δx_i máximo tiende a cero, se define $\int_a^b f(x)\,dx$ como el límite F_a^b. Se mostró que este límite existe

[1] Compárese con Titchmarsh, *The Theory of Functions,* Oxford, 1932, Secciones 11.21 a 11.23, pp. 350-354.

cuando $f(x)$ es continua en $[a, b]$. Sin embargo, a menudo se presenta la necesidad de definir una integral cuando $f(x)$ no está definida o no es continua *en todos* los puntos del intervalo cerrado I, o bien cuando el intervalo de integración se extiende a infinito. Sería deseable, por ejemplo, adjudicar un significado apropiado a expresiones tales como

$$\int_0^1 \frac{1}{\sqrt{x}}\,dx, \qquad \int_0^1 \operatorname{sen}\frac{1}{x}\,dx,$$

$$\int_0^\infty e^{-x}\,dx, \qquad \int_1^\infty \frac{\operatorname{sen} x}{x^2}\,dx, \quad \text{etc.}$$

En primer lugar, se extiende el concepto de integral a funciones que son continuas en el intervalo abierto (a, b) pero que no están necesariamente definidas o son continuas en los puntos extremos a, b. Para números cualesquiera α, β, con $a < \alpha < \beta < b$, la integral ordinaria ("propia") $\int_\alpha^\beta f(x)\,dx$ queda entonces definida. Si

$$F = \lim_{\varepsilon \to 0} \int_{\alpha_\varepsilon}^{\beta_\varepsilon} f(x)\,dx$$

existe cuando $a < \alpha_\varepsilon < \beta_\varepsilon < b$ y $\lim_{\varepsilon\to 0} \alpha_\varepsilon = a$, $\lim_{\varepsilon\to 0} \beta_\varepsilon = b$, y si F es independiente de la elección particular de α_ε y β_ε, se dice que la integral impropia $\int_a^b f(x)\,dx$ converge y posee el valor F.

Integrandos seccionalmente continuos. Si, más generalmente, $f(x)$ está definida y es continua en (a, b) con la posible excepción de un número finito de puntos intermedios $c_1, c_2, \ldots, c_n$, y si f es continua en cada uno de los intervalos abiertos (a, c_1), (c_1, c_2), $\ldots$, (c_n, b) se define $\int_a^b f(x)\,dx$ como la suma de integrales impropias sobre los subintervalos, siempre que cada una de éstas converja.

La integral impropia $\int_a^b f(x)\,dx$ siempre converge cuando f es continua y *acotada* en un intervalo abierto (a, b). Por ejemplo, la integral

$$\int_0^1 \operatorname{sen}\frac{1}{x}\,dx = \lim_{\varepsilon\to 0} \int_\varepsilon^1 \operatorname{sen}\frac{1}{x}\,dx$$

converge. Para probar esta propiedad general puede suponerse, por brevedad, que f es continua en b, pero no necesariamente en a. Entonces, por definición,

$$\int_a^b f(x)\,dx = \lim_{\alpha\to a} F(\alpha),$$

donde $F(\alpha)$, para $a < \alpha < b$, está definida como $\int_a^b f(x)\,dx$. Si M es una cota superior para f y α_n es una sucesión que tiende a a, se tiene, por el teorema del valor medio del cálculo integral, $|F(\alpha_n) - F(\alpha_m)| \leq M\,|\alpha_n - \alpha_m|$; y, consecuentemente, por el criterio de convergencia de Cauchy, $\lim_{n\to\infty} F(\alpha_n)$ existe. Puesto que éste es el caso para cualquier sucesión α_n que converge a a, se sigue que $\lim_{\alpha\to a} F(\alpha)$ existe.

De hecho, cuando f es continua y acotada en (a,b) se pueden asignar a f cualesquiera valores en los puntos extremos a, b y obtener también $\int_a^b f(x)\,dx$ directamente como una integral "propia" definida como el límite de sumas de Riemann. Se ve fácilmente que para funciones continuas y acotadas f ambas definiciones se aplican y conducen al mismo valor, independientemente de la elección de $f(a)$ y $f(b)$. Lo mismo es cierto, más generalmente, para funciones acotadas que están definidas y son continuas en (a, b) con la posible excepción de un número finito de puntos. En particular, $\int_a^b f(x)\,dx$ siempre existe cuando f es continua excepto en un

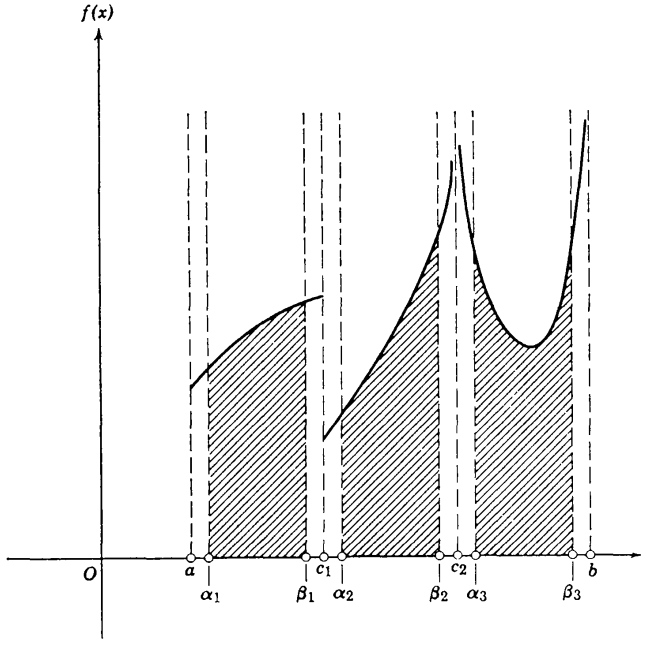


Figura 3.30 La integral de una función con discontinuidades.

número finito de discontinuidades de salto. Por todo lo anterior, la convergencia de una integral impropia de una función sobre un intervalo finito requiere atención solamente cuando f se torna infinita.

Nótese que la interpretación geométrica de la integral como el área bajo la curva es la misma que para una f continua (Fig. 3.30).

b. Funciones con discontinuidades infinitas

Se comenzará con la integral

$$J = \int_0^1 \frac{dx}{x^\alpha},$$

donde α es un número positivo. El integrando $1/x^\alpha$ se torna infinito para $x \to 0$. Por lo tanto, se debe definir J tomando la integral J_ε desde el límite positivo ε hasta el límite 1, y, finalmente, hacer que ε tienda a cero. De acuerdo con las reglas elementales de integración se obtiene, suponiendo $\alpha \neq 1$,

$$J_\varepsilon = \int_\varepsilon^1 \frac{dx}{x^\alpha} = \frac{1}{1-\alpha}(1 - \varepsilon^{1-\alpha}).$$

Se observan de inmediato las siguientes posibilidades: 1) α es mayor que 1; entonces, para $\varepsilon \to 0$ el miembro derecho tiende a infinito. 2) α es menor que 1; entonces el miembro derecho tiende al límite $1/(1-\alpha)$. En el segundo caso, por lo tanto, solamente se tiene que tomar este valor límite como la integral $J = \int_0^1 \frac{dx}{x^\alpha}$. En el primer caso se dice que la integral desde 0 hasta 1 *no existe*, o bien, que *diverge*. 3) En el tercer caso, en el que $\alpha = 1$, la integral es igual a $-\log \varepsilon$ y, por lo tanto, para $\varepsilon \to 0$ no tiende a un límite sino que tiende a infinito; esto es, la integral $\int_0^1 \frac{dx}{x} = J$ no existe, o sea, es divergente.

Otro ejemplo de un integrando con una discontinuidad infinita está dado por $f(x) = 1/\sqrt{1-x^2}$. Se encuentra

$$\int_0^{1-\varepsilon} \frac{dx}{\sqrt{1-x^2}} = \operatorname{arc\,sen}(1-\varepsilon).$$

Para $\varepsilon \to 0$ el miembro derecho converge al límite $\pi/2$; éste es, entonces, el valor de la integral:

$$\frac{\pi}{2} = \int_0^1 \frac{dx}{\sqrt{1-x^2}},$$

aun cuando el integrando se torna infinito en el punto $x = 1$.

c. Interpretación de la integral como un área

Las integrales impropias pueden ser interpretadas como áreas de regiones que se extienden al infinito definidas a partir de regiones acotadas, por medio del paso al límite. Por ejemplo, los resultados anteriores para la función $1/x^\alpha$ establecen que el área acotada por el eje x, la recta $x = 1$, la recta $x = \varepsilon$ y la curva $y = 1/x^\alpha$

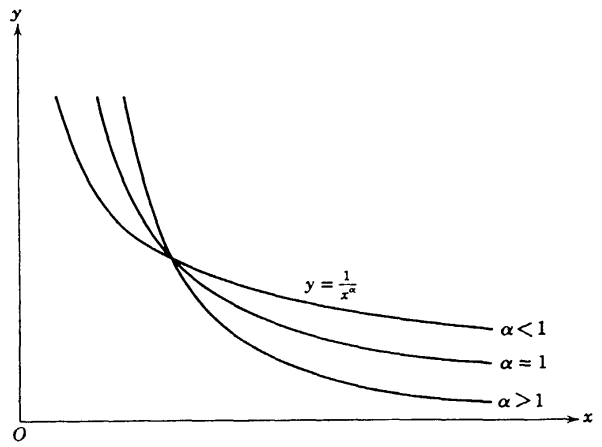


Figura 3.31 Para ilustrar la convergencia o la divergencia de integrales impropias.

tiende a un límite finito cuando $\varepsilon \to 0$, siempre que $\alpha < 1$, y que tiende a infinito si $\alpha \geq 1$. Este hecho puede ser expresado como sigue: El área entre el eje x, el eje y, la curva $y = 1/x^\alpha$ y la recta $x = 1$ es finita o infinita según que $\alpha < 1$ o $\alpha \geq 1$.

La intuición no puede dar, por supuesto, una información segura sobre la finitud o la infinitud del área de una región que se extiende al infinito. La figura 3.31 ilustra el hecho de que para $\alpha < 1$ el área bajo la curva permanece finita, en tanto que para $\alpha \geq 1$ es infinita; un hecho que, ciertamente, no es sugerido por la intuición geométrica.

d. Criterios para determinar la convergencia

Para probar la convergencia de la integral de una función $f(x)$ con una discontinuidad infinita en el punto $x = b$, puede hacerse uso a menudo del siguiente criterio.

Sea la función $f(x)$ continua en el intervalo $a < x < b$, y supóngase que $\lim_{x \to b} f(x) = \infty$. Entonces, la integral $\int_a^b f(x)\,dx$ *converge* si existe un número positivo μ menor que uno y un número fijo M, independientes de x, tal que en todas partes del intervalo $a < x < b$ se cumple la desigualdad $|f(x)| \leq M/(b-x)^\mu$; en en otras palabras, *si en el punto $x = b$ la función $f(x)$ se torna infinita de orden menor que el primero: $f(x) = 0[1/(b-x)^\mu]$* para algún $\mu < 1$. Por otra parte, la integral *diverge* si existen un número $\nu \geq 1$ y un número fijo N tales que en todas partes del intervalo $a \leq x < b$ se cumple la desigualdad $f(x) \geq N/(b-x)^\nu$; en otras palabras, *si en el punto $x = b$ la función positiva $f(x)$ se torna infinita de primer orden al menos.*

La demostración surge casi de inmediato por comparación con el caso muy sencillo que se acaba de discutir. Para probar la primera parte del teorema se observa que para $0 < \varepsilon < b - a$ se tiene

$$0 \leq \frac{M}{(b-x)^\mu} + f(x) \leq \frac{2M}{(b-x)^\mu},$$

y, por lo tanto, también

$$0 \leq \int_a^{b-\varepsilon} \left[\frac{M}{(b-x)^\mu} + f(x) \right] dx \leq \int_a^{b-\varepsilon} \frac{2M}{(b-x)^\mu}\, dx.$$

Si $\varepsilon \to 0$, el integrando en la derecha, el cual se obtiene a partir de la integral $\int dx/x^\mu$ mediante una sencilla substitución de x por $b - x$, posee un límite, y, por ende, permanece acotado. Además, los valores de la integral de en medio crecen monótonamente conforme $\varepsilon \to 0$; puesto que también están acotados, deben poseer un límite, y así la integral

$$\int_a^b \left(\frac{M}{(b-x)^\mu}\, dx + f(x) \right) dx = \lim_{\varepsilon \to 0} \left(\int_a^{b-\varepsilon} \frac{2M}{(b-x)^\mu}\, dx + \int_a^{b-\varepsilon} f(x)\, dx \right)$$

converge. La convergencia de la integral de $M/(b-x)^\mu$ implica entonces también la de $\int_a^b f(x)\,dx$.

La demostración de la segunda parte del teorema se deja como un ejercicio para el lector.

De la misma manera se ve de inmediato que teoremas exactamente análogos se cumplen cuando el límite o extremo *inferior* de integración es un punto de discontinuidad infinita. Si un punto de discontinuidad infinita está situado en el interior del intervalo de integración, se separa simplemente el intervalo en dos subintervalos por medio de este punto y se aplican entonces estas consideraciones a cada uno de esos subintervalos.

Como un ejemplo se considera la integral elíptica

$$\int_0^1 \frac{dx}{\sqrt{(1-x^2)(1-k^2x^2)}} \qquad (k^2 < 1).$$

Por la identidad $1 - x^2 = (1-x)(1+x)$ se ve de inmediato que conforme $x \to 1$ el integrando se torna infinito, pero sólo de orden $\frac{1}{2}$; de lo que se sigue que la integral impropia converge. (Para $k = 1$ la integral diverge.)

e. Intervalo de integración infinito

Otra extensión importante del concepto de integral se refiere a la de un intervalo de integración infinito. Para una formulación precisa se introduce la siguiente notación: Si la integral

$$\int_a^A f(x)\, dx,$$

con a fijo, tiende a un límite definido cuando $A \to \infty$, la integral de $f(x)$ sobre el intervalo infinito $x \geq a$ se define como

$$\lim_{A \to \infty} \int_a^A f(x)\, dx = \int_a^\infty f(x)\, dx.$$

Otra vez, tal integral se denomina *convergente.*

Ejemplos. Ejemplos sencillos de las diversas posibilidades están dados aquí también por las funciones $f(x) = 1/x^\alpha$:

$$\int_1^A \frac{dx}{x^\alpha} = \frac{1}{1-\alpha}\,(A^{1-\alpha} - 1).$$

Aquí se ve, una vez más, que si se excluye el caso $\alpha = 1$, la integral extendida al infinito existe para el caso $\alpha > 1$, y, en efecto,

$$\int_1^\infty \frac{dx}{x^\alpha} = \frac{1}{\alpha - 1}.$$

Cuando $\alpha < 1$ la integral ya no existe. Para el caso $\alpha = 1$, claramente la integral de nuevo deja de existir puesto que $\log x$ tiende a infinito cuando x tiende a infinito. Por ello se ve que, en lo que respecta a la integración sobre un intervalo infinito, las funciones $1/x^\alpha$ no se comportan del mismo modo que para la integración hasta el origen. Esta afirmación se hace también plausible mediante una ojeada a la Fig. 3.31. Pues, obviamente, mientras más grande sea α las curvas se acercarán más rápidamente al eje x para $x \to \infty$; y así es plausible que el área considerada tienda a un límite definido para valores suficientemente grandes de α.

El siguiente criterio para la existencia de una integral con un límite infinito es a menudo útil. (Se supone nuevamente que para valores suficientemente grandes de x, digamos, para $x \geq a$, el integrando es continuo.)

Criterio de convergencia

La integral $\int_a^\infty f(x)\,dx$ *converge si la función* $f(x)$ *se anula en el infinito con un orden superior al primero;* esto es, si existe un número $\nu > 1$ tal que para todos los valores de x suficientemente grandes se cumple la relación $|f(x)| \leq M/x^\nu$, donde M es un número fijo independiente de x. En símbolos: $f(x) = 0\left(\frac{1}{x^\nu}\right)$ Por otra parte, *la integral diverge si la función permanece positiva y se anula en el infinito con un orden no superior al primero,* esto es, si existe un número fijo $N > 0$ tal que $xf(x) \geq N$.

La demostración de estos criterios es exactamente paralela al argumento anterior y puede dejarse al lector.

Un ejemplo muy sencillo es la integral $\int_a^\infty \frac{1}{x^2}\,dx$ $(a > 0)$. El integrando se anula en el infinito como infinitésimo de segundo orden. Se ve de inmediato que la integral converge, pues $\int_a^A \frac{dx}{x^2}\,dx = \frac{1}{a} - \frac{1}{A}$, y, por ende,

$$\int_a^\infty \frac{1}{x^2}\,dx = \frac{1}{a}.$$

Otro ejemplo igualmente simple es

$$\int_0^\infty \frac{1}{1+x^2}\,dx = \lim_{A \to \infty} (\arctan A - \arctan 0) = \frac{\pi}{2}.$$

Entonces, evidentemente también se tiene

$$\int_{-\infty}^{+\infty} \frac{1}{1+x^2}\,dx = \pi,$$

puesto que el integrando es una función par. Es curioso que el área entre la curva $y = 1/(1 + x^2)$ y el eje x (ver Fig. 3.8, p. 238), que se extiende al infinito, resulte ser la misma que la del círculo de radio uno.

f. La función gamma

Otro ejemplo de particular importancia en el análisis es el de la llamada función gamma:

$$\Gamma(n) = \int_0^\infty e^{-x}x^{n-1}\,dx \qquad (n > 0).$$

Partiendo el intervalo de integración en dos partes, una desde $x = 0$ hasta $x = 1$ y otra desde $x = 1$ hasta $x = \infty$, se ve que la integral sobre la primera parte converge, puesto que $0 < e^{-x}x^{n-1} < 1/x^\mu$ con $\mu = 1 - n < 1$. Para la integral sobre la segunda, la parte infinita, el criterio de convergencia también se satisface; por ejemplo, para $\nu = 2$ se tiene $\lim_{x\to\infty} x^2e^{-x}x^{n-1} = 0$, puesto que la función exponencial e^{-x} tiende a cero con un orden superior al de cualquier potencia $1/x^m$ $(m > 0)$ (ver p. 274). Esta función gamma, que se considera como una función del número n (no necesariamente un entero), satisface una relación notable que se obtiene mediante integración por partes como sigue. Primero se tiene (con $f(x) = x^{n-1}$, $g'(x) = e^{-x}$)

$$\int e^{-x}x^{n-1}\,dx = -e^{-x}x^{n-1} + (n-1)\int e^{-x}x^{n-2}\,dx.$$

Si se toma esta relación integral entre 0 y A y entonces se deja crecer A más allá de toda cota, se obtiene inmediatamente

$$\Gamma(n) = (n-1)\int_0^\infty e^{-x}x^{n-2}\,dx = (n-1)\Gamma(n-1) \qquad \text{para} \quad n > 1;$$

y de esta fórmula de recurrencia, siempre que μ sea un entero y $0 < \mu < n$, se sigue que

$$\Gamma(n) = (n-1)(n-2)\cdots(n-\mu)\int_0^\infty e^{-x}x^{n-\mu-1}\,dx.$$

En particular, si n es un entero positivo se tiene, con $\mu = n - 1$, la relación

$$\Gamma(n) = (n-1)(n-2)\cdots 3\cdot 2\cdot 1\int_0^\infty e^{-x}\,dx;$$

y, puesto que

$$\int_0^\infty e^{-x}\,dx = 1,$$

se tiene finalmente

$$\Gamma(n) = (n-1)(n-2)\cdots 2\cdot 1 = (n-1)!,$$

una expresión muy útil de un factorial mediante una integral.

Otros ejemplos. Las integrales

$$\int_0^\infty e^{-x^2}\,dx, \qquad \int_0^\infty x^n e^{-x^2}\,dx$$

también convergen, como puede deducirse fácilmente a partir del criterio antes dado. La primera es idéntica con $\frac{1}{2}\Gamma(\frac{1}{2})$, la segunda con $\frac{1}{2}\Gamma[(n+1)/2]$ para $n > -\frac{1}{2}$, como se ve mediante la substitución $x^2 = u$, $dx = (1/2\sqrt{u})\,du$.

g. La integral de Dirichlet

En muchas aplicaciones se encuentran integrales cuya convergencia no surge directamente a partir de nuestro criterio. Un ejemplo importante es proporcionado por la integral

$$I = \int_0^\infty \frac{\operatorname{sen} x}{x}\,dx$$

investigada por Dirichlet. Si el límite superior no es infinito sino finito, la integral es convergente puesto que la función $(\operatorname{sen} x)/x$ es continua para todo x finito (para $x = 0$, está dada por $\lim \dfrac{\operatorname{sen} x}{x} = 1$ con $x \to 0$). La convergencia de la integral I se debe al cambio periódico del signo del integrando, lo que hace que las contribuciones a la integral debidas a los intervalos vecinos de longitud π casi se cancelen unas con otras (Fig. 3.32). Así, la suma del número infinito de áreas entre el eje x

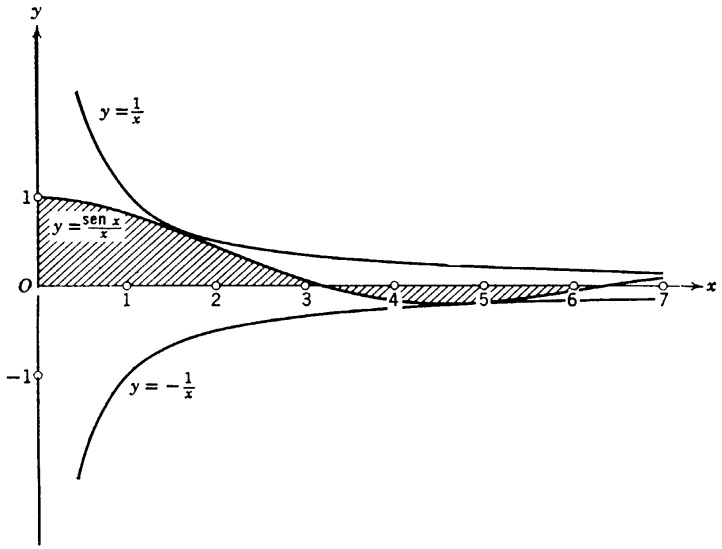


Figura 3.32 Gráfica de $y = \dfrac{\operatorname{sen} x}{x}$.

y la curva $y = \dfrac{\operatorname{sen} x}{x}$ converge, si se cuentan las áreas por encima del eje x como positivas y por debajo del eje x como negativas. (Por otra parte, la suma de los valores numéricos de todas las áreas, esto es, la integral

$$\int_0^\infty \frac{|\operatorname{sen} x|}{x}\,dx$$

como puede fácilmente mostrarse, diverge.)

El carácter alternante de la función sen x es causa del hecho de que su integral indefinida

$$\int \operatorname{sen} x\,dx = 1 - \cos x$$

está acotada para todo x. Se hace uso de esto al calcular el valor de la expresión

$$I_{AB} = \int_A^B \frac{\operatorname{sen} x}{x}\,dx = \int_A^B \frac{1}{x}\frac{d(1-\cos x)}{dx}\,dx$$

La integración por partes muestra que

$$I_{AB} = \frac{1-\cos B}{B} - \frac{1-\cos A}{A} + \int_A^B \frac{1-\cos x}{x^2}\,dx.$$

Por lo tanto,

$$\int_0^\infty \frac{\operatorname{sen} x}{x}\,dx = \lim_{\substack{A\to 0\\ B\to\infty}} I_{AB} = \int_0^\infty \frac{1-\cos x}{x^2}\,dx,$$

donde la integral en el miembro derecho es claramente convergente. En otras palabras, la integral I existe. En la sección 8.4*c* se establecerá, además, el hecho notable de que I posee el valor $\pi/2$.

h. Substitución. Integrales de Fresnel

Evidentemente, todas las reglas para la substitución de nuevas variables, etcétera, permanecen válidas para integrales impropias convergentes. A menudo tales transformaciones pueden conducir a expresiones diferentes más manejables para la integral.

Como un ejemplo, para calcular

$$\int_0^\infty xe^{-x^2}\,dx$$

se introduce la nueva variable $u = x^2$ y se obtiene

$$\int_0^\infty xe^{-x^2}\,dx = \frac{1}{2}\int_0^\infty e^{-u}\,du = \lim_{A\to\infty}\frac{1}{2}(1-e^{-A}) = \frac{1}{2}.$$

Otro ejemplo en la investigación de integrales impropias está dado por las *integrales de Fresnel,* que aparecen en la teoría de difracción de la luz:

$$F_1 = \int_0^\infty \operatorname{sen}(x^2)\,dx, \qquad F_2 = \int_0^\infty \cos(x^2)\,dx.$$

La substitución $x^2 = u$ conduce a

$$F_1 = \frac{1}{2}\int_0^\infty \frac{\operatorname{sen} u}{\sqrt{u}}\,du, \qquad F_2 = \frac{1}{2}\int_0^\infty \frac{\cos u}{\sqrt{u}}\,du.$$

Integrando por partes se encuentra

$$\int_A^B \frac{\operatorname{sen} u}{\sqrt{u}}\,du = \frac{1-\cos B}{\sqrt{B}} - \frac{1-\cos A}{\sqrt{A}} + \frac{1}{2}\int_A^B \frac{1-\cos u}{u^{3/2}}\,du.$$

Si A y B tienden a cero e infinito respectivamente, se ve, mediante el mismo argumento que para la integral de Dirichlet, que la integral F_1 converge. La convergencia de la integral F_2 se prueba exactamente de la misma manera.

Estas integrales de Fresnel muestran que una integral impropia puede existir aún si el integrando no tiende a cero cuando $x \to \infty$. En efecto, una integral impropia puede existir aún cuando el integrando no esté acotado, como se muestra mediante el ejemplo

$$\int_0^\infty 2u\cos(u^4)\,du.$$

Cuando $u^4 = n\pi$, esto es, cuando $u = \sqrt[4]{n\pi}$, $n = 0, 1, 2, \ldots$, el integrando se convierte en $2\sqrt[4]{n\pi}\cos n\pi = \pm 2\sqrt[4]{n\pi}$, de modo que no es acotado. Mediante la substitución $u^2 = x$, sin embargo, la integral se reduce a

$$\int_0^\infty \cos(x^2)\,dx,$$

la cual, justamente, se ha visto que es convergente.

Por medio de una substitución una integral impropia puede ser a menudo transformada en una integral propia. Por ejemplo, la transformación $x = \operatorname{sen} u$ conduce a

$$\int_0^1 \frac{dx}{\sqrt{1-x^2}} = \int_0^{\pi/2} du = \frac{\pi}{2}.$$

Por otra parte, integrales de funciones continuas pueden ser transformadas en integrales impropias. Esto sucede si la transformación $u = \phi(x)$ es tal que en un extremo del intervalo de integración la derivada $\phi'(x)$ se anula, de modo que dx/du es infinita.

3.16 Las ecuaciones diferenciales de las funciones trigonométricas

a. Comentarios introductorios sobre las ecuaciones diferenciales

La integración es simplemente un primer paso en un campo mucho más extenso: En vez de invertir la derivación mediante la integración, esto es, resolver la ecuación $y' = f(x)$ con una $f(x)$ dada, para tener $y = F(x)$, puede desearse encontrar funciones $y = F(x)$ que satisfagan relaciones más generales entre y y las derivadas de y. Tales "ecuaciones diferenciales" aparecen en todas partes en las aplicaciones, así como en contextos estrictamente teóricos. Sobre estas ecuaciones se hacen estudios penetrantes que van mucho más allá del alcance de este libro. Se volverá sobre algunos aspectos elementales de la teoría de las ecuaciones diferenciales posteriormente, en éste y en el siguiente volumen. Por ahora la discusión se restringirá a ejemplos bastante sencillos pero significativos. Se discutirán las ecuaciones diferenciales de las funciones $\operatorname{sen} x$ y $\cos x$, las que han sido mencionadas ya en la p. 192.

Aunque en la trigonometría elemental estas funciones y sus propiedades fueron tomadas desde el punto de vista geométrico, se descarta ahora la confianza en la intuición geométrica y se expresan las funciones trigonométricas en una forma sencilla sobre una base precisa, analítica, de acuerdo con la tendencia general del desarrollo mencionado anteriormente.

b. Sen x y cos x definidos mediante una ecuación diferencial y condiciones iniciales

Se considera la ecuación diferencial

$$u'' + u = 0$$

con el deseo de caracterizar soluciones $u(x)$ que serán identificadas con las funciones seno y coseno. Cualquier función $u = F(x)$ que satisface la ecuación, esto es, para la cual $F''(x) + F(x) = 0$, se denomina una *solución*.[1]

Salta a la vista que, junto con una solución $u = F(x)$, la función $u = F(x + h)$, para una constante arbitraria h, es también una solución, como se verifica de inmediato derivando $F(x + h)$ dos veces con respecto a x. Análogamente, se ve de inmediato que, conjuntamente con $F(x)$, la derivada $F'(x) = u$ es también una solución; como también lo es, claro está, $cF(x)$, con el factor constante c. Además, junto con $F_1(x)$ y $F_2(x)$, cualquier combinación lineal $c_1F_1(x) + c_2F_2(x) = F(x)$, con constantes c_1 y c_2, es una solución.

[1] Por supuesto, siempre se entiende que las funciones consideradas son derivables un número suficiente de veces.

Para destacar de la multitud de soluciones de una ecuación diferencial una específica se imponen "condiciones iniciales" estipulando que para $x = 0$ los valores de $u = F(0)$ y $u' = F'(0)$ estén prefijados como a y b respectivamente. Se establece primero el siguiente *teorema:*

La solución está determinada de manera única por estos valores iniciales.

Para la demostración, se comienza con un comentario general válido para cualquier solución u. Multiplicando la ecuación diferencial por $2u'$ se encuentra, debido a que $2u''u' = (u'^2)'$ y $2u'u = (u^2)'$, la ecuación

$$0 = 2u''u' + 2u'u = [(u')^2 + u^2]',$$

la cual puede ser integrada de inmediato e implica

$$u'^2 + u^2 = c,$$

donde c es una constante, esto es, no depende de x. Por lo tanto, c debe poseer el mismo valor que el miembro izquierdo para $x = 0$. Así se tiene que para cualquier solución u es

$$u'^2(0) + u^2(0) = c.$$

Ahora supóngase que se tienen dos soluciones u_1 y u_2 con las mismas condiciones iniciales: Entonces la diferencia $z = u_1 - u_2$ es una solución con $z'(0) = z(0) = 0$. Por lo tanto, se tiene $c = 0$ y para todo x: $x'^2 + z^2 = 0$. Esto significa que $z = 0$ y $z' = 0$, lo que demuestra el aserto.

Se definen ahora las funciones $\operatorname{sen} x$ y $\cos x$ como aquellas soluciones de la ecuación diferencial $u''(x) + u(x) = 0$ para las cuales las condiciones iniciales son, respectivamente, para $u = \operatorname{sen} x$,

$$u(0) = a = 0, \qquad u'(0) = b = 1,$$

y para $u = \cos x$,

$$u(0) = a = 1, \qquad u'(0) = b = 0.$$

Se toma por cierto aquí que tales soluciones existen y son arbitrariamente derivables cuantas veces se quiera; mientras que la demostración de esto se dará posteriormente en un contexto más general (ver sección 9.2).[1]

La única solución u de $u'' + u = 0$ para la cual es $u = a$, $u' = b$ para $x = 0$ es entonces la función $u = a \cos x + b \operatorname{sen} x$. Esto prueba que toda solución de la ecuación diferencial es una combinación lineal de $\cos x$ y $\operatorname{sen} x$.

Se obtienen ahora las propiedades básicas de las funciones trigonométricas a partir de la ecuación diferencial $u'' + u = 0$ aplicada, por ejemplo,

[1] Obsérvese que estos hechos pueden inferirse inmediatamente a partir de la ecuación $u'^2 + u^2 = 1$, la cual es válida para $\operatorname{sen} x$, así como también para $\cos x$, y, a partir de su forma equivalente $dx/du = 1/\sqrt{1 - u^2}$, las funciones inversas de $\operatorname{sen} x$ y $\cos x$ son obtenidas inmediatamente por integración.

a la función $u = \operatorname{sen} x$. Obviamente, si u es una solución, $v = u'$ también lo es: $v'' + v = 0$. Debido a que $u'' + u = v' + u = 0$, se tiene $v'(0) = -u(0) = 0$, mientras que $v(0) = u'(0) = 1$. Por lo tanto,

$$v(x) = \cos x = \frac{d}{dx} \operatorname{sen} x.$$

Análogamente se deduce $(d/dx) \cos x = -\operatorname{sen} x$.

El teorema central de la trigonometría es el teorema de adición:

$$\cos (x + y) = \cos x \cos y - \operatorname{sen} x \operatorname{sen} y.$$

Se sigue de inmediato, a partir del presente enfoque, que: en primer lugar, la función $\cos (x + y)$, como una función de x, manteniendo por el momento a y constante, es una solución $u(x)$ de la ecuación diferencial $u'' + u = 0$, que satisface para $x = 0$ las condiciones iniciales $u(0) = \cos y = a$ y $u'(0) = -\operatorname{sen} y = b$; en segundo lugar, la solución —que, de acuerdo con el teorema anterior, es la única— para la cual $u(0) = a$ y $u'(0) = b$, es precisamente $a \cos x + b \operatorname{sen} x$, como se verifica de inmediato. Por lo tanto, se tiene de inmediato para nuestra solución $\cos (x + y)$ la expresión

$$\cos (x + y) = \cos x \cos y - \operatorname{sen} x \operatorname{sen} y,$$

como se quería probar.

Los comentarios en esta sección podrían ser suficientes para indicar cómo las funciones trigonométricas pueden ser introducidas de una manera enteramente analítica sin referencia alguna a la geometría.

Sin entrar en más detalles, se menciona lo siguiente.

El número $\frac{1}{2}\pi$ podría ser definido ahora como el menor valor positivo de x para el cual $\cos x = 0$.

La periodicidad de las funciones trigonométricas surge asimismo fácilmente a partir del enfoque analítico.

Se volverá a la construcción analítica de las funciones trigonométricas mediante las series infinitas de potencias (ver sección 5.56).

PROBLEMAS

SECCION 3.1, página 223

1. Sea $P(x) = a_0 + a_1 x + a_2 x^2 + \cdots + a_n x^n$.
(a) Calcular el polinomio $F(x)$ a partir de la ecuación

$$F(x) - F'(x) = P(x).$$

*(b) Calcular $F(x)$ a partir de la ecuación

$$c_0 F(x) + c_1 F'(x) + c_2 F''(x) = P(x).$$

2. Encontrar el límite, para $n \to \infty$, del valor absoluto de la n-ésima derivada de $1/x$ en el punto $x = 2$.

3. Probar que si $f^{(n)}(x) = 0$ para todo x, entonces f es un polinomio de grado $n - 1$ a lo más, y recíprocamente.

4. Determinar la forma de una función racional r para la cual

$$\lim_{x \to \infty} \frac{xr'(x)}{r(x)} = 0.$$

5. Probar por inducción que la n-ésima derivada de un producto puede encontrarse de acuerdo con la siguiente regla (regla de Leibnitz):

$$\frac{d^n}{dx^n}(fg) = f\frac{d^n g}{dx^n} + \binom{n}{1}\frac{df}{dx}\frac{d^{n-1}g}{dx^{n-1}} + \binom{n}{2}\frac{d^2 f}{dx^2}\frac{d^{n-2}g}{dx^{n-2}} + \cdots + \binom{n}{n-1}\frac{d^{n-1}f}{dx^{n-1}}\frac{dg}{dx} + \frac{d^n f}{dx^n} g.$$

Aquí $\binom{n}{1} = n$, $\binom{n}{2} = \frac{n(n-1)}{2!}$, etc., denotan los coeficientes binomiales.

6. Probar que $\sum_{i=1}^{n-1} i x^{i-1} = \frac{(n-1)x^n - nx^{n-1} + 1}{(x-1)^2}$.

SECCION 3.2, página 228

1. Sea $y = e^x(a \operatorname{sen} x + b \cos x)$. Demostrar que y'' puede expresarse como una combinación lineal de y e y', esto es,

$$y'' = py' + qy,$$

donde p y q son constantes. Expresar todas las derivadas superiores como combinaciones lineales de y' e y.

*2. Encontrar la n-ésima derivada de $\operatorname{arc\,sen} x$ en $x = 0$, y luego la de $(\operatorname{arc\,sen} x)^2$ en $x = 0$.

SECCION 3.3, página 239

1. Encontrar la segunda derivada de $f[g\{h(x)\}]$.

2. Derivar la siguiente función: $\log_{v(x)} u(x)$, [esto es, el logaritmo de $u(x)$ en la base $v(x)$; y sea $v(x) > 0$].

3. ¿Qué condiciones deben satisfacer los coeficientes α, β, a, b, c para que

$$\frac{\alpha x + \beta}{\sqrt{(ax^2 + 2bx + c)}}$$

posea en todas partes una derivada finita que en ningún caso se anule?

4. Mostrar que $d^n(e^{x^2/2})/dx^n = u_n(x)e^{x^2/2}$, donde $u_n(x)$ es un polinomio de grado n. Establecer la relación de recurrencia

$$u_{n+1} = xu_n + u_n'.$$

*5. Mediante la aplicación de la regla de Leibnitz a

$$\frac{d}{dx}(e^{x^2/2}) = xe^{x^2/2},$$

obténgase la relación de recurrencia

$$u_{n+1} = xu_n + nu_{n-1}.$$

***6.** Mediante la combinación de las relaciones de recurrencia de los problemas 4 y 5, obtener la ecuación diferencial

$$u_n'' + xu_n' - nu_n = 0$$

satisfecha por $u_n(x)$.

7. Encontrar la solución polinomial

$$u_n(x) = x^n + a_1x^{n-1} + \cdots + a_n$$

de la ecuación diferencial $u_n'' + xu_n' - nu_n = 0$.

***8.** Si $P_n(x) = \dfrac{1}{2^n n!}\dfrac{d^n}{dx^n}(x^2 - 1)^n$, probar las relaciones

(*a*) $P'_{n+1} = \dfrac{x^2 - 1}{2(n+1)}P_n'' + \dfrac{(n+2)x}{n+1}P_n' + \dfrac{n+2}{2}P_n$.

(*b*) $P'_{n+1} = xP_n' + (n+1)P_n$.

(*c*) $\dfrac{d}{dx}[(x^2 - 1)P_n'] - n(n+1)P_n = 0$.

9. Encontrar la solución polinomial

$$P_n = \frac{(2n)!}{2^n(n!)^2}x^n + a_1x^{n-1} + \cdots + a_n$$

de la ecuación diferencial

$$\frac{d}{dx}[(x^2 - 1)P_n'] - n(n+1)P_n = 0.$$

10. Determinar el polinomio $P_n(x) = \dfrac{1}{2^n n!}\dfrac{d^n}{dx^n}(x^2 - 1)^n$ utilizando el teorema del binomio.

***11.** Sea $\lambda_{n,p}(x) = \binom{p}{n}x^n(1 - x)^{p-n}$, $n = 0, 1, 2, \ldots, p$. Mostrar que

$$1 = \sum_{n=0}^{p}\lambda_{n,p}(x).$$

$$x^k = \sum_{n=k}^{p}\frac{\binom{n}{k}}{\binom{p}{k}}\lambda_{n,p}(x).$$

. . . . . .

$$x^p = \lambda_{p,p}(x).$$

. . . . . .

SECCION 3.4, página 244

1. La función $f(x)$ satisface la ecuación

$$f(x + y) = f(x)f(y).$$

a) Si $f(x)$ es derivable, entonces $f(x) \equiv 0$ o bien $f(x) = e^{ax}$.
*****b*) Si $f(x)$ es continua, entonces $f(x) \equiv 0$ ó bien $f(x) = e^{ax}$.

2. Si una función derivable $f(x)$ satisface la ecuación

$$f(xy) = f(x) + f(y),$$

entonces $f(x) = \alpha \log x$.

3. Probar que si $f(x)$ es continua y

$$f(x) = \int_0^x f(t)\,dt,$$

entonces $f(x)$ es idénticamente cero.

SECCION 3.5, página 250

1. Demostrar la fórmula

$$\operatorname{senh} a + \operatorname{senh} b = 2\operatorname{senh}\left(\frac{a+b}{2}\right)\cosh\left(\frac{a-b}{2}\right).$$

Obténganse fórmulas análogas para $\operatorname{sen} a - \operatorname{senh} b$, $\cosh a + \cosh b$, $\cosh a - \cosh b$.

2. Expresar $\tanh(a + b)$ en términos de $\tanh a$ y de $\tanh b$.
Expresar $\coth(a + b)$ en términos de $\coth a$ y $\coth b$.
Expresar $\operatorname{sen} \frac{1}{2}a$ y $\cosh \frac{1}{2}a$ en términos de $\cosh a$.

3. Derivar:
(*a*) $\cosh x + \operatorname{senh} x$; (*b*) $e^{\tanh x + \coth x}$,
(*c*) $\log \operatorname{senh}(x + \cosh^2 x)$; (*d*) $\operatorname{arc\,cosh} x + \operatorname{arc\,senh} x$,
(*e*) $\operatorname{arc\,senh}(a \cos x)$; (*f*) $(\operatorname{arc\,tanh}(2x/(1 + x^2)))$.

4. Calcular el área acotada por la catenaria $y = \cosh x$, las ordenadas $x = a$ y $x = b$ y el eje x.

SECCION 3.6, página 257

1. Determinar los máximos, mínimos y puntos de inflexión de $x^3 + 3px + q$. Discutir la naturaleza de las raíces de $x^3 + 3px + q = 0$.

2. Dada la parábola $y^2 = 2px$, $p > 0$, y un punto $P(x = \xi, y = \eta)$ dentro de ella ($\eta^2 < 2p\xi$), encontrar la trayectoria mínima (consistente en dos segmentos de recta) que va de P a un punto Q sobre la parábola y de ahí al foco $F(x = \frac{1}{2}p, y = 0)$ de la parábola. Demostrar que el ángulo FQP es bisectado por la normal a la parábola y que QP es paralelo al eje de la parábola (principio del espejo parabólico).

3. De todos los triángulos con base y ángulo vertical dados, el triángulo isósceles posee el área máxima.

4. De todos los triángulos con base y área dadas, el triángulo isósceles posee el máximo ángulo vertical.

***5.** De todos los triángulos con área dada, el triángulo equilátero posee el mínimo perímetro.

***6.** De todos los triángulos con perímetro dado, el triángulo equilátero posee el área máxima.

***7.** De todos los triángulos inscritos en una circunferencia, el triángulo equilátero posee el área máxima.

8. Probar que si $p > 1$ y $x > 0$, es $x^p - 1 \geq p(x - 1)$.

9. Probar la desigualdad $1 > (\operatorname{sen} x)/x \geq 2/\pi$, $0 \leq x \leq \pi/2$.

10. Probar que: (*a*) $\tan x \geq x$, $0 \leq \pi/2$;
(*b*) $\cos x \geq 1 - x^2/2$.

***11.** Dados $a_1 > 0$, $a_2 > 0, \ldots, a_n > 0$, determínese el mínimo de

$$\frac{a_1 + \cdots + a_{n-1} + x}{\sqrt[n]{a_1 a_2 \cdots a_{n-1} x}}$$

para $x > 0$. Utilizar el resultado para probar por inducción matemática que (cfr. problema 13, p. 131):

$$\sqrt[n]{a_1 a_2 \cdots a_n} \leq \frac{a_1 + \cdots + a_n}{n}.$$

12. (*a*) Dados n números fijos $a_1, \ldots, a_n$, determinar x de modo que $\sum_{i=1}^{n} (a_i - x)^2$ sea un mínimo.

***(*b*)** Minimizar $\sum_{i=1}^{n} |a_i - x|$.

***(*c*)** Minimizar $\sum_{i=1}^{n} \lambda_i |a_i - x|$. donde $\lambda_i > 0$.

13. Bosquejar la gráfica de la función

$$y = (x^2)^x, \; y(0) = 1.$$

Demuéstrese que la función es continua en $x = 0$. ¿Posee la función máximos, mínimos o puntos de inflexión?

***14.** Encontrar el mínimo valor α tal que

$$\left(1 + \frac{1}{x}\right)^{x+\alpha} > e$$

para todo x positivo. (*Sugerencia:* Es sabido que $[1 + (1/x)]^{x+1}$ decrece y $[1 + (1/x)]^x$ crece, ambas monótonamente, al límite e en el infinito.)

***15.** (*a*) Encontrar el punto para el cual la suma de sus distancias a los tres lados de un triángulo es un mínimo.

(*b*) Encontrar el punto para el cual la suma de sus distancias a los vértices es un mínimo.

16. Probar las siguientes desigualdades:

(*a*) $e^x > 1/(1 + x), x > 0.$

(*b*) $e^x > 1 + \log(1 + x), x > 0.$

(*c*) $e^x > 1 + (1 + x)\log(1 + x), x > 0.$

17. Supóngase que $f''(x) < 0$ en (a, b). Probar que:

(*a*) Todo arco de la gráfica dentro del intervalo está situado por encima de la cuerda que une sus puntos extremos.

(*b*) La gráfica está situada por debajo de la tangente, para cualquier punto dentro de (a, b).

***18.** Sea f una función que tiene segunda derivada en (a, b).

(*a*) Demostrar que cualquiera de las condiciones *a* o *b* del problema 22 es suficiente para garantizar que $f''(x) \leq 0$.

(*b*) Demostrar que la condición

$$f\left(\frac{x + y}{2}\right) \geq \frac{f(x) + f(y)}{2}$$

para todo x e y en (a, b) es suficiente para garantizar que $f''(x) \leq 0$.

***19.** Sean a, b, dos números positivos, y p y q números cualesquiera diferentes de cero y tales que $p < q$. Probar que

$$\frac{[\theta a^p + (1 - \theta) b^p]^{1/p}}{[\theta a^q + (1 - \theta) b^q]^{1/q}} \leq 1$$

para todos los valores de θ en el intervalo $0 < \theta < 1$.

(Esta es la desigualdad de Jensen, que establece que la potencia media p-ésima $[\theta a^p + (1 - \theta) b^p]^{1/p}$ de dos cantidades a, b es una función creciente de p.)

20. Mostrar que el signo de igualdad en la desigualdad anterior se cumple si y sólo si $a = b$.

21. Probar que $\lim_{p \to 0} [\theta a^p + (1 - \theta) b^p]^{1/p} = a^\theta b^{1-\theta}$.

22. Definiendo la potencia media 0-ésima de a, b, como $a^\theta b^{1-\theta}$, demostrar que la desigualdad de Jensen se aplica a este caso y se convierte en (con $a \neq b$):

$$a^\theta b^{1-\theta} \gtrless [\theta a^q + (1 - \theta) b^q]^{1/q} \text{ de acuerdo con que } q \gtrless 0.$$

Para $q = 1$, es

$$a^\theta b^{1-\theta} \leq \theta a + (1 - \theta) b.$$

23. Probar la desigualdad

$$a^\theta b^{1-\theta} \leq \theta a + (1 - \theta) b,$$

$a, b > 0$, $0 < \theta < 1$, sin referencia a la desigualdad de Jensen, y mostrar que la igualdad se cumple sólo si $a = b$. (Esta igualdad establece que la media geométrica de θ, $1 - \theta$, es menor que la correspondiente media aritmética.)

***24.** Sea f continua y positiva en $[a, b]$ y denótese por M su valor máximo. Probar que

$$M = \lim_{n \to \infty} \sqrt[n]{\int_a^b [f(x)]^n \, dx}.$$

SECCION 3.7, página 269

1. Sea $f(x)$ una función continua que se anula, junto con su primera derivada, para $x = 0$. Mostrar que $f(x)$ se anula con un orden superior al de x cuando $x \to 0$.

2. Mostrar que

$$f(x) = \frac{a_0 x^n + a_1 x^{n-1} + \cdots + a_n}{b_0 x^m + b_1 x^{m-1} + \cdots + b_m},$$

donde a_0, $b_0 \neq 0$, es del mismo orden de magnitud que x^{n-m} cuando $x \to \infty$.

***3.** Probar que e^x no es una función racional.

***4.** Probar que e^x no puede satisfacer una ecuación algebraica con polinomios en x como coeficientes.

5. Si el orden de magnitud de la función positiva $f(x)$ cuando $x \to \infty$ es mayor, el mismo o bien menor que el de x^m, probar que $\int_a^x f(\xi)\, d\xi$ posee el correspondiente orden de magnitud relativo a x^{m+1}.

6. Comparar el orden de magnitud, para $x \to \infty$, de $\int_a^x f(\xi)\, d\xi$ relativo a $f(x)$ para las siguientes funciones $f(x)$:

(*a*) $\dfrac{e\sqrt{x}}{\sqrt{x}}$.

(*b*) e^x.

(*c*) xe^{x^2}.

(*d*) $\log x$.

SECCION 3.8, página 284

1. Encontrar el límite, para $n \to \infty$, de $a_n = \dfrac{1}{n+1} + \dfrac{1}{n+2} + \cdots + \dfrac{1}{2n}$.

*2. Encontrar el límite de

$$b_n = \frac{1}{\sqrt{n^2 - 0}} + \frac{1}{\sqrt{n^2 - 1}} + \frac{1}{\sqrt{n^2 - 4}} + \cdots + \frac{1}{\sqrt{n^2 - (n-1)^2}}.$$

*3. Si α es cualquier número real mayor que -1, evaluar

$$\lim_{n\to\infty} \frac{1^\alpha + 2^\alpha + 3^\alpha + \cdots + n^2}{n^{\alpha+1}}.$$

SECCION 3.11, página 295

1. Demostrar que para todos los valores positivos impares de n, la integral $\int e^{-x^2} x^n \, dx$ puede ser evaluada en términos de funciones elementales.

2. Demostrar que si n es par, la integral $\int e^{-x^2} x^n \, dx$ puede ser evaluada en términos de funciones elementales y de la integral $\int e^{-x^2} \, dx$ (para la cual se han construido tablas).

3. Probar que

$$\int_0^x \left[\int_0^u f(t) \, dt \right] du = \int_0^x f(u)(x - u) \, du.$$

*4. El problema 3 da una fórmula para la segunda integral iterada. Probar que la n-ésima integral iterada de $f(x)$ está dada por

$$\frac{1}{(n-1)!} \int_0^x f(u)(x - u)^{n-1} \, du.$$

5. Probar para el coeficiente binomial $\binom{\gamma}{u}$ que

$$\binom{n}{k} = \left[(n+1) \int_0^1 x^k (1 - x)^{n-k} \, dx \right]^{-1}$$

6. Obtener una fórmula recursiva para

$$\int x^p (ax^n + b)^q \, dx$$

y utilizar esta relación para integrar

$$\int x^3 (x^7 + 1)^4 \, dx.$$

*7. (*a*) Sea $P_n(x) = \dfrac{1}{2^n n!} \dfrac{d^n}{dx^n} (x^2 - 1)^n$: Demostrar que

$$\int_{-1}^1 P_n(x) P_m(x) \, dx = 0, \quad \text{si } m \neq \nu.$$

(*b*) Probar que $\displaystyle\int_{-1}^1 P_n^2(x) \, dx = \frac{2}{2n+1}$.

(*c*) Probar que $\displaystyle\int_1^{-1} x^m P_n(x) \, dx = 0$, si $m < n$.

(*d*) Evaluar $\displaystyle\int_{-1}^1 x^n P_n(x) \, dx$.

SECCION 3.12, página 302

*1. Integrar

$$\int \frac{dx}{x^6 + 1}.$$

2. Utilizar el desarrollo en fracciones parciales para probar las fórmulas de Newton:

$$\frac{\alpha_1{}^k}{g'(\alpha_1)} + \frac{\alpha_2{}^k}{g'(\alpha_2)} + \cdots + \frac{\alpha_n{}^k}{g'(\alpha_n)} = \begin{cases} 0 & \text{para } k = 0, 1, 2, \ldots, n-2 \\ 1 & \text{para } k = n - 1, \end{cases}$$

donde $g(x)$ es un polinomio de la forma $x^n + \alpha_1 x^{n-1} + \cdots$ con raíces distintas $\alpha_1, \ldots, \alpha_n$.

SECCION 3.14, página 317

*1. Probar que la substitución $x = (\alpha t + \beta)/(\gamma t + \delta)$, con $\alpha\delta - \gamma\beta \neq 0$, transforma la integral

$$\int \frac{dx}{\sqrt{ax^4 + bx^3 + cx^2 + dx + e}}$$

en una integral de un tipo similar, y que si la expresión bicuadrática

$$ax^4 + bx^3 + cx^2 + dx + e$$

no posee factores repetidos, tampoco los posee la nueva expresión bicuadrática en t que toma su lugar. Probar que lo mismo se cumple para

$$\int R(x, \sqrt{ax^4 + bx^3 + cx^2 + dx + e}) \, dx,$$

donde R es una función racional.

2. La función

$$\phi(x) = \int_0^x \frac{du}{\sqrt{1 - k^2 \operatorname{sen}^2 u}}$$

es conocida como la *integral elíptica de primer orden.*

(*a*) Mostrar que ϕ es continua y creciente, y que, por lo tanto, posee una inversa continua.

(*b*) Denótese por $am(x)$ la inversa de $\phi(x)$. Probar que $sn(x) = \operatorname{sen}[am(x)]$, donde $sn(x)$ está definida en la p. 319, nota al pie número 3.

SECION 3.15, página 320

*1. Probar que $\displaystyle\int_0^\infty \operatorname{sen}^2\left[\pi \left(x + \frac{1}{x} \right) \right] dx$ *no existe* (diverge).

*2. Probar que $\displaystyle\lim_{k\to\infty} \int_0^\infty \frac{dx}{1 + kx^{10}} = 0$.

3. ¿Para qué valores de s es $(a)\int_0^\infty \frac{x^{s-1}}{1+x}\,dx$, $(b)\int_0^\infty \frac{\operatorname{sen} x}{x^s}\,dx$ convergente?

***4.** ¿Converge la integral $\int_0^\infty \frac{\operatorname{sen} t}{1+t}\,dt$?

***5.** (a) Si a es un número positivo fijo, probar que

$$\lim_{h\to 0+}\int_{-a}^{a} \frac{h}{h^2+x^2}\,dx = \pi.$$

(b) Si $f(x)$ es continua en el intervalo $-1 \leq x \leq 1$, probar que

$$\lim_{h\to 0}\int_{-1}^{1} \frac{h}{h^2+x^2} f(x)\,dx = \pi f(0).$$

***6.** Probar que $\lim_{x\to\infty} e^{-x^2}\int_0^x e^{t^2}\,dt = 0$.

7. Suponiendo que $|\alpha| \neq |\beta|$, probar que

$$\lim_{T\to\infty} \frac{1}{T}\int_0^T \operatorname{sen}\alpha x \operatorname{sen}\beta x\,dx = 0.$$

***8.** Si la integral $\int_a^\infty \frac{f(x)}{x}\,dx$ converge para cualquier valor positivo a, y si $f(x)$ tiende a un límite L cuando $x \to 0$, mostrar que $\int_0^\infty \frac{f(\alpha x) - f(\beta x)}{x}\,dx$ converge para α y β positivos y posee el valor $L \log\frac{\beta}{\alpha}$.

9. Refiriéndose al problema 8, probar que

(b) $\int_0^\infty \frac{e^{-\alpha x} - e^{-\beta x}}{x}\,dx = \log\frac{\beta}{\alpha}$.

(a) $\int_0^\infty \frac{\cos\alpha x - \cos\beta x}{x}\,dx = \log\frac{\beta}{\alpha}$.

***10.** Si la integral $\int_a^b \frac{f(x)}{x}\,dx$ converge para cualesquiera valores positivos de a y b, y si $f(x)$ tiende a un límite M cuando $x \to \infty$ y a un límite L cuando $x \to 0$, mostrar que

$$\int_0^\infty \frac{f(\alpha x) - f(\beta x)}{x}\,dx = (L - M)\log\frac{\beta}{\alpha}.$$

11. Obténganse las expresiones siguientes para la función gamma:

$$\Gamma(n) = 2\int_0^\infty x^{2n-1} e^{-x^2}\,dx,$$

$$\Gamma(n) = \int_0^1 \left(\log\frac{1}{x}\right)^{n-1} dx.$$

SECCION 3.16, página 331

1. Obtener la fórmula de adición para sen $(x + y)$.

2. Sin utilizar las fórmulas de adición, probar que $\cos x$ es una función par y en x una función impar.

3. (a)* Probar que para algún h positivo es $\cos x < 1$ para $0 < x < h$.
(b) Si $\cos z > 0$ para $0 \leq z \leq 2^n x$, probar que

$$\cos(2^{n+1}x) < 2^n(\cos x - 1) + 1.$$

(c) Combinando los resultados (a) y (b), probar que $\cos x$ posee un cero.

4. Sea a el menor cero positivo de $\cos x$. Probar que

$$\operatorname{sen}(x + 4a) = \operatorname{sen} x,$$

$$\cos(x + 4a) = \cos x.$$

5. Completar en los pasos siguientes la demostración indirecta de que $\cos x$ posee un cero:
(a) Si $\cos x$ no posee ceros, entonces sen x es monótona creciente para $x \geq 0$.
(b) Las funciones sen x y $\cos x$ están acotadas por arriba y por abajo.
(c) El límite de sen x cuando x tiende a infinito existe y es positivo.
(d) La ecuación

$$\cos x = 1 - \int_0^x \operatorname{sen} t\,dt$$

está en contradicción con el inciso (b).

PROBLEMAS DIVERSOS

1. Probar que

$$\frac{d^n}{dx^n} f(\log x) = x^{-n}\frac{d}{dt}\left(\frac{d}{dt} - 1\right)\left(\frac{d}{dt} - 2\right)\cdots\left(\frac{d}{dt} - n + 1\right) f(t)$$

donde $t = \log x$. Aquí se emplea

$$\left(\frac{d}{dt} - k\right)\phi = \frac{d\phi}{dt} - k\phi,$$

donde ϕ es cualquier función de t y k es una constante.

2. Una curva suave cerrada C se dice que es *convexa* si está situada por completo de un lado de cada tangente. Demostrar que en el triángulo de área mínima circunscrito a C cada lado es tangente a C en el punto medio del lado.

4

Aplicaciones en física y geometría

4.1 Teoría de curvas planas

a. Representación paramétrica

Definición

La representación de una curva por medio de una ecuación $y = f(x)$ impone una seria restricción geométrica: Una curva representada de esa manera no debe ser intersectada en más de un punto por cualquier paralela al eje y. Usualmente esta restricción puede superarse descomponiendo la curva en porciones, cada una representable en la forma $y = f(x)$. Así, una circunferencia de radio a alrededor del origen está dada por las dos funciones $y = \sqrt{a^2 - x^2}$ e $y = -\sqrt{a^2 - x^2}$ definidas para $-a \leq x \leq a$. Sin embargo, para una curva tan simple como una recta paralela al eje y este artificio no sirve.

Mayor flexibilidad se obtiene usando una representación *implícita* mediante una ecuación $\phi(x, y) = 0$ que involucra una función ϕ de dos variables independientes. Por ejemplo, la circunferencia de radio a alrededor del origen está completamente descripta por $\phi(x, y) = x^2 + y^2 - a^2 = 0$. Cualquier línea recta en el plano tiene una ecuación implícita de la forma $ax + by + c = 0$, donde a, b y c son constantes y a y b no se anulan simultáneamente. Para $b = 0$ se obtiene una paralela al eje y.

La descripción implícita de una curva tiene la desventaja de que para encontrar puntos (x, y) de la curva explícitamente, es decir, el valor de y para un valor de x dado, se debe *resolver* la ecuación $\phi(x, y) = 0$. Este problema se discutirá en detalle en el Volumen II.

La descripción más directa y más flexible de una curva es una *representación paramétrica*. En vez de considerar una de las coordenadas rectangulares en particular, y o x, como una función de la otra, se consideran ambas coordenadas x e y como funciones de una *tercera* variable independiente t, el llamado *parámetro* [1]; y el punto con coordenadas x e y describe entonces la curva conforme t recorre un cierto intervalo. Tal representación paramétrica ya ha sido encontrada antes; por ejemplo, la circunferencia $x^2 + y^2 = a^2$ tiene la representación paramétrica $x = a \cos t$, $y = a \operatorname{sen} t$. Aquí t denota el ángulo en el centro de la circunferencia.

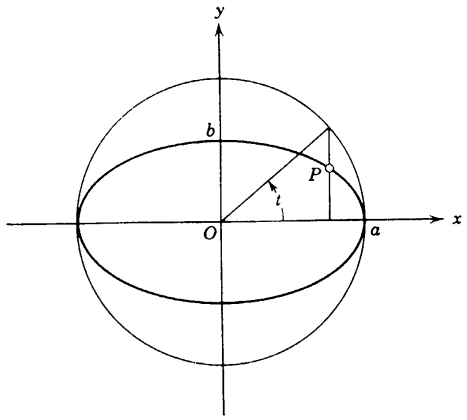


Figura 4.1

Para el elipse $x^2/a^2 + y^2/b^2 = 1$ se tiene una representación paramétrica similar: $x = a \cos t$, $y = b \operatorname{sen} t$, donde t es el llamado ángulo excéntrico, esto es, el ángulo en el centro de la elipse asociado al punto de la circunferencia circunscrita que está verticalmente por encima o por debajo del punto $P = (a \cos t, b \operatorname{sen} t)$ de la elipse. Se supone aquí que $b < a$ (véase Fig. 4.1). En ambos casos el punto con coordenadas x, y describe por completo la circunferencia o la elipse conforme el parámetro t recorre el intervalo $0 \leq t < 2\pi$.

En general, una curva C es representada paramétricamente mediante dos funciones de un parámetro t,

$$x = \phi(t) = x(t), \qquad y = \psi(t) = y(t);$$

la notación abreviada $x(t)$ e $y(t)$ se usará siempre que no dé lugar a confusión.*

En lo que sigue se supondrá que ϕ y ψ tienen derivadas continuas, a menos que se diga lo contrario

[1] Esta palabra denota una variable auxiliar que no se desea señalar en especial.

* La notación $x = \phi(t)$, etc., pone de relieve la conexión funcional entre la variable dependiente y la independiente; la notación $x(t)$, etc., significa justamente que t va a ser considerada como la variable independiente que determina el valor funcional de x según una forma prescrita.

Transformación del intervalo paramétrico en la curva. Sentido de dirección

Dada una curva, las dos funciones $\phi(t)$ y $\psi(t)$ deben determinarse de manera que el conjunto de pares de valores funcionales, $x(t)$ e $y(t)$, correspondientes a un cierto intervalo de valores de t definan todos los puntos de la curva y ningún otro punto más. Se tiene entonces una correspondencia entre los puntos de la curva y los valores de t en un intervalo del eje t. La representación paramétrica define una *transformación de un intervalo del eje t en la curva,*[2] siendo el punto t original del eje t transformado en el punto $x = \phi(t)$, $y = \psi(t)$ de C.

Puesto que $x(t)$ e $y(t)$ se han considerado continuas, puntos vecinos sobre el eje t corresponden a puntos vecinos sobre la curva. Además, como los puntos del eje t están ordenados, de manera obvia se puede asignar un orden o "sentido" a los puntos de C diciendo simplemente que el punto en el cual se transforma el número t_1 *precede* al punto en el cual se transforma t_2 si $t_1 < t_2$ (ver p. 352). La representación paramétrica da así un significado preciso a la vaga noción intuitiva de curva como un conjunto de puntos dispuestos en el mismo orden que sobre una línea recta.

b. Cambios de parámetros

Los valores del parámetro t sirven para distinguir los diferentes puntos de la curva C. Dichos valores desempeñan el papel de "nombres" para los puntos individuales de la curva.

La misma curva C admite muchas representaciones paramétricas diferentes. Cualquier cantidad que varíe continuamente a lo largo de la curva y que tenga diferentes valores en puntos diferentes de ella pueed servir como parámetro.

Si, por ejemplo, la curva está dada originalmente por una ecuación $y = f(x)$, se puede escoger como parámetro t la variable x y describir la curva mediante las funciones $x = t$, $y = f(t)$. Análogamente, para una curva descripta dando x como una función de y, por ejemplo $x = g(y)$, se puede usar y como parámetro t y escribir $x = g(t)$, $y = t$.

Para una curva dada por una ecuación $r = h(\theta)$ en *coordenadas polares* τ, θ (véase capítulo 1, p. 124), se puede escoger θ como parámetro t y obtener la representación paramétrica

$$x = r\cos\theta = h(t)\cos t = \phi(t),$$

$$y = r\,\mathrm{sen}\,\theta = h(t)\,\mathrm{sen}\,t = \psi(t).$$

[2] Más exactamente, se dice que la transformación es *sobre* la curva C (en inglés, *"onto"*) porque los puntos imagen son *todos* los puntos de la curva *(N. del R.)*.

Partiendo de una representación paramétrica dada, $x = \phi(t)$, $y = \psi(t)$, para una curva C, siempre se pueden obtener muchas otras representaciones paramétricas. Para ello se toma una función arbitraria $\tau = \chi(t)$ que sea monótona y continua en el intervalo t correspondiente a los puntos de C. La función χ tiene entonces una inversa monótona y continua $t = \sigma(\tau)$ en un correspondiente intervalo de τ. Las coordenadas de los puntos (x, y) de C pueden entonces representarse en la forma

$$x = \phi[\sigma(\tau)] = \alpha(\tau), \quad y = \psi[\sigma(\tau)] = \beta(\tau).$$

Las funciones $\alpha(\tau)$ y $\beta(\tau)$ son también continuas; además, puntos diferentes de C corresponden a valores diferentes de t y, por lo tanto, debido al carácter monótono de la función σ, a diferentes valores de τ. El efecto total del cambio de parámetro de t a τ es el de "rebautizar" los puntos de C.

Así, la recta $y = x$ tiene la representación paramétrica $x = t$, $y = t$, donde $-\infty < t < \infty$. La substitución $\tau = t^3$ da lugar a la representación paramétrica $x = \tau^{1/3}$, $y = \tau^{1/3}$ para la misma recta.

Análogamente, la elipse $x^2/a^2 + y^2/b^2 = 1$ admite la representación paramétrica $x = a\cos t$, $y = b\,\mathrm{sen}\,t$, donde $0 \le t < 2\pi$. Definiendo $t = c\zeta + d$, para números reales c y d $(c \neq 0)$, se obtiene otra representación, $x(\zeta) = a\cos(c\zeta + d)$, $y(\zeta) = b\,\mathrm{sen}\,(c\zeta + d)$, para la misma elipse, con ζ variando en los intervalos $-d/c \le \zeta < (2\pi - d)/c$ para $c > 0$ y $(2\pi - d)/c < \zeta \le -d/c$ para $c < 0$. La substitución $\tau = \tan(t/2)$ conduce a la *representación paramétrica "racional"* (ver pp. 311-312),

$$x = \frac{a(1-\tau^2)}{1+\tau^2}, \quad y = \frac{2b\tau}{1+\tau^2},$$

para la elipse. Conforme τ recorre todos los valores reales, se obtienen todos los puntos de la elipse con la excepción del punto $S = (-a, 0)$.

Las singularidades en una representación ordinaria pueden desaparecer si se usa un parámetro adecuado. Por ejemplo, la curva $y = \sqrt[3]{x^2}$ puede ser representada por las funciones suaves $x = t^3$, $y = t^2$. El punto con coordenadas x, y describe entonces la curva completa (parábola semicúbica) conforme t varía de $-\infty$ a $+\infty$.

Esta flexibilidad en la elección del parámetro permite a menudo simplificar el estudio de las propiedades geométricas que, por supuesto, no dependen de representaciones específicas.

En particular, puede resultar conveniente algunas veces el uso de una representación $y = f(x)$ para C o parte de C. Tal representación es siempre posible para un pedazo de la curva $(t_0 \le t \le t_1)$ en el cual una de las funciones ϕ, ψ (por ejemplo, $x = \phi(t)$), es monótona. En efecto, para este pedazo se tiene una función inversa única $t = \gamma(x)$, y así es $y = \psi[\gamma(x)]$.[1]

[1] Esto es, por supuesto, sólo una afirmación relativa a una propiedad "local" de una curva, o sea, una afirmación que se hace solamente para una parte pequeña

c. Movimiento a lo largo de una curva. El tiempo como un parámetro. El ejemplo de la cicloide

Movimiento a lo largo de una curva

Muy a menudo el parámetro t tiene el natural significado físico de tiempo. Cualquier movimiento de un punto en el plano puede describirse representando sus coordenadas x e y como funciones del tiempo de manera que al tiempo t el punto (x, y) esté en $(x(t), y(t))$. Por consiguiente, estas dos funciones determinan en forma paramétrica el movimiento a lo largo de un camino o trayectoria C; y ellas constituyen una *transformación* de la escala del tiempo sobre la trayectoria.[2]

Las cicloides y las trocoides

Un ejemplo lo proporcionan las *cicloides,* o sea, las trayectorias de puntos sobre un círculo que rueda uniformemente sin deslizar a lo largo de una línea recta u otro círculo. En el caso más simple un círculo de radio a rueda a lo largo del eje x; y la trayectoria de un punto P sobre su circunferencia es una cicloide "común". Escójanse el origen de coordenadas del sistema y el tiempo inicial de manera que para el tiempo $t = 0$ el punto P esté en el origen y que en el tiempo t el círculo haya girado un cierto ángulo t a partir de su posición original. Esto significa que el círculo gira en el sentido de las manecillas de un reloj "con velocidad angular" igual a uno. Se supone que el círculo rueda uniformemente sin deslizar a lo largo del eje x, de manera que para el tiempo t la distancia del punto de contacto al origen es exactamente igual a la longitud de arco medida desde el punto de contacto hasta P. Así, en el tiempo t el centro M del círculo giratorio debe estar en el punto (at, a); y el centro se mueve con velocidad constante a hacia la derecha. Para las coordenadas de P en el tiempo t se encuentra entonces (ver Fig. 4.2) la representación paramétrica

$$x = a(t - \operatorname{sen} t), \quad y = a(1 - \cos t). \tag{1}$$

Eliminando el parámetro t puede obtenerse la ecuación de la curva en forma no paramétrica a costa, sin embargo, de la claridad de la expresión. Se tiene

$$\cos t = \frac{a - y}{a}, \quad t = \arccos\frac{a - y}{a}, \quad \operatorname{sen} t = \pm\sqrt{1 - \frac{(a - y)^2}{a^2}},$$

de la curva, escogida adecuadamente. Generalmente (por ejemplo, en el caso de la circunferencia), la variable x no puede usarse como parámetro para la curva completa sino sólo para una porción de ella.

[2] A un cambio del parámetro t correspondería entonces un cambio en la escala del tiempo según la cual la curva queda descripta por el punto móvil.

y, por lo tanto,

$$x = a\arccos\frac{a - y}{a} \mp \sqrt{y(2a - y)}, \tag{1a}$$

obteniéndose así x como una función de y.

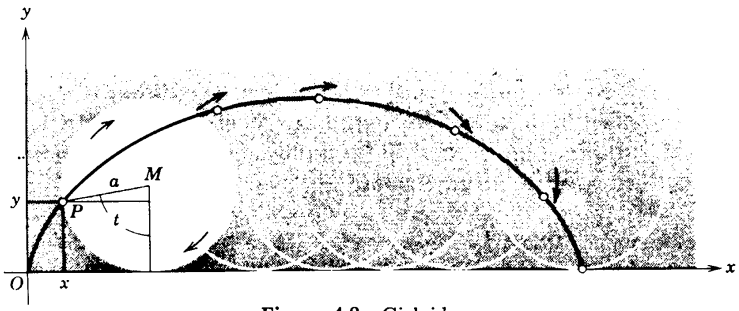


Figura 4.2 Cicloide.

Epicicloide

Nuestro siguiente ejemplo es el de una *epicicloide,* definida como la trayectoria de un punto P fijo sobre la circunferencia de un círculo de radio c, conforme éste rueda a una velocidad uniforme a lo largo de la circunferencia y fuera de un segundo círculo de radio a. Supóngase el círculo fijo centrado en el origen del plano x, y. Supóngase además que el círculo que se mueve rueda a lo largo del círculo fijo de manera que su centro ha girado alrededor del origen a un ángulo t en el tiempo t (Fig. 4.3). Entonces se encuentra que en el tiempo t la posición del punto $P = (x(t), y(t))$, el cual en el tiempo $t = 0$ es el punto de contacto $(a, 0)$, está dada por las ecuaciones paramétricas

$$\begin{aligned} x(t) &= (a + c)\cos t - c\cos\left(\frac{a + c}{c}t\right), \\ y(t) &= (a + c)\operatorname{sen} t - c\operatorname{sen}\left(\frac{a + c}{c}t\right). \end{aligned} \tag{2}$$

Cuando $a = c$ la curva formada se llama *cardioide* (Fig. 4.4) y está dada por las ecuaciones paramétricas

$$\begin{aligned} x(t) &= 2a\cos t - a\cos(2t), \\ y(t) &= 2a\operatorname{sen} t - a\operatorname{sen}(2t). \end{aligned} \tag{3}$$

Una tercera variedad de cicloides se obtiene como el lugar geométrico de un punto sujeto a la circunferencia de un círculo que rueda sobre la circunferencia de otro círculo fijo, pero por su interior. Para encontrar las ecuaciones paramétricas de esta *"hipocicloide"*, sea a el radio del círculo fijo y c el del círculo que gira. Supóngase que el punto P de la circunferencia del círculo que se mueve está localizado en $(a, 0)$ en el tiempo $t = 0$. Supóngase, además, que el círculo que gira se mueve a lo largo del círculo fijo de manera que en el tiempo t su centro ha girado alrededor del origen en un ángulo t (Fig. 4.5). Se encuentra entonces que las ecuaciones paramétricas para la hipocicloide son

$$(4)\qquad \begin{aligned} x(t) &= (a-c)\cos t + c\cos\left(\frac{a-c}{c}t\right), \\ y(t) &= (a-c)\operatorname{sen} t - c\operatorname{sen}\left(\frac{a-c}{c}t\right). \end{aligned}$$

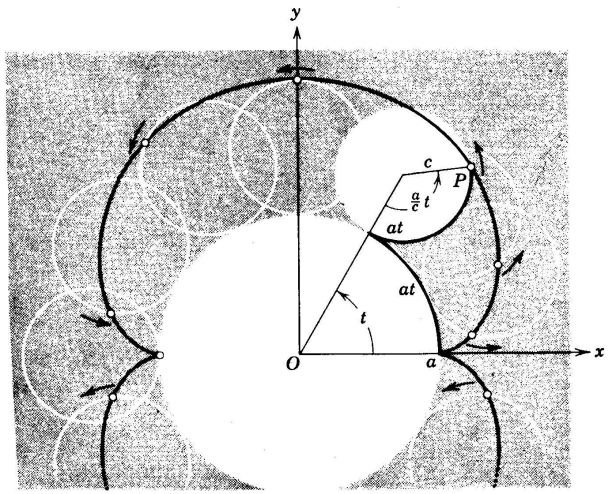


Figura 4.3 Epicicloide.

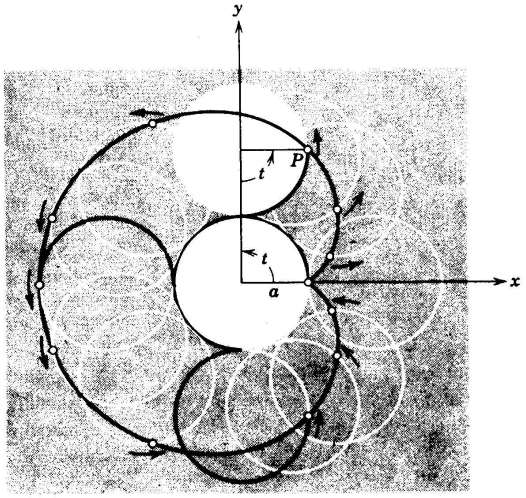


Figura 4.4 Cardioide.

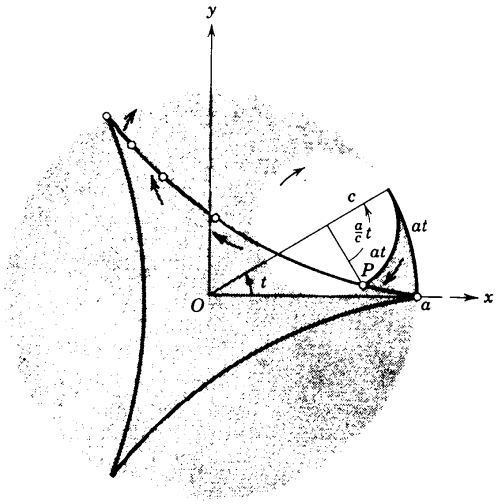


Figura 4.5 Hipocicloide.

En el caso especial en el que el círculo fijo tiene un radio igual a dos veces el radio del círculo que se mueve, $c = \frac{1}{2}a$, se obtiene

$$x(t) = a\cos t,$$
$$y(t) = 0,$$

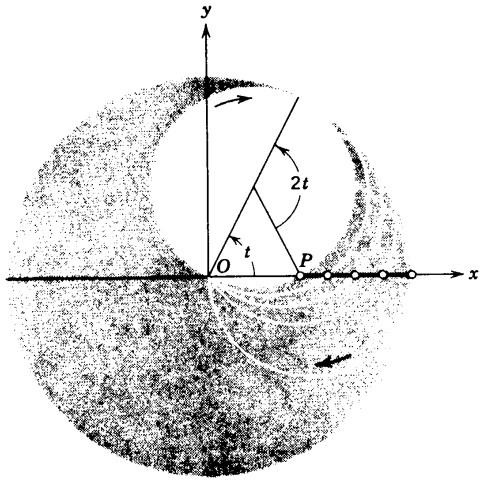


Figura 4.6 Un punto P sobre la periferia de un círculo que rueda interiormente sobre una circunferencia de radio doble describe un segmento de recta.

y la hipocicloide degenera en el diámetro del círculo fijo, recorrido hacia atrás y hacia adelante. La característica interesante de este ejemplo es que proporciona una solución mecánica al problema de dibujar una línea recta usando únicamente movimientos circulares (Fig. 4.6).

Si el radio del círculo fijo es tres veces el radio del círculo móvil, entonces $c = a/3$ y

$$x(t) = \tfrac{3}{4}a \cos t + \tfrac{1}{4}a \cos (2t),$$
$$y(t) = \tfrac{3}{4}a \operatorname{sen} t - \tfrac{1}{4}a \operatorname{sen} (2t).$$

Mediante un cálculo elemental se encuentra

$$x^2 + y^2 = \tfrac{5}{16}a^2 + \tfrac{1}{16}a^2 \cos (3t),$$

de manera que la hipocicloide encuentra al círculo fijo en exactamente tres puntos y la curva aparece como se muestra en la Fig. 4.5.

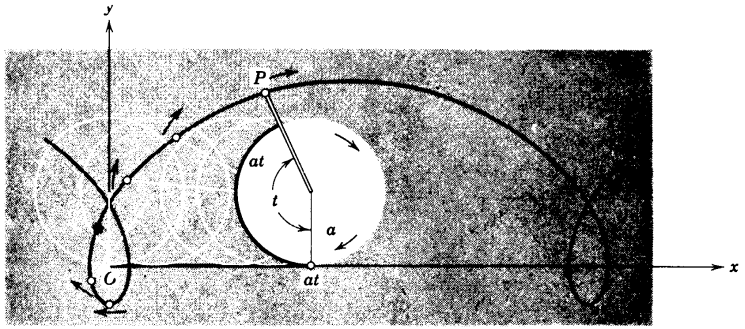


Figura 4.7 Trocoide.

Trocoides

Curvas más generales llamadas *trocoides* (epitrocoides, hipotrocoides) se obtienen al considerar el movimiento de un punto P sujeto a un círculo (pero no necesariamente sobre su periferia) cuando dicho círculo rueda a lo largo de una línea recta o a lo largo de otro círculo por su parte exterior o interior (ver Fig. 4.7). El mismo tipo de curva aparece como la trayectoria de un punto que se mueve uniformemente sobre un círculo en tanto que el centro del círculo mismo se mueve uniformemente a lo largo de una línea recta o de un círculo. Estas curvas juegan un papel muy importante en la descripción ptolomeica del movimiento aparente de los planetas.

Algunas de las propiedades notables de las cicloides serán discutidas más adelante en este capítulo (p. 446).

d. Clasificación de curvas. Orientación

Definiciones

Entre las características más obvias de una curva están el número de partes separadas o *ramas* y el número de espiras que ellas tengan. Una hipérbola es un ejemplo de una curva que consiste de dos ramas ajenas;

otro de tales ejemplos es la curva $y^2 = (4 - x^2)(x^2 - 1)$ que consiste de dos óvalos separados. Nos interesarán particularmente las curvas consistentes de una sola parte, las curvas *conexas*. Una curva conexa puede *intersectarse* a sí misma, como la trocoide (Fig. 4.7), o la "lemniscata" de la Fig. 1.S.3, p. 126.

Una curva conexa sin autointersecciones se denomina *simple*. Entre las curvas simples todavía se distinguirán las curvas *cerradas*, como los círculos o las elipses, de las curvas que no son cerradas, como parábolas o segmentos de recta. No se intentará aquí dar una clasificación ni rigurosa ni completa de las curvas, sino únicamente señalar ciertas características *"topológicas"* de una curva, importantes para la representación paramétrica.

Arcos simples

Una representación paramétrica de una curva C por medio de dos funciones continuas, $x = \phi(t)$, $y = \psi(t)$, define una transformación del eje t, o de una porción de él, en C. C se denominará un *arco simple* si puede representarse de tal manera que el parámetro t recorra un intervalo cerrado $[a, b]$ sobre el eje t, formando así el dominio de las funciones $\phi(t)$, $\psi(t)$, y si, además, diferentes t en el intervalo corresponden a diferentes puntos P sobre C. Un ejemplo es el arco parabólico $x = t$, $y = t^2$ para $0 \leq t \leq 1$.

El mismo arco C (esto es, los mismos puntos del plano) puede ser representado paramétricamente de muchas maneras. Cualquier función monótona continua $\tau = \chi(t)$, para $a \leq t \leq b$, define un parámetro τ de manera que x e y sean funciones continuas de τ en un intervalo cerrado $[\alpha, \beta]$ conveniente, correspondiendo valores diferentes de τ a diferentes P. De hecho, como es fácil concluir, las substituciones monótonas continuas $\tau = \chi(t)$ proporcionan las representaciones paramétricas continuas *más generales* de un arco simple que asignan diferentes puntos del arco a diferentes valores del parámetro. (Ver notas en la p. 78 relativas a transformaciones continuas biunívocas.)

Una representación paramétrica especial, $x = x(t)$, $y = y(t)$, de un arco simple C tiene un *sentido* definido sobre C correspondiente al sentido de las t *crecientes*. Dados dos puntos distintos cualesquiera, P_0, P_1, se dice que P_1 sigue a P_0 si P_1 pertenece al mayor valor del parámetro t. Si se introduce un nuevo parámetro τ mediante una función continua *creciente* $\tau = \chi(t)$, el orden de los pares de puntos con respecto a τ es el mismo; y el parámetro τ define el *mismo sentido* sobre C. Si $\chi(t)$ es decreciente, el sentido se invierte.

Dirección u orientación de arcos

Un arco simple *dirigido* u *orientado* es aquél en el cual se ha escogido un sentido definido (por ejemplo, el sentido que corresponde a un incre-

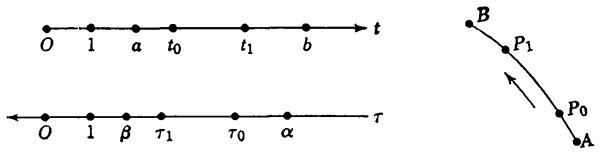


Figura 4.8 Sentido y representación paramétrica.

mento para una elección particular del parámetro t); y ese sentido se llama entonces el *sentido positivo* del arco. El sentido positivo queda completamente especificado si se sabe cuál de los dos puntos extremos del arco *sigue* al otro. Se denominará punto *final* del arco al punto extremo que sigue al otro, y punto *inicial* a éste. Dada cualquier representación paramétrica $x = x(\tau)$, $y = y(\tau)$ del arco orientado, donde $a \leq \tau \leq b$, el sentido positivo será aquél que corresponde al sentido de las τ crecientes si el valor paramétrico $\tau = a$ corresponde al punto inicial y el $\tau = b$, al punto final; en caso contrario, el sentido de las τ crecientes corresponderá al sentido negativo del arco (Fig. 4.8).

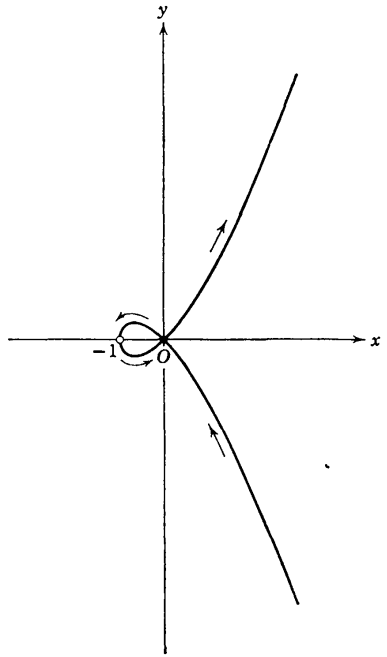


Figura 4.9 Una curva con una espira: $x = t^2 - 1$, $y = t^3 - t$, cuyo sentido es el de las t crecientes.

Dos puntos distintos cualesquiera, P_0, P_1, sobre un arco simple C definen un subarco con puntos extremos P_0, P_1, el cual consiste de los puntos con valores paramétricos comprendidos entre los correspondientes para P_0 y P_1. Si C es un arco dirigido y P_1 sigue a P_0 en el sentido positivo de C, se obtiene un subarco dirigido con punto inicial P_0 y punto final P_1. Un número finito de puntos de subdivisión sobre un arco simple dirigido C parte dicho arco en una sucesión de subarcos dirigidos, siendo el punto inicial de un subarco el punto final del que le precede.

A menudo resulta impráctico restringirse a arcos simples e insistir en que diferentes valores del parámetro t pertenezcan a puntos diferentes sobre la curva. Si, por ejemplo, las ecuaciones $x = x(t)$, $y = y(t)$ dan la posición de una partícula móvil P en el tiempo t, no existe razón para que la partícula no pueda permanecer en reposo durante algún tiempo o para que su trayectoria no se corte a sí misma de manera que la partícula regrese a la misma posición un tiempo después.

Un ejemplo lo constituye la curva $x = t^2 - 1$, $y = t^3 - t$ [la cual también podría describirse completamente mediante la ecuación cúbica $y^2 - x^2(1 + x) = 0$]. Conforme t varía desde $-\infty$ hasta $+\infty$, la curva pasa por el origen dos veces, para $t = -1$ y para $t = +1$ (Fig. 4.9). Fácilmente se verifica que los demás puntos de la curva pertenecen a un único valor de t. Geométricamente, el intervalo $-1 < t < +1$ corresponde a una *espira* de la curva. Aquí, nuevamente, el sentido de las t crecientes define un cierto *orden entre los puntos de la curva*, al menos si se interpretan de alguna manera los puntos correspondientes a $t = -1$ y a $t = +1$ como diferentes, estando uno de ellos situado "encima" del otro. La curva cúbica completa y orientada puede descomponerse en arcos simples dirigidos, por ejemplo, en los arcos correspondientes a $n \leq t \leq n + 1$, donde n recorre todos los enteros.

Curvas cerradas

El ejemplo usual de una representación paramétrica en la cual valores de t diferentes corresponden al mismo punto de la curva está dado por las fórmulas

$$x = a \cos t, \quad y = a \operatorname{sen} t,$$

que describen el movimiento uniforme de un punto sobre una circunferencia, siendo t el tiempo. Conforme t varía desde $-\infty$ hasta $+\infty$ el punto $P = (x, y)$ recorre la circunferencia un número infinito de veces en el sentido contra reloj. Podemos hacer que los puntos de la circunferencia sean recorridos exactamente una vez restringiendo t a cualquier intervalo semiabierto de longitud 2π:

$$\alpha \leq t < \alpha + 2\pi.$$

Los puntos extremos α y $\alpha + 2\pi$ del intervalo corresponden al mismo punto de la circunferencia. Aquí los puntos extremos del intervalo paramétrico no tienen un significado geométrico especial para la curva.

Generalmente, un par de funciones continuas, $x = \phi(t)$, $y = \psi(t)$, definidas en un intervalo cerrado $a \leq t \leq b$ representan una curva *cerrada* si $\phi(a) = \phi(b)$ y $\psi(a) = \psi(b)$. La curva cerrada será simple si valores t diferentes, con $a \leq t < b$, corresponden a puntos diferentes (x, y).

El punto correspondiente a $t = a$ y $t = b$ podría ser cualquier punto de la curva; y es justamente el punto en el cual la curva se "rompe" para hacer que sus puntos correspondan a los de un intervalo sobre el eje.

Curvas cerradas representadas por funciones periódicas

Justamente como en el ejemplo de la circunferencia, se puede evitar el distinguir cualquier punto de división particular tomando para $\phi(t)$ y $\psi(t)$ funciones *periódicas* con período $p = b - a$. Es sumamente valioso hacer aquí algunos comentarios generales acerca de las funciones periódicas, a las cuales regresaremos para tratarlas en forma más intensiva en el capítulo 8.

Una función $f(t)$ se llama *periódica* con período p si está definida para todo t y satisface la ecuación $f(t) = f(t + p)$. Así, por ejemplo, las funciones trigonométricas $\operatorname{sen} t$ y $\cos t$ son periódicas con período 2π. (Cualquier múltiplo $2n\pi$, donde n es un entero, es entonces también un período.) Interpretada geométricamente, $f(t)$ tiene período p si un corrimiento de su gráfica en p unidades hacia la derecha lleva nuevamente a la misma gráfica.

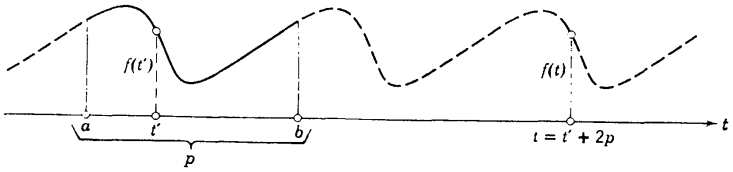


Figura 4.10 Gráfica de una función periódica. $f(t)$.

Puesto que $f(t)$ "se repite" a sí misma, una función $f(t)$ de período p está determinada para todo t si se conoce dicha función en un intervalo $a \leq t < b$ de longitud $p = b - a$ (Fig. 4.10). En efecto, para cada t existe un valor t' en el intervalo $a \leq t' < b$ tal que $t - t' = np$, donde n es un entero [sólo se tiene que tomar para n el mayor entero que no exceda a $(t - a)/p$]. Entonces $f(t) = f(t')$ es conocida.

De hecho, podemos partir con *cualquier* función $f(t)$ *continua* en un intervalo semiabierto $a \leq t < b$; y la función extendida será, claramente, continua para todo t que no sea de la forma $t = a + np$ con n entero (Fig. 4.11).

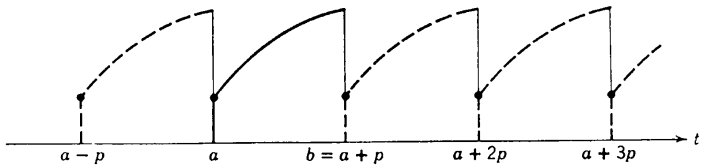


Figura 4.11 Continuación periódica de una función $f(t)$ partiendo del intervalo $a \leq t < b$.

Por ejemplo, la extensión en forma periódica de la función $f(t)$ definida por $f(t) = t$ para $0 \leq t < 1$, conduce a una función de período $p = 1$, que llamaremos "parte fraccional de t", y la cual es discontinua en los puntos t que son enteros (Fig. 4.12*a*). Generalmente, en $t = a + np$ la función periódicamente extendida tendrá el valor $f(a)$; y éste también será el límite de f al aproximarse al punto desde la derecha, en tanto que el límite de f desde la izquierda será el mismo que el correspondiente al punto b. En el caso de mayor interés para nosotros por ahora, comenzamos con una función definida y continua en el intervalo *cerrado* $a \leq t \leq b$, la cual tiene además el mismo valor en los puntos extremos: $f(a) = f(b)$. Continuando periódicamente tal función se llega siempre a una función $f(t)$ de período $p = b - a$, la cual es continua para todo t. (Fig. 4.12*b*).

Las funciones periódicas continuas son ideales para representar curvas cerradas C. Sea C dada paramétricamente por $x = \phi(t)$, $y = \psi(t)$, para

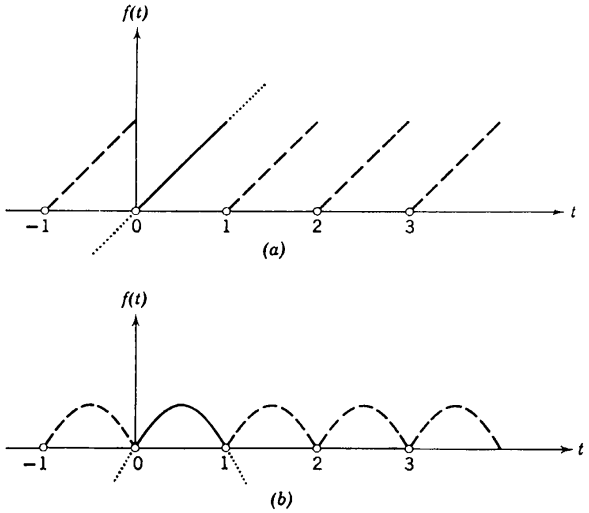


Figura 4.12 Continuación periódica de las funciones $f(t)$ definidas inicialmente en el intervalo $0 \leq t < 1$. Aquí (*a*) $f(t) = t$, (*b*) $f(t) = 2t - 2t^2$.

ϕ, ψ continuas en el intervalo $a \leq t \leq b$ y teniendo los mismos valores en ambos puntos extremos. Se puede extender la definición de estas funciones a todos los valores de t de manera que ϕ y ψ tengan período $b - a = p$ y sean continuas para todo t. Para cualquier t la representación paramétrica extendida solamente conduce a puntos de C puesto que se tiene $t = t' + np$ con n entero y $a \leq t' \leq b$. El punto correspondiente a t es entonces el mismo que el correspondiente a t', el cual está sobre C. Conforme t varía desde $-\infty$ hasta $+\infty$ el punto (x, y) recorre la curva C un número in-

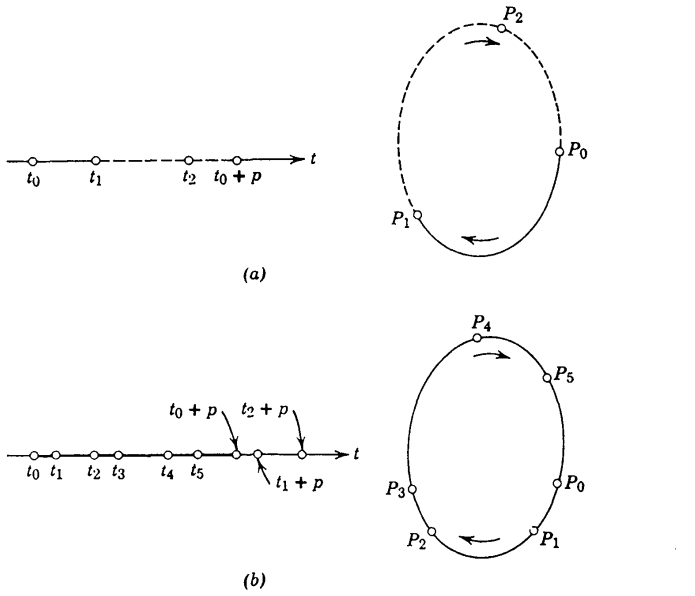


Figura 4.13 Orientación de curvas cerradas en el sentido de las t crecientes.

finito de veces, justamente como en el caso de la circunferencia $x = a \cos t$, $y = a \operatorname{sen} t$. Aquí el papel sobresaliente del valor paramétrico $t = a$ es eliminado. Para cualquier α la curva completa queda ahora representada por $x = \phi(t)$, $y = \psi(t)$ cuando t varía desde α hasta $\alpha + p$.

Una porción de la curva cerrada C correspondiente a los valores del parámetro t en un intervalo $\alpha \leq t \leq \beta$ forma un arco simple si valores t diferentes en ese intervalo conducen a puntos diferentes (x, y). La curva cerrada C completa es una curva simple si diferentes t en el mismo intervalo $\alpha \leq t < \alpha + p$ siempre conducen a puntos diferentes sobre C. Así, cualquier intervalo paramétrico cerrado de longitud menor que p origina un arco simple.

Curvas cerradas compuestas de arcos simples. Orden de puntos

Las curvas cerradas que se considerarán pueden todas ellas descomponerse en arcos simples. Si la curva cerrada C completa es simple, puede descomponerse en dos arcos simples $t_0 \leq t \leq t_1$ y $t_1 \leq t \leq t_0 + p$ que tengan solamente sus puntos extremos, P_0, P_1, en común. El sentido de las t crecientes determina un sentido positivo u orientación sobre C, fijando una dirección positiva sobre cada arco simple de C. Dos puntos distintos cualesquiera, P_0, P_1, sobre la curva cerrada simple C dividen a ésta en dos arcos simples. En el sentido de las t crecientes exactamente uno de los dos arcos tendrá a P_0 como punto inicial y a P_1 como punto final; lo llamaremos P_0P_1. La inversa se cumple para el otro arco.

Orientación y orden

La orientación positiva de C puede también caracterizarse mediante una terna ordenada de puntos $P_0P_1P_2$ de C, si se especifica que P_2 no está sobre el arco dirigido simple con punto inicial P_0 y punto final P_1. Las ternas $P_1P_2P_0$ y $P_2P_0P_1$ obtenidas por permutación cíclica a partir de $P_0P_1P_2$ describen la misma orientación (Fig. 4.13*a*).

*En forma completamente general, cualesquier n puntos distintos sobre la curva cerrada simple C, orientada, siempre se suceden uno al otro en un orden determinado $P_1P_2 \cdots P_n$, excepto por permutaciones cíclicas,[1] y dividen C en arcos simples dirigidos $P_1P_2, \ldots, P_{n-1}P_n$, P_nP_1. Pueden siempre escogerse valores paramétricos $t_1, t_2, \ldots, t_n$ para los puntos $P_1, P_2, \ldots, P_n$ de manera tal que los t_i formen una sucesión monótona creciente y estén todos contenidos en uno y el mismo intervalo paramétrico de longitud igual al período p (Fig. 4.13*b*).

Orientación de curvas y ángulos

Como ya se ha señalado en el capítulo 1, estamos obligados a hacer uso de los signos más y menos para establecer relaciones satisfactorias entre los entes geométricos y los conceptos analíticos expresados por números. Los ejemplos más simples los contituyen las *rectas dirigidas*, tales como el eje numérico. El sentido que se define como positivo en una recta es arbitrario al principio. Un sentido positivo correspondiente a valores crecientes de t puede asociarse con cualquier representación paramétrica particular $x = at + b$, $y = ct + d$ de la recta. Una recta orientada de esta manera *apunta* en una cierta *dirección*. Dos rectas paralelas dirigidas tienen el mismo sentido o bien sentidos opuestos. Una dirección puede también determinarse mediante un *rayo* que parte de un punto P_0, esto es, mediante una semirrecta que consiste de todos los puntos de la recta que "siguen" en el sentido positivo a un punto dado P_0.

[1] Esto es, $P_2P_3 \cdots P_nP_1$, $P_3P_4 \cdots P_nP_1P_2$, $\ldots$, $P_nP_1 \cdots P_{n-1}$ dan la misma orientación.

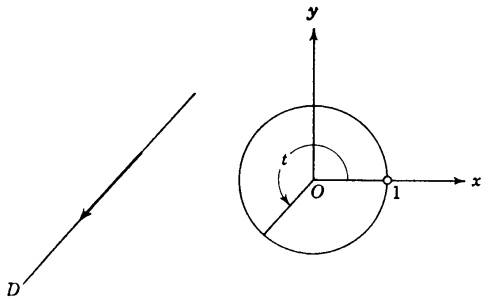


Figura 4.14 Angulo de inclinación t de una dirección D.

Cualquier dirección en el plano puede representarse mediante un rayo que parte del origen, y también mediante el punto P que es común al rayo y a la circunferencia de radio 1 con centro en el origen. Si representamos esta circunferencia unitaria en forma paramétrica mediante $x = \cos t$, $y = \operatorname{sen} t$, habremos asociado con cada dirección ciertos valores t que difieren uno de otro por múltiplos de 2π. Estos son los llamados *ángulos de inclinación* de la dirección o ángulos que la dirección forma con el eje x positivo. Existe siempre exactamente un ángulo de inclinación t para el cual $0 \leq t < 2\pi$ (Fig. 4.14).

Los *ángulos entre dos direcciones* son simplemente las diferencias entre sus ángulos de inclinación. Más precisamente, puesto que el orden en el cual se toman las dos direcciones es importante, decimos que *una dirección de inclinación t' forma con una dirección de inclinación t'' un ángulo* $\alpha = t' - t''$ (Fig. 4.15). Como t y t' pueden cambiar en múltiplos enteros de 2π, el mismo cambio es permisible para el ángulo que una dirección forma con otra.

Sentido de rotación

También se dice que la dirección con ángulo de inclinación t'' se convierte en la dirección t' mediante una *rotación* a través de un ángulo α. La

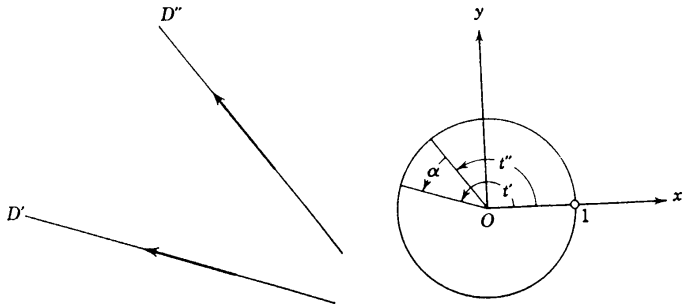


Figura 4.15 Angulo α que la dirección D' forma con la dirección D''.

idea intuitiva de rotación es aquí la de un *movimiento continuo* mediante el cual la dirección con inclinación t'' se transforma en la dirección con inclinación t' pasando por las direcciones con todas las inclinaciones posibles t intermedias entre t'' y t'. La rotación se denomina *positiva* o *contra reloj* si $\alpha = t' - t''$ es positivo, y *negativa* o *en el sentido del reloj* para el caso contrario. Por supuesto, existen muchas rotaciones diferentes, tanto en el sentido del reloj como en el opuesto, que convertirán una cierta dirección en otra dirección dada, a menos que se especifique que el ángulo de rotación satisface $-\pi < \alpha \leq \pi$.

Entonces, finalmente, el sentido positivo de rotación está asociado con una representación paramétrica particular $x = \cos t$, $y = \operatorname{sen} t$ de la circunferencia que se ha escogido. Si, como es costumbre, el eje x apunta hacia la derecha y el eje y hacia arriba, entonces el sentido positivo de rotación coincide con el sentido opuesto al de las manecillas de un reloj convencional.[1]

Lados positivo y negativo de una curva

Una curva separa los puntos del plano cercanos a uno de sus puntos P en dos clases. Por lo menos localmente pueden distinguirse dos "lados" de la curva. Si la curva C está orientada podemos definir un lado *positivo* (o "izquierdo")[2] y un lado *negativo* (o "derecho") de la siguiente manera. Considérese un rayo que parte de P. Se dice que este rayo apunta hacia el lado positivo de una curva si existen puntos Q sobre la curva, arbitrariamente próximos a P y que siguen a P en el sentido dado a la curva, tales que la recta PQ deba girarse en el sentido contra reloj un ángulo comprendido entre 0 y π para llegar al rayo dado (Fig. 4.16). Los puntos sobre el rayo cercanos a P se dice que están sobre el lado positivo de la curva. En el caso opuesto se dice que el rayo apunta hacia el lado negativo de C y que los puntos sobre él están en el lado negativo de la curva. Si C es una curva cerrada simple, ella divide todos los puntos del plano en dos clases, los *interiores* a C y los *exteriores* a C.[3] Se dice que C tiene una *orientación*

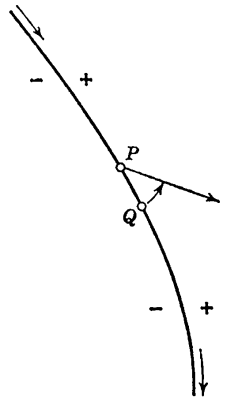


Figura 4.16 Lados positivo y negativo de un arco orientado.

[1] Este sentido, a su vez, es sugerido por el movimiento en el suelo de la sombra de un reloj de Sol en el hemisferio norte.

[2] Los términos "lado derecho" e "izquierdo" corresponden al uso ordinario de las palabras "ribera izquierda" y "ribera derecha" para un río orientado según la dirección de su flujo.

[3] Estos conceptos, así como la división del plano en dos partes por medio de una curva continua, cerrada, simple, se analizan en forma precisa en la topología. Aquí deben aceptarse sobre una base intuitiva.

contra reloj si su interior está sobre el lado positivo (esto es, izquierdo) (Fig. 4.17).

Pero si la curva cerrada C consiste de varias espiras, entonces no siempre es posible recorrer la curva de manera que todas las regiones encerradas estén en el lado positivo de ella (véase Fig. 4.18).

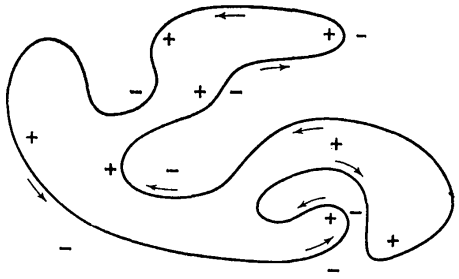


Figura 4.17 Curva cerrada simple con orientación contra reloj.

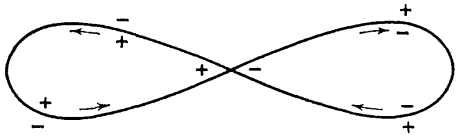


Figura 4.18

c. Derivadas, tangentes y normales en representación paramétrica

Dirección y velocidad

Para una curva C dada en representación paramétrica usando el tiempo como parámetro t,

$$x = x(t) = \phi(t), \quad y = y(t) = \psi(t),$$

las derivadas se denotan, como lo hizo Newton, mediante un punto:

$$\dot{x} = \frac{d\phi}{dt} = \dot{\phi}, \quad \dot{y} = \frac{d\psi}{dt} = \dot{\psi}.$$

Las derivadas $\dot{x}$ e $\dot{y}$ son a menudo visualizadas convenientemente como las "componentes de velocidad" o las "velocidades" de las coordenadas de un punto P que se mueve a lo largo de C.

Siempre que $\dot{x} \neq 0$, será posible representar la correspondiente porción de C por medio de una ecuación $y = f(x)$ calculando primero t como una función de x a partir de la primera ecuación y luego substituyendo la expresión resultante para t en la segunda ecuación. Por la regla de la cadena

de la derivación y la regla para la derivada de la inversa de una función (ver p. 228), se encuentra que la pendiente de la tangente a la curva está dada por

$$\frac{dy}{dx} = \frac{dy}{dt}\frac{dt}{dx} = \frac{\dfrac{dy}{dt}}{\dfrac{dx}{dt}} = \frac{\dot{y}}{\dot{x}}.$$

La fórmula equivalente $dx/dy = \dot{x}/\dot{y}$ se cumple si $\dot{y} \neq 0$.

A menos que se especifique lo contrario, siempre se supondrá que $\dot{x}$ e $\dot{y}$ no se anulan simultáneamente, o sea, escrito en forma concisa, se supondrá que

$$\dot{x}^2 + \dot{y}^2 \neq 0.$$

Entonces la tangente siempre existe;[1] es horizontal si $\dot{y} = 0$ y vertical si $\dot{x} = 0$.

Para la *cicloide,* por ejemplo [ver Ec. (1), p. 347] se tiene

$$\dot{x} = a(1 - \cos t) = 2a \operatorname{sen}^2 \frac{t}{2},$$

$$\dot{y} = a \operatorname{sen} t = 2a \operatorname{sen} \frac{t}{2} \cos \frac{t}{2},$$

$$\frac{dy}{dx} = \cot \frac{t}{2}.$$

Estas fórmulas muestran que $\dot{x}^2 + \dot{y}^2 \neq 0$ excepto para $t = 0, \pm 2\pi, \pm 4\pi, \ldots$. Además, la cicloide tiene una cúspide (esto es, un punto en el que se invierte la dirección) con una tangente vertical en esos puntos excepcionales, en los que también encuentra al eje x, esto es, para los cuales $y = 0$. En efecto, al aproximarse a estos puntos la derivada $y' = \dot{y}/\dot{x} = \cot(t/2)$ se vuelve infinita.

Tangente, normal y cosenos directores

La ecuación de la tangente a la curva en el punto x, y es

$$\eta - y = \frac{dy}{dx}(\xi - x),$$

donde ξ y η son las coordenadas "variables" correspondientes a un punto arbitrario sobre la tangente, en tanto que x, y y dy/dx tienen los valores

[1] Observamos que la condición $\dot{x}^2 + \dot{y}^2 \neq 0$, aunque suficiente, no es necesaria para garantizar una representación no paramétrica. Así, podemos definir la curva $y = x^2$ por medio de las ecuaciones paramétricas $x = t^3$, $y = t^6$. En el origen del eje t la condición de positividad para $\dot{x}^2 + \dot{y}^2$ falla, pero aún la curva tiene una representación no paramétrica bien definida.

fijos que corresponden al punto de contacto. Substituyendo dy/dx por $\dot{y}/\dot{x}$ podemos escribir la ecuación de la tangente en la forma

$$(\xi - x)\dot{y} - (\eta - y)\dot{x} = 0. \tag{5}$$

La misma ecuación exactamente se obtiene con la hipótesis $\dot{y} \neq 0$; y lo único que se necesita es expresar x como función de y. En los puntos excepcionales para los que tanto $\dot{x}$ como $\dot{y}$ se anulan para el mismo valor t, la ecuación pierde todo significado, pues entonces es satisfecha por ξ, η cualesquiera.

La *normal a la curva,* esto es, la recta que pasa por un punto de la curva y es perpendicular a la tangente en ese punto, tiene la pendiente $-dx/dy$. Esto conduce a la ecuación

$$(\xi - x)\dot{x} + (\eta - y)\dot{y} = 0 \tag{6}$$

para la normal.

Si un punto de C corresponde a varios valores de t, entonces en general existe una tangente diferente para cada una de las ramas de la curva que pasan por el punto, o para cada valor de t. Por ejemplo, la curva $x = t^2 - 1$, $y = t^3 - t$ (Fig. 4.9, p. 353) pasa por el origen para $t = -1$ y $t = +1$. Para $t = -1$ se encuentra que la ecuación de la tangente es $\xi + \eta = 0$, en tanto que para $t = +1$ la tangente está dada por $\xi - \eta = 0$.

De la definición de derivada se tiene

$$\frac{dy}{dx} = \frac{\dot{y}}{\dot{x}} = \tan\alpha,$$

donde α es el ángulo que la tangente forma con el eje x. Esto significa que una rotación del eje x a través del ángulo α (contra reloj si $\alpha > 0$, en el sentido del reloj si $\alpha < 0$) hará que dicho eje sea paralelo a la tangente. Entonces también rotaciones a través de ángulos $\alpha \pm \pi$, $\alpha \pm 2\pi$, ... harán el eje x paralelo a dicha tangente. Por lo tanto, el ángulo α está determinado solamente salvo un múltiplo de π, en tanto que $\tan\alpha$ está determinada en forma única. De las relaciones $\dot{y}/\dot{x} = (\operatorname{sen}\alpha)/(\cos\alpha)$ y $\dot{x}^2 + \dot{y}^2 \neq 0$ se obtiene

$$\cos\alpha = \pm\frac{\dot{x}}{\sqrt{\dot{x}^2 + \dot{y}^2}}, \qquad \operatorname{sen}\alpha = \pm\frac{\dot{y}}{\sqrt{\dot{x}^2 + \dot{y}^2}}$$

(debiendo usarse el mismo signo en ambas fórmulas). A $\cos\alpha$ y $\operatorname{sen}\alpha$ se les llama los *cosenos directores* de la tangente.[1]

Asignación de direcciones a la tangente y a la normal

Las dos posibles elecciones para los cosenos directores corresponden a las dos direcciones en las cuales puede recorrerse la tangente; y los correspondientes ángulos α difieren entre sí en un múltiplo impar de π. Una de las dos direcciones de la tangente corresponde a t creciente; la otra, a t decreciente. Supóngase que la dirección de la curva es la asociada a las t crecientes. Entonces, por definición, la dirección positiva de la tangente, o la que corresponde a valores crecientes de t, es aquélla que forma con el eje x positivo un ángulo α para el cual $\cos\alpha$ tiene el mismo signo que $\dot{x}$, y $\operatorname{sen}\alpha$ el mismo signo que $\dot{y}$. Los cosenos directores correspondientes a esa dirección de la tangente son entonces, sin ambigüedad:

$$\cos\alpha = \frac{\dot{x}}{\sqrt{\dot{x}^2 + \dot{y}^2}}, \qquad \operatorname{sen}\alpha = \frac{\dot{y}}{\sqrt{\dot{x}^2 + \dot{y}^2}}. \tag{7}$$

Si, por ejemplo, $\dot{x} = dx/dt > 0$, la dirección de la tangente que corresponde a las t crecientes es también la que corresponde a las x crecientes;

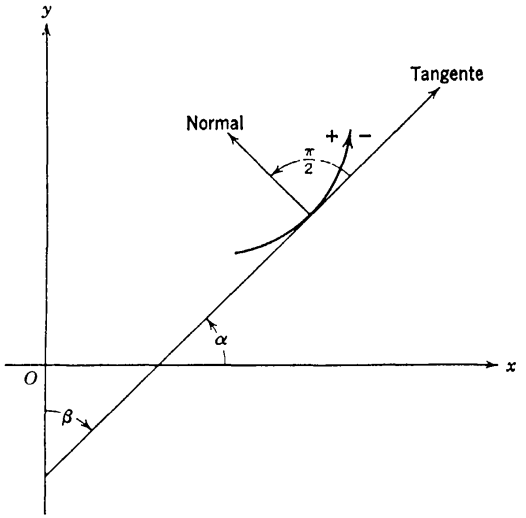


Figura 4.19 Tangente y normal positiva de una curva orientada.

y el ángulo que esa dirección forma con el eje x positivo tiene entonces un coseno positivo. Análogamente, la dirección normal que se obtiene haciendo girar un ángulo $\pi/2$ en el sentido positivo (contra reloj) la dirección de la tangente positiva, que corresponde a valores crecientes de t, tiene como cosenos directores

$$\cos\left(\alpha + \frac{\pi}{2}\right) = \frac{-\dot{y}}{\sqrt{\dot{x}^2 + \dot{y}^2}}, \qquad \operatorname{sen}\left(\alpha + \frac{\pi}{2}\right) = \frac{\dot{x}}{\sqrt{\dot{x}^2 + \dot{y}^2}}$$

Esta es la llamada dirección normal positiva y apunta hacia el "lado positivo" de la curva (Fig. 4.19).

[1] Aquí se concibe $\operatorname{sen}\alpha$ como $\cos\beta$, donde $\beta = \pi/2 - \alpha$ es el ángulo que el eje y forma con la tangente.

Si se introduce un nuevo parámetro $\tau = \chi(t)$ para la curva, los valores de $\cos\alpha$ y $\operatorname{sen}\alpha$ no cambian si $d\tau/dt > 0$ y cambian de signo si $d\tau/dt < 0$; esto es, si cambia el sentido de la curva cambian también los sentidos positivos de la tangente y de la normal.

Puntos críticos

Si $\dot{x}$ e $\dot{y}$ son continuas y $\dot{x}^2 + \dot{y}^2 > 0$, las cantidades $\cos\alpha$ y $\operatorname{sen}\alpha$ que determinan la dirección de la tangente variarán continuamente con t. La tangente, cuya ecuación es

$$(\xi - x)\operatorname{sen}\alpha - (\eta - y)\cos\alpha = 0,$$

cambia entonces en forma continua a lo largo de la curva; lo mismo que la normal.

Si $\dot{x}$ e $\dot{y}$ se anulan para cierto valor de t, los cosenos directores de la tangente no quedan definidos por nuestras fórmulas; y la tangente puede no existir en absoluto o no estar definida en forma única. A tales puntos se les llama puntos "críticos" o puntos "estacionarios". Mediante ejemplos se ilustrarán varias posibilidades en relación con los puntos críticos.

Uno de tales ejemplos lo proporciona la curva $y = |x|$ con la representación paramétrica $x = t^3$, $y = |t|^3$; esta curva tiene una esquina para $t = 0$, aunque tanto $\dot{x}$ como $\dot{y}$ son continuas. En el ejemplo de la cicloide, discutido en la p. 362, los puntos "estacionarios" en los cuales $\dot{x} = \dot{y} = 0$ corresponden a cúspides. Por otra parte, en algunos casos la anulación de $\dot{x}$ y de $\dot{y}$ es sólo consecuencia de la representación paramétrica específica y no tiene relación con el comportamiento de la curva, como ocurre para la recta representada por $x = t^3$, $y = t^3$, para el valor paramétrico $t = 0$.

Esquinas

Las curvas que consisten de varios arcos suaves que se unen en esquinas son representadas convenientemente en forma paramétrica mediante

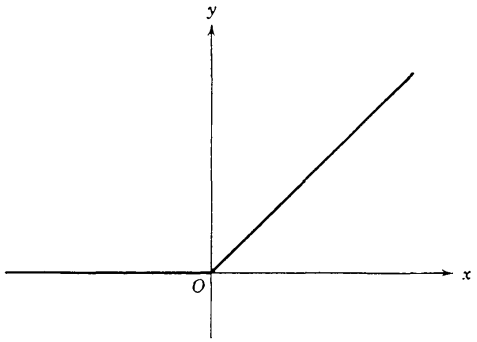


Figura 4.20 Gráfica de $x = t$, $y = \frac{1}{2}(t + |t|)$.

funciones $x(t)$, $y(t)$ continuas pero con derivadas $\dot{x}$, $\dot{y}$ que presentan discontinuidades de salto. Esto queda ilustrado por el ejemplo trivial de la línea quebrada con representación

$$x = t, \quad y = 0 \quad \text{para} \quad t \leq 0$$

y

$$x = t, \quad y = t \quad \text{para} \quad t \geq 0.$$

Aquí $\dot{x} = 1$, $\dot{y} = 0$ para $t < 0$, y $\dot{x} = 1$, $\dot{y} = 1$ para $t > 0$. En $t = 0$ la tangente está indeterminada (ver Fig. 4.20).

f. La longitud de una curva

La longitud como una integral

Existen dos tipos diferentes de cantidades o propiedades geométricas asociadas con una curva. El primer tipo depende sólo del comportamiento *local* de la curva, esto es, en la inmediata vecindad de un punto; tales propiedades son aquéllas que pueden representarse mediante derivadas en el punto en cuestión. Propiedades del segundo tipo, las *propiedades globales,* dependen de la configuración total de la curva o de una porción de ella y son usualmente expresadas en forma analítica mediante el concepto de integral. Empezaremos considerando una cantidad del segundo tipo, la longitud de una curva.

Por supuesto, se tiene ya una noción intuitiva de lo que se entiende por longitud de una curva. Sin embargo, como en el caso clásico de los arcos circulares, debe darse un significado matemático preciso al concepto intuitivo. Guiados por la intuición definimos la longitud de una curva arbitraria como el límite de las longitudes de polígonos que se le aproximan, en particular, polígonos inscritos. Las longitudes de los polígonos, a su vez, son definidas no bien se escoja una unidad de longitud. El resultado final será la expresión de la longitud mediante una integral.

Considérese la curva dada en la forma $x = x(t)$, $y = y(t)$, $\alpha \leq t \leq \beta$. En el intervalo comprendido entre α y β se escogen puntos intermedios $t_1, t_2, \ldots, t_{n-1}$ tales que

$$\alpha = t_0 < t_1 < t_2 < \cdots < t_{n-1} < t_n = \beta.$$

Los puntos $P_0, P_1, \ldots, P_n$ de la curva que corresponden a estos valores t_i en el mismo orden se unen mediante segmentos de recta, obteniéndose así un polígono inscrito. La longitud del perímetro de este polígono inscrito depende de la manera en que los puntos t_i o los vértices P_i del polígono sean escogidos. Déjese ahora que el número de puntos t_i crezca más allá de toda cota, de manera que, al mismo tiempo, la longitud del subintervalo (t_i, t_{i+1}) más largo tienda a cero. La longitud de la curva se define entonces

como el límite de los perímetros de estos polígonos inscritos, *siempre que* tal límite exista y sea independiente de la manera particular como se escojan los polígonos. Cuando esta hipótesis (hipótesis de *rectificabilidad)* se satisface, podemos hablar de la longitud de la curva.

Se supondrá que las funciones $x(t)$ e $y(t)$ tienen derivadas $\dot{x}(t)$ e $\dot{y}(t)$ continuas para $\alpha \leq t \leq \beta$. El polígono inscrito correspondiente a la subdivisión del intervalo t mediante los puntos t_i, con $\Delta t_i = t_{i+1} - t_i$, tiene vértices $P_i = (x(t_i), y(t_i))$; su longitud total está dada por la expresión

$$S = \sum_{i=0}^{n-1} \overline{P_i P_{i+1}} = \sum_{i=0}^{n-1} \sqrt{[x(t_{i+1}) - x(t_i)]^2 + [y(t_{i+1}) - y(t_i)]^2},$$

de acuerdo con el teorema de Pitágoras (cfr. Fig. 4.21, p. 374). Por el teorema del valor medio del cálculo diferencial,

$$x(t_{i+1}) - x(t_i) = \dot{x}(\xi_i)\,\Delta t_i, \qquad y(t_{i+1}) - y(t_i) = \dot{y}(\eta_i)\,\Delta t_i,$$

donde ξ_i y η_i son valores intermedios en el intervalo $t_i < t < t_{i+1}$. Esto conduce a la expresión

$$S_n = \sum_{i=0}^{n-1} \sqrt{[\dot{x}^2(\xi_i)]^2 + [\dot{y}^2(\eta_i)]^2}\,\Delta t_i$$

para la longitud del polígono, donde se ha usado el hecho de que las diferencias Δt_i son positivas. Si el número n de puntos t_i de subdivisión crece más allá de toda cota en tanto que, simultáneamente, el máximo valor Δt_i tiende a cero, la suma S_n tiende a la integral

$$L = \int_\alpha^\beta \sqrt{\dot{x}^2 + \dot{y}^2}\,dt.$$

Este hecho es una consecuencia directa de los teoremas de existencia para integrales, del capítulo 2.[1]

Lo anterior prueba que, para $\dot{x}$, $\dot{y}$ continuas, la curva realmente tiene una longitud y que esta longitud está dada por la expresión

$$L = \int_\alpha^\beta \sqrt{\dot{x}^2 + \dot{y}^2}\,dt. \tag{8}$$

La misma regla es válida si se admite que $\dot{x}$ e $\dot{y}$ sean discontinuas en puntos aislados, en los que la curva podría no tener, entonces, una tangente única; y, por supuesto, la integral debe entonces ser considerada como "impropia" (ver capítulo 3, p. 320). Curvas "rectificables" más generales, para las cuales la integral anterior sigue teniendo significado, no serán discutidas en este volumen.

[1] Puesto que los puntos intermedios ξ_i y η_i no necesariamente coinciden, hacemos uso de las sumas de aproximación más generales cuya convergencia a la integral se demostró en la p. 216.

Definición alternativa de longitud

Hacemos una observación interesante: El perímetro S de cualquier polígono inscrito π nunca puede exceder la longitud L de la curva. (En particular, la distancia entre los puntos extremos de la curva no puede exceder a L, pues la línea recta que une los puntos extremos es la curva *más corta* que va de uno de tales puntos al otro.) En efecto, L puede obtenerse como el límite de los perímetros de una sucesión especial de polígonos inscritos, en la cual se parte con el polígono π de perímetro S y se obtienen los siguientes agregando en forma sucesiva más y más vértices. Insertando un vértice adicional entre dos vértices sucesivos de un polígono inscrito nunca puede llegarse a una disminución de los perímetros, pues un lado de un triángulo jamás excede la suma de los otros dos. Así, L es el límite de una sucesión no decreciente de perímetros que comienza con S. Por tanto, $S \leq L$. Por consiguiente, en vez de definir L como el límite de los perímetros de una sucesión de polígonos inscritos correspondientes a subdivisiones cada vez más finas del intervalo t, podría también haberse definido L como el *extremo superior* [1] de los perímetros de *todos* los polígonos inscriptos. Es interesante el hecho de que la longitud pueda definirse sin recurrir formalmente a ningún paso al límite.

Invariancia de la longitud respecto de cambios de parámetros

Resulta claro de la definición que la longitud L de una curva C no puede depender de la representación paramétrica particular usada para C. Por lo tanto, si se introduce un nuevo parámetro $\tau = \chi(t)$, donde $d\tau/dt > 0$, nuestra fórmula integral para L debe dar el mismo valor independiente de que el parámetro usado sea t o τ. Esto puede verificarse inmediatamente a partir de la regla de la cadena de la derivación y de las leyes de substitución para integrales. En efecto, se tiene

$$\sqrt{\dot{x}^2 + \dot{y}^2} = \sqrt{\left(\frac{dx}{dt}\right)^2 + \left(\frac{dy}{dt}\right)^2} = \sqrt{\left(\frac{dx}{d\tau}\frac{d\tau}{dt}\right)^2 + \left(\frac{dy}{d\tau}\frac{d\tau}{dt}\right)^2}$$

$$= \sqrt{\left(\frac{dx}{d\tau}\right)^2 + \left(\frac{dy}{d\tau}\right)^2}\frac{d\tau}{dt};$$

por lo tanto, si $\chi(\alpha) = a$, $\chi(\beta) = b$,

$$L = \int_\alpha^\beta \sqrt{\dot{x}^2 + \dot{y}^2}\,dt = \int_\alpha^\beta \sqrt{\left(\frac{dx}{d\tau}\right)^2 + \left(\frac{dy}{d\tau}\right)^2}\frac{d\tau}{dt}\,dt$$

$$= \int_a^b \sqrt{\left(\frac{dx}{d\tau}\right)^2 + \left(\frac{dy}{d\tau}\right)^2}\,d\tau,$$

de manera que la expresión de la longitud basada en el parámetro τ conduce al mismo valor de L. Si, por otra parte, $d\tau/dt < 0$, análogamente se encuentra

[1] Definido como *la menor de las cotas superiores,* o *supremum,* y denotado frecuentemente por los símbolos "*extr*" y "*sup*". (Cfr. pp. 120-121). *(N. del R.)*

$$\int_\alpha^\beta \sqrt{\dot{x}^2 + \dot{y}^2}\, dt = -\int_a^b \sqrt{\left(\frac{dx}{d\tau}\right)^2 + \left(\frac{dy}{d\tau}\right)^2}\, d\tau$$
$$= \int_b^a \sqrt{\left(\frac{dx}{d\tau}\right)^2 + \left(\frac{dy}{d\tau}\right)^2}\, d\tau;$$

y el miembro derecho es nuevamente la integral correcta para la longitud de C referida al parámetro τ, puesto que ahora $b < a$ debido a que $\chi(t)$ es una función decreciente.

Para una curva dada en forma no paramétrica mediante la función $y = f(x)$, $a \leq x \leq b$, podemos introducir x como parámetro t. Entonces $\dot{x} = 1$, $\dot{y} = dy/dx$. La longitud de la curva queda entonces dada por

$$L = \int_a^b \sqrt{1 + \left(\frac{dy}{dx}\right)^2}\, dx. \tag{9}$$

Ejemplos. Como un ejemplo se encuentra la longitud de un segmento de la parábola $y = \frac{1}{2}x^2$ correspondiente al intervalo $a \leq x \leq b$:

$$L = \int_a^b \sqrt{1 + x^2}\, dx.$$

En este caso la substitución $x = \operatorname{senh} t$ (véase capítulo 3, p. 293) conduce a

$$\int_{ar\,\operatorname{senh} a}^{ar\,\operatorname{senh} b} \cosh^2 t\, dt = \tfrac{1}{2}\int_{ar\,\operatorname{senh} a}^{ar\,\operatorname{senh} b} (1 + \cosh 2t)\, dt$$
$$= \tfrac{1}{2}(t + \operatorname{senh} t \cosh t)\Big|_{ar\,\operatorname{senh} a}^{ar\,\operatorname{senh} b}$$
$$= \tfrac{1}{2}(ar\,\operatorname{senh} b + b\sqrt{1 + b^2} - ar\,\operatorname{senh} a\sqrt{1 + a^2}).$$

Para una curva dada por la ecuación $r = r(\theta)$, $\alpha \leq \theta \leq \beta$, en coordenadas polares, se tiene la representación $x = r(\theta)\cos\theta$, $y = r(\theta)\operatorname{sen}\theta$. Escogiendo θ como parámetro se tiene

$$\dot{x} = \dot{r}\cos\theta - r\operatorname{sen}\theta, \qquad \dot{y} = \dot{r}\operatorname{sen}\theta + r\cos\theta, \qquad \dot{x}^2 + \dot{y}^2 = r^2 + \dot{r}^2.$$

Esto conduce a la expresión

$$L = \int_\alpha^\beta \sqrt{r^2 + \left(\frac{dr}{d\theta}\right)^2}\, d\theta \tag{10}$$

para la longitud de una curva en coordenadas polares. Por ejemplo, para la circunferencia de radio a alrededor del origen se tiene la ecuación $r =$ constante $= a$, $0 \leq \theta \leq 2\pi$. Esto da para la longitud total de la circunferencia

$$L = \int_0^{2\pi} a\, d\theta = 2\pi a.$$

Aditividad de la longitud

Sea C una curva dada por $x = x(t)$, $y = y(t)$, $\alpha \leq t \leq \beta$, donde $\dot{x}$ e $\dot{y}$ son continuas. Sea γ cualquier valor entre α y β. Por las reglas generales para integrales se tiene

$$\int_\alpha^\beta \sqrt{\dot{x}^2 + \dot{y}^2}\, dt = \int_\alpha^\gamma \sqrt{\dot{x}^2 + \dot{y}^2}\, dt + \int_\gamma^\beta \sqrt{\dot{x}^2 + \dot{y}^2}\, dt.$$

Las integrales de la derecha representan respectivamente las longitudes de las porciones en las que C es dividida por el punto correspondiente a $t = \gamma$. Por lo tanto, la longitud de la curva completa es igual a la suma de las longitudes de sus partes.

No es necesario que $\dot{x}$ e $\dot{y}$ sean continuas. Las integrales existen también cuando $\dot{x}$ e $\dot{y}$ tienen un número finito de discontinuidades de salto, como ocurriría para una curva con esquinas. La longitud total de la curva es entonces la suma de las longitudes de las porciones suaves comprendidas entre esquinas. Comportamientos aún más singulares de $\dot{x}$ e $\dot{y}$ son permitidos siempre y cuando la expresión de la longitud tenga sentido como integral impropia.

g. La longitud de arco como parámetro

Hemos visto que una y la misma curva admite muchas representaciones paramétricas diferentes $x = x(t)$, $y = y(t)$. Cualquier función monótona de t puede usarse como parámetro en lugar de t. Sin embargo, para muchos propósitos es ventajoso tomar como referencia para C algún "parámetro usual" que de alguna manera sea distinguido geométricamente. La abscisa x o el ángulo polar θ no son adecuados para ese propósito si las curvas han de describirse globalmente; además, dichos parámetros dependen de la elección del sistema de coordenadas. La posibilidad de medir longitudes a lo largo de una curva proporciona un parámetro natural definido geométricamente al cual pueden ser referidos los puntos P de una curva rectificable. Este parámetro es la longitud de la porción de la curva comprendida entre P y algún punto fijo P_0.

Comenzamos con una representación paramétrica arbitraria $x = x(t)$, $y = y(t)$, $\alpha \leq t \leq \beta$ de C. La derivación con respecto a t es indicada mediante un punto. Introducimos la "longitud de arco" s por medio de la integral indefinida

$$s = \int \sqrt{\dot{x}^2 + \dot{y}^2}\, dt, \tag{11}$$

o, en forma más precisa, s como una función de t por

$$s = s(t) = c + \int_{t_0}^t \sqrt{\dot{x}^2(\tau) + \dot{y}^2(\tau)}\, d\tau, \tag{11a}$$

donde c es una constante, t_0 es un valor entre α y β, y con τ se designa la variable de integración para distinguirla del límite superior t. Claramente, para cualesquier valores t_1 y t_2 del intervalo paramétrico la diferencia

$$s(t_2) - s(t_1) = \int_{t_1}^{t_2} \sqrt{\dot{x}^2 + \dot{y}^2}\, d\tau \tag{12}$$

es igual a la longitud del pedazo de curva limitado por los puntos correspondientes a $t = t_1$ y $t = t_2$, suponiéndose $t_1 < t_2$. Para $t_1 > t_2$ la diferencia $s(t_2) - s(t_1)$ es la negativa de la longitud de esa porción de curva. Así, el conocimiento de cualquier integral indefinida s permite calcular la longitud de cualquier parte de la curva.

El signo de la longitud de arco

Si la constante c tiene el valor 0, puede interpretarse $s(t)$ como la longitud de arco de la curva (o "distancia a lo largo de la curva") entre el punto P_0 con parámetro t_0 y el punto P con parámetro t. Aquí la longitud se cuenta positiva en el caso en que el arco con punto inicial P_0 y punto final P tenga la orientación correspondiente a las t crecientes.[1]

La forma integral de la definición de s es equivalente a la relación

$$\frac{ds}{dt} = \sqrt{\left(\frac{dx}{dt}\right)^2 + \left(\frac{dy}{dt}\right)^2}. \tag{12a}$$

Usando la notación simbólica para diferenciales (p. 201), $ds = (ds/dt)\,dt$, etc., podemos escribir esa relación en la sugestiva forma

$$ds = \sqrt{dx^2 + dy^2}$$

para el "elemento de longitud" ds.

Velocidad de movimiento a lo largo de una curva

Si t se interpreta como el tiempo y $x(t)$, $y(t)$ como las coordenadas de posición en el tiempo t de un punto o partícula puntual móvil, se tiene que

$$\dot{s} = \frac{ds}{dt} = \lim_{h \to 0} \frac{s(t+h) - s(t)}{h}$$

da la medida de la variación respecto al tiempo de la distancia recorrida por el punto *a lo largo de su trayectoria,* esto es, la *velocidad* de la partícula en el instante t.[2] Para una partícula que se mueve con velocidad

[1] Nótese que la variable s no es única; depende de la elección de P_0 y c y también de la orientación inducida por el parámetro t sobre la curva. Sin embargo, cualquier otra longitud de arco es expresable en términos de s en la forma (s + constante) o bien ($-s$ + constante).

[2] O sea, el límite de la razón incremental $\Delta s/h$ de espacio sobre tiempo h. El cociente $\Delta s/h$ se llama *velocidad media* del punto en el intervalo $(t, t+h)$. *(N. del R.)*

uniforme a lo largo de la curva $\dot{s}$ es constante y s es entonces una función lineal del tiempo t.

Si nuestra hipótesis usual de que

$$\dot{x}^2 + \dot{y}^2 \neq 0$$

se satisface, tenemos $ds/dt \neq 0$ y podemos entonces introducir s como parámetro. De esta manera se simplifican una gran cantidad de fórmulas y cálculos. Las cantidades

$$\frac{dx}{ds} = \frac{dx}{dt}\frac{dt}{ds} = \frac{\dot{x}}{\sqrt{\dot{x}^2 + \dot{y}^2}} \qquad \frac{dy}{ds} = \frac{dy}{dt}\frac{dt}{ds} = \frac{\dot{y}}{\sqrt{\dot{x}^2 + \dot{y}^2}}$$

son entonces justamente los cosenos directores de la tangente que apunta en la dirección de s creciente (ver (7), p. 364). La relación

$$\left(\frac{dx}{ds}\right)^2 + \left(\frac{dy}{ds}\right)^2 = 1 \tag{13}$$

caracteriza el parámetro s como la longitud de arco a lo largo de la curva.

h. Curvatura

Definición mediante la medida del cambio de dirección

Se discute a continuación un concepto básico que se refiere solamente al *comportamiento local* de una curva en la vecindad de un punto, el concepto de curvatura.

Conforme se recorre la curva, el ángulo de inclinación α de la tangente a la curva variará en una medida definida por unidad de longitud de arco recorrida. Llamaremos a esta medida del cambio de α *la curvatura de la curva.* De acuerdo con esto, la curvatura está definida como

$$\kappa = \frac{d\alpha}{ds}. \tag{14}$$

Expresiones paramétricas. Supóngase que la curva está dada paramétricamente mediante las funciones $x = x(t)$, $y = y(t)$ que tienen primera y segunda derivadas con respecto a t continuas, para las cuales se cumple $\dot{x}^2 + \dot{y}^2 \neq 0$. Al calcular $d\alpha/ds$, o sea, la medida de cambio del ángulo de dirección α en el punto P, se tiene que tomar en cuenta que α no está definido en forma única. Sin embargo, la función trigonométrica de α, $\tan\alpha = \dot{y}/\dot{x}$ (o bien $\cot\alpha = \dot{x}/\dot{y}$ para $\dot{x} = 0$) tiene un valor definido. Para formar $d\alpha/ds$ puede siempre suponerse que los valores paramétricos pertenecientes a puntos en la vecindad de P están todos ellos en un intervalo en todo el cual una de las cantidades $\dot{x}$, $\dot{y}$ es diferente de cero. Si, por ejemplo, $\dot{x} \neq 0$, se

puede asignar a α un valor que varíe continuamente con t en el intervalo tomando simplemente

$$\alpha = \alpha(t) = \arctan\frac{\dot{y}}{\dot{x}} + n\pi,$$

donde n es un entero fijo, posiblemente negativo, y "arc tan" indica el valor principal de la función (cfr. p. 236), el cual está entre $-\pi/2$ y $\pi/2$. Análogamente, si $\dot{y} \neq 0$ en el intervalo, se puede tomar para α la expresión

$$\alpha(t) = \operatorname{arc\,cot}\frac{\dot{x}}{\dot{y}} + n\pi = \frac{\pi}{2} - \arctan\frac{\dot{x}}{\dot{y}} + n\pi.^{1}$$

En cualquier caso, se encuentra por derivación directa para cualquier representación paramétrica:

$$\dot{\alpha} = \frac{d\alpha}{dt} = \frac{\dot{x}\ddot{y} - \ddot{x}\dot{y}}{\dot{x}^2 + \dot{y}^2}.$$

Puesto que también (ver (12*a*), p. 371)

$$\dot{s} = \frac{ds}{dt} = \sqrt{\dot{x}^2 + \dot{y}^2},$$

se obtiene para la curvatura $d\alpha/ds = \dot{\alpha}/\dot{s}$ de la curva la expresión

$$\kappa = \frac{d\alpha}{ds} = \frac{\dot{\alpha}}{\dot{s}} = \frac{\dot{x}\ddot{y} - \dot{y}\ddot{x}}{(\dot{x}^2 + \dot{y}^2)^{3/2}}. \tag{15}$$

Escogiendo en particular la longitud de arco s como el parámetro t se tiene

$$\dot{x}^2 + \dot{y}^2 = 1$$

[véase Ec. (13), p. 372] y, por lo tanto, se obtiene el resultado simplificado

$$\kappa = \dot{x}\ddot{y} - \dot{y}\ddot{x}. \tag{15a}$$

Signo y valor absoluto de la curvatura

Introduciendo un nuevo parámetro $\tau = \tau(t)$ en lugar de t no se afecta la dirección de la tangente y, por lo tanto, no ocurre un cambio en α. Análogamente, el valor absoluto de la diferencia de los valores de s en dos puntos tiene un significado geométrico que es independiente de la elección

[1] Podríamos definir $\alpha(t)$ como una función continua para *todos* los valores del parámetro t dividiendo el intervalo paramétrico total en subintervalos en cada uno de los cuales $\dot{x} \neq 0$ o $\dot{y} \neq 0$. En cada subintervalo podemos definir entonces $\alpha(t)$ por medio de una de las expresiones de arriba, eligiendo el entero constante n en cada subintervalo de manera que los valores de α en el extremo común de dos intervalos adyacentes, calculados usando las expresiones correspondientes a esos intervalos, coincidan.

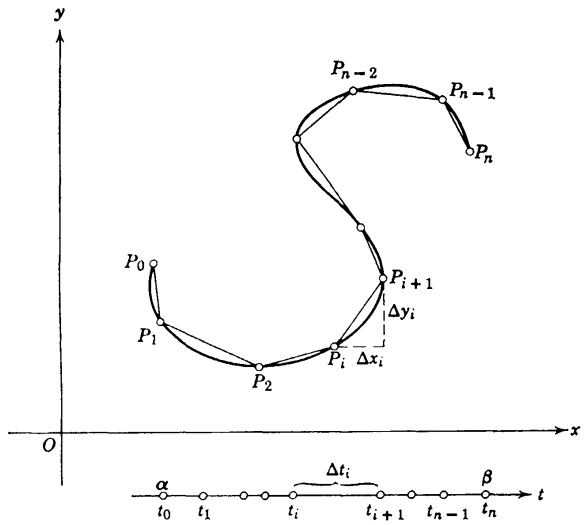


Figura 4.21 Rectificación de curvas.

del parámetro, a saber, el significado de distancia medida a lo largo de la curva. Sin embargo, el *signo* de la diferencia debe siempre tomarse como el signo de la diferencia entre los correspondientes valores paramétricos, puesto que se ha definido s como una función creciente de t. Así, el valor absoluto de la curvatura $|\kappa| = |d\alpha/ds|$ no depende de la elección del pará-

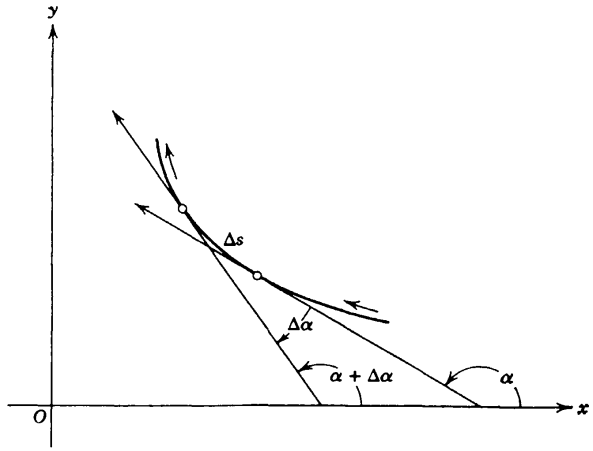


Figura 4.21(*a*) Curvatura $\kappa = \lim \Delta\alpha/\Delta s$ de una curva. (En el caso ilustrado se tiene $\kappa < 0$.)

metro, en tanto que el signo de κ depende del sentido sobre la curva que corresponde a t creciente. Obviamente, $\kappa > 0$ significa que α crece con s, esto es, que la tangente gira contra reloj conforme se recorre la curva con s o t crecientes (ver Fig. 4.21*a*). En este caso la orientación de la curva C es tal que el lado positivo de ella también es su lado "interno", esto es, el lado hacia el cual C se dobla.

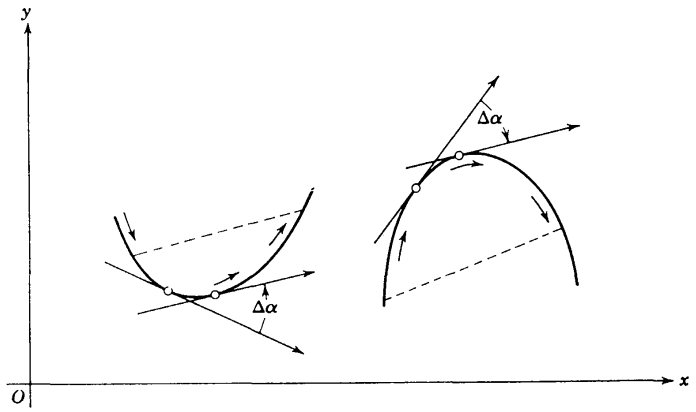


Figura 4.22 Gráfica de una función $f(x)$ convexa (izquierda) y cóncava (derecha).

Si la curva está dada mediante una ecuación $y = f(x)$, se tiene, usando x como parámetro,

$$\kappa = \frac{y''}{(1 + y'^2)^{3/2}}, \tag{16}$$

donde y' e y'' son las derivadas de y con respecto a la variable x. Aquí el signo de la curvatura es el que corresponde a las x crecientes. Obviamente, κ es positiva para $y'' > 0$; en este caso la tangente gira contra reloj conforme x crece; y la función $f(x)$ se llama *convexa*. La porción de la curva que une dos puntos cualesquiera está por debajo de la línea recta que une los mismos puntos. Para $y'' < 0$ la tangente gira en el sentido del reloj para x creciente, y la función f se llama *cóncava*. (Fig. 4.22.) Aquí la curva está por encima de la cuerda que une dos de sus puntos. El caso intermedio en que la curvatura tiene el valor cero corresponde (en términos generales) a un *punto de inflexión*, en el cual $y'' = 0$ (ver p. 258).

Ejemplos. Para la curvatura de la circunferencia de radio a dada por $x = a\cos t$, $y = a \operatorname{sen} t$ se encuentra el valor constante $1/a$ a partir de la fórmula general (15). Así, la curvatura de una circunferencia recorrida en el sentido contra reloj es la recíproca del radio. Este resultado nos asegura que nuestra definición de curvatura es realmente adecuada; pues en

el caso de una circunferencia pensamos como cosa natural que la recíproca del radio sea una medida de su curvatura.

Un segundo ejemplo es la curva definida mediante la función $y = x^3$. La curvatura es

$$\kappa = \frac{6x}{(1 + 9x^4)^{3/2}}.$$

Para $x < 0$ la función $y = x^3$ es cóncava puesto que $\kappa < 0$, y la tangente gira en el sentido del reloj, en tanto que para $x = 0$ se tiene un punto de inflexión, y para $x > 0$ la función se vuelve convexa.

Una función cuya curvatura es idénticamente igual a cero es una línea recta, como puede verse fácilmente por nuestra definición, y la línea recta es la única curva que satisface esa condición.

Circunferencia de curvatura y centro de curvatura

Introducimos $\rho = 1/\kappa$. La cantidad $|\rho| = 1/|\kappa|$ se llama *radio de curvatura* en el punto en cuestión. (Es infinito en un punto de inflexión, en el que $\kappa = 0$.) Para una circunferencia el radio de curvatura en cualquier punto es justamente el radio de la circunferencia.

A cualquier punto $P = (x, y)$ de la curva C se le asigna una circunferencia tangente a C en P y que tiene la misma curvatura que C cuando la curva y la circunferencia se recorren en el mismo sentido en P. Esta circunferencia se llama *circunferencia de curvatura* de la curva C en el punto P.[1] Su centro es el *centro de curvatura* de la curva C correspondiente al punto P. (Fig. 4.23). Puesto que C y el círculo tienen el mismo radio de curvatura, el radio de la circunferencia debe ser el radio de curvatura $|\rho|$ de C, y el centro (ξ, η) de la misma debe estar sobre la normal a C en el punto P y a una distancia $|\rho|$ de este punto. Además, puesto que C y la circunferencia se curvan hacia el mismo lado, el centro está a lo largo de la dirección normal en P, sobre el lado positivo o el negativo según que la curvatura κ sea positiva o negativa.

La dirección que va de P al centro de curvatura forma un ángulo $\alpha + \pi/2$ con el eje x positivo si $\kappa > 0$. Así, si ξ, η son las coordenadas del centro de curvatura y x, y las de P, se tiene [Ver Ec. (7), p. 364]:

$$\frac{\xi - x}{\rho} = \cos\left(\alpha + \frac{\pi}{2}\right) = -\operatorname{sen}\alpha = \frac{-\dot{y}}{\sqrt{\dot{x}^2 + \dot{y}^2}},$$

$$\frac{\eta - y}{\rho} = \operatorname{sen}\left(\alpha + \frac{\pi}{2}\right) = \cos\alpha = \frac{\dot{x}}{\sqrt{\dot{x}^2 + \dot{y}^2}}.$$

[1] Algunos dicen "*círculo* de curvatura", aunque es claro que se trata de una *circunferencia*. Tal lenguaje erróneo aparece en varias lenguas. *(N. del R.)*

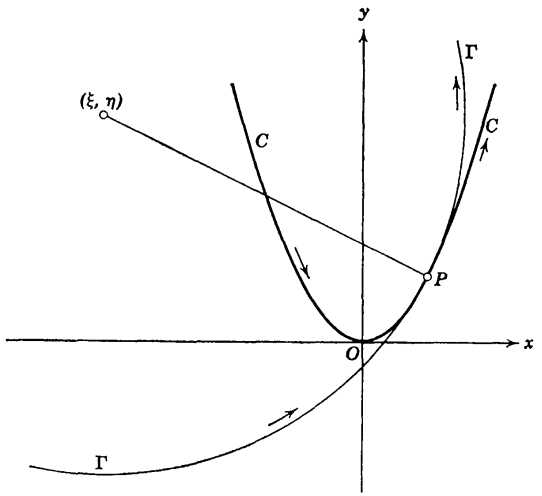


Figura 4.23. Círculo de curvatura Γ y centro de curvatura (ξ, η) correspondientes a un punto P de la curva C.

Por lo tanto, para $\kappa > 0$ resulta

$$(17) \qquad \xi = x - \frac{\rho\dot{y}}{\sqrt{\dot{x}^2 + \dot{y}^2}}, \qquad \eta = y + \frac{\rho\dot{x}}{\sqrt{\dot{x}^2 + \dot{y}^2}}.$$

Si el arco cuya longitud es s se toma como parámetro t, se obtienen las sencillas expresiones

$$(17a) \qquad \xi = x - \rho\dot{y}, \quad \eta = y + \rho\dot{x}.$$

Las mismas fórmulas para ξ, η se obtienen cuando $\kappa < 0$, en cuyo caso el radio de curvatura es $-\rho$ y, además, la dirección que va de P al centro forma un ángulo $\alpha - \pi/2$ con el eje x positivo.

La circunferencia de curvatura como circunferencia osculatriz

Las fórmulas (17) dan una expresión para el centro de curvatura en términos del parámetro t correspondiente al punto P de la curva. Conforme t recorre todos los valores del intervalo paramétrico, el centro de curvatura describe una curva, la llamada *evoluta* de la curva dada; y, puesto que, junto con x e y, se tienen que considerar $\dot{x}$, $\dot{y}$ y ρ como funciones conocidas de t, las fórmulas anteriores dan las ecuaciones paramétricas de esta evoluta. En el Apéndice I, p. 442, se encontrarán ejemplos y una discusión de las propiedades geométricas de la evoluta.

Dos curvas cualesquiera se dice que "*osculan*" en un punto P, o que tienen "*contacto de orden dos*" en P, si dichas curvas pasan por P, tienen

la misma tangente en P y tienen también la misma curvatura cuando se orientan de la misma manera. Obviamente, dos curvas osculantes tienen la misma circunferencia y el mismo centro de curvatura en P. Si las curvas están dadas mediante ecuaciones $y = f(x)$ e $y = g(x)$ en forma no paramétrica, es fácil expresar la condición de que ellas tengan un punto de contacto P y las mismas tangentes y curvaturas en ese punto. Si x es la abscisa del punto de contacto P, se tiene $f(x) = g(x)$, $f'(x) = g'(x)$; y la igualdad de las curvaturas es expresada mediante

$$\frac{f''(x)}{[1 + f'^2(x)]^{3/2}} = \frac{g''(x)}{[1 + g'^2(x)]^{3/2}},$$

y, por lo tanto, también es $f''(x) = g''(x)$. Así, la condición para que exista un punto de contacto con curvaturas iguales es que los valores de f y g, junto con los valores de las derivadas primera y segunda, coincidan en ese punto.

Considérense una curva C: $y = f(x)$ y su circunferencia de curvatura Γ en P representada mediante $y = g(x)$ en una vecindad de P. Puesto que la circunferencia Γ coincide con su circunferencia de curvatura, vemos que C y Γ tienen la misma circunferencia de curvatura y, por lo tanto, osculan en P. Consecuentemente, en el punto de contacto es $f(x) = g(x)$, $f'(x) = g'(x)$, $f''(x) = g''(x)$. Se dice que esta circunferencia es "la circunferencia que más se aproxima" a la curva en el punto P de contacto, puesto que ninguna otra circunferencia que intersecta a la curva en el punto de contacto tiene "contacto de orden dos" con C en ese punto. La circunferencia de curvatura es la *circunferencia osculante* u *osculatriz*. (Véase también capítulo 5, p. 478.)

Por otra parte, de la misma manera que la tangente a una curva es el límite, para $P_1 \to P$, de una recta que pasa por dos puntos consecutivos P y P_1 de C, se puede demostrar que la circunferencia de curvatura en P es el límite de las circunferencias que pasan por tres puntos P, P_1, P_2, para $P_1 \to P$ y $P_2 \to P$. La demostración se deja al lector. (Ver problema 4, p. 455.)

i. Cambio de ejes de coordenadas. Invariancia

Las propiedades inherentes a una situación geométrica o física no dependen del sistema de coordenadas específico o "marco de referencia" con respecto al cual dichas propiedades son formuladas. El carácter intrínseco de propiedades tales como distancia, longitud o ángulo debe reflejarse en enunciados que muestren que las respectivas fórmulas permanecen inalterables o son *invariantes* si se pasa de un sistema de coordenadas a otro. Algunos comentarios breves referentes a este punto son oportunos en esta sección.

Se usarán las ecuaciones generales que relacionan las coordenadas x, y de un punto P en un sistema de coordenadas con las coordenadas ξ, η del mismo punto P en cualquier otro sistema. La posición relativa del segundo conjunto de ejes de coordenadas respecto al primer conjunto está caracterizada mediante las coordenadas a, b que el origen del segundo sistema tiene respecto al primer sistema, y mediante el ángulo γ que el eje positivo ξ forma con el eje x positivo.[1] Las coordenadas (x, y) y (ξ, η) del mismo punto en los dos sistemas están relacionadas (cfr. Fig. 4.24) por medio de la transformación

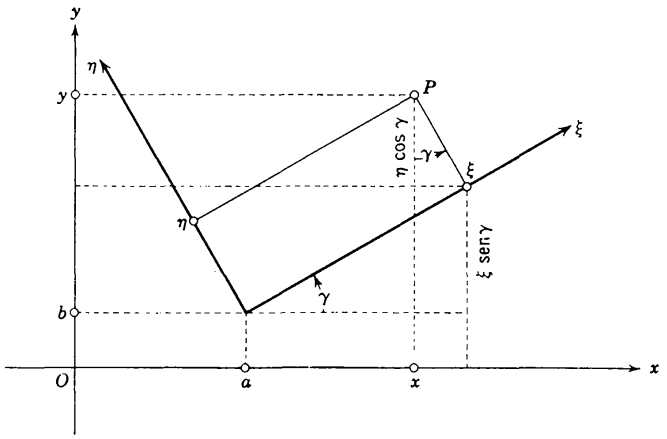


Figura 4.24 Cambio de ejes coordenados.

$$\begin{aligned} x &= \xi \cos\gamma - \eta \operatorname{sen}\gamma + a, \\ y &= \xi \operatorname{sen}\gamma + \eta \cos\gamma + b. \end{aligned} \tag{18}$$

Para $\gamma = 0$ no se tiene rotación de ejes sino solamente un *desplazamiento paralelo* o *traslación*, y las fórmulas toman la forma sencilla $x = \xi + a$, $y = \eta + b$.

Resolviendo las ecuaciones para ξ, η en términos de x, y se encuentra

$$\begin{aligned} \xi &= (x - a)\cos\gamma + (y - b)\operatorname{sen}\gamma, \\ \eta &= -(x - a)\operatorname{sen}\gamma + (y - b)\cos\gamma. \end{aligned} \tag{18a}$$

Si x e y son funciones de un parámetro t que define una curva, a partir de estas fórmulas se obtienen inmediatamente expresiones para ξ y η como funciones de t, dando la representación paramétrica de la misma curva

[1] Nos restringimos a sistemas coordenados "dextrógiros", en los cuales la dirección positiva del segundo sistema de ejes se obtiene mediante una rotación de 90° contra reloj del primer sistema.

en el sistema ξ, η. Derivando con respecto a t (las cantidades a, b, γ, que fijan la posición relativa entre los sistemas de coordenadas, no dependen de t) se obtiene la transformación de las "componentes de velocidad", esto es, la transformación para las derivadas de las coordenadas con respecto a t:

$$\dot{x} = \dot{\xi}\cos\gamma - \dot{\eta}\operatorname{sen}\gamma, \qquad \dot{y} = \dot{\xi}\operatorname{sen}\gamma + \dot{\eta}\cos\gamma.^{1}$$

Se confirma que

$$\dot{x}^2 + \dot{y}^2 = \dot{\xi}^2 + \dot{\eta}^2.$$

Entonces, la expresión $\sqrt{\dot{x}^2 + \dot{y}^2}$ *tiene el mismo valor en todos los sistemas de coordenadas.* Esta propiedad de invariancia es, por supuesto, obvia a partir de la interpretación de esta cantidad como la medida de cambio respecto a t de la longitud a lo largo de la curva, ds/dt. El lector puede verificar mediante un cálculo sencillo que también la expresión $\kappa = (\dot{x}\ddot{y} - \dot{y}\ddot{x})(\dot{x}^2 + \dot{y}^2)^{-3/2}$ para la *curvatura es invariante.* (Esto, por supuesto, se obtiene también directamente a partir del hecho de que los ángulos que la tangente forma con los ejes ξ y x, respectivamente, difieren sólo en el valor constante γ, de manera que $\kappa = d\alpha/ds$ no puede cambiar.)

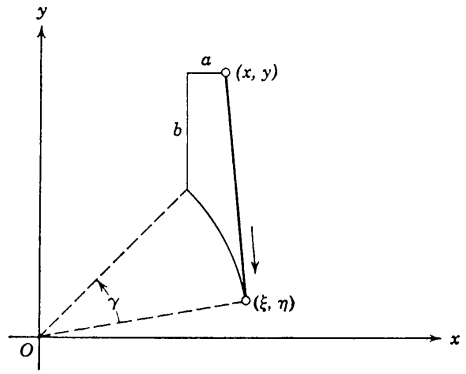


Figura 4.25 Desplazamiento del punto P de la posición (x, y) a la posición (ξ, η).

Las ecuaciones (18) que relacionan las coordenadas x, y con las coordenadas ξ, η son a menudo interpretadas en una forma diferente que corresponde a un desplazamiento. En esta interpretación, los *punto P* son des-

[1] En algunas aplicaciones físicas, donde t representa el tiempo, la posición relativa de los dos sistemas de coordenadas también depende de t. Supóngase que x, y representan las coordenadas de una partícula en un sistema coordenado en reposo, en tanto que ξ, η son las coordenadas de la misma partícula respecto a un sistema móvil, por ejemplo, uno sujeto a la Tierra. Las funciones $x(t)$, $y(t)$ describen la trayectoria de la partícula según la ve un observador en reposo, en tanto que $\xi(t)$, $\eta(t)$ describen la trayectoria vista por un observador móvil. Las fórmulas que relacionan $\dot{x}$, $\dot{y}$ con $\dot{\xi}$, $\dot{\eta}$ tienen entonces que incluir también los términos que provienen de derivar a, b y γ.

plazados en vez de serlo los ejes coordenados (Fig. 4.25). Solamente se usa un sistema de coordenadas. El punto con coordenadas (x, y) en ese sistema es transformado en el punto con coordenadas (ξ, η) del mismo sistema. La *invariancia de la longitud* o de la *curvatura* de una curva significan ahora que esas cantidades no cambian cuando la curva completa sufre un *movimiento rígido.*

*j. Movimiento uniforme en la teoría especial de la relatividad

Como se señaló en la p. 255, existen profundas analogías entre las funciones trigonométricas y las hiperbólicas, que tienen su contraparte geométrica en la correspondencia existente entre las propiedades de las elipses y de las hipérbolas. La relación se hará clara cuando se esté en condiciones de definir las funciones trigonométricas para un argumento imaginario y verificar que $\cos(it) = \cosh t$, $\operatorname{sen}(it) = i \operatorname{senh} t$ (sección 7.7*a*). Como una aplicación de esta analogía, considérense las "rotaciones hiperbólicas" del plano, que pueden ser identificadas con las *transformaciones de Lorentz* de una recta en la teoría especial de la relatividad de Einstein.

Se encontró en (18*a*) de p. 379 que una rotación de los ejes de coordenadas en un ángulo γ, en la que el origen queda fijo, puede describirse mediante las ecuaciones

$$(18b) \qquad \xi = x\cos\gamma + y\operatorname{sen}\gamma, \qquad \eta = -x\operatorname{sen}\gamma + y\cos\gamma,$$

que relacionan las coordenadas x, y de un punto P en el primer sistema con sus coordenadas ξ, η en el segundo sistema. La distancia de P al origen está dada por la misma expresión en ambos sistemas:

$$OP = \sqrt{x^2 + y^2} = \sqrt{\xi^2 + \eta^2}.$$

Esto surge también inmediatamente a partir de las ecuaciones de transformación si se hace uso de la identidad $\cos^2\gamma + \operatorname{sen}^2\gamma = 1$.

Se considera ahora la transformación análoga con coeficientes que son funciones hiperbólicas en lugar de funciones trigonométricas:

$$(19) \qquad \xi = x\cosh\alpha - t\operatorname{senh}\alpha, \qquad \tau = -x\operatorname{senh}\alpha + t\cosh\alpha.$$

Estas fórmulas pueden obtenerse a partir de las fórmulas (18*b*) para rotaciones, tomando para el ángulo de rotación γ y las coordenadas y y η cantidades imaginarias puras:

$$\gamma = i\alpha, \qquad y = it, \qquad \eta = i\tau.$$

Se observa que para un valor real de α (que significaría un ángulo imaginario de rotación γ en la interpretación original) las fórmulas (19) definen ξ y τ como funciones lineales reales de x y t. Estas funciones tienen la propiedad especial de que

$$\begin{aligned}\xi^2 - \tau^2 &= (x\cosh\alpha - t\operatorname{senh}\alpha)^2 - (-x\operatorname{senh}\alpha + t\cosh\alpha)^2 \\ &= x^2 - t^2,\end{aligned}$$

como una consecuencia de la identidad $\cosh^2\alpha - \operatorname{senh}^2\alpha = 1$. (Esto surge, por supuesto, también a partir de la observación de que $x^2 - t^2 = x^2 + y^2$ es el cuadrado de la distancia al origen en el plano x, y.) Ahora interprétese t como el tiempo y x como una coordenada espacial que describe la posición de un punto en un espacio unidimensional, esto es, en una línea recta. Cualquier evento ocurre en un cierto punto y en un cierto tiempo. Estas dos partes de la información las proporcionan los dos números x y t que dan respectivamente la distancia x (con su signo) del punto al origen 0 y el tiempo t que ha transcurrido a partir del instante $t = 0$. En la teoría de la relatividad se adopta el punto de vista de que los valores medidos de esta distancia y los tiempos transcurridos dependen del sistema de referencia utilizado por el observador, esto es, del sistema de coordenadas especial en el continuo espacio-tiempo. Las cantidades ξ, τ obtenidas a partir de las fórmulas (19) describirán el mismo evento en un sistema de referencia diferente en el cual las distancias e intervalos de tiempo pueden tener valores también diferentes. La cantidad que permanece inalterable al pasar de un sistema de referencia a otro (lo que se conoce como "transformación de Lorentz") es

$$\sqrt{x^2 - t^2} = \sqrt{\xi^2 - \tau^2},$$

la "distancia espacio-tiempo" del evento al origen. Para un observador que use el segundo sistema, la cantidad ξ es la distancia espacial medida a partir del origen $\xi = 0$. Ese origen es un punto para el cual

$$x\cosh\alpha - t\operatorname{senh}\alpha = 0,$$

o bien $x/t = \tanh\alpha$. Así, el origen del segundo sistema es un punto que en el primer sistema parece moverse con velocidad uniforme $v = dx/dt = \tanh\alpha$ relativa al origen de este sistema. Por lo tanto, la transformación de Lorentz relaciona los valores de distancias y tiempos según los registran observadores en los dos sistemas que se mueven con velocidad relativa constante v. Aquí

$$v = \frac{\operatorname{senh}\alpha}{\cosh\alpha} = \frac{e^{\alpha} - e^{-\alpha}}{e^{\alpha} + e^{-\alpha}}$$

está comprendida necesariamente entre -1 y $+1$, de manera que las velocidades relativas de los dos sistemas están restringidas a valores numéricamente menores que 1. El valor 1 representa aquí, para una elección adecuada de unidades, la velocidad c de la luz, que no puede ser excedida por v.

Para una u constante la ecuación $x = ut$ corresponde a un punto que en el primer sistema se mueve con velocidad u, partiendo de $x = 0$ en el tiempo $t = 0$. En el segundo sistema el mismo punto tendrá la velocidad

$$\omega = \frac{d\xi}{d\tau} = \left(\frac{d\xi}{dt}\right) \Big/ \left(\frac{d\tau}{dt}\right) = \frac{u - \tanh\alpha}{1 - u\tanh\alpha} = \frac{u - v}{1 - uv}.$$

Este resultado, válido en la teoría especial de la relatividad de Einstein, difiere del que se obtendría en la cinemática clásica, donde la velocidad ω de un punto con respecto a un sistema que se mueve con velocidad v estaría dada simplemente por $\omega = u - v$. La fórmula relativista muestra que $\omega = u$ cuando $u = +1$ o -1; esto corresponde al hecho, sugerido por el famoso experimento de Michelson-Morley, de que la velocidad de la luz es la misma para observadores que se mueven con velocidades diferentes.

k. Integrales que expresan áreas dentro de curvas cerradas

En el capítulo 2 el concepto de integral fue motivado por medio de referencia al "área bajo una curva", esto es, el área de una franja de forma especial. Esta reducción al caso especial de áreas bajo curvas no es completamente satisfactoria, pues las áreas que realmente se encuentran con más frecuencia son las de dominios limitados por curvas cerradas C, los que son de una forma más general que las franjas cuyas áreas pueden representarse mediante integrales de la forma $\int_a^b f(x)\,dx$.

La fórmula fundamental

Se derivará ahora una elegante representación integral de validez general para el área limitada por una curva cerrada C que está dada en forma paramétrica, descomponiendo simplemente el área en cuestión en áreas de franjas especiales. Esta representación será independiente de la representación paramétrica, así como también del sistema de coordenadas. Además, expresará el *área orientada* dentro de la curva de acuerdo con el sentido asignado a la frontera C; esto es, asignará al área dentro de una curva cerrada simple C el signo negativo o el positivo según que la frontera esté orientada en el sentido del reloj o contra reloj.

Supóngase que la curva simple C cerrada y orientada se da mediante $x = x(t)$, $y = y(t)$, donde t varía en el intervalo $\alpha \leq t \leq \beta$ y el sentido de t creciente determina el sentido de C. Supóngase también que x e y son funciones continuas de t (con el mismo valor en $t = \alpha$ que en $t = \beta$) y que sus primeras derivadas $\dot{x}$ e $\dot{y}$ son continuas, con la posible excepción de un número finito de discontinuidades de salto en el caso de que C tenga esquinas. Con estas hipótesis se demostrará la fórmula fundamental

$$(20) \qquad A = -\int_\alpha^\beta y\dot{x}\,dt = \int_\alpha^\beta x\dot{y}\,dt = \frac{1}{2}\int_\alpha^\beta (x\dot{y} - y\dot{x})\,dt$$

para el área orientada A encerrada por C.

Que las tres representaciones integrales en la fórmula son equivalentes se sigue de inmediato integrando por partes la primera de ellas y usando las condiciones de periodicidad $x(\alpha) = x(\beta)$, $y(\alpha) = y(\beta)$. La tercera representación, más simétrica, es justamente la media aritmética de las dos primeras.

La expresión (20) no depende de la posición del sistema de coordenadas en el plano. En efecto, la expresión simétrica

$$A = \frac{1}{2}\int_\alpha^\beta (x\dot{y} - y\dot{x})\,dt$$

muestra claramente que el valor de A es independiente de la elección del sistema de coordenadas. Como se vio en la p. 379, un cambio de coordenadas de un sistema xy a un sistema $\xi\eta$ se realiza mediante una substitución de la forma

$$x = \xi \cos\gamma - \eta \operatorname{sen}\gamma + a,$$
$$y = \xi \operatorname{sen}\gamma + \eta \cos\gamma + b,$$

con a, b y γ constantes. Derivando estas fórmulas respecto a t se llega a

$$\dot{x} = \dot{\xi}\cos\gamma - \dot{\eta}\operatorname{sen}\gamma, \qquad \dot{y} = \dot{\xi}\operatorname{sen}\gamma + \dot{\eta}\cos\gamma$$

y, en consecuencia,

$$x\dot{y} - y\dot{x} = \xi\dot{\eta} - \eta\dot{\xi} + a\dot{y} - b\dot{x}.$$

Por tanto, la expresión $x\dot{y} - y\dot{x}$ es invariante bajo rotaciones alrededor del origen (esto es, cuando $a = b = 0$). Aún cuando a y b no se anulen, el valor de la integral para A no resulta afectado ya que

$$\int_\alpha^\beta (a\dot{y} - b\dot{x})\,dt = (ay - bx)\Big|_\alpha^\beta = 0$$

para la curva cerrada C.

Prueba de la fórmula fundamental (20). *Integrales de línea sobre arcos simples.* La fórmula básica (20) se prueba con algunos pasos sencillos.

Primero, sea C un arco simple *orientado*, con punto inicial P_0 y punto final P_1. Sea $x = x(t)$, $y = y(t)$ cualquier representación paramétrica de C con P_0 y P_1 correspondiendo respectivamente a $t = t_0$, y $t = t_1$. (Aquí t_0 puede ser mayor o menor que t_1.) Entonces la integral

$$A = -\int_{t_0}^{t_1} y\frac{dx}{dt}\,dt$$

depende solamente de C y no de la representación paramétrica particular. Esta es una consecuencia obvia de la regla de substitución. Si se introduce un nuevo parámetro τ mediante la función monótona $\tau = \chi(t)$, donde $\tau_0 = \chi(t_0)$, $\tau_1 = \chi(t_1)$, la integral correspondiente es [1]

$$-\int_{\tau_0}^{\tau_1} y\frac{dx}{d\tau}\,d\tau = -\int_{t_0}^{t_1} y\frac{dx}{d\tau}\frac{d\tau}{dt}\,dt = -\int_{t_0}^{t_1} y\frac{dx}{dt}\,dt = A.$$

Queda, por consiguiente, justificado el omitir en la expresión para la integral A la referencia a cualquier parámetro especial t, pudiendo escribirse simplemente

$$A = A_C = -\int_C y\,dx.$$

Aquí, A_C para un arco simple y orientado C debe calcularse refiriendo el arco a un parámetro t, usando $dx = (dx/dt)\,dt$ y tomando como límites para la integración sobre t los valores paramétricos correspondientes a los puntos extremos de C en el orden determinado por la orientación del arco.*

Si C' es el arco que se obtiene de C cambiando la orientación de éste, esto es, el arco con punto inicial P_1 y punto final P_0, usando la misma representación paramétrica para C' se tiene

[1] Suponemos continuas no sólo $x(t)$ e $y(t)$ sino también $\tau(t)$ y sus derivadas, con la posible excepción, para estas últimas, de un número finito de discontinuidades de salto.

* La integral $\int_C y\,dx$ es un ejemplo de las *integrales de línea* generales $\int_C p\,dx + q\,dy$, que serán discutidas en el Volumen II.

$$A_{C'} = -\int_{t_1}^{t_0} y\frac{dx}{dt}dt = +\int_{t_0}^{t_1} y\frac{dx}{dt}dt = -A_C.$$

Por lo tanto, cambiando la orientación del arco se cambia el signo de la integral A_C.

Si el arco simple y orientado C se divide en subarcos orientados C_1, C_2, ..., C_n, cada uno con la misma orientación que C, se tendrá obviamente

$$A_C = A_{C_1} + A_{C_2} + \cdots + A_{C_n}.$$

Pues, en efecto, en una representación paramétrica de C en la que, por ejemplo, el sentido de C sea el de t creciente, esta descomposición corresponde a una subdivisión del intervalo paramétrico $t_0 \leq t \leq t_n$ de C en subintervalos $t_0 \leq t \leq t_1$, $t_1 \leq t \leq t_2$, ..., $t_{n-1} \leq t \leq t_n$ correspondientes a C_1, ..., C_n. El resultado se obtiene entonces por la aditividad de las integrales.

La aditividad de las integrales A_C facilita mucho el cálculo de A_C en casos en que C consiste de varios arcos suaves C_1, C_2..., cada uno con su propia representación paramétrica. No se necesita construir artificialmente una representación paramétrica única para la curva C completa, sino calcular cada A_{C_i} separadamente a partir de su representación paramétrica y luego efectuar la suma. Además, los A_{C_i} pueden sumarse en cualquier orden; sólo se necesita asegurarse de que todos los C_i tienen la misma orientación que C.

La integral de línea fundamental para curvas cerradas

Se puede ahora definir A_C para cualquier curva simple C, cerrada y orientada, descomponiendo C en arcos simples C_1, ..., C_n con orientaciones que coincidan con la de C y formando la suma de los A_{C_i}.[1] Si la curva C completa tiene la representación paramétrica $x = x(t)$, $y = y(t)$, para $\alpha \leq t \leq \beta$, donde el sentido de las t crecientes da la orientación de C y donde $t = \alpha$ y $t = \beta$ corresponden al mismo punto, entonces A_C está dada nuevamente por

$$A_C = -\int_{\alpha}^{\beta} y\frac{dx}{dt}dt.$$

De la misma manera se puede definir A_C para curvas orientadas C no simples mediante la descomposición en arcos simples orientados, aún cuando C consista de varias partes ajenas pero con la condición de que cada porción de C tenga un sentido definido.

[1] Que el valor de A_C así obtenido no depende de la manera particular en que C es dividida en arcos simples, se demuestra fácilmente: primero, la propiedad aditiva de A para arcos simples muestra que el refinamiento de una subdivisión dada por la adición de puntos de división no cambia el valor resultante de A_0; además, dos subdivisiones cualesquiera pueden reemplazarse por aquélla que es un refinamiento de ambas, sin que cambie el valor de A_C.

La integral fundamental como un área

Regresamos ahora al punto principal; esto es, identificamos las expresiones A_C para una curva cerrada con la intuitiva cantidad geométrica que representa el área orientada que C encierra.

Consideramos primero un dominio G limitado en su parte superior por un arco C_1: $y = g(x)$ para $a \leq x \leq b$; en su parte inferior por un arco C_3: $y = f(x)$ para $a \leq x \leq b$; y a los lados por los segmentos de recta C_2,

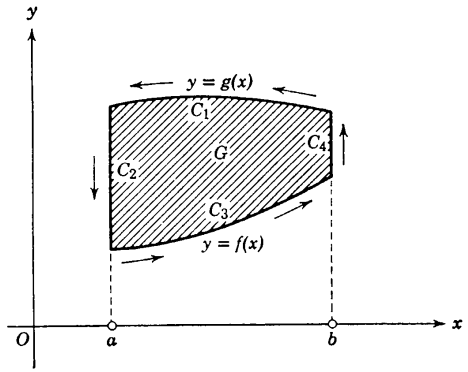


Figura 4.25(*a*) Area de una "celda".

C_4 dados por $x = a$ y $x = b$ (Fig. 4.25*a*). Aquí se permite que C_2 y C_4 se reduzcan a puntos. Si damos a C la orientación contra reloj, el arco C_1 será recorrido en el sentido de las x decrecientes y el arco C_3 en el de las x crecientes. Al formar A_C como la suma de las cuatro A_{C_i}, las porciones C_2 y C_4 a lo largo de las cuales x es constante no aportan contribución alguna pues $dx/dt = 0$. Usando x como parámetro sobre los arcos C_1 y C_3 encontramos

$$A_C = A_{C_1} + A_{C_3} = -\int_b^a g(x)\,dx - \int_a^b f(x)\,dx$$
$$= \int_a^b g(x)\,dx - \int_a^b f(x)\,dx.$$

Esto es, claramente, el área positiva del dominio G en el caso en que G quede completamente por encima del eje x, siendo igual a la diferencia entre las áreas que están respectivamente por debajo de las curvas C_1 y C_3. Podemos siempre garantizar que G queda por encima del eje x reemplazando simplemente y por $y + c$, donde c es una constante adecuada, esto es, mediante una traslación en la dirección y. Esto no altera las áreas y tampoco afecta el valor $A_C = -\int_C y\,dx$ para una curva cerrada C,

como vimos antes. Por lo tanto, para dominios G del tipo descripto que tengan fronteras C intersectadas a lo más en dos puntos por paralelas al eje y, la integral A_C representa el área, tomada positiva si C está orientada contra reloj y negativa para la orientación opuesta. El mismo resultado se obtiene para áreas limitadas por curvas C que sean intersectadas a lo más en dos puntos por paralelas al eje x. Lo único que se necesita es escribir A_C en la forma $\int_C x\,dy$ e intercambiar x e y en el argumento anterior. A los dominios G pertenecientes a uno de estos dos tipos los llamaremos "celdas". Hablaremos de "celdas orientadas" cuando las curvas de frontera correspondientes tengan una orientación o la otra.

Consideremos ahora un dominio G con la frontera orientada C y que esté compuesto por cierto número de celdas simples $G_1, G_2, \ldots, G_n$ con fronteras $C_1, \ldots, C_n$ respectivamente; y se supone que todas estas celdas tienen la misma orientación, digamos contra reloj. Entonces, como se indica en la Fig. 4.26, las partes de las fronteras que son comunes a dos celdas adyacentes se recorren en sentidos diferentes al considerar sucesivamente dichas partes como arcos de frontera de una celda y de la celda vecina Por consiguiente, si sumamos las integrales $A_{C_i} = -\int y\,dx$ que se tienen para las diferentes celdas las contribuciones de todas las fronteras interiores de las celdas se anulan, y así obtenemos

$$A = \sum_{i=1}^{n} A_{C_i} = \sum_{i=1}^{n} \left(-\int_{C_i} y\,dx \right) = -\int_C y\,dx = A_C,$$

donde A es el área orientada del dominio total G.

Así, las fórmulas (20) para el área A de un dominio orientado G interior a una curva cerrada quedan demostradas para todos los dominios que puedan descomponerse en celdas simples, por ejemplo, trazando paralelas a los ejes coordenados.

Para todos los dominios que encontraremos, esta hipótesis se satisfará obviamente; como ocurre, por ejemplo, para dominios poligonales.

Comentarios adicionales

Finalmente, podría agregarse que de la misma manera se demuestra la validez de la fórmula para el área aún en el caso de dominios múltiplemente conexos, tales como los dominios en forma de anillo, que puedan descomponerse en un número finito de celdas simples. En estos casos todas las curvas de frontera tienen que ser recorridas sistemáticamente en un sentido tal que el interior de G esté siempre en el lado "izquierdo" o siempre en el lado "derecho".

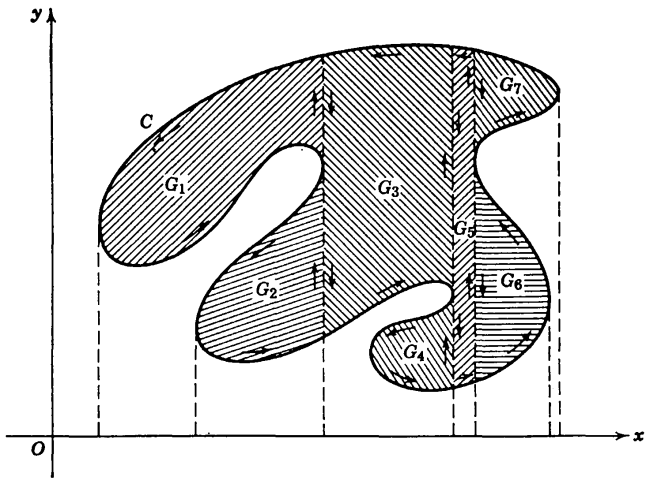


Figura 4.26 Descomposición de un dominio orientado en celdas orientadas.

Las fórmulas para A conservan su validez aún cuando C no sea una curva simple sino una que se intersecta a sí misma dividiendo el plano en más de dos regiones. En este caso podemos considerar la fórmula como una guía para interpretar adecuadamente el área como una combinación aditiva de las áreas orientadas que corresponden a las diferentes partes conexas del plano limitadas por C. Este asunto se discutirá en el Apéndice II de este capítulo.

Ejemplos. Como un ejemplo, podemos encontrar el *área encerrada por la elipse* $x^2/a^2 + y^2/b^2 = 1$. Usando la orientación contra reloj para la elipse y a partir de la representación paramétrica

$$x = a\cos t, \qquad y = b\,\text{sen}\,t, \qquad \text{para} \quad 0 \le t \le 2\pi,$$

encontramos que

$$A = \frac{1}{2}\int_0^{2\pi} (x\dot{y} - y\dot{x})\,dt = \frac{1}{2}\int_0^{2\pi} ab\,dt = \pi ab.$$

Area en coordenadas polares. Para expresar el área en coordenadas polares r y θ consideremos primero el área A de la región limitada por un segmento de curva $r = f(\theta)$ y los radios $\theta = \alpha$ y $\theta = \beta$. Suponemos que $\alpha < \beta$ y que θ puede usarse como parámetro a lo largo de la curva (esto es, que puntos diferentes tienen ángulos polares diferentes). Usamos para A la expresión

$$A = \frac{1}{2}\int (x\,dy - y\,dx) = \frac{1}{2}\int (x\dot{y} - y\dot{x})\,dt.$$

la cual tiene entonces que ser extendida sobre la parte curva de la frontera y sobre los dos radios. Sobre los radios $\theta = \alpha$ y $\theta = \beta$ podemos usar r como parámetro y, a partir de $x = r\cos\theta$, $y = r\,\text{sen}\,\theta$ y $\theta =$ constante, encontrar que $\dot{x} = \cos\theta$, $\dot{y} = \text{sen}\,\theta$, de manera que $x\dot{y} - y\dot{x} = 0$. Sobre la parte curva usamos θ como parámetro. Entonces,

$$\dot{x} = \frac{dr}{d\theta}\cos\theta - r\,\text{sen}\,\theta, \qquad \dot{y} = \frac{dr}{d\theta}\,\text{sen}\,\theta + r\cos\theta,$$

y así $x\dot{y} - y\dot{x} = r^2$. En consecuencia,

$$A = \frac{1}{2}\int_\alpha^\beta r^2\,d\theta = \frac{1}{2}\int_\alpha^\beta f^2(\theta)\,d\theta. \tag{21}$$

Para una curva cerrada simple C que contenga el origen en su interior y sea intersectada por cada rayo que parta del origen en exactamente un punto, podemos usar θ como parámetro, con $0 \leq \theta \leq 2\pi$, y encontrar así para el área encerrada

$$A = \frac{1}{2}\int_0^{2\pi} r^2\,d\theta. \tag{22}$$

La fórmula (21) para el área en coordenadas polares puede también derivarse directamente a partir de la definición de integral. Para ello se divide el dominio en sectores trazando radios desde el origen (Fig. 4.27). Cada sector queda descripto por las desigualdades

$$\theta_{i-1} < \theta < \theta_i, \quad 0 < r < f(\theta).$$

Obviamente, el área del sector queda entre las áreas de los sectores circulares inscripto y circunscripto; el área de un sector del dominio es entonces igual a $\frac{1}{2}r^2(\theta_i - \theta_{i-1})$, donde r está comprendido entre el valor más grande y el más pequeño de $f(\theta)$ para el intervalo $\theta_{i-1} < \theta < \theta_i$. Conforme se refina la subdivisión la suma de las áreas de los sectores del dominio evidentemente converge a la integral $\frac{1}{2}\int_\alpha^\beta r^2\,d\theta$.

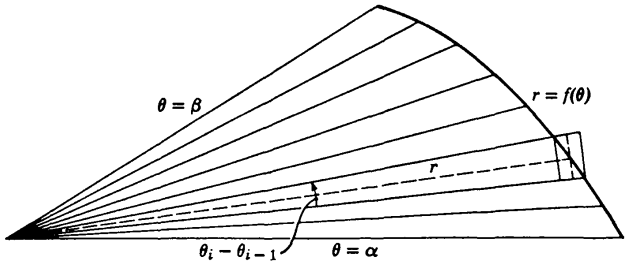


Figura 4.27 Area en coordenadas polares.

Area en un lemniscato

Como un ejemplo de la Ec. (21), consideremos el área limitada por una espira de una leminscata. La ecuación de la lemniscata (cfr. pp. 125-126) es $r^2 = 2a^2\cos 2\theta$; y una espira es obtenida haciendo variar θ de $-\pi/4$ a $+\pi/4$. Esto da la expresión

$$a^2\int_{-\pi/4}^{\pi/4}\cos 2\theta\,d\theta = a^2$$

para el área. Por supuesto, la otra espira tiene un área del mismo valor absoluto pero de signo negativo.

Area limitada por un hipérbola

Consideremos ahora el área de un sector limitado por la hipérbola $x^2 - y^2 = 1$, la cual hemos calculado ya en la p. 255 de una manera más bien complicada (véase Fig. 3.12). Para la hipérbola (más bien, para su rama derecha) se tiene la representación paramétrica $x = \cosh t$, $y = \text{senh}\,t$. Para el doble del área limitada por la hipérbola y los radios que van a los puntos con parámetros 0 y t se encuentra el valor

$$2A = \int_0^t (x\dot{y} - y\dot{x})\,d\tau = \int_0^t (\cosh^2\tau - \text{senh}^2\tau)\,d\tau$$
$$= \int_0^t d\tau = t.$$

(Nuevamente, no hay contribución a la integral por parte de los radios.)

l. Centro de masa y momento de una curva

Veamos ahora algunas ideas que surgen en la mecánica. Consideramos un sistema de n partículas en un plano cuyas masas son $m_1, m_2, \ldots, m_n$ y las respectivas ordenadas $y_1, y_2, \ldots, y_n$. Entonces llamamos a

$$T = \sum_{\nu=1}^{n} m_\nu y_\nu = m_1y_1 + m_2y_2 + \cdots + m_ny_n$$

el *momento del sistema de partículas con respecto al eje x*. La expresión $\eta = T/M$, donde M denota la masa total $m_1 + m_2 + \cdots + m_n$ del sistema, define la *altura del centro de masa* del sistema de partículas sobre el eje x, o sea su ordenada. Es, justamente, la *media pesada* de $y_1, y_2, \ldots, y_n$ usando los "factores peso" $m_1, m_2, \ldots, m_n$ (ver. p. 164). Por lo tanto, η es la *altura promedio* de las masas. Análogamente definimos el momento con respecto al eje y y la abscisa del centro de masa.

Podemos extender ahora fácilmente estas definiciones de momento al caso de una curva a lo largo de la cual hay una masa distribuida unifor-

memente, y definir así las coordenadas ξ y η del centro de masa de tal curva. (La hipótesis de una densidad constante, digamos μ, a lo largo de la curva no es esencial: cualquier distribución continua podría discutirse igualmente.)

En un procedimiento típico en la mecánica se comienza con un sistema de un número finito n de partículas y se pasa luego a un límite para $n \to \infty$. Para hacer esto, introducimos la longitud de arco s como un parámetro sobre la curva y subdividimos ésta en arcos de longitudes $\Delta s_1, \Delta s_2, \ldots, \Delta s_n$, usando $(n - 1)$ puntos de división. Representamos la masa $\mu\,\Delta s_i$ de cada arco Δs_i como si estuviese concentrada en un punto arbitrario del arco, por ejemplo, el punto con ordenada y_i.

Por definición, el momento con respecto al eje x de este sistema de partículas es

$$T = \mu \sum y_i\,\Delta s_i.$$

Si ahora la mayor de las cantidades Δs_i tiende a cero, esta suma tiende a un límite dado por la integral

$$T = \mu \int_{s_0}^{s_1} y\,ds = \mu \int_{x_0}^{x_1} y\sqrt{1 + y'^2}\,dx, \tag{23}$$

la cual es, por consiguiente, aceptada en forma natural como la definición del momento de la curva con respecto al eje x. Puesto que la masa total de la curva es igual a su longitud multiplicada por μ, es decir,

$$\mu \int_{s_0}^{s_1} ds = \mu(s_1 - s_0),$$

llegamos inmediatamente a las siguientes expresiones para las coordenadas del centro de masa de la curva:

$$\xi = \frac{\displaystyle\int_{s_0}^{s_1} x\,ds}{s_1 - s_0}, \qquad \eta = \frac{\displaystyle\int_{s_0}^{s_1} y\,ds}{s_1 - s_0}. \tag{24}$$

Estos asertos son realmente *definiciones* del momento y del centro de masa de una curva; pero son extensiones tan directas del caso más simple de un número finito de partículas, que uno naturalmente espera —como es, precisamente, el caso— que cualquier afirmación en mecánica sobre el centro de masa o el momento de un sistema de partículas sea válida también para distribuciones continuas de masa a lo largo de curvas.

m. Area y volumen de una superficie de revolución

Regla de Guldin

Si se hace girar una curva $y = f(x)$, para la cual $f(x) \geq 0$, alrededor del eje x, se describe una superficie llamada *superficie de revolución*. El

área de esta superficie, cuyas abscisas suponemos comprendidas entre las cotas x_0 y $x_1 > x_0$, se obtiene mediante una discusión análoga a la anterior. En efecto, si reemplazamos la curva por un polígono inscripto, en vez de una superficie curva tendremos una figura compuesta por un cierto número de conos truncados. La intuición sugiere que podríamos definir el área de la superficie de revolución como el límite de las áreas de estas superficies cónicas cuando la longitud del lado mayor del polígono inscripto tiende a cero. De la geometría elemental sabemos que el área de cada cono truncado es igual a la longitud de la generatriz multiplicada por el perímetro de la sección circular de radio medio. (Fig. 4.28.) Si sumamos estas expresiones y después realizamos el paso al límite obtenemos para el área la expresión

$$A = 2\pi \int_{s_0}^{s_1} y\,ds = 2\pi \int_{x_0}^{x_1} y\sqrt{1 + y'^2}\,dx = 2\pi\eta(s_1 - s_0). \tag{25}$$

Expresado en palabras, este resultado dice que el área de una superficie de revolución es igual a la longitud de la curva generadora multiplicada por la distancia recorrida por el centro de masa *(regla de Guldin)*.

Del mismo modo encontramos que el volumen interior a la superficie de revolución, limitado en los extremos por los planos $x = x_0$ y $x = x_1 > x_0$, está dado por la expresión

$$V = \pi \int_{x_0}^{x_1} y^2\,dx. \tag{26}$$

Esta fórmula se obtiene siguiendo la idea intuitiva de que el volumen en cuestión es el límite de los volúmenes de las figuras antes mencionadas consistentes en conos truncados. El resto de la demostración se deja al lector.

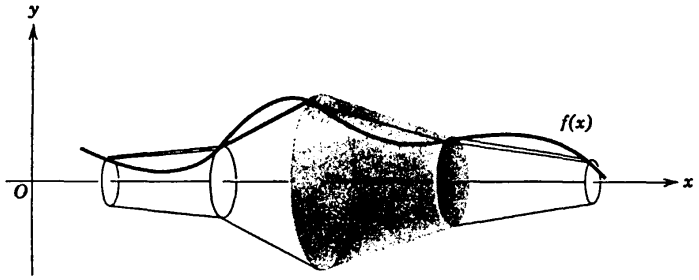


Figura 4.28 Area de una superficie de revolución.

n. Momento de inercia

En el estudio de la rotación de un objeto juegan un papel importante ciertas cantidades llamadas *momentos de inercia*. Estas serán mencionadas aquí brevemente.

Supongamos que una partícula de masa m a una distancia y del eje x gira uniformemente alrededor de ese eje con velocidad angular ω (esto es, en la unidad de tiempo gira un ángulo ω). La *energía cinética* de la partícula, expresada como la mitad del producto de la masa por el cuadrado de la velocidad, es obviamente

$$\frac{m}{2}(y\omega)^2:$$

Al coeficiente de $\frac{1}{2}\omega^2$, esto es, a la cantidad my^2 le llamamos el *momento de inercia de la partícula alrededor del eje x.*

Análogamente, si tenemos n partículas con masas $m_1, m_2, \ldots, m_n$ y ordenadas $y_1, y_2, \ldots, y_n$, llamamos a la expresión

$$T = \sum_i m_i y_i^2$$

el *momento de inercia del sistema de masas* alrededor del eje x. El momento de inercia es una cantidad que pertenece al propio sistema de masas, sin referencia a su estado de movimiento. Su importancia radica en el hecho de que en la rotación rígida del sistema alrededor de un eje, sin que cambie la distancia mutua entre pares de partículas, la energía cinética se obtiene multiplicando el momento de inercia alrededor de ese eje por la mitad del cuadrado de la velocidad angular. Así, el momento de inercia alrededor de un eje juega el mismo papel en la rotación alrededor de ese eje que la masa en el movimiento rectilíneo.

Supóngase que se tiene ahora una curva arbitraria $y = f(x)$ comprendida entre las abscisas x_0 y $x_1 > x_0$, a lo largo de la cual una masa está distribuida uniformemente con densidad unitaria. Para definir el momento de inercia de esta curva se procede justamente como en la sección anterior, llegándose a la siguiente expresión para el momento de inercia alrededor del eje x

$$T_x = \int_{s_2}^{s_1} y^2\,ds = \int_{s_0}^{x_1} y^2\sqrt{1+y'^2}\,dx. \tag{27}$$

Para el momento de inercia alrededor del eje y se tiene, en forma semejante,

$$T_y = \int_{s_0}^{s_1} x^2\,ds = \int_{x_0}^{x_1} x^2\sqrt{1+y'^2}\,dx. \tag{28}$$

4.2 Ejemplos

De la gran variedad de curvas planas escogemos unos pocos ejemplos típicos para ilustrar los conceptos discutidos.

a. La cicloide común

De las ecuaciones $x = a(t - \operatorname{sen} t)$, $y = a(1 - \cos t)$ (cfr. (1), p. 347) se obtiene $\dot{x} = a(1 - \cos t)$, $\dot{y} = a \operatorname{sen} t$, y para la longitud de arcos se tiene

$$s = \int_0^\alpha \sqrt{\dot{x}^2 + \dot{y}^2}\,dt = \int_0^\alpha \sqrt{2a^2(1-\cos t)}\,dt.$$

Puesto que $1 - \cos t = 2\operatorname{sen}^2 t/2$, el integrando es igual a $2a \operatorname{sen} t/2$ y, por consiguiente, con $0 \leq \alpha \leq 2\pi$, se obtiene

$$s = 2a\int_0^\alpha \operatorname{sen}(t/2)\,dt = -4a\cos\frac{t}{2}\Big|_0^\alpha = 4a\left(1 - \cos\frac{\alpha}{2}\right) = 8a\operatorname{sen}\frac{\alpha}{4}.$$

Si, en particular, se considera la longitud de arco entre dos cúspides sucesivas, debe tomarse $\alpha = 2\pi$ ya que el intervalo $0 \leq t \leq 2\pi$ de los valores del parámetro corresponde a una revolución del círculo giratorio. Así se obtiene el valor $8a$; esto es, la longitud de arco de la cicloide entre dos cúspides sucesivas es igual a *cuatro* veces el diámetro del círculo giratorio.

De manera análoga se calcula el área limitada por un arco de la cicloide y el eje x:

$$\begin{aligned} I &= \int_0^{2\pi} y\dot{x}\,dt = a^2\int_0^{2\pi}(1-\cos t)^2\,dt \\ &= a^2\int_0^{2\pi}(1 - 2\cos t + \cos^2 t)\,dt \\ &= a^2\left(t - 2\operatorname{sen} t + \frac{t}{2} + \frac{\operatorname{sen} 2t}{4}\right)\Big|_0^{2\pi} = 3a^2\pi. \end{aligned}$$

Esta área es, por consiguiente, tres veces el área del círculo giratorio.

Para el radio de curvatura $|\rho| = 1/|\kappa|$ se tiene, por la Ec. (15) de p. 373,

$$\rho = \frac{(\dot{x}^2+\dot{y}^2)^{3/2}}{\ddot{y}\dot{x} - \dot{y}\ddot{x}} = -2a\sqrt{2(1-\cos t)} = -4a\left|\operatorname{sen}\frac{t}{2}\right|.$$

En los puntos $t = 0$, $t = \pm 2\pi, \ldots$ esta expresión tiene el valor cero. Estos puntos son las cúspides; en ellos la cicloide intersecta el eje x en ángulo recto.

El área de la superficie de revolución que se forma al hacer girar un arco de la cicloide alrededor del eje x está dado, según nuestra fórmula (25), p. 392, por

$$\begin{aligned} A &= 2\pi\int_0^{8a} y\,ds = 2\pi\int_0^{2\pi} a(1-\cos t)\cdot 2a\operatorname{sen}\frac{t}{2}\,dt \\ &= 8a^2\pi\int_0^{2\pi}\operatorname{sen}^3\frac{t}{2}\,dt = 16a^2\pi\int_0^{\pi}\operatorname{sen}^3 u\,du \end{aligned}$$

$$= 16a^2\pi \int_0^{\pi} (1 - \cos^2 u) \operatorname{sen} u \, du.$$

La última integral puede evaluarse mediante la substitución $\cos u = v$; y se encuentra que

$$A = 16a^2\pi \left(-\cos u + \frac{1}{3}\cos^2 u \right) \Big|_0^{\pi} = \frac{64a^2\pi}{3}.$$

Como ejercicio, el lector puede calcular la altura por encima del eje x, η, del centro de masa de la cicloide, y también el momento de inercia T_x. Los resultados son

$$\eta = \frac{4}{3}a = \frac{A}{2\pi s} \quad \text{y} \quad T_x = \frac{256}{15}a^3.$$

b. La catenaria

La catenaria [1] es la curva definida por la ecuación $y = \cosh x$. La longitud de la catenaria entre las abscisas $x = a$ y $x = b$ es

$$s = \int_a^b \sqrt{1 + \operatorname{sen}^2 x}\, dx = \int_a^b \cosh x \, dx = \operatorname{senh} b - \operatorname{senh} a.$$

Para el área de la superficie de revolución obtenida haciendo girar la catenaria alrededor del eje x, la llamada superficie *catenoide*, se encuentra

$$A = 2\pi \int_a^b \cosh^2 x \, dx = 2\pi \int_a^b \frac{1 + \cosh 2x}{a}\, dx$$
$$= \pi(b - a + \tfrac{1}{2}\operatorname{senh} 2b - \tfrac{1}{2}\operatorname{senh} 2a).$$

A partir de aquí se obtiene a continuación la altura del centro de masa del arco de a a b:

$$\eta = \frac{A}{2\pi s} = \frac{b - a + \tfrac{1}{2}\operatorname{senh} 2b - \tfrac{1}{2}\operatorname{senh} 2a}{2(\operatorname{senh} b - \operatorname{senh} a)}.$$

Finalmente, para la curvatura se tiene

$$\kappa = \frac{y''}{(1 + y'^2)^{3/2}} = \frac{\cosh x}{\cosh^3 x} = \frac{1}{\cosh^2 x}.$$

[1] El nombre proviene del hecho de que una cadena suspendida de sus extremos adopta la forma de esta curva. Es curioso que la misma curva surja en una aplicación física completamente diferente. Una película de jabón, limitada por dos círculos en el espacio que están en planos paralelos y con sus centros sobre la misma perpendicular a los planos, tiene la forma de una superficie de revolución obtenida al hacer girar la catenaria alrededor del eje x.

c. La elipse y la lemniscata

Las longitudes de arco de estas dos curvas no pueden reducirse a funciones elementales, sino que pertenecen a la clase de "integrales elípticas" mencionadas en la p. 318.

Para la elipse $y = (b/a)\sqrt{a^2 - x^2}$ se obtiene

$$s = \frac{1}{a}\int \sqrt{\frac{a^4 - (a^2 - b^2)x^2}{a^2 - x^2}}\, dx = a\int \sqrt{\frac{1 - \eta^2\xi^2}{1 - \xi^2}}\, d\xi,$$

donde se ha hecho $x/a = \xi$, $1 - b^2/a^2 = \eta^2$. Mediante la substitución $\xi = \operatorname{sen}\phi$ esta integral puede escribirse en la forma

$$s = a\int \sqrt{1 - \eta^2 \operatorname{sen}^2 \phi}\, d\phi.$$

Aquí, para obtener el semiperímetro de la elipse debe hacerse que x recorra el intervalo de $-a$ a $+a$, que corresponde al intervalo

$$-1 \le \xi \le +1 \quad \text{o} \quad -\frac{\pi}{2} \le \phi \le +\frac{\pi}{2}.$$

Para la lemniscata, cuya ecuación en coordenadas polares r, t es $r^2 = 2a^2 \cos 2t$, de manera análoga se obtiene

$$s = \int \sqrt{r^2 + \dot{r}^2}\, dt = \int \sqrt{2a^2 \cos 2t + 2a^2 \frac{\operatorname{sen}^2 2t}{\cos 2t}}\, dt$$
$$= a\sqrt{2}\int \frac{dt}{\sqrt{\cos 2t}} = a\sqrt{2}\int \frac{dt}{\sqrt{1 - 2\operatorname{sen}^2 t}}$$

Si se introduce $u = \tan t$ como variable independiente en la última integral, se tiene

$$\operatorname{sen}^2 t = \frac{u^2}{1 + u^2}, \qquad dt = \frac{du}{1 + u^2},$$

y, en consecuencia,

$$s = a\sqrt{2}\int \frac{du}{\sqrt{1 - u^4}}.$$

En una espira completa de la lemniscata, u corre de -1 a $+1$, y la longitud de arco es entonces igual a

$$a\sqrt{2}\int_{-1}^{+1} \frac{du}{\sqrt{1 - u^4}}$$

una integral elíptica especial que jugó un papel importante en las investigaciones de Gauss.

4.3 Vectores en dos dimensiones

Para las discusiones de curvas planas y muchos otros tópicos en geometría, mecánica y física, la notación vectorial constituye una herramienta conveniente y casi indispensable. En este capítulo desarrollaremos y aplicaremos el concepto de vector en dos dimensiones, dejando para el Volumen II las extensiones a dimensiones superiores.

Explicación intuitiva

Muchos entes matemáticos y físicos quedan caracterizados por un solo número, llamado "escalar", ya que mide el ente con relación a una escala. Como ejemplos tenemos ángulos, longitudes, áreas, tiempos, masas y temperaturas. Sin embargo, existen otros entes para los cuales tal caracterización no es posible; por ejemplo, la forma de un triángulo, la posición de un punto en el espacio, la aceleración o la dirección del movimiento de una partícula y el estado de tensión de un cuerpo. Para identificar cada una de estas entidades se requieren varios números. Gradualmente se han ido desarrollando conceptos matemáticos que van más allá del continuo de números reales y que permiten representar tales entidades con un solo símbolo.[1] Los *vectores* en un plano son objetos matemáticos cuya descripción requiere dos componentes de información: una *longitud* y una *dirección*. A este tipo pertenecen, por ejemplo, la posición relativa y orden de dos puntos, la velocidad y la aceleración de una partícula, la fuerza que actúa sobre una partícula, etc.[2]

Geométrica o intuitivamente, un vector es esencialmente un segmento de recta *dirigido,* del plano o del espacio, caracterizado por su longitud o magnitud y por su dirección. Ordinariamente los vectores se indican con *flechas* de la longitud dada y apuntando en la dirección que se especifica. A menos que explícitamente se imponga una restricción, el *vector es "libre"*, esto es, la localización del extremo inicial del segmento dirigido (o flecha) no es parte inherente de las especificaciones que se dan para el vector.

A pesar de que conceptos físicos tales como velocidad, aceleración y fuerza son ejemplos primarios de vectores en las aplicaciones, definiremos los vectores geométricamente mediante 'traslaciones" o "desplazamientos paralelos".

El análisis vectorial comienza simplemente dando un nombre, "vector", a tales segmentos de recta dirigidos o desplazamientos paralelos. Sin embargo, la característica decisiva de los vectores no consiste en tener un nombre que los unifique sino en que (de la misma manera que los números complejos) pueden combinarse entre ellos o con escalares (esto es, números ordinarios) mediante un conjunto de reglas, dando origen al álgebra o análisis vectorial, de manera tal que tienen interpretaciones naturales en las diferentes aplicaciones, como, por ejemplo, la superposición de dos velocidades o el trabajo realizado en un desplazamiento venciendo una fuerza. En el lenguaje vectorial que surge intuitivamente, muchas relaciones matemáticas y físicas pueden expresarse en forma concisa y clara.

[1] Por supuesto, los números complejos $a + ib = z$ son entes de este tipo, siendo símbolos que representan pares de números reales a, b. Es, sin duda, conveniente algunas veces usar números complejos en vez de vectores.

[2] Los vctores son insuficientes para algunos propósitos; por ejemplo, para describir tensiones o curvaturas de espacios se usan entes más generales llamados "tensores".

a. Definición de vectores mediante traslaciones. Notaciones

El tipo más simple de transformación en el plano es una *traslación* o *desplazamiento paralelo.* Una traslación desplaza o transforma cualquier punto $P = (x, y)$ en el punto $P' = (x', y')$ de coordenadas

$$x' = x + a, \quad y' = y + b,$$

donde a y b son constantes. La traslación está completamente determinada por las constantes a y b, que llamaremos *componentes* de la traslación. Usaremos el término "vector" como otro nombre para "traslación". Usando negritas para denotar vectores o traslaciones, escribimos $\mathbf{R} = (a, b)$ para el vector con componentes a, b (Fig. 4.29).

Las componentes del vector $\mathbf{R}$ están determinadas por un par de puntos correspondientes $P = (x, y)$ y $P' = (x', y')$, ya que entonces

$$a = x' - x, \quad b = y' - y.$$

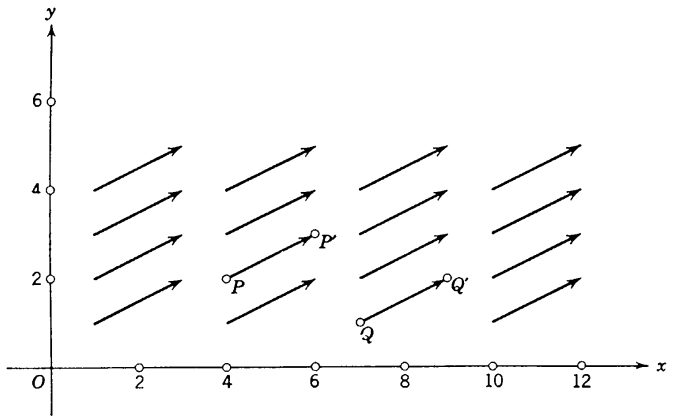


Figura 4.29 La traslación $x' = x + 2$, $y' = y + 1$ correspondiente al vector $\mathbf{R} = \overrightarrow{PP'} = \overrightarrow{QQ'} = (2, 1)$.

Claramente, para puntos P y P' cualesquiera siempre es posible encontrar una traslación $\mathbf{R}$ que transforme P en P'. A este vector lo denotamos por $\mathbf{R} = \overrightarrow{PP'}$. Cualquier par ordenado de puntos $P = (x, y)$, $P' = (x', y')$, esto es, cualquier segmento orientado determina así el vector $\mathbf{R} = \overrightarrow{PP'} = (x' - x,\ y' = y)$. Obsérvese que un segundo par de punto $Q = (\xi,\ \eta)$, $Q' = (\xi', \eta')$ define el mismo vector si $\xi' - \xi = x' - x$ y $\eta' - \eta = y' - y$; la misma traslación $\mathbf{R}$ transforma entonces P en P' y Q en Q'. Los vectores $\mathbf{R}$ están determinados por dos números, sus *componentes*, de la misma manera que los puntos de un plano lo están por sus dos coordenadas. La diferencia básica consiste en que un vector está representado geométricamente por un *par* de puntos. En la representación $\mathbf{R} = \overrightarrow{PP'}$ llamamos a P el *punto inicial* y a P' el *punto final*. Para un $\mathbf{R}$ dado, uno de los puntos, digamos, el punto inicial $P = (x, y)$, puede escogerse arbitrariamente; entonces el punto final $P' = (x', y')$ queda determinado en forma única por las relaciones $x' = x + a$, $y' = y + b$. Intercambiando los puntos inicial y final se obtiene el vector *opuesto* $\overrightarrow{P'P} = (-a, -b)$.

Si escogemos como punto inicial el origen $O = (0, 0)$, podemos asociar unívocamente un vector $\mathbf{R}$ con todo punto $Q = (x, y)$ tomando simplemente $\mathbf{R} = \overrightarrow{OQ}$. El vector $\mathbf{R}$ con punto inicial fijo O se llama entonces *vector de posición* de Q. Las componnentes de la posición del punto Q son simplemente las coordenadas x, y de ese punto.

El vector $\mathbf{R}$ de componentes $a = 0$, $b = 0$ se llama *vector nulo* y se denota por $\mathbf{O}$. Corresponde a una traslación que deja fijo cualquier punto:

$$\mathbf{O} = (0, 0) = \overrightarrow{PP}.$$

La distancia r entre dos puntos $P = (x, y)$ y $P' = (x', y')$ después sólo del vector $\mathbf{R} = (a, b) = \overrightarrow{PP'}$, puesto que

$$r = \sqrt{(x' - x)^2 + (y' - y)^2} = \sqrt{a^2 + b^2}.$$

A esta distancia le llamamos la *longitud del vector* $\mathbf{R}$, y escribimos $r = \mathbf{R}$. La longitud de $\mathbf{R}$ es siempre un número positivo a menos que $\mathbf{R} = \mathbf{O}$ (ver Fig. 4.30).

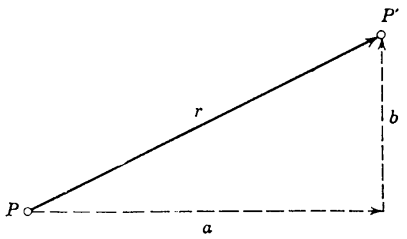


Figura 4.30 Componentes a, b y longitud r de un vector $\mathbf{R} = \overrightarrow{PP'}$.

Definimos el *producto* de un vector $\mathbf{R} = (a, b)$ por un número o "escalar" λ como el vector

$$\mathbf{R}^* = \lambda\mathbf{R} = (\lambda a, \lambda b).$$

Con $\lambda = -1$ tenemos en $\mathbf{R}^* = (-a, -b)$ el vector *opuesto* a $\mathbf{R}$ (Fig. 4.31).

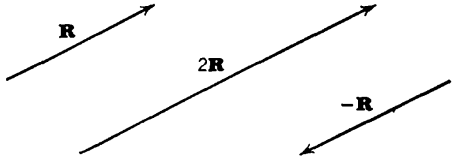


Figura 4.31 Múltiplos escalares de un vector R.

Si $\mathbf{R} = \overrightarrow{PP'} = (a, b)$, con $P = (x, y)$, $P' = (x', y')$, podemos representar $\mathbf{R}^* = \lambda\mathbf{R}$ como $\overrightarrow{PP''}$, donde $P'' = (x'', y'') = (x + \lambda a, y + \lambda b)$ (ver Fig. 4.32). Para $a = b = 0$ se tiene, por supuesto, $P'' = P' = P$. Para a y b no simultáneamente nulos, el punto $P'' = (x'', y'') = (x + \lambda a, y + \lambda b)$ recorre completamente, para λ variable, la recta

$$x''b - y''a = xb - ya.$$

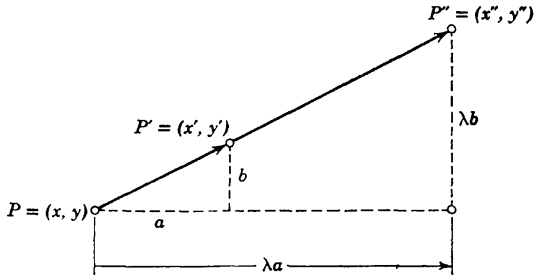


Figura 4.32 La relación vectorial $\mathbf{R}^* = \overrightarrow{PP''} = \lambda\overrightarrow{PP'}$ para $\lambda = \frac{3}{2}$.

El valor $\lambda = 0$ da $P'' = P$, en tanto que $\lambda = 1$ da $P'' = P'$. Así, P'' está sobre la recta que pasa por P y P'. Para $\lambda > 0$ los puntos P'' y P' están sobre el *mismo lado* de P; para $\lambda < 0$, están en *lados opuestos*.

Los dos vectores $\mathbf{R} = (a, b)$ y $\mathbf{R}^* = (a^*, b^*)$ se dice que tienen el *mismo sentido* si $\mathbf{R}^* = \lambda\mathbf{R}$ con λ positivo, y sentidos opuestos si $\lambda < 0$. Si $\mathbf{R} = \mathbf{O}$, esto significa que también $\mathbf{R}^* = \mathbf{O}$. Si $\mathbf{R} \neq \mathbf{O}$, la condición necesaria y suficiente para que $\mathbf{R}^*$ tenga el mismo sentido que $\mathbf{R}$ es que

$$\frac{a}{\sqrt{a^2+b^2}} = \frac{a^*}{\sqrt{a^{*2}+b^{*2}}}, \qquad \frac{b}{\sqrt{a^2+b^2}} = \frac{b^*}{\sqrt{a^{*2}+b^{*2}}}.$$

A las cantidades

$$\xi = \frac{a}{\sqrt{a^2+b^2}} = \frac{a}{|\mathbf{R}|} = \frac{a}{r}, \qquad \eta = \frac{b}{\sqrt{a^2+b^2}} = \frac{b}{|\mathbf{R}|} = \frac{b}{r},$$

que determinan la dirección del vector, $\mathbf{R}$ les llamamos *cosenos directores* de $\mathbf{R}$. No están definidos, por supuesto, para $\mathbf{R} = \mathbf{O}$. Puesto que $\xi^2 + \eta^2 = 1$, podemos siempre encontrar un ángulo α y un correspondiente ángulo $\beta = \pi/2 - \alpha$ tales que

$$\xi = \cos\alpha, \qquad \eta = \operatorname{sen}\alpha = \cos\beta.$$

El ángulo α se llama *ángulo de dirección* de $\mathbf{R}$ (Fig. 4.33). Está determinado sólo a menos de un múltiplo par de π. Para $\mathbf{R} = \overrightarrow{PP'}$ tenemos

$$\cos\alpha = \frac{x'-x}{r}, \qquad \operatorname{sen}\alpha = \frac{y'-y}{r}.$$

Obviamente, α es el ángulo entre el eje x positivo y la recta que va de P a P'. Más precisamente, una rotación del eje x positivo alrededor del origen en un ángulo α (considerado positivo si la rotación es contra reloj, negativo en el caso opuesto) dará a dicho eje la dirección de P a P'. El vector opuesto $-\mathbf{R} = (-a, -b)$ tiene cosenos directores $-\xi$ y $-\eta$, y ángulos de dirección que difieren de α en múltiplos impares de π. Si el punto inicial P del vector $\mathbf{R} = \overrightarrow{PP'}$ es el origen, el ángulo de dirección α de $\mathbf{R}$ es simplemente el ángulo polar θ de P'.

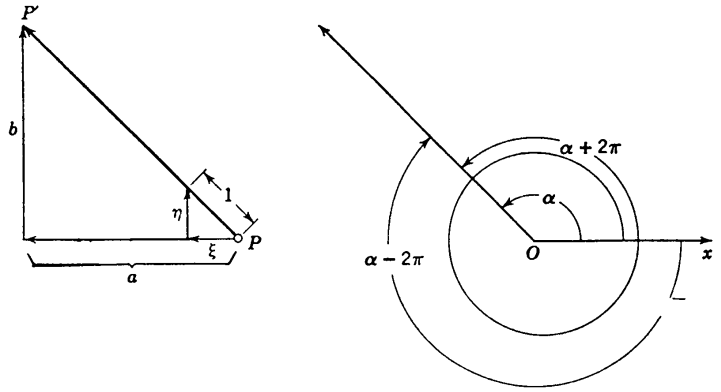


Figura 4.33 Cosenos directores ξ, η y ángulos de dirección para un vector $\overrightarrow{PP'}$.

b. Adición y multiplicación de vectores

Suma de vectores

Los vectores se han definido mediante traslaciones, esto es, como ciertas transformaciones de puntos en el plano. Existe una manera perfectamente general de combinar dos transformaciones cualesquiera aplicándolas *sucesivamente*. Si la primera transformación lleva el punto P al punto P' y la segunda lleva P' a P'', la transformación combinada es aquélla que lleva P a P''. En el caso de dos vectores $\mathbf{R} = (a, b)$ y $\mathbf{R}^* = (a^*, b^*)$, el vector $\mathbf{R}$ transformará el punto $P = (x, y)$ en el punto $P' = (x + a, y + b)$ y $\mathbf{R}^*$ transformará P' en $P'' = (x + a + a^*, y + b + b^*)$. La transformación de P en P'' resultante es también una traslación; la llamaremos la *suma* o la *resultante de los vectores* $\mathbf{R} = \overrightarrow{PP'}$ y $\mathbf{R}^* = \overrightarrow{P'P''}$, y la denotaremos por $\mathbf{R} + \mathbf{R}^*$ (Fig. 4.34).[1] Las componentes de la suma son $a + a^*$ y $b + b^*$. Así, nuestra definición de suma de dos vectores es

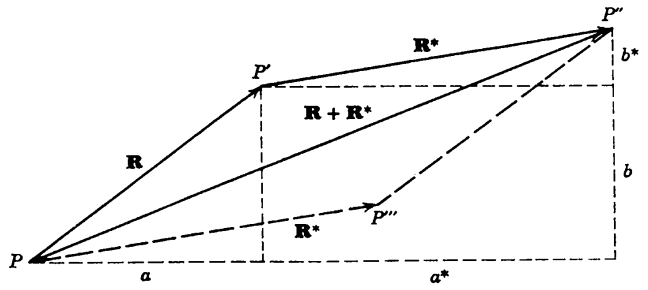


Figura 4.34 Adición de los vectores $\overrightarrow{PP'} = (a, b)$ y $\overrightarrow{P'P''} = (a^*, b^*)$.

$$\overrightarrow{PP'} + \overrightarrow{P'P''} = \overrightarrow{PP''},$$

o bien, si describimos los vectores por sus componentes,

$$(a, b) + (a^*, b^*) = (a + a^*, b + b^*).$$

Si $\mathbf{R}^*$ se toma con el mismo punto inicial que $\mathbf{R}$, digamos $\mathbf{R}^* = \overrightarrow{PP'''}$, los puntos P, P''', P'' y P' constituyen los vértices de un paralelogramo. Los dos lados que parten de P representan los vectores $\mathbf{R}$ y $\mathbf{R}^*$. La suma $\mathbf{R} + \mathbf{R}^*$ está representada por la *diagonal* que parte de P (*"construcción del paralelogramo" para la suma de vectores*).

[1] Esta "suma" es realmente el "producto símbólico" de las dos transformaciones, como se define en la p. 76. La notación de suma es aquí más natural debido a que corresponde a la adición de las componentes.

Las sumas vectoriales obedecen las leyes conmutativa y asociativa de la aritmética, ya que la adición de vectores se reduce a la adición de las componentes correspondientes (Fig. 4.35). Dichas sumas obedecen además las *leyes distributivas para la multiplicación de una suma de dos vectores por un número* λ *y de un vector por la suma de dos números* λ, μ:

$$\lambda(\mathbf{R} + \mathbf{R}^*) = \lambda\mathbf{R} + \lambda\mathbf{R}^*, \qquad (\lambda + \mu)\mathbf{R} = \lambda\mathbf{R} + \mu\mathbf{R}.^2$$

Estas reglas nos permiten expresar un vector $\overrightarrow{PP'}$ en términos de los vectores de posición $\overrightarrow{OP}$ y $\overrightarrow{OP'}$ de los puntos P y P' (Fig. 4.36):

$$\overrightarrow{PP'} = \overrightarrow{PO} + \overrightarrow{OP'} = \overrightarrow{OP'} + \overrightarrow{PO} = \overrightarrow{OP'} - \overrightarrow{OP}.$$

Es importante darse cuenta de que, generalmente, si se va de un punto P a un punto Q pasando por los puntos $A, B, C, \ldots, E, F$, entonces el vector $\overrightarrow{PQ}$ es la suma de los vectores $\overrightarrow{PA}, \overrightarrow{AB}, \overrightarrow{BC}, \ldots, \overrightarrow{EF}, \overrightarrow{FQ}$ (Fig. 4.37).

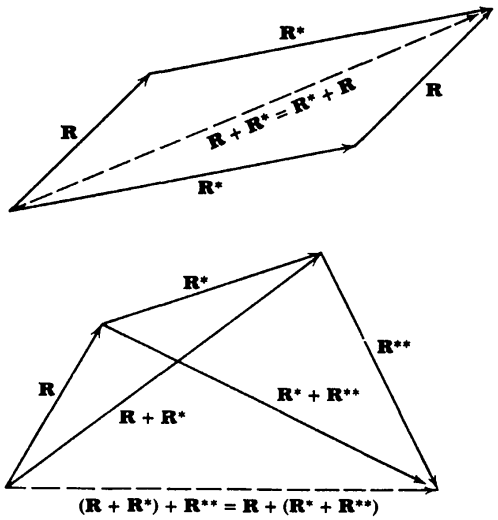


Figura 4.35 Leyes conmutativa y asociativa de la adición vectorial.

Angulo entre vectores

El ángulo θ formado por un vector $\mathbf{R}^* = (a^*, b^*)$ con el vector $\mathbf{R} = (a, b)$ se define como la diferencia entre sus ángulos de dirección:

[2] Para distinguir entre vectores y números en una ecuación, siempre ponemos el número antes del vector al escribir productos. La combinación $\mathbf{R}\lambda$ no se usará, aunque podría definirse por $\lambda\mathbf{R} = \mathbf{R}\lambda$.

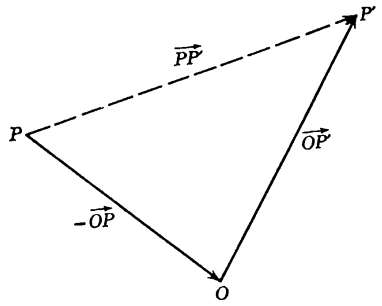


Figura 4.36 $\overrightarrow{PP'} = \overrightarrow{OP'} - \overrightarrow{OP}$.

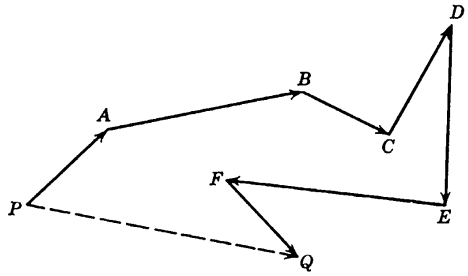


Figura 4.37 $\overrightarrow{PQ} = \overrightarrow{PA} + \overrightarrow{AB} + \overrightarrow{BC} + \cdots + FQ$.

$\theta = \alpha^* - \alpha$. (Aquí se supone que ninguno de los vectores $\mathbf{R}$ y $\mathbf{R}^*$ es el vector cero.) Nuevamente, el ángulo θ está determinado solamente a menos de múltiplos enteros de 2π (Fig. 4.38). Una rotación de ángulo θ (con el signo de θ indicando el sentido de la rotación) transformará la dirección de $\mathbf{R}$ en la de $\mathbf{R}^*$. Las cantidades $\cos\theta$ y $\operatorname{sen}\theta$, que están determinadas unívocamente, pueden inmediatamente expresarse en términos de los cosenos directores de $\mathbf{R}$ y $\mathbf{R}^*$:

$$\cos\theta = \cos(\alpha^* - \alpha) = \cos\alpha\cos\alpha^* + \operatorname{sen}\alpha\operatorname{sen}\alpha^*$$

$$= \frac{aa^* + bb^*}{\sqrt{a^2 + b^2}\sqrt{a^{*2} + b^{*2}}}$$

$$\operatorname{sen}\theta = \operatorname{sen}(\alpha^* - \alpha) = \cos\alpha\operatorname{sen}\alpha^* - \operatorname{sen}\alpha\cos\alpha^*$$

$$= \frac{ab^* - a^*b}{\sqrt{a^2 + b^2}\sqrt{a^{*2} + b^{*2}}}$$

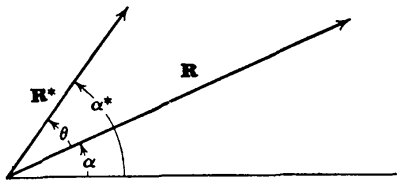


Figura 4.38. Angulo θ que el vector $\mathbf{R}^*$ forma con $\mathbf{R}$.

El denominador de cada expresión es justamente el producto rr^* de las longitudes de los vectores. Las expresiones que aparecen en los numeradores se introducen como "productos" de los dos vectores.

Producto interior y producto exterior de dos vectores

Se define el producto "escalar", producto 'interno" o producto "punto" de dos vectores, $\mathbf{R} = (a, b)$ y $\mathbf{R}^* = (a^*, b^*)$, mediante la expresión

$$\mathbf{R} \cdot \mathbf{R}^* = aa^* + bb^* = rr^* \cos\theta,$$

y el producto "exterior" o producto "cruz", por la expresión

$$\mathbf{R} \times \mathbf{R}^* = ab^* - a^*b = rr^* \operatorname{sen}\theta.^1$$

Como puede fácilmente verificarse, *los productos interior y exterior obedecen las leyes distributiva y asociativa:*

$$\mathbf{R} \cdot (\mathbf{R}^* + \mathbf{R}^{**}) = \mathbf{R} \cdot \mathbf{R}^* + \mathbf{R} \cdot \mathbf{R}^{**},$$

$$\mathbf{R} \times (\mathbf{R}^* + \mathbf{R}^{**}) = \mathbf{R} \times \mathbf{R}^* + \mathbf{R} \times \mathbf{R}^{**},$$

$$\lambda(\mathbf{R} \cdot \mathbf{R}^*) = (\lambda\mathbf{R}) \cdot \mathbf{R}^* = \mathbf{R} \cdot (\lambda\mathbf{R}^*),$$

$$\lambda(\mathbf{R} \times \mathbf{R}^*) = (\lambda\mathbf{R}) \times \mathbf{R}^* = \mathbf{R} \times (\lambda\mathbf{R}^*).$$

La ley conmutativa de la multiplicación también se cumple para productos interiores:

$$\mathbf{R} \cdot \mathbf{R}^* = \mathbf{R}^* \cdot \mathbf{R}.$$

Sin embargo, para productos exteriores el intercambio de los factores cambia el signo:

$$\mathbf{R} \times \mathbf{R}^* = -\mathbf{R}^* \times \mathbf{R}.$$

Dados $\mathbf{R}$ y $\mathbf{R}^*$ con el mismo punto inicial, $\mathbf{R} = \overrightarrow{PQ}$, $\mathbf{R}^* = \overrightarrow{PQ^*}$, podemos interpretar $\mathbf{R} \cdot \mathbf{R}^*$ como el producto de la proyección $r^* \cos\theta$ del segmento PQ^* sobre el segmento PQ y la longitud r de este segmento. El producto exterior $\mathbf{R} \times \mathbf{R}^*$ es simplemente el doble del área del triángulo orientado PQQ^*, tomada con signo positivo si los vértices PQQ^* están ordenados contra reloj y con signo negativo para el orden inverso (Fig. 4.39).

[1] Con nuestra definición, tanto el producto interior como el exterior son realmente "escalares". El término "producto escalar" se reserva para el producto interior debido a que en tres dimensiones el análogo del producto exterior es un vector (en este caso, el producto exterior también se llama *"producto vectorial"*).

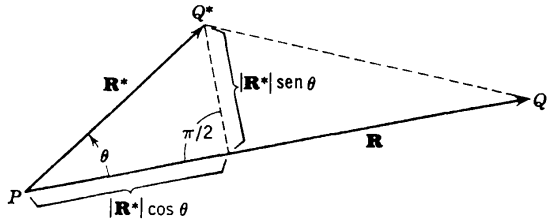


Figura 4.39 El producto vectorial $\mathbf{R} \times \mathbf{R}^* = |\mathbf{R}|\,|\mathbf{R}^*| \operatorname{sen}\theta$ como el doble del área del triángulo PQQ^*.

Para cualquier vector $\mathbf{R} = (a, b)$,

$$\mathbf{R} \cdot \mathbf{R} = a^2 + b^2 = |\mathbf{R}|^2$$

es el cuadrado de la longitud del vector. Así, $\mathbf{R} \cdot \mathbf{R}$ es positivo a menos que $\mathbf{R} = \mathbf{O}$. Por otra parte, $\mathbf{R} \times \mathbf{R}$ es siempre cero. La condición para que dos vectores no nulos sean mutuamente ortogonales es que $\mathbf{R} \cdot \mathbf{R}^* = 0$, y serán paralelos (esto es, tendrán la misma dirección y sentidos iguales u opuestos) si $\mathbf{R} \times \mathbf{R}^* = 0$.

Ecuación de la recta

Usando la notación vectorial podemos escribir fácilmente la ecuación de una recta que pasa por dos puntos. Si $P = (x, y)$, $P_0 = (x_0, y_0)$ y $P_1 = (x_1, y_1)$ son tres puntos tales que $P_0 \neq P_1$, entonces P estará sobre la recta que pasa por P_0 y P_1 si los vectores $\overrightarrow{P_0P}$ y $\overrightarrow{P_0P_1}$ son paralelos, esto es, si

$$\overrightarrow{P_0P} \times \overrightarrow{P_0P_1} = \mathbf{0}.$$

Si $\mathbf{R} = \overrightarrow{OP}$, $\mathbf{R}_0 = \overrightarrow{OP_0}$ y $\mathbf{R}_1 = \overrightarrow{OP_1}$ son los vectores de posición de los tres puntos, la condición toma la forma

$$(\mathbf{R} - \mathbf{R}_0) \times (\mathbf{R}_1 - \mathbf{R}_0) = \mathbf{0}$$

o bien

$$(\mathbf{R}_1 - \mathbf{R}_0) \times \mathbf{R} = \mathbf{R}_1 \times \mathbf{R}_0.$$

Substituyendo los vectores de posición por las coordenadas de los puntos se obtiene la ecuación de la recta en la forma usual (Fig. 4.40):

$$(x_1 - x_0)y - (y_1 - y_0)x = x_1y_0 - y_1x_0.$$

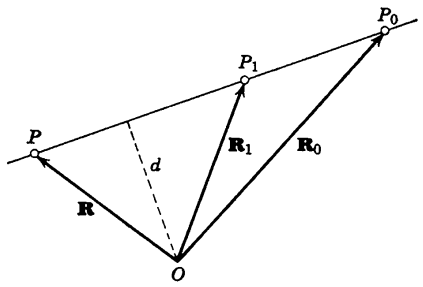


Figura 4.40 La recta en notación vectorial.

En vez de especificar dos puntos de la recta, podemos especificar sólo uno, P_0, y requerir que la recta sea paralela al vector $\mathbf{S} = (a, b)$. Evidentemente, la ecuación de la recta es entonces

$$(\mathbf{R} - \mathbf{R}_0) \times \mathbf{S} = 0,$$

o bien

$$(x - x_0)b - (y - y_0)a = 0.$$

Para $\mathbf{S} = \overrightarrow{P_0P_1}$ se obtiene la ecuación primera.

La distancia d de la recta al origen puede también expresarse en notación vectorial. Es claro que el producto de d y la longitud del vector $\overrightarrow{P_0P_1}$ es el doble del área del triángulo OP_0P_1. Por lo tanto,

$$d = \frac{1}{|P_0P_1|}(\overrightarrow{OP_0} \times \overrightarrow{OP_1}) = \frac{\mathbf{R}_0 \times \mathbf{R}_1}{|\mathbf{R}_1 - \mathbf{R}_0|}$$

$$= \frac{x_0y_1 - x_1y_0}{\sqrt{(x_1 - x_0)^2 + (y_1 - y_0)^2}}.$$

Aquí se toma d con signo positivo si los puntos O, P_0 y P_1 se suceden uno al otro en el sentido contra reloj.

Vectores coordenados. Un vector $\mathbf{R} = (a, b)$ puede representarse en la forma

$$\mathbf{R} = a\mathbf{i} + b\mathbf{j}, \tag{29}$$

donde se denotan por $\mathbf{i}$ y $\mathbf{j}$ los *"vectores coordenados"*

$$\mathbf{i} = (1, 0), \qquad \mathbf{j} = (0, 1). \tag{30}$$

De esta manera se divide $\mathbf{R}$ en dos vectores $a\mathbf{i}$ y $b\mathbf{j}$ que apuntan respectivamente en las direcciones positivas de los ejes x e y. Las componentes a y b de $\mathbf{R}$ son justamente las longitudes (con sus signos) de estos dos vectores.

En las aplicaciones se tiene a menudo necesidad de representar un vector $\mathbf{R}$ como la resultante de vectores con dos direcciones ortogonales (esto es, mutuamente perpendiculares) dadas. Con este fin, lo mejor es introducir dos *vectores unitarios* (o sea, vectores de longitud 1) $\mathbf{I}$ y $\mathbf{J}$ con las direcciones prescritas. La descomposición de $\mathbf{R}$ requerida se realiza si puede representarse $\mathbf{R}$ en la forma

$$\mathbf{R} = A\mathbf{I} + B\mathbf{J} \tag{31}$$

con escalares adecuados A, B (cfr. Fig. 4.40). Es fácil encontrar los valores de A y B en el caso en que tal representación de $\mathbf{R}$ exista. Pues, en efecto, por hipótesis, los vectores $\mathbf{I}$ y $\mathbf{J}$ son vectores unitarios de longitud 1, de manera que

$$\mathbf{I} \cdot \mathbf{I} = \mathbf{J} \cdot \mathbf{J} = 1, \qquad \mathbf{I} \cdot \mathbf{J} = 0. \tag{32}$$

Formando el producto escalar de la Ec. (31) con $\mathbf{I}$, $\mathbf{J}$, respectivamente, se encuentra de inmediato que A y B deben tener los valores

$$A = \mathbf{R} \cdot \mathbf{I}, \qquad B = \mathbf{R} \cdot \mathbf{J}; \tag{33}$$

es decir, A y B son las longitudes (con sus signos respectivos) de las proyecciones en las direcciones dadas del segmento que representa a $\mathbf{R}$.

La posibilidad de expresar $\mathbf{R}$ como una combinación lineal (31) de $\mathbf{I}$ y $\mathbf{J}$ se deriva de la representación (29) de $\mathbf{R}$ en términos de $\mathbf{i}$ y $\mathbf{j}$, siempre que pueda demostrarse que $\mathbf{i}$ y $\mathbf{j}$ mismos pueden ser expresados en términos de $\mathbf{I}$ y $\mathbf{J}$. Sin embargo, $\mathbf{I} = (\alpha, \beta)$ y $\mathbf{J} = (\gamma, \delta)$ pueden escribirse de la manera siguiente

$$\mathbf{I} = \alpha\mathbf{i} + \beta\mathbf{j}, \qquad \mathbf{J} = \gamma\mathbf{i} + \delta\mathbf{j}. \tag{34}$$

Como consecuencia de (32), las cantidades α, β, γ, δ deben satisfacer las llamadas relaciones de ortogonalidad

$$\alpha^2 + \beta^2 = \gamma^2 + \delta^2 = 1, \qquad \alpha\gamma + \beta\delta = 0. \tag{35}$$

Si se multiplica la primera de las ecuaciones (34) por δ, la segunda por β, y se restan, se obtiene

$$(\alpha\delta - \beta\gamma)\mathbf{i} = \delta\mathbf{I} - \beta\mathbf{J}; \tag{36}$$

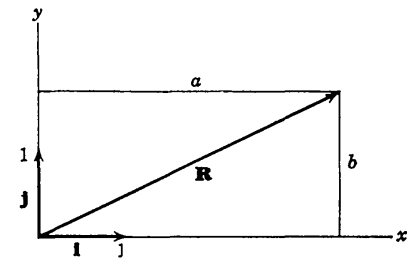
y, en forma análoga,

$$(\alpha\delta - \beta\gamma)\mathbf{j} = -\gamma\mathbf{I} + \alpha\mathbf{J}. \tag{37}$$

Aquí se tiene para los vectores unitarios mutuamente perpendiculares $\mathbf{I}$ y $\mathbf{J}$ que

$$\alpha\delta - \beta\gamma = \mathbf{I} \times \mathbf{J} = \pm 1, \tag{38}$$

donde se escoge un signo u otro dependiendo de que el sentido de la rotación en 90° que transforme **I** en **J** sea contra reloj o en el sentido del reloj. En uno u otro caso las fórmulas (36) y (37) expresan **i** y **j** en términos de **I** y **J**. Substituyendo estas expresiones en (29) se obtiene la justificación de la fórmula de representación (31) para un vector arbitrario **R**.



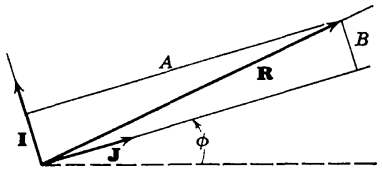


Figura 4.40

La fórmula (31) puede también interpretarse como la representación del vector **R** en un nuevo sistema de coordenadas cuyos ejes apuntan respectivamente en las direcciones de **I** y **J**. Las componentes de un vector unitario son al mismo tiempo los cosenos directores del ángulo de dirección ϕ y ψ, respectivamente. Entonces,

$$\alpha = \cos\phi, \quad \beta = \operatorname{sen}\phi, \quad \gamma = \cos\psi, \quad \delta = \operatorname{sen}\psi.$$

Aquí se tiene $\psi = \phi + \frac{1}{2}\pi$, o bien $\psi = \phi - \frac{1}{2}\pi$. En el primer caso (que corresponde a un sistema dextrógiro con vectores coordenados **I** y **J**) se tiene $\gamma = -\beta$, $\delta = \alpha$, $\alpha\delta - \beta\gamma = +1$; de manera que

$$\mathbf{I} = (\cos\phi, \operatorname{sen}\phi), \quad \mathbf{J} = (-\operatorname{sen}\phi, \cos\phi). \tag{39}$$

Las fórmulas (33) que dan las componentes de **R** referidas a los vectores coordenados **I** y **J** toman entonces la forma

$$A = a\cos\phi + b\operatorname{sen}\phi, \quad B = -a\operatorname{sen}\phi + b\cos\phi. \tag{40}$$

Estas fórmulas expresan las relaciones entre las componentes de uno y el mismo vector **R** en dos sistemas de coordenadas dextrógiros que se obtienen uno del otro mediante una rotación de ejes en un ángulo ϕ. Si se supone que los sistemas de coordenadas tienen el mismo origen O y que **R** es el vector de posición $\overrightarrow{OP}$ de un punto arbitrario P, las fórmulas (40) expresan la transformación de los sistemas de coordenadas obtenidas ya en la p. 379, ecuación (18). Las componentes a, b y A, B son entonces respectivamente las coordenadas de P en los dos sistemas.

c. Vectores variables, sus derivadas e integrales

Es natural considerar vectores $\mathbf{R} = (a, b)$ cuyas componentes a, b sean funciones de una variable t, digamos $a = a(t)$, $b = b(t)$. Para cualquier t se tiene entonces un vector

$$\mathbf{R} = \mathbf{R}(t) = (a(t), b(t))$$

y se dice que $\mathbf{R}(t)$ es una *función vectorial* de t. Un ejemplo lo proporciona el vector de posición de un punto que se mueve al variar el tiempo t.

Se dice que $\mathbf{R}(t)$ tiene el límite $\mathbf{R}^* = (a^*, b^*)$ para $t \to t_0$ si $a(t)$ tiene el límite a^* y $b(t)$ el límite b^* para $t \to t_0$. En ese caso la longitud de $\mathbf{R}(t)$ tiende hacia la de $\mathbf{R}^*$ y, si $\mathbf{R}^* \neq 0$, la dirección de $\mathbf{R}(t)$ tiende hacia la de $\mathbf{R}^*$ (esto significa que los cosenos directores de **R** tienden hacia los de $\mathbf{R}^*$). Se dice que el vector $\mathbf{R}(t)$ depende continuamente de t si

$$\lim_{t \to t_0} \mathbf{R}(t) = \mathbf{R}(t_0),$$

esto es, si las componentes de **R** son funciones continuas de t. La longitud y, si $\mathbf{R}(t_0) \neq \mathbf{O}$, también la dirección de un vector continuo varían continuamente con t.

Para introducir la *derivada de un vector* se forma para dos valores t y $t + h$ del parámetro el cociente incremental

$$\frac{1}{h}[\mathbf{R}(t+h) - \mathbf{R}(t)] = \left[\frac{a(t+h) - a(t)}{h}, \frac{b(t+h) - b(t)}{h}\right],$$

y se define la derivada de **R** como el límite del cociente incremental para $h \to 0$:

$$\dot{\mathbf{R}} = \frac{d\mathbf{R}}{dt} = \lim_{h \to 0} \frac{1}{h}[\mathbf{R}(t+h) - \mathbf{R}(t)] = \left(\frac{da}{dt}, \frac{db}{dt}\right) = (\dot{a}, \dot{b}).$$

La derivada de un vector se forma derivando sus componentes.

Las derivadas de productos de vectores, como se ve fácilmente, obedecen las reglas ordinarias de las derivadas de productos de funciones reales o "escalares":

$$(\mathbf{RS})^{\cdot} = \frac{d\mathbf{R}\cdot\mathbf{S}}{dt} = \frac{d\mathbf{R}}{dt}\cdot\mathbf{S} + \mathbf{R}\cdot\frac{d\mathbf{S}}{dt} = \dot{\mathbf{R}}\mathbf{S} + \mathbf{R}\dot{\mathbf{S}}$$

$$(\mathbf{R}\times\mathbf{S})^{\cdot} = \frac{d\mathbf{R}\times\mathbf{S}}{dt} = \dot{\mathbf{R}}\times\mathbf{S} + \mathbf{R}\times\dot{\mathbf{S}},$$

donde para los productos exteriores es necesario conservar el orden original de los factores.

En forma análoga se define la integral del vector $\mathbf{R}(t)$ en términos de las integrales de sus componentes:

$$\int_\alpha^\beta \mathbf{R}(t)\,dt = \left(\int_\alpha^\beta a(t)\,dt, \int_\alpha^\beta b(t)\,dt\right).$$

El teorema fundamental del cálculo implica que

$$\frac{d}{dt}\int_\alpha^t \mathbf{R}(s)\,ds = \mathbf{R}(t).$$

d. Aplicación a curvas planas. Dirección, "rapidez" y aceleración

Vector velocidad

En la Sección 4.1 se representó una curva C mediante dos funciones $x = \phi(t)$ e $y = \psi(t)$. Cada valor de t en el dominio de estas funciones determina un punto $P = (x, y)$ sobre C; en este caso t puede considerarse como el tiempo y P como un punto móvil cuya posición en el tiempo t está dada por $x(t)$ e $y(t)$. Si se identifican x e y con las componentes del vector de posición $\mathbf{R} = \overrightarrow{OP}$ de P, entonces C queda descripta por el punto final del vector de posición

$$\mathbf{R} = \mathbf{R}(t) = (x(t), y(t))$$

(Fig. 4.41). Para dos puntos P y P' de C correspondientes a los valores paramétricos t y $t + \Delta t$, la expresión

$$\overrightarrow{PP'} = \overrightarrow{OP'} - \overrightarrow{OP} = \mathbf{R}(t + \Delta t) - \mathbf{R}(t) = \Delta\mathbf{R}$$

da el vector representado por la secante dirigida de C con puntos extremos P, P'. Aquí Δt es positivo, esto es, si el punto P' sigue a P sobre C en el sentido de las t crecientes, entonces el vector

$$\frac{t}{\Delta t}(\mathbf{R}(t + \Delta t) - \mathbf{R}(t))$$

tiene la misma dirección que el vector $\mathbf{R}(t + \Delta t) - \mathbf{R}(t) = \overrightarrow{PP'}$; su longitud es la distancia entre los puntos P y P' dividida por Δt. Cuando Δt tiende a cero se obtiene en el límite el vector

$$\dot{\mathbf{R}} = \dot{\mathbf{R}}(t) = (\dot{x}(t), \dot{y}(t)),$$

donde, nuevamente, el punto se usa para denotar derivación con respecto al parámetro t. La dirección de $\dot{\mathbf{R}}$ es el límite de las direcciones de las secantes PP', y, por lo tanto, es la dirección de la tangente en el punto P. Más precisamente, $\dot{\mathbf{R}}$ apunta en aquella dirección de la tangente que co-

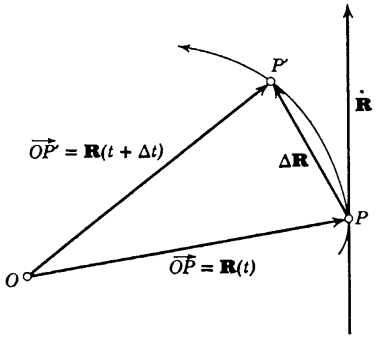


Figura 4.41 Derivada del vector de posición para una curva.

rresponde a valores crecientes de t sobre C, suponiendo que $\mathbf{R} \neq \mathbf{O}$. Los cosenos directores de $\dot{\mathbf{R}}$ son las cantidades

$$\cos\alpha = \frac{\dot{x}}{\sqrt{\dot{x}^2 + \dot{y}^2}}, \qquad \operatorname{sen}\alpha = \frac{\dot{y}}{\sqrt{\dot{x}^2 + \dot{y}^2}},$$

introducidas en la p. 364 como cosenos directores de la tangente. La longitud de $\dot{\mathbf{R}}$,

$$|\dot{\mathbf{R}}| = \sqrt{\dot{x}^2 + \dot{y}^2},$$

puede interpretarse como ds/dt, esto es, la medida de cambio, con respecto al parámetro t, de la longitud s a lo largo de la curva. Si t representa el tiempo, $|\dot{\mathbf{R}}|$ representa entonces la velocidad con la que el punto se mueve a lo largo de la curva.

En Mecánica uno debe considerar la velocidad de una partícula no sólo como algo que tiene una cierta magnitud (la "rapidez") sino también una cierta *dirección*. La *velocidad* está entonces representada por el *vector* $\dot{\mathbf{R}} = (\dot{x}, \dot{y})$, cuya longitud es la velocidad (escalar) y cuya dirección es la dirección instantánea de movimiento, esto es, la dirección de la tangente en el sentido de las t crecientes.

Aceleración

Análogamente, la *aceleración* de la partícula se define como el vector $\ddot{\mathbf{R}} = (\ddot{x}, \ddot{y})$. Si la aceleración se anula, esto significa que $\ddot{x} = \ddot{y} = 0$; si $\ddot{\mathbf{R}} = \mathbf{O}$ a lo largo de un intervalo t completo, las componentes de la velocidad tienen valores constantes $\dot{x} = a$, $\dot{y} = b$; y las componentes del vector de posición mismo son entonces funciones lineales de t: $x = at + c$, $y = bt + d$. En este caso la partícula se mueve con velocidad constante a lo largo de una línea recta.

Todos nuestros resultados anteriores relativos a curvas son fácilmente expresables en notación vectorial si se describe la curva por el vector de posición $\mathbf{R} = \mathbf{R}(t) = (x(t), y(t))$, con $\alpha \leq t \leq \beta$. Se encuentra para la longitud [cfr. Ec. (8), p. 367] la expresión

$$\int_\alpha^\beta |\dot{\mathbf{R}}|\, dt,$$

en tanto que para que el área, con su signo, encerrada por una curva [cfr. Ec. (20), p. 383]:

$$A = \tfrac{1}{2} \int_\alpha^\beta \mathbf{R} \times \dot{\mathbf{R}}\, dt$$

(el signo de esta cantidad depende de la orientación de la curva). Finalmente, para la curva κ se tiene la fórmula [cfr. Ec. (15), p. 373]

$$\kappa = \frac{\dot{\mathbf{R}} \times \ddot{\mathbf{R}}}{|\dot{\mathbf{R}}|^3}.$$

Componentes tangencial y normal de la aceleración

Las fórmulas anteriores tienen implicaciones interesantes si nuevamente se interpreta t como el tiempo. Sea γ el ángulo formado por el vector $\ddot{\mathbf{R}}$ con el vector $\dot{\mathbf{R}}$, esto es, con la dirección instantánea de movimiento. La cantidad $|\ddot{\mathbf{R}}| \cos \gamma$ representa la proyección de $\ddot{\mathbf{R}}$ en la dirección de $\dot{\mathbf{R}}$; la llamamos *componente tangencial* de la aceleración. Análogamente, $|\ddot{\mathbf{R}}| \operatorname{sen} \gamma$ es la proyección de $\ddot{\mathbf{R}}$ sobre la normal (más precisamente, sobre la normal que se obtiene a partir de $\dot{\mathbf{R}}$ mediante una rotación contra reloj de 90°); ésta es la *componente normal* de la aceleración (ver Fig. 4.42). Por definición de productos interior y exterior,

$$|\ddot{\mathbf{R}}| \cos \gamma = \frac{\dot{\mathbf{R}} \cdot \ddot{\mathbf{R}}}{|\dot{\mathbf{R}}|} \qquad |\ddot{\mathbf{R}}| \operatorname{sen} \gamma = \frac{\dot{\mathbf{R}} \times \ddot{\mathbf{R}}}{|\dot{\mathbf{R}}|}.$$

Ahora bien,

$$\dot{\mathbf{R}} \cdot \ddot{\mathbf{R}} = \frac{1}{2}(\dot{\mathbf{R}} \cdot \ddot{\mathbf{R}} + \ddot{\mathbf{R}} \cdot \ddot{\mathbf{R}}) = \frac{1}{2}\frac{d}{dt}(\dot{\mathbf{R}} \cdot \dot{\mathbf{R}}) = \frac{1}{2}\frac{dv^2}{dt} = v\frac{dv}{dt},$$

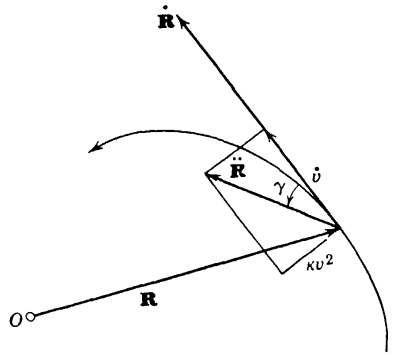


Figura 4.42 Aceleración tangencial y aceleración normal.

donde $v = ds/dt = |\dot{\mathbf{R}}| = \sqrt{\dot{\mathbf{R}} \cdot \dot{\mathbf{R}}}$ es la *velocidad* (escalar) del punto. Por lo tanto

$$|\ddot{\mathbf{R}}| \cos \gamma = \frac{dv}{dt} = \dot{v}; \tag{41}$$

De esta manera, la componente tangencial de la aceleración es idéntica a la medida de cambio, o sea, la derivada, con respecto al tiempo, de la velocidad. Para la aceleración normal la fórmula de la curvatura conduce a

$$|\ddot{\mathbf{R}}| \operatorname{sen} \gamma = \kappa |\dot{\mathbf{R}}|^2 = \kappa v^2, \tag{42}$$

esto es, el producto del cuadrado de la velocidad por la curvatura.

Para una partícula que se mueve con velocidad *constante* v a lo largo de una curva, la aceleración tangencial $\dot{v}$ se anula. El vector aceleración es entonces perpendicular a la curva. Más precisamente, dicho vector apunta hacia el lado "interno" de la curva, o sea, el lado hacia el cual la curva se dobla (esto se ve, por ejemplo, del hecho de que $\operatorname{sen} \gamma > 0$ cuando $\kappa > 0$, o, lo que es lo mismo, cuando la tangente gira contra reloj). Al moverse a lo largo de una curva con rapidez constante un punto experimenta, por lo tanto, una aceleración hacia el interior de la curva, que es proporcional a la curvatura y también al cuadrado de la velocidad. Este hecho es de obvia significación debido a que, como resultado de la ley de Newton (discutida más adelante), se necesita una fuerza proporcional a la aceleración para mantener el punto P sobre la curva.

4.4 Movimiento de una partícula bajo la acción de fuerzas especificadas

Los primeros desarrollos del cálculo fueron estimulados en forma decisiva no solamente por la geometría sino también en una gran medida por

los conceptos de la Mecánica. Esta ciencia descansa sobre ciertos principios básicos que fueron establecidos primeramente por Newton. El establecimiento de estos principios involucra el concepto de derivada, y su aplicación requiere de la teoría de la integración. Sin analizar en detalle los principios de Newton, ilustraremos con algunos ejemplos sencillos cómo el cálculo se aplica en Mecánica.

a. Ley de movimiento de Newton

Nos restringiremos a la consideración de una única partícula, esto es, un punto en el cual se supone concentrada una masa m. Supondremos además que el movimiento se efectúa en el plano x, y, en el cual la posición de la partícula en el tiempo t está especificada por sus coordenadas $x = x(t)$, $y = y(t)$, o, equivalentemente, por su "vector de posición" $\mathbf{R} = \mathbf{R}(t) = (x(t), y(t))$. Un punto sobre una cantidad indica derivación con respecto al tiempo t. La *velocidad* y la *aceleración* de la partícula quedan entonces representadas por los vectores

$$\dot{\mathbf{R}} = (\dot{x}, \dot{y}) \qquad \text{y} \quad \ddot{\mathbf{R}} = (\ddot{x}, \ddot{y}).$$

En Mecánica se relaciona el movimiento de un punto con el concepto de *fuerzas* de dirección y magnitud definida que actúan sobre este punto. También puede describirse una fuerza mediante un vector $\mathbf{F} = (\rho, \sigma)$. El efecto de varias fuerzas $\mathbf{F}_1$, $\mathbf{F}_2, \ldots$ que actúan sobre la misma partícula es el mismo que el de una única fuerza $\mathbf{F}$, la *fuerza resultante,* que es simplemente la suma vectorial $\mathbf{F} = \mathbf{F}_1 + \mathbf{F}_2 + \cdots$ de las fuerzas individuales.

La ley fundamental de Newton establece que: *la masa m multiplicada por la aceleración es igual a la fuerza que actúa sobre la partícula;* en símbolos,

$$m\ddot{\mathbf{R}} = \mathbf{F}. \tag{43}$$

Si escribimos esta ecuación vectorial, que expresa la ley fundamental, en términos de las componentes de esos vectores, obtenemos el par equivalente de ecuaciones

$$m\ddot{x} = \rho, \qquad m\ddot{y} = \sigma. \tag{44}$$

Puesto que aceleración y fuerza difieren solamente en el factor positivo m, *la dirección de la aceleración es la misma que la de la fuerza.* Si ninguna fuerza actúa, esto es, $\mathbf{F} = \mathbf{O}$, la aceleración se anula, la velocidad es constante y x, y se vuelven funciones lineales de t. Esta es la *primera ley de Newton:* una partícula sobre la que no actúan fuerzas se mueve con velocidad constante a lo largo de una línea recta.

La ley de Newton $m\ddot{\mathbf{R}} = \mathbf{F}$, en el primer ejemplo, no es más que la definición cuantitativa del concepto de fuerza. El lado izquierdo de esta

relación puede determinarse observando el movimiento, mediante lo cual se obtiene entonces la fuerza.

Sin embargo, la ley de Newton tiene un significado más profundo debido al hecho de que en muchos casos podemos determinar la fuerza actuante a partir de otras consideraciones físicas, sin conocer el movimiento correspondiente. Esta ley fundamental no es entonces una *definición de fuerza* sino una relación a partir de la cual podemos esperar la determinación del movimiento. Esta manera de considerar las cosas al usar la ley de Newton muestra su utilidad en los numerosos casos en los que consideraciones físicas permiten expresar la fuerza $\mathbf{F}$ o sus componentes ρ, σ de manera explícita como funciones de la posición y la velocidad de la partícula y del tiempo t. La ley de movimiento no es entonces una tautología sino que proporciona dos ecuaciones que expresan $m\ddot{x}, m\ddot{y}$ en términos de x, y, $\dot{x}$, $\dot{y}$ y t, las llamadas *ecuaciones de movimiento.* Son éstas *ecuaciones diferenciales,* esto es, relaciones entre funciones y sus derivadas. La solución de estas ecuaciones diferenciales, o sea, el encontrar todas los pares de funciones $x(t)$, $y(t)$ para las cuales las ecuaciones de movimiento son válidas, dan todos los movimientos posibles de una partícula bajo la acción de la fuerza prescrita.

b. Movimiento de cuerpos en caída

El ejemplo más sencillo de una fuerza conocida es el de la *gravedad* que actúa sobre una partícula cercana a la superficie de la Tierra. Se sabe por observación directa que (despreciando los efectos de la resistencia del aire) todo cuerpo que cae tiene una aceleración dirigida verticalmente hacia abajo y que tiene la misma magnitud g para todos los cuerpos. Medida en metros por segundo (m.s^{-2}), g tiene aproximadamente el valor 9.80.[1] Si escogemos un sistema de coordenadas x, y en el cual el eje y apunte verticalmente hacia arriba en tanto que el eje x es horizontal, la aceleración $\ddot{\mathbf{R}} = (\ddot{x}, \ddot{y})$ tiene las componentes

$$\ddot{x} = 0, \qquad \ddot{y} = -g.$$

Por la ley fundamental de Newton, el vector $\mathbf{F}$ que representa la fuerza de gravedad que actúa sobre una partícula de masa m debe entonces ser

$$\mathbf{F} = (0, -mg).$$

Este vector fuerza está de la misma manera dirigido verticalmente hacia abajo; su magnitud, el *peso* del cuerpo cerca de la superficie de la Tierra, es mg.

[1] El valor preciso de g, que además de la atracción gravitacional incluye los efectos de la rotación de la Tierra, depende de la posición sobre ésta.

Cuando se cancela el factor m, las ecuaciones de movimiento de una partícula bajo la acción de la gravedad toman la forma

$$\ddot{x} = 0, \qquad \ddot{y} = -g.$$

A partir de estas ecuaciones podemos obtener fácilmente una descripción del movimiento más general posible para un cuerpo que cae. Integrando con respecto a t se obtiene

$$\ddot{x} = a, \qquad \ddot{y} = -gt + b,$$

donde a y b son constantes. Una integración posterior muestra entonces que

$$x = at + c, \qquad y = -\tfrac{1}{2}gt^2 + bt + d,$$

donde c y d son también constantes. Así, la solución general de nuestras ecuaciones de movimiento depende de cuatro constantes no especificadas, a, b, c, d. Podemos relacionar inmediatamente los valores de estas constantes para un movimiento particular con las *condiciones iniciales* de este movimiento. Si al tiempo inicial $t = 0$ la partícula está en el punto (x_0, y_0), entonces, haciendo $t = 0$, se encuentra

$$c = x_0, \qquad d = y_0.$$

La velocidad $\dot{\mathbf{R}} = (\dot{x}, \dot{y}) = (a, -gt + b)$ se reduce a (a, b) para $t = 0$. Así (c, d) y (a, b) representan respectivamente la posición y la velocidad iniciales de la partícula. Cualquier elección de estas condiciones iniciales origina en forma unívoca un movimiento.

En el caso $a \neq 0$, esto es, el caso en que la velocidad inicial no es vertical, podemos eliminar t y obtener una representación no paramétrica para la órbita de la partícula. Despejando t de la primera ecuación y substituyendo el resultado en la segunda, se llega a

$$y = -\frac{g}{2a^2}(x - c)^2 + \frac{b}{a}(x - c) + d.$$

Por lo tanto, la trayectoria es una parábola. Para $a = 0$ tenemos $x = c =$ constante, y todo el movimiento se efectúa a lo largo de una recta vertical.

c. Movimiento de una partícula restringido a una curva dada

En la mayoría de los problemas de mecánica las fuerzas que actúan sobre una partícula dependen de la posición y de la velocidad de dicha partícula. En general, las ecuaciones de movimiento son demasiado complicadas para permitirnos determinar todos los movimientos posibles. Una simplificación grande se logra si puede considerarse como conocida la curva C descrita por la partícula, de manera que sólo tenga que determinarse el

movimiento a lo largo de la curva. En una variedad muy grande de problemas mecánicos la partícula está *restringida* (mediante algún artificio mecánico) a moverse sobre una curva dada C. El ejemplo más simple lo constituye el péndulo plano, en el que una masa m está sujeta a un punto P_0 por medio de un hilo inextensible de longitud L y se mueve bajo la acción de la gravedad sobre una circunferencia de radio L con centro en P_0.

Usando el arco de longitud s como parámetro a lo largo de la curva C, ésta queda determinada por las ecuaciones $x = x(s)$, $y = y(s)$. Encontrar el movimiento de la partícula a lo largo de C equivale entonces a encontrar s como función de t. Una ecuación de movimiento válida a lo largo de la curva se encuentra de la manera siguiente.

Multiplíquense escalarmente ambos lados de la fórmula de Newton $m\ddot{\mathbf{R}} = \mathbf{F}$ por un vector ξ:

$$m\ddot{\mathbf{R}} \cdot \xi = \mathbf{F} \cdot \xi.$$

Si se escoge ξ igual al vector de longitud 1 cuya dirección es la de la tangente a C en el sentido de las s crecientes, esto es, $\xi = d\mathbf{R}/ds$, la expresión $\mathbf{F} \cdot \xi = f$ dará la *componente tangencial* de la fuerza, o *fuerza que actúa en la dirección del movimiento.* De acuerdo con la ecuación (41) de la p. 414, la componente tangencial $\ddot{\mathbf{R}} \cdot \xi$ de la aceleración es justamente $dv/dt = d^2s/dt^2$, esto es, la aceleración de la partícula a lo largo de la curva. La ley de Newton conduce entonces a la fórmula

$$m\ddot{s} = f, \tag{45}$$

o sea, *la masa de la partícula multiplicada por su aceleración a lo largo de la trayectoria es igual a la fuerza que actúa en la dirección del movimiento.*

Al aplicar esta ecuación a una partícula *restringida* a moverse sobre una curva C, suponemos que las restricciones no contribuyen a la fuerza f.[1] Para una fuerza $\mathbf{F} = (\rho, \sigma)$ la ecuación (44) de la p. 415 conduce a

$$f = \rho\frac{dx}{ds} + \sigma\frac{dy}{ds}; \tag{46}$$

puesto que el vector ξ tiene como componentes $\frac{dx}{ds}, \frac{dy}{ds}$ (ver p. 412). Para una curva C conocida los cosenos directores $\frac{dx}{ds}$ y $\frac{dy}{ds}$ de la tangente pueden considerarse como funciones conocidas de s. De la misma manera, si la fuerza $\mathbf{F} = 2(\rho, \sigma)$ depende solamente de la posición de la partícula, f será una función conocida de s. El movimiento de la partícula a lo largo de C

[1] Realmente, el mecanismo de restricción tiene que proporcionar una fuerza que mantenga a la partícula sobre C (en el péndulo simple la proporciona la tensión del hilo). Suponemos que esta fuerza de "reacción" es perpendicular a la curva y así no tiene componente tangencial; y éste es el caso cuando la partícula se desliza sin fricción sobre la curva.

tiene entonces que ser determinado a partir de la ecuación diferencial relativamente sencilla $m\ddot{s} = f(s)$.

Específicamente, para la fuerza gravitacional $\mathbf{F} = (0, -mg)$ se tiene

$$f = -mg\frac{dy}{ds}. \tag{46a}$$

Así, la ecuación de movimiento de una partícula restringida a moverse bajo la acción de la gravedad sobre una curva C se transforma en

$$\frac{d^2s}{dt^2} = -g\frac{dy}{ds}. \tag{47}$$

Si α denota el ángulo de inclinación de la curva, tenemos $dy/ds = \operatorname{sen}\alpha$ (ver Fig. 4.43), y la ecuación de movimiento se convierte en

$$\frac{d^2s}{dt^2} = -g\operatorname{sen}\alpha$$

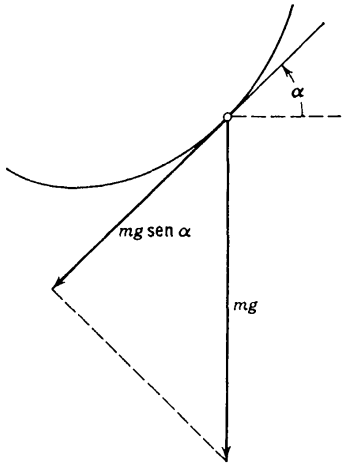


Figura 4.43 Movimiento sobre una curva dada, bajo la acción de la gravedad.

Para una partícula restringida a moverse sobre una circunferencia de radio L con centro en el origen ("péndulo simple"), es

$$x = L\operatorname{sen}\theta, \quad y = -L\cos\theta,$$

donde $\theta = s/L$ es el ángulo polar medido a partir de la vertical que apunta hacia abajo. En este caso (ver Fig. 4.44) es $\alpha = \theta$, por lo que

$$\frac{d^2s}{dt^2} = -g\frac{dy}{d\theta}\frac{d\theta}{ds} = -g\operatorname{sen}\theta,$$

o bien

$$\frac{d^2\theta}{dt^2} = -\frac{g}{L}\operatorname{sen}\theta.$$

4.5 Caída libre de un cuerpo venciendo la resistencia del aire

Comenzaremos con dos ejemplos relativos al movimiento de una partícula a lo largo de una recta. Consideraremos solamente los casos en que la fuerza actúa en la dirección de la recta, de manera que no es necesario un mecanismo de restricción.

La trayectoria de un cuerpo que cae libremente hacia abajo puede describirse mediante las expresiones $x =$ constante, $y = s$. Si la única fuerza presente es la de gravedad, se tiene la ecuación de movimiento

$$m\ddot{s} = -mg.$$

Para una partícula que se suelta en el tiempo $t = 0$ desde la altura $y_0 = s_0$ con velocidad inicial v_0 (considerada positiva si es hacia arriba), se encuentra por integración que

$$s = -\tfrac{1}{2}gt^2 + v_0t + s_0.$$

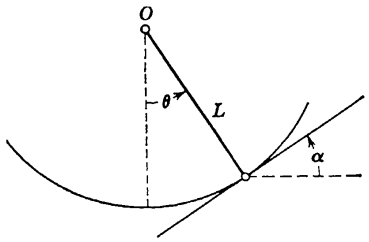


Figura 4.44 El péndulo simple.

Si se desea tomar en cuenta el efecto de la *fricción* o *resistencia del aire* que actúa sobre la partícula, se la tendrá que considerar como una fuerza de sentido opuesto al del movimiento y respecto a la cual pueden hacerse ciertas consideraciones de tipo físico.[1] Se obtendrán resultados para diferentes hipótesis físicas: *a)* la resistencia es proporcional a la velocidad y está dada por una expresión de la forma $-r\dot{s}$, donde r es una constante positiva; *b)* la resistencia es proporcional al cuadrado de la velocidad y es de la forma $-r\dot{s}^2$ para $\dot{s}$ positiva, y de la forma $r\dot{s}^2$ para $\dot{s}$ negativa. De acuerdo con la ley de Newton se obtienen las ecuaciones de movimiento

$$(a) \qquad m\ddot{s} = -mg - r\dot{s},$$

$$(b) \qquad m\ddot{s} = -mg + r\dot{s}^2,$$

[1] Estas hipótesis deben escogerse adecuadas al sistema físico particular que se esté considerando; por ejemplo, la ley de resistencia para bajas velocidades no es la misma que para velocidades altas (tales como velocidades de bala).

donde se ha supuesto para (b) que el cuerpo está cayendo $(\dot{s} < 0)$. Si se considera, en primer lugar, $\dot{s} = v(t)$ como la función buscada, se tiene

$$(a) \qquad m\dot{v} = -mg - rv,$$

$$(b) \qquad m\dot{v} = -mg + rv^2.$$

En vez de usar estas ecuaciones para determinar v como función de t, se determina t como función de v escribiendo dichas ecuaciones diferenciales en la forma

$$(a) \qquad \frac{dt}{dv} = -\frac{1}{g(1 + k^2v)},$$

$$(b) \qquad \frac{dt}{dv} = -\frac{1}{g(1 - k^2v^2)}$$

donde se ha hecho $\sqrt{r/mg} = k$. Con la ayuda de los métodos dados en el capítulo 3 pueden realizarse inmediatamente las integraciones, obteniéndose

$$(a) \qquad t = -\frac{1}{gk^2}\log(1 + k^2v) + t_0,$$

$$(b) \qquad t = \frac{1}{2gk}\log\frac{1 - kv}{1 + kv} + t_0.$$

Resolviendo para v estas ecuaciones se tiene

$$(a) \qquad v = -\frac{1}{k^2}(1 - e^{-gk^2(t-t_0)}).$$

$$(b) \qquad v = -\frac{1}{k}\frac{1 - e^{-2gk(t-t_0)}}{1 + e^{-2gk(t-t_0)}} = -\frac{1}{k}\tanh[gk(t - t_0)]$$

Estas ecuaciones revelan ya una propiedad importante del movimiento. La velocidad no crece con el tiempo más allá de toda cota, sino que tiende a un límite definido que depende de la masa m y de la constante r (la que, a su vez, depende de la forma del cuerpo que cae y de la densidad del aire). En efecto,

$$(a) \qquad \lim_{t\to\infty} v(t) = -\frac{1}{k^2} = -\frac{mg}{r},$$

$$(b) \qquad \lim_{t\to\infty} v(t) = -\frac{1}{k} = -\sqrt{\frac{mg}{r}}$$

La resistencia de fricción para la velocidad límite contrarresta justamente al atracción gravitacional. Una segunda integración realizada en las expresiones para $v(t) = \dot{s}$, nuevamente con la ayuda de los métodos del capítulo 3, conduce a los resultados (que pueden verificarse por derivación)

$$(a) \qquad s(t) = -\frac{1}{k^2}(t - t_0) - \frac{1}{gk^4}(e^{-gk^2(t-t_0)} - 1) + c$$

$$(b) \qquad s(t) = -\frac{1}{gk^2}\log[\cosh kg(t - t_0)] + c,$$

donde c es una constante de integración. Aquí t_0 es el instante para el cual la velocidad de la partícula tendría que ser 0, y c es la altura correspondiente. Las dos constantes c y t_0 pueden también ser fácilmente relacionadas con la velocidad y la posición en cualquier otro instante t_1 si se consideran aquellas cantidades como condiciones iniciales.

4.6 El tipo más simple de vibración elástica: movimiento de un resorte

Como un segundo ejemplo —de mayor significación— considérese el caso de una partícula que se mueve sobre el eje x bajo la acción de una fuerza elástica que la atrae hacia el origen. Por lo que respecta a la fuerza elástica se considerará que su magnitud es proporcional a la distancia desde el origen y que está siempre dirigida hacia ese punto. En otras palabras, se tomará la fuerza igual a $-kx$, donde el coeficiente k es una medida de la rigidez de la conexión elástica. Puesto que k es positiva, la fuerza es negativa cuando x es positiva y positiva cuando x es negativa. La ley de Newton en este caso establece que

$$(48) \qquad m\ddot{x} = -kx.$$

Esta ecuación diferencial por sí sola no determina completamente el movimiento, sino que para un instante dado, por ejemplo $t = 0$, pueden fijarse arbitrariamente la posición $x(0) = x_0$ y la velocidad $\dot{x}(0) = v_0$ iniciales; esto es, en lenguaje físico, que la partícula puede iniciar su movimiento desde cualquier posición y con cualquier velocidad, quedando determinado el movimiento subsecuente por la ecuación diferencial. Matemáticamente esto queda expresado por el hecho de que la solución general de la ecuación diferencial contiene dos constantes de integración cuyos valores, al principio indeterminados, pueden encontrarse mediante las condiciones iniciales. Esto se demostrará de inmediato.

Podemos establecer fácilmente dicha solución de una manera directa. Si hacemos $\omega = \sqrt{k/m}$, nuestra ecuación diferencial se convierte en $d^2x/dt^2 = -\omega^2x$. La substitución $\tau = \omega t$ de la variable independiente reduce esta ecuación a la forma $d^2x/d\tau^2 = -x$, discutida en el capítulo 3, p. 331. Así la ecuación diferencial es satisfecha por todas las funciones

$$x(t) = c_1\cos\omega t + c_2\operatorname{sen}\omega t,$$

lo cual puede también verificarse de inmediato por derivación (c_1 y c_2 denotan constantes arbitriamente escogidas). En el capítulo 3, p. 332, vimos que no existen otras soluciones para dicha ecuación diferencial, y, por lo tanto, cualquier movimiento originado por la acción de la fuerza

elástica está dado por esa expresión. Esta puede fácilmente ponerse en la forma

$$x(t) = a \operatorname{sen} \omega(t - \delta) = -a \operatorname{sen} \omega\, \delta \cos \omega t + a \cos \omega\, \delta \operatorname{sen} \omega t.$$

Sólo se necesita escribir $-a \operatorname{sen} \omega\, \delta = c_1$ y $a \cos \omega\, \delta = c_2$, introduciendo así las nuevas constantes a y δ en vez de c_1 y c_2. Movimientos de este tipo son llamados *sinusoidales* o *armónicos simples*. Estos movimientos son periódicos: cualquier *estado* [esto es, posición $x(t)$ y velocidad $\dot{x}(t)$] se repite después de un tiempo $T = 2\pi/\omega$, llamado *período*, pues las funciones $\operatorname{sen} \omega t$ y $\cos \omega t$ tienen período T. El número a se llama el *desplazamiento máximo* o *amplitud* de la oscilación. El número $1/T = \omega/2\pi$ se llama la *frecuencia* de la oscilación; mide el número de oscilaciones por unidad de tiempo. En el capítulo 8 regresaremos a la teoría de las oscilaciones.

4.7 Movimiento sobre una curva dada

a. La ecuación diferencial y su solución

Volvamos ahora a la forma general del problema relativo al movimiento a lo largo de una cierta curva bajo la acción de una fuerza $mf(s)$ arbitraria pero prefijada. Determinaremos $s(t)$ como función de t usando la ecuación diferencial [Ec. (45), p. 419]

$$\ddot{s} = f(s),$$

donde $f(s)$ es una función dada.[1] Esta ecuación diferencial en s puede resolverse completamente mediante el siguiente artificio.

Consideremos cualquier función primitiva $F(s)$ de $f(s)$, esto es $F'(s) = f(s)$, y multipliquemos ambos lados de la ecuación $\ddot{s} = f(s) = F'(s)$ por $\dot{s}$. Podemos entonces escribir el miembro de la izquierda en la forma $d(\dot{s}^2/2)\,dt$, como se ve de inmediato derivando la expresión $\dot{s}^2$. Por otra parte, si en $F(s)$ consideramos la cantidad s como una función de t, por la regla de la cadena de la derivación, el lado de la derecha, $F'(s)\dot{s}$, es la derivada de $F(s)$ con respecto al tiempo t. Por lo tanto, se tiene inmediatamente

$$\frac{d}{dt}\left(\frac{1}{2}\dot{s}^2\right) = \frac{d}{dt}F(s),$$

o bien, integrando,

$$\tfrac{1}{2}\dot{s}^2 = F(s) + c,$$

donde c denota una constante por determinar.

Hemos, en este punto, llegado a una ecuación que sólo involucra la función $s(t)$ y su *primera* derivada. (Más adelante interpretaremos esta

[1] Nuestra ecuación original de movimiento sobre una curva era $m\ddot{s} = f(s)$; sin embargo, podemos siempre escribir la función $f(s)$ en la forma $mf(s)$, obteniendo la forma más simple usada aquí para la ecuación.

ecuación como la que expresa la conservación de la energía durante el movimiento.) Escríbase esta ecuación en la forma $ds/dt = \sqrt{2[F(s) + c]}$. Vemos que no puede encontrarse inmediatamente por integración s como función de t. Sin embargo, obtenemos una solución del problema si primero nos concretamos a encontrar la función inversa $t(s)$, esto es, el tiempo que tarda la partícula en alcanzar una posición definida s. Para $t(s)$ tenemos la ecuación

$$\frac{dt}{ds} = \frac{1}{\sqrt{2[F(s) + c]}};$$

y así, la derivada de la función $t(s)$ es conocida y se tiene

$$t = \int \frac{ds}{\sqrt{2[F(s) + c]}} + c_1,$$

donde c_1 es otra constante de integración. El problema se habrá resuelto tan pronto como se realice esta última integración, pues aunque no se determine la posición s como una función de t, se encuentra, por el contrario, el tiempo t como una función de la posición s. El hecho de que todavía se disponga de dos constantes de integración, c y c_1, nos permite hacer que la solución general satisfaga condiciones iniciales específicas.

La discusión general puede ilustrarse con nuestro primer ejemplo sobre vibraciones elásticas si identificamos x con s. En este caso, $f(s) = -\omega^2 s$ y, por ejemplo, $F(s) = -\frac{1}{2}\omega^2 s^2$. Obtenemos entonces

$$\frac{dt}{ds} = \frac{1}{\sqrt{2c - \omega^2 s^2}},$$

y además,

$$t = \int \frac{ds}{\sqrt{2c - \omega^2 s^2}} + c_1.$$

Esta integral puede evaluarse fácilmente introduciendo $\omega s/\sqrt{2c}$ como una nueva variable: así obtenemos

$$t = \frac{1}{\omega} \operatorname{arc\,sen} \frac{\omega s}{\sqrt{2c}} + c_1,$$

o bien, tomando la función inversa,

$$s = \frac{\sqrt{2c}}{\omega} \operatorname{sen} \omega(t - c_1).$$

Llegamos así a exactamente la misma fórmula obtenida anteriormente para la solución.

Este ejemplo ilustra también lo que significan las constantes de integración y cómo se determinan. Si, por ejemplo, se requiere que en el

tiempo $t = 0$ la partícula esté en el punto $s = 0$ y que su velocidad en ese instante sea $\dot{s}(0) = 1$, se obtienen las dos ecuaciones

$$0 = \frac{\sqrt{2c}}{\omega} \operatorname{sen} \omega c_1, \qquad 1 = \sqrt{2c} \cos \omega c_1,$$

a partir de las cuales se encuentra que las constantes tienen los valores $c_1 = 0$, $c = \frac{1}{2}$. Las constantes de integración c y c_1 pueden determinarse exactamente de la misma manera cuando se prescriben arbitrariamente la posición s_0 y la velocidad $\dot{s}_0$ iniciales (en el tiempo $t = 0$).

b. Partícula que se desliza hacia abajo sobre una curva

El caso de una partícula que se desliza sin fricción sobre una curva bajo la acción de la gravedad puede tratarse en forma muy simple mediante el método descrito anteriormente. Encontramos ya en la p. 419 la ecuación de movimiento correspondiente a este caso:

$$\ddot{s} = -g \frac{dy}{ds},$$

donde los puntos indican derivación con respecto a t. El miembro derecho de esta ecuación es una función conocida de s, puesto que conocemos la curva y podemos por consiguiente considerar x e y como funciones conocidas de s.

Multipliquemos ambos miembros de esta ecuación por $\dot{s}$, como en la última sección. El miembro izquierdo se convierte entonces en la derivada con respecto a t de $\frac{1}{2}\dot{s}^2$. Si en la expresión $y(s)$ consideramos s como una función de t, el miembro derecho de nuestra ecuación es la derivada con respecto a t de $-gy$. Integrando tendremos entonces

$$\tfrac{1}{2}\dot{s}^2 = -gy + c,$$

donde c es una constante de integración. Para encontrar el significado de esta constante supongamos que en el tiempo $t = 0$ nuestra partícula está en el punto de la curva con coordenadas x_0 e y_0, y que en ese instante su velocidad es cero, esto es, $\dot{s}(0) = 0$. Entonces, haciendo $t = 0$ tenemos inmediatamente $-gy_0 + c = 0$, de manera que

$$\tfrac{1}{2}\dot{s}^2 = g(y_0 - y).$$

Puesto que $\dot{s}^2$ nunca es negativo, la altura y de la partícula jamás excede al valor y_0 y sólo es igual a él cuando la velocidad es cero. Mientras más abajo se encuentre la partícula, mayor será su velocidad. Ahora, en vez de considerar s como una función de t, consideraremos la función inversa $t(s)$. Para esta función obtenemos inmediatamente

$$\frac{dt}{ds} = \pm \frac{1}{\sqrt{2g(y_0 - y)}}$$

que es equivalente a

$$t = c_1 \pm \int \frac{ds}{\sqrt{2g(y_0 - y)}},$$

donde c_1 es una nueva constante de integración. En relación con el signo de la raíz cuadrada, que es el mismo que el de $\dot{s}$, observamos que si la partícula se mueve a lo largo de un arco que en todas partes está por debajo de y_0 excepto en los extremos, el signo no puede cambiar. Pues el signo de $\dot{s}$ sólo puede cambiar donde $\dot{s} = 0$, esto es, donde $y - y_0 = 0$. De esta manera, la partícula solamente puede "regresar" en puntos de la curva de máxima elevación. En vez de usar la longitud de arco s como parámetro de referencia para la curva, puede usarse cualquier otro parámetro θ tal que $x = \phi(\theta)$, $y = \psi(\theta)$. Introduciendo θ como variable independiente se obtiene

$$t = c_1 \pm \int \frac{ds}{d\theta} \frac{d\theta}{\sqrt{2g(y_0 - y)}} = c_1 \pm \int \frac{x'^2 + y'^2}{2g(y_0 - y)}\, d\theta,$$

donde las funciones $x' = \phi'(\theta)$, $y' = \psi'(\theta)$, e $y = \psi(\theta)$ son conocidas. Para determinar la constante de integración c_1 observamos que para $t = 0$ el parámetro θ tendrá un valor θ_0. Esto nos da inmediatamente la solución en la forma

$$t = \pm \int_{\theta_0}^{\theta} \sqrt{\frac{x'^2 + y'^2}{2g(y_0 - y)}}\, d\theta. \tag{49}$$

Nótese que esta función representa el tiempo que tarda la partícula para moverse del punto con valor paramétrico θ_0 al punto con valor paramétrico θ. La función inversa $\theta(t)$ correspondiente nos permite describir com-

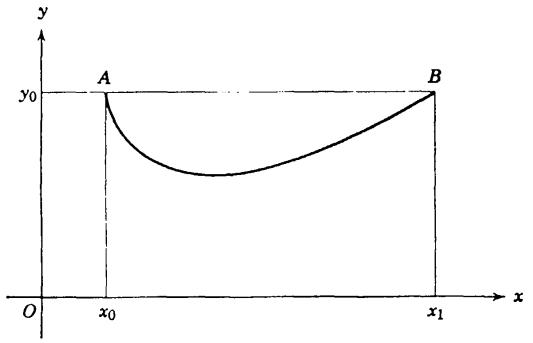


Figura 4.45

pletamente el movimiento: para cada instante t podemos determinar el punto $x = \phi[\theta(t)]$, $y = \psi[\theta(t)]$ por el que la partícula está pasando.

c. Discusión del movimiento

De las ecuaciones recién encontradas puede deducirse, aún sin una expresión explícita para el resultado de la integración, la naturaleza general del movimiento mediante un simple razonamiento intuitivo. Supongamos que nuestra curva es del tipo mostrado en la Fig. 4.45, esto es, que consiste de un arco convexo hacia abajo; y tomemos s creciente de izquierda a derecha. Si inicialmente soltamos la partícula desde el punto A de coordenadas $x_0 = \phi(\theta_0)$, $y_0 = \psi(\theta_0)$, correspondientes a $\theta = \theta_0$, la velocidad aumenta, pues la aceleración $\ddot{s}$ es positiva. La partícula viaja de A al punto más bajo con velocidad siempre creciente. Sin embargo, después de pasar por el punto más bajo la aceleración es negativa ya que el miembro derecho, $-g\,dy/ds$, de la ecuación de movimiento es negativo. Por lo tanto, la velocidad decrece. De la ecuación $\dot{s}^2 = 2g(y_0 - y)$ se ve de inmediato que la velocidad alcanza el valor cero cuando la partícula llega al punto B cuya altura es la misma que la de la posición inicial A. Puesto que la aceleración es todavía negativa, el movimiento debe invertirse en ese punto y la partícula oscilará de regreso hacia A; y esta acción se repetirá indefinidamente. (El lector recordará que la fricción no se ha tomado en cuenta.) En este movimiento oscilatorio el tiempo que tarda el punto en regresar de B a A debe claramente ser el mismo que el que tarda en moverse de A a B, pues para alturas iguales se tienen valores iguales de $|\dot{s}|$. Si denotamos con T el tiempo requerido para un viaje completo de A a B y viceversa, el movimiento obviamente será periódico con período T. Si θ_0 y θ_1 son los valores del parámetro correspondientes a los puntos A y B, respectivamente, el semiperíodo está dado por la expresión

$$\frac{T}{2} = \frac{1}{\sqrt{2g}}\left|\int_{\theta_0}^{\theta_1} \frac{x'^2 + y'^2}{y_0 - y}\,d\theta\right| = \frac{1}{\sqrt{2g}}\left|\int_{\theta_0}^{\theta_1} \frac{\phi'^2(\theta) + \psi'^2(\theta)}{\psi(\theta_0) - \psi(\theta)}\,d\theta\right|. \tag{50}$$

Si θ_2 es el valor del parámetro correspondiente al punto más bajo de la curva, el tiempo que la partícula tarda en caer de A hasta ese punto es

$$\frac{1}{\sqrt{2g}}\left|\int_{\theta_0}^{\theta_2} \sqrt{\frac{x'^2 + y'^2}{y_0 - y}}\,d\theta\right|.$$

d. El péndulo ordinario

El ejemplo más sencillo está dado por el llamado péndulo simple. En este caso la curva a considerar es una circunferencia de radio fijo L:

$$x = L\,\mathrm{sen}\,\theta, \qquad y = -L\cos\theta,$$

donde el ángulo θ se mide en el sentido positivo a partir de la posición de reposo. De la expresión general (50), y usando el teorema de adición para el coseno, se obtiene inmediatamente

$$T = \sqrt{\frac{2L}{g}}\int_{-\theta_0}^{\theta_0} \frac{d\theta}{\sqrt{\cos\theta - \cos\theta_0}} = \sqrt{\frac{L}{g}}\int_{-\theta_0}^{\theta_0} \frac{d\theta}{\sqrt{\mathrm{sen}^2\frac{\theta_0}{2} - \mathrm{sen}^2\frac{\theta}{2}}},$$

donde θ_0 $(0 < \theta_0 < \pi)$ denota la amplitud de oscilación del péndulo, esto es, la posición angular desde la que se suelta la partícula con velocidad cero en el tiempo $t = 0$.[1] Mediante la substitución

$$u = \frac{\mathrm{sen}\,(\theta/2)}{\mathrm{sen}\,(\theta_0/2)}, \qquad \frac{du}{d\theta} = \frac{\cos(\theta/2)}{2\,\mathrm{sen}\,(\theta_0/2)},$$

nuestra expresión para el período de oscilación del péndulo se transforma en

$$T = 2\sqrt{\frac{L}{g}}\int_{-1}^{1} \frac{du}{\sqrt{(1-u^2)\left(1 - u^2\,\mathrm{sen}^2\left(\frac{\theta_0}{2}\right)\right)}}.$$

Tenemos así expresado el período de oscilación del péndulo mediante una *integral elíptica* (ver p. 318).

Si se supone que la amplitud de la oscilación es pequeña, de manera que con suficiente aproximación se puede reemplazar el segundo factor del subradical por 1, se obtiene la expresión

$$2\sqrt{\frac{L}{g}}\int_{-1}^{1} \frac{du}{\sqrt{1-u^2}}$$

como una aproximación para el período de oscilación. Esta última integral puede evaluarse usando la fórmula 13 de nuestra tabla de integrales (p. 284); así se obtiene la expresión $2\pi\sqrt{L/g}$ como un valor aproximado para T. Hasta este orden de aproximación, el período es independiente de θ_0, o sea, de la amplitud de oscilación del péndulo. Es claro que el

[1] Hemos supuesto aquí que la velocidad se hace cero en algún momento durante el movimiento. Esto excluye el tipo de movimiento de salto del péndulo, para el cual θ no es periódica y varía monótonamente para todo t.

período exacto es mayor y aumenta con θ_0. Puesto que en el intervalo de integración

$$1 \geq 1 - u^2 \operatorname{sen}^2 \frac{\theta_0}{2} \geq 1 - \operatorname{sen}^2 \frac{\theta_0}{2} = \cos^2 \frac{\theta_0}{2},$$

se encuentran las siguientes estimaciones para el período

$$2\pi \sqrt{\frac{L}{g}} \leq T \leq \frac{1}{\cos(\theta_0/2)} 2\pi \sqrt{\frac{L}{g}}.$$

Para ángulos $\theta_0 < 10°$ se tiene $1/(\cos \theta_0/2) \leq \operatorname{seg} 5° < 1.004$, de manera que el período estará dado por la fórmula $2\pi\sqrt{L/g}$ con un error relativo menor que el $\frac{1}{4}\%$. Para una aproximación más fina de la integral elíptica que expresa T véase la sección 7.6*f*.

e. El péndulo cicloidal

El hecho de que el período de oscilación de un péndulo ordinario no es estrictamente independiente de la amplitud de oscilación hizo que Christian Huygens, en sus constantes esfuerzos por construir relojes precisos, buscase una curva C para la cual el período de oscilación fuese independiente de la posición sobre C desde la cual la partícula oscilante inicia su movimiento.[1] Huygens reconoció que la cicloide es tal curva.

Para que una partícula sea realmente capaz de oscilar sobre una cicloide, las cúspides de esta curva deben apuntar en dirección opuesta a la de la gravedad; esto es, la ciclode considerada anteriormente (p. 347) debe ser reflejada con respecto del eje x (cfr. Fig. 4.2, p. 348). Por consiguiente, escribimos las ecuaciones de la cicloide en la forma

$$x = a(\theta + \pi + \operatorname{sen} \theta),$$
$$y = -a(1 + \cos \theta),$$

la cual involucra también un cambio del parámetro t por $\theta + \pi$ (Fig. 4.46). El tiempo que la partícula tarda en ir de un punto con altura

$$y_0 = -a(1 + \cos \theta_0) \qquad (0 < \theta_0 < \pi)$$

al punto más bajo y luego subir nuevamente a la altura y_0 es, por la fórmula (50) de la p. 428,

$$\frac{T}{2} = \sqrt{\frac{1}{2g}} \int_{-\theta_0}^{\theta_0} \sqrt{\frac{x'^2 + y'^2}{y_0 - y}}\, d\theta = \sqrt{\frac{2a}{g}} \int_{-\theta_0}^{\theta_0} \frac{\cos(\theta/2)}{\sqrt{\cos\theta - \cos\theta_0}}\, d\theta.$$

[1] Las oscilaciones se dice entonces que son *isócronas*.

Usando exactamente las mismas substituciones que para el período del péndulo simple se llega a la integral

$$\frac{T}{2} = 2\sqrt{\frac{a}{g}} \int_{-1}^{1} \frac{du}{\sqrt{1 - u^2}},$$

y, por consiguiente,

$$T = 4\pi \sqrt{\frac{a}{g}}.$$

El período de oscilación T, por lo tanto, es, sin duda, independiente de la amplitud θ_0. En la p. 446 se describirá una manera sencilla de lograr, mediante un hilo, que una partícula se mueva realmente sobre una cicloide.

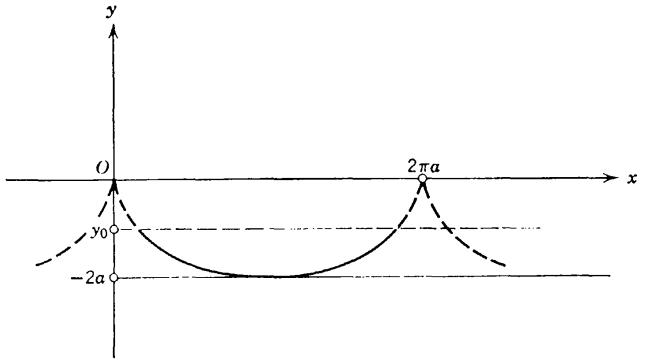


Figura 4.46 Trayectoria descrita por un péndulo cicloidal.

*4.8 Movimiento de un campo gravitacional

Como un ejemplo de movimiento no restringido, consideremos una partícula que se mueve en el campo gravitacional de una masa atractiva.

a. Ley universal de la gravitación de Newton

La descripción del movimiento de los planetas dada por Kepler, basada en las precisas observaciones de Tycho Brahe, condujo a Newton a formular su ley general de la atracción gravitacional entre dos partículas cualesquiera. Sean $P_0 = (x_0, y_0)$ y $P = (x, y)$ dos partículas de masas m_0 y m, respectivamente. Sea $r = \sqrt{(x - x_0)^2 + (y - y_0)^2}$ la distancia entre las partículas. Entonces P_0 ejerce sobre P una fuerza $\mathbf{F}$ que tiene la dirección de $\overrightarrow{PP_0}$ y la magnitud $|\mathbf{F}| = \gamma m_0 m / r^2$, donde γ es la *"constante universal de la gravitación*. Puesto que entonces $\mathbf{F}$ solamente puede diferir en un factor positivo del vector $\overrightarrow{PP_0}$, cuya magnitud es r, debe tenerse

$$\mathbf{F} = \frac{\gamma m_0 m}{r^3}\overrightarrow{PP_0} = \left(\frac{\gamma m_0 m(x_0 - x)}{r^3} \quad \frac{\gamma m_0 m(y_0 - y)}{r^3} \right).$$

Esta ley de atracción se refiere a *partículas*, esto es, a cuerpos que pueden considerarse concentrados en puntos, despreciando la extensión real de los mismos (Fig. 4.47). La validez de tal hipótesis es lo suficientemente plausible para cuerpos celestes, cuyas distancias mutuas son enormes comparadas con sus diámetros. Newton extendió considerablemente el rango de aplicación de esta ley demostrando que la misma también describe la

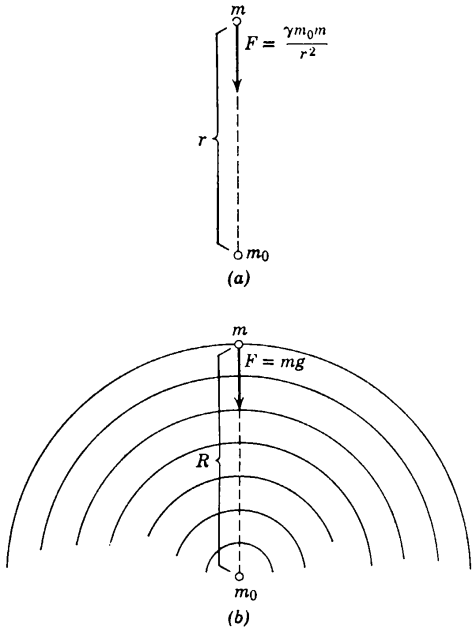


Figura 4.47 *a*) Atracción newtoniana entre dos partículas. *b*) Atracción gravitacional de la Tierra.

atracción que un cuerpo de masa m_0 y extensión no nula ejerce sobre una partícula de masa m, suponiendo que el cuerpo es una esfera cuya densidad es constante o, más generalmente, suponiendo que el cuerpo está constituido por capas esféricas concéntricas de densidad constante. En este caso la atracción del cuerpo sobre una partícula P situada fuera de él es la misma que si toda la masa m_0 del cuerpo estuviese concentrada en su centro P_0 (Fig. 4.47). La Tierra puede, con bastante aproximación, considerarse como constituida por capas concéntricas de densidad constante, de manera que su atracción sobre una partícula de masa m situada en la superficie terrestre está diriigda hacia el centro P_0 de la Tierra (esto es, verticalmente hacia abajo para un observador) y tiene una magnitud igual a $\gamma m_0 m/R^2$, donde R es el radio de la Tierra y m_0 su masa. Se puede identificar $\gamma m_0 m/R^2$ con mg, siendo g la aceleración gravitacional (ver p. 416). En otras palabras, $g = \gamma m_0/R^2$.

A partir de la ley fundamental de Newton se encuentran las ecuaciones de movimiento para una partícula P de masa m que se mueve bajo la influencia de la atracción de una masa m_0 situada en P_0:

$$x = \frac{\gamma m_0(x_0 - x)}{r^3}, \qquad y = \frac{\gamma m_0(y_0 - y)}{r^3}.$$

Supóngase ahora, para simplificar, que m_0 es mucho mayor que m, que los efectos de la atracción de P sobre P_0 pueden despreciarse y que P_0 puede considerarse en reposo. Esta podría ser, por ejemplo, la situación de una pareja de cuerpos como el Sol y un planeta, o bien la Tierra y un objeto sobre su superficie. Tomando el origen de coordenadas en P_0 se tienen para $P = (x, y)$ las ecuaciones de movimiento

$$\ddot{x} = -\frac{\gamma m_0 x}{r^3}, \qquad \ddot{y} = -\frac{\gamma m_0 y}{r^3} \tag{51}$$

con $r = \sqrt{x^2 + y^2}$.

b. Movimiento circular en torno al centro de atracción

No se intentará encontrar la solución más general de estas ecuaciones diferenciales (lo que correspondería, como es bien sabido, al movimiento sobre una trayectoria en forma de sección cónica con un foco en el centro de atracción). En vez de ello, se considerarán solamente los tipos más simples de movimiento compatibles con estas ecuaciones, como son los movimientos circulares uniformes en torno al origen y los movimientos a lo largo de un radio que parte de ese mismo punto. Para un movimiento circular uniforme de P sobre la circunferencia de radio a con centro en el origen, se tiene $r = a$ y

$$x = a\cos\omega t, \qquad y = a\,\mathrm{sen}\,\omega t,$$

donde ω es una constante. El *período* T del movimiento, esto es, el tiempo que tarda P en regresar a la misma posición, es $T = 2\pi/\omega$. Para las componentes de la velocidad se encuentra

$$\dot{x} = -a\omega\,\mathrm{sen}\,\omega t, \qquad \dot{y} = a\omega\cos\omega t,$$

de manera que la velocidad de P sobre su órbita es

$$v = \sqrt{\dot{x}^2 + \dot{y}^2} = a\omega = \frac{2\pi a}{T}. \tag{52}$$

La aceleración de P tiene por componentes

$$\ddot{x} = -a\omega^2 \cos \omega t = -\omega^2 x, \qquad \ddot{y} = -a\omega^2 \operatorname{sen} \omega t = -\omega^2 y.$$

Es claro entonces que las ecuaciones de movimiento (51) son satisfechas si

$$\omega^2 = \frac{\gamma m_0}{a^3},$$

o bien

$$a^3 = \frac{\gamma m_0}{\omega^2} = \frac{\gamma m_0}{4\pi^2} T^2. \tag{53}$$

Esta es precisamente la *tercera ley de Kepler* para el caso especial de movimiento circular, según la cual *los cubos de las distancias de los planetas al Sol son proporcionales a los cuadrados de sus períodos.*

Pueden darse algunas ilustraciones sencillas de la ley de Kepler para el caso en que el cuerpo atractivo es la Tierra, con m_0 su masa y R su radio. Observando que en este caso $\gamma m_0 = gR^2$, se tiene

$$a^3 = \frac{gR^2}{4\pi^2} T^2.$$

Para un satélite que circunvale la Tierra al nivel de la cima de un árbol (despreciando, por supuesto, la resistencia del aire) se tiene $a = R \sim 3963$ millas. A partir de nuestra fórmula se encuentra para el período del satélite el valor

$$T = 2\pi \sqrt{\frac{R}{g}} \sim 1.4 \text{ horas}$$

y para su velocidad orbital

$$v = \frac{2\pi R}{T} = \sqrt{Rg} \sim 8.231 \text{ metros/seg.} \tag{54}$$

Se puede comparar el valor de T para el satélite que circunvala la Tierra con el período de 27.32 días de la Luna, o sea, el tiempo después del cual ésta regresa a la misma posición entre las estrellas ("mes sideral"). Por la ley de Kepler, la razón de la distancia a entre la Luna y el centro de la Tierra al radio de ésta sería igual a la potencia $\frac{2}{3}$ de la relación de períodos. Esto da para dicha distancia el valor

$$a = \left(\frac{27.32 \times 24}{1.4}\right)^{2/3} R \sim 60R \sim 400{,}000 \text{ Km.},$$

que concuerda bien con el valor promedio real de esa distancia.

c. Movimiento radial — Velocidad de escape

El segundo tipo de movimiento que consideraremos es el de una partícula que se mueve a lo largo de un rayo, por ejemplo, el eje x, desde el centro de atracción. Aquí $y = 0$, $x = r$, de manera que las ecuaciones de movimiento se reducen a

$$\ddot{x} = -\frac{\gamma m_0}{x^2}.$$

Siguiendo el procedimiento general para ecuaciones del tipo $\ddot{s} = f(s)$, multiplicamos ambos miembros de esa ecuación por $\dot{x}$ y obtenemos

$$\dot{x}\ddot{x} = -\gamma m_0 \frac{\dot{x}}{x^2}$$

o bien

$$\frac{d}{dt}\left(\frac{1}{2}\dot{x}^2\right) = \frac{d}{dt}\left(\frac{\gamma m_0}{x}\right).$$

Así, la expresión

$$\frac{1}{2}\dot{x}^2 - \frac{\gamma m_0}{x}$$

tiene un valor constante h durante el movimiento. (Más adelante reconoceremos este hecho como un ejemplo de la ley de la conservación de la energía.) Si introducimos x, en vez de t, como variable independiente, obtenemos

$$\frac{dt}{dx} = \frac{1}{\dot{x}} = \pm \frac{1}{\sqrt{2h + (2\gamma m_0/x)}},$$

que, por integración, conduce a

$$t = t_0 \pm \int_{x_0}^{x} \frac{d\xi}{\sqrt{2h + (2\gamma m_0/\xi)}}.$$

No intentaremos realizar la integración, la cual puede efectuarse fácilmente con ayuda de los métodos desarrollados en el capítulo 3. Para una partícula que se suelta en el tiempo $t_0 = 0$ desde la distancia x_0 con velocidad nula, se tiene $h = -\gamma m_0/x_0$. El tiempo que tarda dicha partícula en caer sobre la partícula atractiva ($x = 0$) es entonces

$$t = \int_0^{x_0} \frac{d\xi}{\sqrt{2\gamma m_0(1/\xi - 1/x_0)}} = \frac{\pi}{2}\sqrt{\frac{{x_0}^3}{2\gamma m_0}}.$$

Por la ley de Kepler, esto es $\sqrt{\frac{1}{32}}$ veces el tiempo que necesitaría la partícula para dar una vuelta alrededor del centro de atracción a la distancia x_0 [ver Ec. (53), p. 434].

La relación

$$\frac{1}{2}\dot{x}^2 - \frac{\gamma m_0}{x} = h$$

tiene una consecuencia interesante cuando se investigan las condiciones en las que una partícula puede escapar al infinito. Puesto que $\frac{1}{2}\dot{x}^2 \geq 0$, se encuentra que para $x \to \infty$ la constante h debe ser no negativa y, por lo tanto, que $\frac{1}{2}\dot{x}^2 - \gamma m_0/x \geq 0$ durante todo el movimiento. Como caso especial, una partícula que inicia su movimiento desde la distancia $x = a$ con velocidad v puede escapar al infinito sólo si $\frac{1}{2}v^2 - \gamma m_0/a \geq 0$. El valor más bajo posible de la velocidad v que permitirá a la partícula escapar al infinito es entonces $v = \sqrt{2\gamma m_0/a}$. Esta es la *velocidad de escape* v_e. Para una masa que parte de la superficie de la Tierra para escapar al infinito, esto es, escapar de la atracción gravitacional, se tiene $a = R$, $\gamma m_0 = gR^2$, de manera que

$$v_e = \sqrt{2gR} \sim 11{,}300 \text{ metros/seg.}$$

Por consiguiente [cfr. (54), p. 434], la velocidad de escape es justamente $\sqrt{2}$ veces la velocidad requerida para mantener un satélite en órbita circular cerca de la Tierra. Un meteorito que cayese a la Tierra desde el infinito también tendría la velocidad v_e en el momento del impacto, si se desprecian la resistencia del aire y el movimiento de la Tierra en su órbita.

4.9 Trabajo y energía

a. Trabajo realizado por fuerzas durante un movimiento

El *concepto de trabajo* arroja nueva luz sobre las consideraciones de la última sección y sobre muchos otros tópicos en mecánica y física.

Consideremos otra vez una partícula que se mueve sobre una curva bajo la acción de un fuerza que actúa a lo largo de dicha curva, y supongamos que la posición de la partícula está dada por la longitud de arco medida desde cualquier punto inicial fijo. En general, la fuerza que actúa en la dirección del movimiento será a su vez una función de s. Esta función tendrá valores positivos cuando el sentido de la fuerza sea el mismo que el de los valores crecientes de s, y tendrá valores negativos en el caso opuesto.

Si la magnitud de la fuerza es constante a lo largo de toda la trayectoria, entonces se llama *trabajo realizado por la fuerza* al producto de la fuerza por la distancia $(s_1 - s_0)$ recorrida, siendo s_1 el punto final del movimiento y s_0 el punto inicial. Si la fuerza no es constante, definimos el trabajo mediante un proceso de límite. Subdividimos el intervalo de s_0 a s_1 en n subintervalos iguales o diferentes y observamos que si esos subintervalos son pequeños, la fuerza en cada uno de ellos es casi constante; si σ_ν

es un punto escogido arbitrariamente en el ν-ésimo subintervalo, entonces en este subintervalo la fuerza será aproximadamente $f(\sigma_\nu)$. Si la fuerza en el ν-ésimo subintervalo fuese exactamente $f(\sigma_\nu)$, el trabajo realizado por ella sería precisamente

$$\sum_{\nu=1}^{n} f(\sigma_\nu)\, \Delta s_\nu,$$

donde Δs_ν denota, como es costumbre, la longitud del ν-ésimo subintervalo. Si pasamos ahora al límite dejando que n crezca más allá de toda cota en tanto que la longitud del subintervalo mayor tiende a cero, por la definición de integral, la anterior suma tiende entonces a

$$W = \int_{s_0}^{s_1} f(s)\, ds,$$

cantidad que llamamos, en forma natural, el *trabajo* realizado por la fuerza.

Si la fuerza y el movimiento tienen el mismo sentido, el trabajo realizado por ella es positivo; y decimos entonces que *la fuerza efectúa trabajo*. Por otra parte, si la fuerza y el movimiento tienen sentidos opuestos, el trabajo realizado por ella es negativo; y en este caso decimos que *se efectúa trabajo contra la fuerza.*[1]

Si consideramos la coordenada de posición s como una función del tiempo t, de manera que la fuerza $f(s) = p$ sea también una función de t, entonces en una plano con coordenadas rectangulares s y p podemos representar el punto de coordenadas $s = s(t)$ y $p = p(t)$ como una función del tiempo. Este punto describirá una curva que podemos llamar el *diagrama de trabajo* del movimiento. Si se trata de un movimiento periódico, como el de cualquier máquina, entonces después de un cierto tiempo T (un período) el punto móvil $(s(t), p(t))$ regresará a la misma posición; esto es, el diagrama de trabajo será una curva cerrada. En este caso la curva puede consistir simplemente de uno y el mismo arco, recorrido primero hacia adelante y después hacia atrás. Esto ocurre, por ejemplo, en el caso de oscilaciones elásticas. Sin embargo, también es posible que la curva sea de un tipo más general de curva cerrada, limitando un área. Este es el caso, por ejemplo, de las máquinas en que la presión sobre un pistón no es la misma durante la carrera hacia adelante que durante la carrera hacia atrás. El trabajo realizado en un *ciclo*, esto es, en el tiempo T, estará entonces dado simplemente por el negativo del área, con signo negativo, encerrada por el diagrama de trabajo, o sea, en otras palabras, por la integral

[1] Obsérvese que aquí debemos caracterizar cuidadosamente la fuerza de la que hablamos. Por ejemplo, al levantar un peso el trabajo realizado por la fuerza de gravedad es negativo: Se hace trabajo contra la gravedad. Pero desde el punto de vista de la persona que levanta el peso, el trabajo realizado es positivo, pues deberá ejercer una fuerza opuesta a la de gravedad.

$$\int_{t_0}^{t_0+T} p(t)\,\frac{ds}{dt}\,dt,$$

donde el intervalo de tiempo de t_0 a $t_0 + T$ representa exactamente un período del movimiento. Si la frontera del área se recorre en el sentido positivo, el trabajo realizado es negativo; y si se recorre en el sentido negativo, el trabajo realizado será positivo. Si la curva consiste de varias espiras, algunas recorridas positivamente y otras negativamente, el trabajo realizado es la suma de las áreas de las espiras, con su signo cambiado.

Estas consideraciones son ilustradas en la práctica por el *diagrama indicador* de una máquina de vapor antigua. Mediante un dispositivo mecánico diseñado convenientemente, se hace mover un lápiz sobre una hoja de papel; el movimiento horizontal del lápiz relativo al papel es proporcional a la distancia s del pistón a su posición extrema, en tanto que el movimiento vertical es proporcional a la presión del vapor y, por consiguiente, a la fuerza total p del vapor sobre el pistón. Este describe entonces el diagrama de trabajo de la máquina sobre una escala conocida. El área de este diagrama se mide (generalmente usando un planímetro), y así se encuentra el trabajo realizado por el vapor sobre el pistón. Aquí vemos también que nuestra convención para el signo de un área, como se discutió en la p. 382, es definitivamente de interés práctico. Pues ocurre a veces, cuando la máquina funciona baja, que el vapor muy expandido en el extremo de la carrera del pistón tiene una presión menor que la requerida para su expulsión hacia el recorrido de regreso. En el diagrama esto aparece como una espira recorrida positivamente; y la máquina extrae energía del volante en vez de proporcionársela.

b. Trabajo y energía cinética. Conservación de la energía

La ley de movimiento

$$m\ddot{s} = f$$

conduce a una relación fundamental entre los cambios en velocidad durante el movimiento de una partícula a lo largo de una curva y el trabajo realizado por la fuerza f en la dirección del movimiento. Usando el mismo artificio que en los ejemplos anteriores, multiplicamos ambos miembros de la ecuación de movimiento por $\dot{s}$:

$$m\dot{s}\ddot{s} = f(s)\dot{s}.$$

Ahora bien $m\dot{s}\ddot{s} = (d/dt)\,\frac{1}{2}m\dot{s}^2 = (d/dt)\,\frac{1}{2}mv^2$, donde $v(t) = \dot{s}$ es la velocidad de la partícula. Integrando ambos lados de la ecuación con respecto a t entre los límites t_0 y t_1, obtenemos

$$\frac{1}{2}mv^2(t_1) - \frac{1}{2}mv^2(t_0) = \int_{t_0}^{t_0} f(s)\,\frac{ds}{dt}\,dt$$

$$\int_{t_0}^{s_1} f(s)\,ds = W.$$

La cantidad $\frac{1}{2}mv^2$ se llama la *energía cinética, K,* de la partícula. Por lo tanto: *El cambio en energía cinética de una partícula durante su movimiento es igual al trabajo realizado por la fuerza que actúa sobre ella en la dirección del movimiento.*

La cantidad f representa la fuerza que actúa en la dirección del movimiento, o *componente tangencial* de la fuerza. Para una fuerza $\mathbf{F} = (\rho, \sigma)$, su componente en la dirección del movimiento es

$$f = \mathbf{F} \cdot \frac{d\mathbf{R}}{ds} = \rho\,\frac{dx}{ds} + \sigma\,\frac{dy}{ds}.$$

Si ρ y σ son funciones conocidas de x e y, y si se sabe que la partícula se mueve sobre una curva $x = x(s)$, $y = y(s)$, entonces f también se vuelve una función conocida de s. En consecuencia, para calcular el trabajo

$$W = \int_{s_0}^{s_1} f(s)\,ds \tag{55}$$

cuando la partícula se mueve de una posición (x_0, y_0) a otra (x_1, y_1), en general tiene que conocerse la trayectoria sobre la cual la partícula se mueve.

En una clase importante de casos el trabajo W depende sólo de las posiciones inicial y final, pudiendo expresarse en la forma

$$W = V(x_0, y_0) - V(x_1, y_1) \tag{56}$$

en la que $V(x, y)$ es una función adecuada que se llama *energía potencial.* La fórmula que expresa que el cambio en energía cinética es igual al trabajo realizado por la fuerza puede entonces escribirse también en la forma

$$\tfrac{1}{2}mv^2(t_1) + V(x_1, y_1) = \tfrac{1}{2}mv^2(t_0) + V(x_0, y_0). \tag{57}$$

Así la cantidad $K + V$, suma de las energías mecánicas cinética y potencial, es decir, la *energía total,* no cambia durante el movimiento. Este es un ejemplo de la ley física general de la *conservación de la energía.*

Una función de energía potencial V puede fácilmente construirse para algunos de los movimientos discutidos antes. Así, para una partícula sujeta a la acción de la gravedad se tiene $\mathbf{F} = (0, -mg)$ y $f = -mg(dy/ds)$. El trabajo realizado por la fuerza de gravedad cuando la partícula pasa de la posición (x_0, y_0) a la posición (x_1, y_1) es entonces

$$W = \int_{x_0}^{s_1} -mg\,\frac{dy}{ds}\,ds = \int_{y_0}^{y_1} -mg\,dy = mgy_0 - mgy_1.$$

Vemos que W es proporcional a la diferencia de alturas entre las posiciones inicial y final. Como función de energía potencial V, podemos escoger $V = mgy$ (o, más generalmente, $V = mgy + c$, donde c es cualquier constante). La ley de la conservación de la energía establece entonces que la cantidad

$$\tfrac{1}{2}v^2 + gy$$

es constante durante el movimiento. Este hecho ha sido ya observado al investigar el movimiento de una partícula que se desliza hacia abajo sobre una curva (p. 426).

c. La atracción mutua de dos masas

Otro ejemplo de una fuerza con la cual puede asociarse una función de energía potencial V lo proporciona la atracción gravitacional $\mathbf{F}$ que una partícula $P_0 = (x_0, y_0)$ de masa m_0 ejerce sobre otra partícula $P = (x, y)$ de masa m. Aquí es

$$\mathbf{F} = \left[\frac{-\mu(x - x_0)}{r^3}, \frac{-\mu(y - y_0)}{r^3}\right],$$

donde $\mu = \gamma m_0 m$ y $r = \sqrt{(x - x_0)^2 + (y - y_0)^2}$. (Según la ley de Coulomb, el mismo tipo de fórmula describe la interacción de dos cargas eléctricas.)

La fuerza en la dirección del movimiento es entonces

$$f = -\frac{\mu}{r^3}\left[(x - x_0)\frac{dx}{ds} + (y - y_0)\frac{dy}{ds}\right] = -\frac{\mu}{r^2}\frac{dr}{ds} = \frac{d}{ds}\frac{\mu}{r},$$

puesto que

$$(x - x_0)\frac{dx}{ds} + (y - y_0)\frac{dy}{ds} = \frac{1}{2}\frac{d}{ds}[(x - x_0)^2 + (y - y_0)^2]$$

$$= \frac{1}{2}\frac{dr^2}{ds} = r\frac{dr}{ds}.$$

El trabajo realizado por la fuerza de atracción cuando la partícula P se mueve de la posición (x_1, y_1) a la posición (x_2, y_2) es

$$W = \int_{s_1}^{s_2}\left(\frac{d}{ds}\frac{\mu}{r}\right)ds = \frac{\mu}{r_2} - \frac{\mu}{r_1} = V(x_1, y_1) - V(x_2, y_2),$$

donde $V(x, y) = -\mu/r = -\mu/\sqrt{(x - x_0)^2 + (y - y_0)^2}$ es la energía potencial.

Si la partícula se mueve de la posición (x_1, y_1) hasta el infinito (lo que correspondería a $r_2 = \infty$), el trabajo realizado por la fuerza de atracción

es $-\mu/r_1$. El trabajo realizado por una fuerza opuesta que mueva la partícula hasta el infinito tiene el mismo valor numérico pero signo contrario. Por lo tanto, $\mu/r_1 = -V(x_1, y_1)$ es el trabajo que tiene que hacerse *contra* la fuerza de atracción para mover la partícula de la posición (x_1, y_1) hasta el infinito. Esta importante expresión se llama *potencial mutuo* de las dos partículas. Aquí, por consiguiente, el potencial se define como el trabajo requerido para separar completamente las dos masas atractivas, por ejemplo, el trabajo requerido para arrancar por completo un electrón de un átomo (potencial de ionización).

Si la masa atractiva P_0 se considera fija, la ley de la conservación de la energía implica que la partícula atraída P se mueve de manera tal que la expresión

$$\frac{1}{2}v^2 - \frac{\gamma m_0}{r} = h$$

(la energía total por unidad de masa m) tenga un valor constante durante el movimiento. Este hecho ha sido derivado ya para el caso especial de un movimiento puramente radial; y ahora vemos que se cumple para cualquier tipo de movimiento bajo la influencia de la atracción gravitacional. Nuevamente podemos concluir que $h \geq 0$ para una partícula que escapa al infinito; su órbita es entonces no acotada (parábola o hipérbola) en vez de acotada (elipse). La velocidad de escape es

$$v_e = \sqrt{\frac{2\gamma m_0}{r}}$$

la cual corresponde a $h = 0$; es la mínima velocidad que hará que la partícula escape al infinito desde una distancia dada r. No depende de la *dirección* en la cual se suelte la partícula sino sólo de la distancia r al centro de atracción.

d. El estiramiento de un resorte

Como un tercer ejemplo, considérese el trabajo que se hace al estirar un resorte. Con las hipótesis relativas a las propiedades elásticas del resorte que se hicieron en la p. 423, la fuerza actuante es $f = -kx$, donde k es constante. El trabajo que debe hacerse contra esta fuerza para estirar el resorte desde la posición sin deformación $x = 0$ a la posición final $x = x_1$ está dado, en consecuencia, por la integral

$$\int_0^{x_1} kx\,dx = \frac{1}{2}kx_1^2.$$

*e. La carga de un condensador

El concepto de trabajo en otras ramas de la física puede tratarse de manera análoga. Por ejemplo, considérese el problema de cargar un condensador. Si se denota por Q la cantidad de electricidad en el condensador, por C su capacidad y por V la diferencia de potencial (voltaje) entre sus placas, entonces por la física se sabe que $Q = CV$. Además, el trabajo hecho al mover una carga Q a través de una diferencia de potencial V es igual a QV. Puesto que durante el proceso de carga del condensador la diferencia de potencial V no es constante sino que crece con Q, se efectúa un paso al límite de manera análoga a como se hizo en la p. 436, y así se obtiene como expresión para el trabajo realizado al cargar el condensador

$$\int_0^{Q_1} V\,dQ = \frac{1}{C}\int_0^{Q_1} Q\,dQ = \frac{1}{2}\frac{Q_1^2}{C} = \frac{1}{2}Q_1V_1,$$

donde Q_1 es la cantidad total de electricidad que pasa al condensador y V_1 es la diferencia de potencial entre sus placas al final del proceso de carga.

Apéndice

*A.1 Propiedades de la evoluta

En la p. 377 se definió la evoluta E de una curva C como el lugar geométrico de los centros de curvatura de C. Si C está representada por: $x = x(s)$, $y = y(s)$, usando como parámetro la longitud de arco s, entonces el centro de curvatura (ξ, η) del punto de C con parámetro s está dado por [cfr. (17*a*), p. 377]

$$\xi = x - \rho\dot{y}, \qquad \eta = y + \rho\dot{x}, \tag{58}$$

donde

$$\kappa = \frac{1}{\rho} = \dot{x}\ddot{y} - \dot{y}\ddot{x}.$$

Las cantidades κ y $|\rho|$ son, respectivamente, la *curvatura* y el *radio de curvatura* de C.

A partir de estas fórmulas pueden deducirse algunas propiedades geométricas interesantes de la evoluta.

Derivando la relación $\dot{x}^2 + \dot{y}^2 = 1$ se llega a $\dot{x}\ddot{x} + \dot{y}\ddot{y} = 0$. Puesto que también $\dot{x}\ddot{y} - \dot{y}\ddot{x} = 1/\rho$, se tiene

$$x = -\frac{1}{\rho}\dot{y}, \qquad \ddot{y} = \frac{1}{\rho}\dot{x}. \tag{59}$$

Derivando con respecto a s las fórmulas (58), resulta

$$\dot{\xi} = \dot{x} - \rho\ddot{y} - \dot{\rho}\dot{y} = -\dot{\rho}\dot{y}, \qquad \dot{\eta} = \dot{y} + \rho\ddot{x} + \dot{\rho}\dot{x} = \dot{\rho}\dot{x},$$

y, en consecuencia,

$$\dot{\xi}\dot{x} + \dot{\eta}\dot{y} = 0.$$

Puesto que los cosenos directores de la normal a la curva están dados por $-\dot{y}$ y $\dot{x}$, *la normal a la curva C es tangente a la evoluta E en el centro de curvatura;* o bien, la tangente a la evoluta es la normal a la curva dada; o, lo que es lo mismo, *la evoluta es la "envolvente" de las normales* (cfr. Fig. A.1).

Además, si se denota por σ la longitud de arco de la evoluta medida desde un punto fijo arbitrario, usando s como parámetro se tendrá

$$\dot{\sigma}^2 = \left(\frac{d\sigma}{ds}\right)^2 = \dot{\xi}^2 + \dot{\eta}^2.$$

Y como $\dot{x}^2 + y = 1$, de las fórmulas (59) se obtiene

$$\dot{\sigma}^2 = \dot{\rho}^2.$$

Si el sentido en que se mide σ se escoge adecuadamente, se sigue que

$$\dot{\sigma} = \dot{\rho},$$

siempre $\rho \neq 0$. La integración da

$$\sigma_1 - \sigma_0 = \rho_1 - \rho_0.$$

Esto es, *la longitud de arco entre dos puntos de la evoluta es igual a la diferencia entre los correspondientes radios de curvatura, suponiendo que ρ permanece diferente de cero para el arco en cuestión.*

Esta última condición no es superflua. Pues si $\dot{\rho}$ cambia de signo entonces la fórmula $\dot{\sigma} = \dot{\rho}$ muestra que al pasar al correspondiente punto de la evoluta la longitud de arco σ tiene un máximo o un mínimo; o sea, al pasar a este punto no se continúa considerando σ hacia adelante sino que debe invertirse el sentido en que se mide. Si se desea evitar esta inversión debe cambiarse el signo en la fórmula anterior, esto es, poner $\dot{\sigma} = -\dot{\rho}$ al pasar a tal punto.

También puede notarse que los centros de curvatura que corresponden a máximos o mínimos del radio de curvatura son *cúspides de la evoluta.* [La prueba no se da aquí.] (Ver Figs. A.4, A.6.)

La relación geométrica recién encontrada puede expresarse aún en otra forma: Imaginemos un hilo flexible e inextensible colocado a lo largo de un arco de la evoluta E y estirado de manera que una parte de él se extienda tangencialmente fuera de la curva; si, además, el punto extremo Q del hilo queda inicialmente sobre la curva original C, entonces al desenrollarlo

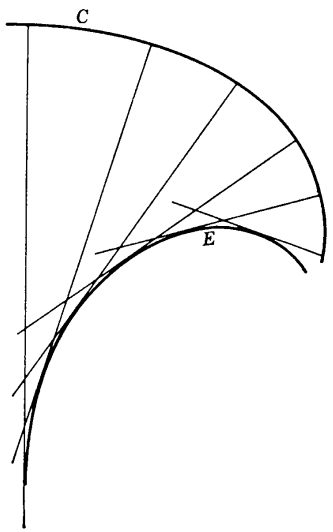


Figura A.1 Evoluta (E).

Q describirá dicha curva C. De aquí proviene el nombre de evoluta *(evolvere,* desenrollar). La curva C se llama una *involuta* de la evoluta E. Por otra parte, podemos comenzar con una curva arbitraria E, y mediante este proceso de desenrollado construir su involuta C. Entonces, en forma inversa se ve que E será la evoluta de C (Fig. A.2).

Para la demostración, considérese la curva E, que ahora es la curva dada, descripta en la forma $\xi = \xi(\sigma), \eta = \eta(\sigma)$, donde las coordenadas rectangulares corrientes se denotan por ξ y η y el parámetro σ es la lon-

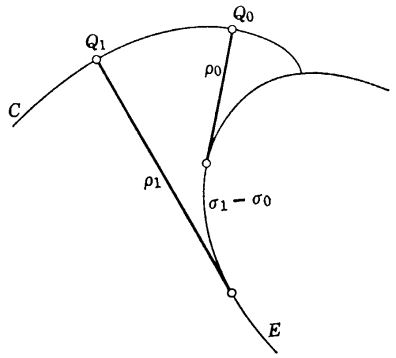


Figura A.2 Construcción de la involuta C de una curva E: $\rho_1 = \rho_0 + \sigma_1 - \sigma_0$, mediante un hilo.

gitud de arco sobre E. El enrollado se hace según lo indica la Fig. A.3. Cuando el hilo está completamente enrollado sobre la evoluta E, el extremo Q de aquél coincide con el punto A de E que corresponde a un arco de longitud a. Si ahora se desenrolla el hilo hasta que sea tangente a la evoluta en el punto P, correspondiente a la longitud de arco $\sigma > a$, la longitud del segmento PQ será $(\sigma - a)$ y sus cosenos directores $-\xi$ y $-\eta$, donde el punto denota ahora derivación con respecto a σ. Así se obtienen para las coordenadas x, y del punto Q las expresiones

$$x = \xi - (\sigma - a)\dot{\xi}, \qquad y = \eta - (\sigma - a)\dot{\eta}, \tag{60}$$

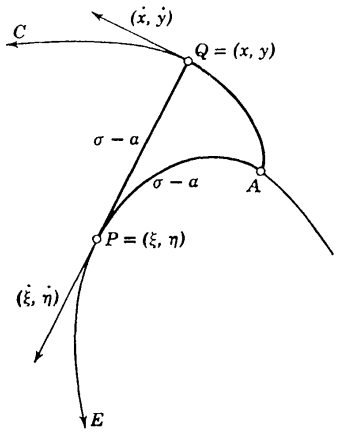


Figura A.3

que son las ecuaciones de la involuta descrita por Q, en términos del parámetro σ. Derivando con respecto a σ resulta

$$\begin{aligned} \dot{x} &= \dot{\xi} - \dot{\xi} + (a - \sigma)\ddot{\xi} = (a - \sigma)\ddot{\xi}, \\ \dot{y} &= \dot{\eta} - \dot{\eta} + (a - \sigma)\ddot{\eta} - (a - \sigma)\ddot{\eta}. \end{aligned} \tag{61}$$

Puesto que $\dot{\xi}\ddot{\xi} + \dot{\eta}\ddot{\eta} = 0$, se encuentra de inmediato que

$$\dot{\xi}\dot{x} + \eta\dot{y} = 0,$$

lo cual muestra que la recta PQ es normal a la involuta C. Por consiguiente, puede afirmarse que las normales a la curva C son tangentes a la curva E. Los cosenos directores de la tangente a E son $\dot{\xi}$ y η, por lo que para los cosenos directores de la tangente a C se encuentran las expresiones

$$\frac{\dot{x}}{\sqrt{\dot{x}^2 + \dot{y}^2}} = \dot{\eta}, \qquad \frac{\dot{y}}{\sqrt{\dot{x}^2 + \dot{y}^2}} = -\dot{\xi}. \tag{62}$$

Derivando con respecto a σ la relación $\dot{\xi}\dot{x} + \dot{\eta}\dot{y} = 0$ y substituyendo para $\dot{\xi}$, $\dot{\eta}$, $\ddot{\xi}$, $\ddot{\eta}$ las expresiones que resultan de las ecuaciones (61), (62) anteriores, se llega a

$$0 = \ddot{\xi}\dot{x} + \ddot{\eta}\dot{y} + \dot{\xi}\ddot{x} + \dot{\eta}\ddot{y} = \frac{\dot{x}^2 + \dot{y}^2}{a - \sigma} + \frac{-\ddot{x}\dot{y} + \dot{x}\ddot{y}}{\sqrt{\dot{x}^2 + \dot{y}^2}}.$$

Por lo tanto, el radio de curvatura de la curva C correspondiente al punto $Q = (x, y)$ vuelve a ser [ver fórmula (15), p. 373]

$$\rho = \frac{1}{\kappa} = \frac{(\dot{x}^2 + \dot{y}^2)^{3/2}}{\dot{x}\ddot{y} - \ddot{y}\dot{x}} = \sigma - a.$$

Esto representa también la distancia de Q al punto $P = (\xi, \eta)$. Debido a que P también está sobre la normal a C en Q, P resulta ser el centro de curvatura de C correspondiente al punto Q. Así, *cualquier curva E es la evoluta de todas sus involutas.*

Ejemplos. Consideremos la evoluta de la cicloide

$$x = \pi + t + \operatorname{sen} t, \qquad y = -1 - \cos t.$$

Por la Ec. (17) de p. 377, el centro de curvatura (ξ, η) de una curva referida a un parámetro t arbitrario es

$$\xi = x - \dot{y}\frac{\dot{x}^2 + \dot{y}^2}{\dot{x}\ddot{y} - \ddot{y}\dot{x}}, \qquad \eta = y + \dot{x}\frac{\dot{x}^2 + \dot{y}^2}{\dot{x}\ddot{y} - \ddot{y}\dot{x}}.$$

Un breve cálculo da el siguiente resultado para la evoluta de la cicloide

$$\xi = \pi + t - \operatorname{sen} t, \qquad \eta = 1 + \cos t.$$

Si se pone $t = \tau - \pi$, entonces

$$\xi + \pi = \pi + \tau + \operatorname{sen} \tau, \qquad \eta - 2 = -1 - \cos \tau;$$

y estas ecuaciones muestran que la evoluta es a su vez una cicloide similar a la curva original, pudiendo obtenerse a partir de ésta por traslación, como lo indica la Fig. A.4.

Lo anterior nos proporciona un método simple para construir un *péndulo cicloidal* (ver p. 430). Si una masa P se sujeta mediante una cuerda de longitud 4 a una de las cúspides de la evoluta, por efecto de la tensión, parte de la cuerda coincidirá con la evoluta y la parte restante estará a lo largo de una tangente. La masa P es entonces forzada a caer sobre la involuta, o sea, sobre la cicloide original. Bajo la acción de la gravedad P debe efectuar sobre alguna porción de la cicloide un movimiento isócrono, con un período que no depende de la posición en la que P indica su movimiento. (El parámetro t al cual es referida la cicloide *no* corresponde al

tiempo en el movimiento isócrono.) La longitud de la porción recta de la cuerda en un péndulo de este tipo varía durante el movimiento (ver Fig. A.4).

Como un ejemplo adicional se obtiene la ecuación para la involuta de una circunferencia. Se comienza con la circunferencia $\xi = \cos \sigma$, $\eta = -\operatorname{sen} \sigma$ y se desenrolla la tangente como lo indica la Fig. A.5. La involuta de esta curva queda entonces dada en la forma

$$x = \cos \sigma + \sigma \operatorname{sen} \sigma, \qquad y = -\operatorname{sen} \sigma + \sigma \cos \sigma.$$

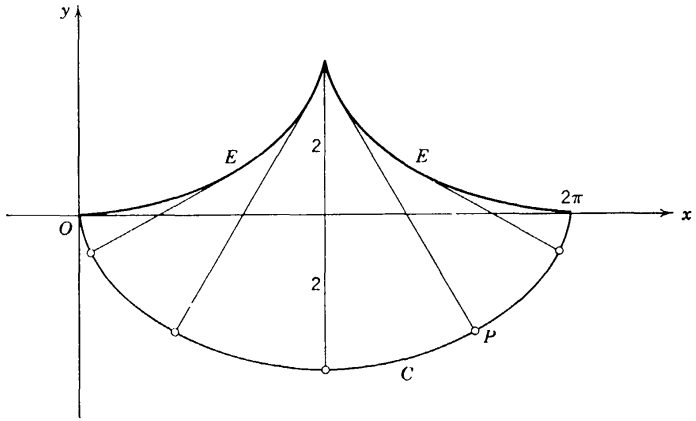


Figura A.4 El péndulo cicloidal.

(usando la ecuación (60) de la p. 445 con $a = 0$).

Por último, se determina la evoluta de la elipse $x = a \cos t$, $y = b \operatorname{sen} t$. De inmediato resulta

$$\xi = x - \dot{y}\frac{\dot{x}^2 + \dot{y}^2}{\dot{x}\ddot{y} - \ddot{y}\dot{x}} = \frac{a^2 - b^2}{a}\cos^3 t.$$

y

$$\eta = y + \dot{x}\frac{\dot{x}^2 + \dot{y}^2}{\dot{x}\ddot{y} - \ddot{y}\dot{x}} = -\frac{a^2 - b^2}{b}\operatorname{sen}^3 t,$$

como representación paramétrica de la evoluta. Si de estas ecuaciones se elimina t de la manera usual, se llega a la ecuación de la evoluta en forma no paramétrica:

$$(a\xi)^{2/3} + (b\eta)^{2/3} = (a^2 - b^2)^{2/3}.$$

Esta curva es una *astroide*. Su gráfica se da en la Fig. A.6. Por medio de las ecuaciones paramétricas podemos fácilmente cerciorarnos de que los centros de curvatura que corresponden a los vértices de la elipse son efectivamente las cúspides de la astroide.

*A.2 Areas limitadas por curvas cerradas. Indices

En la sección 4.2 se representó el área limitada por una curva cerrada $x = x(t)$, $y = y(t)$, $\alpha \leq t \leq \beta$, la cual en ninguna parte se intersecta a sí misma *(curva cerrada simple)*, por medio de la integral

$$A = -\int_{\alpha}^{\beta} y(t)\dot{x}(t)\, dt;$$

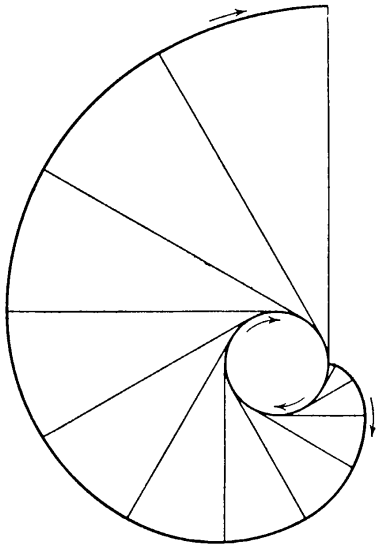


Figura A.5 Involuta de la circunferencia.

el valor que se obtiene es positivo o negativo dependiendo de que la frontera se recorrra en el sentido contra reloj o en el opuesto. Esta fórmula sigue siendo válida como definición de A aun para curvas que presentan autointersecciones. Falta, sin embargo, ver cómo A se relaciona con las áreas en tales casos. Supóngase que la curva C, dada por las ecuaciones $x = x(t)$, $y = y(t)$, se corta a sí misma en un número finito de puntos y de esta manera divide el plano en un número finito de porciones $R_1, R_2, \ldots$. Supóngase además que las derivadas son continuas y que $\dot{x}^2 + \dot{y}^2 \neq 0$, excepto quizá para un número finito de discontinuidades de salto (que pueden corresponder a esquinas o no). Por último, supóngase que la curva tiene un número finito de *rectas de apoyo* $x =$ constante, es decir, rectas verticales que son tangentes a dicha curva o pasan por uno de sus puntos de autointersección.

A cada región R_i se le asigna entonces un entero, el *índice* μ_i, definido de la siguiente manera: Se escoge un punto arbitrario Q de R_i que no esté sobre una recta de apoyo y se traza la semirrecta que se extiende desde Q hacia arriba en la dirección del eje y positivo. Se cuenta el número de veces que la curva C orientada según las t crecientes cruza la semirrecta de derecha a izquierda y se resta del número de veces en que el cruce se

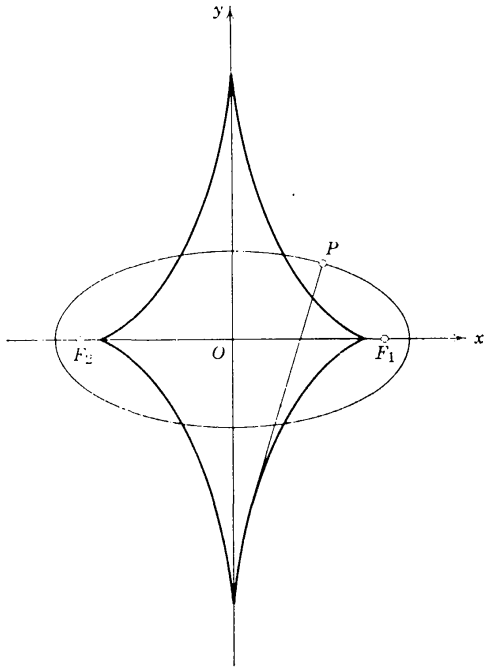


Figura A.6. Evoluta de la elipse.

efectúa de izquierda a derecha; la diferencia es el índice μ_i. Por ejemplo, el interior de la curva representada en la Fig. 4.17, p. 361, tiene el índice $\mu = +1$; y en la Fig. A.7 las regiones $R_1, \ldots, R_5$, R_6 tienen los índices $\mu_1 = -1$, $\mu_2 = -2$, $\mu_3 = -1$, $\mu_4 = 0$, $\mu_5 = 1$ y $\mu_6 = 0$. Este número μ_i efectivamente depende sólo de las regiones R_i y no del punto particular Q escogido en R_i, como puede verse fácilmente de la siguiente manera: Escójase cualquier otro punto Q' en R_i que no esté sobre una recta de apoyo y únase Q con Q' por medio de una línea quebrada contenida completamente en R_i (Fig. A.8). Conforme esta línea se recorre de Q a Q', la diferencia entre el número de cruces de derecha a izquierda y el número de cruces de izquierda a derecha es constante. Esto es así porque entre dos

rectas de apoyo el número de cruces de cada tipo no cambia, mientras que al atravesar la curva una recta de apoyo el número de cruces de ambos tipos sigue siendo el mismo, o bien ambos se incrementan en uno, o bien decrecen en uno; y así, de cualquier manera, la diferencia no se altera. En este proceso una recta de apoyo que encuentra a la curva en varios puntos, digamos $A, B, \ldots, H$, se considera como varias rectas de apoyo diferentes, $FA, FB, \ldots, FH$, donde F es un punto situado verticalmente

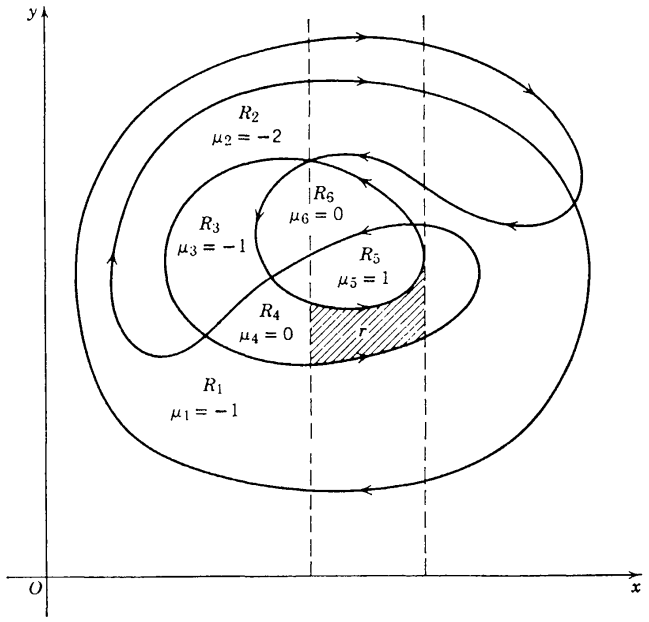


Figura A.7 Indices μ_i de las regiones R_i que forman las curvas cerradas y orientadas. Figura A.8.

por encima de todos los puntos $A, B, \ldots, H$. Nuestro argumento se aplica entonces a cada una de estas rectas. Por lo tanto, el número μ_i tiene el mismo valor independientemente de que para determinarlo se use Q o Q'.

En particular, si la curva no se corta a sí misma, su interior consiste de una única región R cuyo índice es $+1$ o -1 según que la frontera se recorra en el sentido contra reloj o en el opuesto. Para hacer esto evidente, trácese una recta vertical (no de apoyo) que intersecte a la curva; y sobre ella determínese el punto más alto de intersección P con la curva, y escójase un punto Q en R situado debajo de P y tan cercano a él que entre P y Q ningún otro punto de intersección exista. Entonces, por encima de Q se tiene un cruce de la curva que, si ésta se recorre en el sentido contra

reloj, es un cruce de derecha a izquierda, de manera que $\mu = +1$; y en caso contrario $\mu = -1$. Según lo que se vio antes, este mismo valor de μ resulta para cualquier otro punto de R. Para una curva como la considerada y, de hecho, para todas las curvas cerradas, una de las regiones, la "exterior" a la curva, se extiende sin límite en todas direcciones; y vemos de inmediato que esta región tiene índice 0, y será ignorada en lo que sigue. Por lo anterior, la relación entre la integral A y las áreas de las regiones R_i está dada por el siguiente teorema:

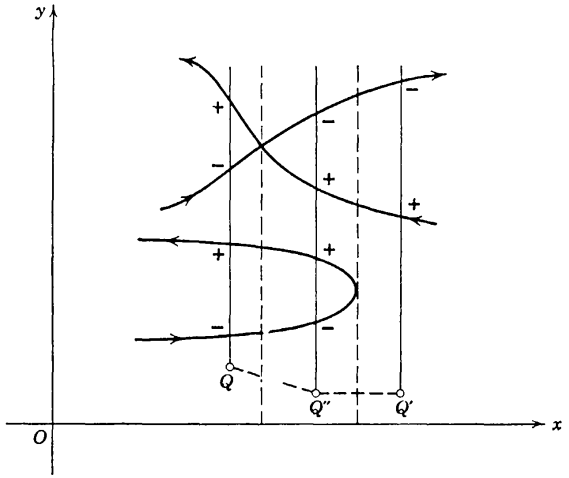


Figura A.8.

TEOREMA. *El valor de la integral* $-\int_\alpha^\beta y\dot{x}\,dt$ *es igual a la suma de las áreas absolutas de las regiones* R_i, *donde cada área* R_i *se cuenta* μ_i *veces. En símbolos:*

$$-\int_\alpha^\beta y\dot{x}\,dt = \sum \mu_i\,|\text{área } R_i|.$$

DEMOSTRACIÓN. La demostración es sencilla. Supongamos, como tenemos derecho a hacerlo, que la curva completa está por encima del eje x. (Agregando una constante a y no se altera el valor de la integral A para una curva cerrada.) Las rectas de apoyo cortan R_i en un número finito de porciones. Sea r una de estas porciones. Entonces, al tomar la integral $-\int y\dot{x}\,dt = -\int y\,dx$ para cada rama univaluada de la función $y = y(x)$ e interpretarla como el área entre la curva y el eje x, encontramos que el valor absoluto del área de r es contada $+1$ veces por cada rama de derecha a izquierda encima de r, y -1 veces por cada rama de izquierda a derecha

encima de r; en total, μ_i veces. Lo mismo es cierto para cada una de las demás porciones de R_i; por lo tanto, R_i se cuenta μ_i veces. Así, la integral sobre la curva completa tiene el valor $\sum \mu_i\,|\text{área } R_i|$, como se estableció (Ver Fig. A.7). Esta fórmula concuerda con la que se encontró para curvas simples, como se ve a partir de la discusión sobre los valores de μ para tales curvas.

La definición dada para el índice μ_i tiene la desventaja de haber sido formulada en términos de un sistema coordenado particular. De hecho, no obstante, puede demostrarse que el valor de μ_i asignado a cada región R_i es independiente del sistema de coordenadas y depende solamente de la curva. Esto puede verse fácilmente identificando μ_i con el número total ν_i de veces que una porción de la curva, para t creciente de α a β, rodea cualquier punto fijo Q_i de R_i en el sentido contra reloj; esto es, el número de veces que C se *enrolla* alrededor de Q. Probaremos la identidad de μ_i con ν_i.

Sea C dada paramétricamente por $x = x(t)$, $y = y(t)$ donde $\alpha \leq t \leq \beta$. Sea $Q = (\xi, \eta)$ un punto que no está sobre una recta de apoyo de C. Tomemos Q como el origen de un sistema de coordenadas polares r, θ, en el cual

$$r = \sqrt{(x-\xi)^2 + (y-\eta)^2}, \qquad \cos\theta = \frac{x-\xi}{r}, \ \operatorname{sen}\theta = \frac{y-\eta}{r}.$$

El ángulo polar θ está determinado solamente a menos de múltiplos enteros de 2π; y, sin embargo, θ está determinado unívocamente como una función de t, por su valor θ_0 para $t = \alpha$, si pedimos que $\theta = \theta(t)$ varíe continuamente con t a lo largo de C. En $t = \beta$ el ángulo θ tendrá el valor $\theta(\beta) = \theta_0 + 2\nu\pi$, donde ν es un entero. El número

$$\nu = \frac{1}{2\pi}[\theta(\beta) - \theta(\alpha)] = \frac{1}{2\pi}\int_\alpha^\beta \frac{d\theta}{dt}\,dt = \frac{1}{2\pi}\int_\sigma d\theta$$

representa el número de veces que la curva orientada C rodea a Q.

Así, la curva C cruza la semirrecta vertical que pasa por Q, para aquellos valores de t en los que la expresión $(1/2\pi)[\theta(t) - \pi/2]$ tiene un valor entero n. Considérense para un n fijo los valores t en el intervalo paramétrico para los cuales $(1/2\pi)(\theta - \pi/2) = n$. Sean σ_n y τ_n los números de tales valores t para los cuales $d\theta/dt > 0$ y $d\theta/dt < 0$, respectivamente. Es obvio que el *índice* en el punto Q es

$$\mu = \sum_n \sigma_n - \sum_n \tau_n = \sum_n (\sigma_n - \tau_n).$$

Por otra parte, $\sigma_n - \tau_n$ puede solamente tener uno de los valores 1, 0, -1, pues la gráfica de $\theta(t)$ en el plano θ, t debe cruzar la recta $\theta = \pi/2 + 2n\pi$ alternadamente desde arriba y desde abajo. Precisamente, se tiene $\sigma_n - \tau_n =$

$\operatorname{sign}[\theta(\beta) - \theta(\alpha)]$ si $\pi/2 + 2n\pi$ está entre $\theta(\alpha)$ y $\theta(\beta)$; de lo contrario, es $\sigma_n - \tau_n = 0$.

En consecuencia, μ es igual al número de valores de la forma $\pi/2 + 2n\pi$, con un n entero que está entre $\theta(\alpha)$ y $\theta(\beta)$, tomado con el signo de $\theta(\beta) - \theta(\alpha)$; o sea, μ es igual a ν.

Puesto que $\theta = \arctan[(y-\eta)/(x-\xi)]$, se tiene

$$\frac{d\theta}{dt} = \frac{\dot{y}(x-\xi) - \dot{x}(y-\eta)}{(x-\xi)^2 + (y-\eta)^2}$$

Esto conduce a la siguiente representación integral para el índice de la curva C cerrada, orientada, con respecto al punto (ξ, η):

$$\mu = \frac{1}{2\pi}\int_\alpha^\beta \frac{\dot{y}(x-\xi) - \dot{x}(y-\eta)}{(x-\xi)^2 + (y-\eta)^2}\,dt,$$

que puede escribirse en forma simple (ver pp. 384-385) sin referencia explícita al parámetro t, como

$$\mu = \frac{1}{2\pi}\int_C \frac{(x-\xi)\,dy - (y-\eta)\,dx}{(x-\xi)^2 + (y-\eta)^2}.$$

La característica notable de estos resultados es que el entero μ, o el ν, que describe una relación topológica entre el punto Q y la curva C, puede determinarse analíticamente a partir de la representación paramétrica de C, evaluando una integral.

PROBLEMAS

SECCION 4.1*c*, página 328

1. Bosquejar la hipocicloide para $a = 4c$ (astroide) y encontrar su ecuación no paramétrica.

2. Demostrar que si c/a es racional la hipocicloide general se cierra después de que el círculo móvil ha girado un número entero de veces, en tanto que si c/a es irracional la curva encuentra a la circunferencia del círculo fijo en una infinidad de puntos y nunca se cierra.

3. Obtener la representación paramétrica

$$x = at - b\operatorname{sen} t, \qquad y = a - b\cos t$$

para la trocoide ordinaria, esto es, para la trayectoria de un punto P sujeto a un disco de radio a que rueda a lo largo de una recta, siendo b la distancia de P al centro del disco (ver Fig. 4.7).

4. Encontrar las ecuaciones paramétricas de la curva $x^3 + y^3 = 3axy$ (la hoja de Descartes), escogiendo como parámetro t la tangente del ángulo entre el eje x y el rayo que va del origen al punto (x, y).

SECCION 4.1*e*, página 361

1. El ángulo α entre dos curvas en un punto de intersección se define como el ángulo entre sus tangentes en ese punto. Encontrar una fórmula para $\cos\alpha$ en términos de las representaciones paramétricas de las curvas.

2. Sean $x = f(t)$ e $y = g(t)$. Encontrar fórmulas para d^2y/dx^2 y d^3y/dx^3 en términos de las derivadas con respecto al parámetro t.

3. Encontrar la fórmula en coordenadas polares para el ángulo α entre dos curvas $r = f(\theta)$ y $r = g(\theta)$.

4. Encontrar las ecuaciones de las curvas que en todas partes se intersectan bajo el mismo ángulo α con las rectas que pasan por el origen.

5. Demostrar lo siguiente: si $x = f(t)$ e $y = g(t)$ son continuas en el intervalo cerrado $[a, b]$ y derivables en el intervalo abierto (a, b), con $x'^2 + y'^2 > 0$, entonces existe al menos un punto sobre el arco abierto

$$x = f(t), \qquad y = g(t), \qquad (a < t < b),$$

en que la tangente es paralela a la cuerda que une los puntos extremos.

6. Sea P el punto de un círculo que describe una cicloide conforme el círculo rueda sobre una recta dada. Sea, además, Q el punto de contacto del círculo con la recta. Demostrar que en cualquier instante la normal a la cicloide en P pasa por Q. ¿Qué propiedad análoga se tiene para la tangente en P?

7. Probar que la longitud del segmento de la tangente a la astroide

$$x = 4c\cos^3\theta, \qquad y = 4c\,\mathrm{sen}^3\,\theta,$$

cortado por los ejes de coordenadas es constante.

*8. Demostrar que las dos familias de elipses e hipérbolas, $(0 < a < b)$,

$$\frac{x^2}{a^2 - \lambda^2} + \frac{y^2}{b^2 - \lambda^2} = 1, \quad \text{para } 0 < \lambda < a,$$

$$\frac{x^2}{a^2 - \tau^2} + \frac{y^2}{b^2 - \tau^2} = 1, \quad \text{para } a < \tau < b,$$

son cofocales (esto es, tienen los mismos focos) y se intersectan a ángulos rectos.

9. (*a*) Demostrar que para la elipse el ángulo entre los dos rayos que van de cada foco a un punto sobre la curva es bisectado por la normal en ese punto.
(*b*) Demostrar que para la hipérbola ese ángulo es bisectado por la tangente.

SECCION 4.1*f*, página 366

1. Probar que la curva definida por

$$y = \begin{cases} x^2\,\mathrm{sen}\,\dfrac{1}{x}, & 0 < x \le 1 \\ 0, & x = 0 \end{cases}$$

tiene longitud finita, pero que la curva continua definida por

$$y = \begin{cases} x\,\mathrm{sen}\,\dfrac{1}{x}, & 0 < x \le 1 \\ 0, & x = 0 \end{cases}$$

no es rectificable.

2. Probar que si la función f está definida y es monótona en el intervalo cerrado $[a, b]$, entonces el arco definido por

$$y = f(x), \qquad (a \le x \le b),$$

es rectificable.

SECCION 4.1*g*, página 370

1. Una *integral elíptica de segunda clase* tiene la forma

$$\int_0^\phi \sqrt{1 - k^2\,\mathrm{sen}^2\,\theta}\,d\theta.$$

(*a*) Demostrar que la longitud de arco de la elipse $x = a\cos\theta$, $y = b\,\mathrm{sen}\,\theta$ puede expresarse en términos de una integral elíptica de segunda clase,
(*b*) Hacer lo mismo para lo trocoide

$$x = at - b\,\mathrm{sen}\,t, \qquad y = a - b\cos t.$$

*(*c*) Demostrar que la longitud de arco de la hipérbola puede expresarse en términos de integrales elípticas de primera y segunda clases.

SECCION 4.1*h*, página 372

1. Sea P un punto del círculo que al rodar genera una cicloide, y sea Q el punto más bajo de dicho círculo en cualquier instante dado.
Demostrar que Q bisecta el segmento que une a P con el centro de la circunferencia osculatriz de la cicloide en P.

2. Encuéntrese el centro de curvatura para $y = x^2$ cuando $x = 0$. Determinar el punto de intersección de las normales a la curva para $x = 0$ y $x = \varepsilon$. Calcular la distancia de la intersección al centro de curvatura. Sugerir una definición alternativa para el centro de curvatura. Demostrar que esta definición es equivalente a la definición dada en el texto.

3. Considérese la cuestión de si la circunferencia osculatriz cruza la curva en el punto de contacto.

*4. Demostrar que la circunferencia de curvatura de la curva C en un punto P es el límite de las circunferencias que pasan por los tres puntos P, P_1, P_2, conforme P_1 y P_2 se acercan a P.

5. Sea $r = f(\theta)$ la ecuación de una curva en coordenadas polares. Probar que la curvatura está dada por la fórmula

$$k = \frac{2r'^2 - rr'' + r^2}{(r'^2 + r^2)^{3/2}},$$

donde

$$r' = \frac{df}{d\theta}, \qquad r'' = \frac{d^2f}{d\theta^2}.$$

6. La curva para la cual el segmento de tangente definido por el punto de contacto y la intersección con el eje y tiene longitud siempre igual a 1, se llama *tractriz*. Encontrar su ecuación. Demostrar que el radio de curvatura en cualquier punto de la curva es inversamente proporcional a la longitud del segmento de normal definido por el punto sobre la curva y la intersección con el eje y. Calcular la longitud de arco de la tractriz y encontrar las ecuaciones paramétricas en términos de la longitud de arco.

7. Sea $x = x(t)$, $y = y(t)$ una curva cerrada. Una longitud constante p es medida sobre la normal desde el punto de su intersección con la curva. El extremo de este segmento describe una curva que se llama "curva paralela" a la original. Encontrar el área, la longitud de arco y el radio de curvatura de la "curva paralela".

8. Demostrar que las únicas curvas cuya curvatura es una constante fija k son circunferencias de radio $1/k$.

***9.** Si la curvatura de una curva en el plano xy es una función monótona de la longitud de arco, demostrar que la curva no es cerrada y que no tiene puntos dobles.

SECCION 4.1*i*, página 378

1. Demostrar que la expresión para la curvatura de una curva $x = x(t)$, $y = y(t)$ no es alterada por una rotación de ejes y tampoco por un cambio de parámetro dado por $t = \phi(\tau)$, donde $\phi'(\tau) > 0$.

SECCION 4.3*d*, página 411

1. Demostrar que si la aceleración es siempre perpendicular a la velocidad, la velocidad (escalar) es constante.

2. El vector velocidad, considerado como un vector de posición, describe una curva conocida como odógrafa. Para una partícula que se mueve sobre una curva cerrada, investigar si la odógrafa de la partícula puede o no ser una recta.

3. Suponiendo que el círculo que rueda se mueve con velocidad constante, encontrar la velocidad y la aceleración del punto P que genera la cicloide.

4. Sea A un punto fijo en el plano y supóngase que el vector aceleración de un punto móvil P está siempre dirigido hacia A y es proporcional a $1/|AP|^2$. Demostrar que la odógrafa es una circunferencia (ver problema 2).

5. Sea A un punto fijo sobre un círculo. Sea P el punto del círculo móvil que se mueve de manera que el vector aceleración apunta hacia A. Probar que la aceleración es proporcional a $|AP|^{-5}$.

SECCION 4.5, página 420

1. Una partícula que se mueve sobre una recta está sujeta a una resistencia que produce una retardación ku^3, donde u es la velocidad y k una constante. Encontrar expresiones para la velocidad (u) y el tiempo (t) en términos de s, la distancia desde la posición inicial, y v_0, la velocidad inicial.

2. Una partícula de masa unitaria se mueve sobre el eje x sujeta a una fuerza $f(x) = -\operatorname{sen} x$.

(a) Determinar el movimiento de la partícula si en el tiempo $t = 0$ se encuentra en $x = 0$ y tiene velocidad $v_0 = 2$. Demostrar que cuando $t \to \infty$ la partícula se aproxima a una posición límite, y encontrar esta posición.

(b) Si las condiciones son las mismas, excepto que v_0 puede tener cualquier valor, demostrar que para $v_0 > 2$ el punto se mueve hacia una distancia infinita cuando $t \to \infty$, y que para $v_0 < 2$ el punto oscila en torno al origen.

3. Escójase un sistema de ejes con origen en el centro de la Tierra, cuyo radio se denota por R. De acuerdo con la ley de la gravitación de Newton, una partícula de masa unitaria que esté sobre el eje y será atraída por la Tierra con una fuerza $-\mu M/y^2$, donde μ es la "constante gravitacional" y M es la masa terrestre.

(a) Determinar el movimiento de la partícula después de que ésta se suelta desde el punto $y_0 (> R)$; esto es, si en el tiempo $t = 0$ está en el punto $y = y_0$ y tiene velocidad $v_0 = 0$.

(b) Encontrar la velocidad de la partícula en el momento de su impacto con la Tierra.

(c) Usando el resultado de (b), calcular la velocidad de una partícula que cae a la Tierra desde el infinito.[1]

***4.** Una partícula en reposo sobre el punto más alto de una circunferencia es ligeramente perturbada y se desliza hacia abajo por acción de la gravedad. ¿En qué punto dejará de moverse sobre la circunferencia?

[1] Esta es la misma que la mínima velocidad con la que debería dispararse un proyectil para que dejase la Tierra sin regresar jamás.

***5.** Una partícula de masa m se mueve sobre una elipse $r = k/(1 - e\cos\theta)$. La fuerza sobre la partícula es cm/r^2, dirigida hacia el origen. Describir el movimiento de la partícula, encontrar su período y demostrar que el radio vector de la partícula barre áreas iguales en tiempos iguales ("Ley de las Areas", de Kepler, para el movimiento planetario).

SECCION 4A.1, página 442

1. Demostrar que la evoluta de una epicicloide (ejemplo, p. 348) es otra epicicloide similar a la original que puede obtenerse de ésta por rotación y contracción.

2. Demostrar que la evoluta de una hipocicloide (ejemplo, p. 348) es otra hipocicloide que puede obtenerse de la primera por rotación y dilatación.

5

Desarrollo de Taylor

5.1 Introducción: Series de potencias

Fue un gran triunfo en los primeros años del Cálculo el descubrimiento hecho por Newton y otros de que muchas funciones conocidas podían expresarse como "polinomios de orden infinito" o "series de potencias", con coeficientes calculados mediante leyes elegantes y claras. Las series geométricas para $1/(1-x)$ y para $1/(1+x^2)$,

$$(1) \qquad \frac{1}{1-x} = 1 + x + x^2 + \cdots + x^n + \cdots$$

$$(1a) \qquad \frac{1}{1+x^2} = 1 - x^2 + x^4 - x^6 + \cdots + (-1)^n x^{2n} + \cdots,$$

válidas en el intervalo abierto $|x| < 1$, son prototipos (ver capítulo 1, página 90).

Desarrollos análogos de la forma

$$f(x) = a_0 + a_1 x + \cdots + a_n x^n + \cdots = \sum_{\nu=0}^{\infty} a_\nu x^\nu,$$

con coeficientes numéricos a_ν, serán deducidos en este capítulo para muchas otras funciones.

Los siguientes son ejemplos notables:

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots + \frac{x^n}{n!} + \cdots;$$

$$\operatorname{sen} x = x - \frac{x^3}{3!} + \frac{x^5}{5!} - + \cdots + \frac{(-1)^n x^{2n+1}}{(2n+1)!} + \cdots;$$

$$\cos x = 1 - \frac{x^2}{2!} + \frac{x^4}{2!} + \cdots + \frac{(-1)^n x^{2n}}{(2n)!} + \cdots$$

Estos desarrollos en series son válidos para todo x.

Teorema general del binomio de Newton. El desarrollo

$$(1+x)^\alpha = 1 + \frac{\alpha}{1!}x + \frac{\alpha(\alpha-1)}{2!}x^2 + \cdots = \sum_{\nu=0}^{\infty} \binom{\alpha}{\nu} x^\nu$$

es válido para $|x| < 1$ y cualquier exponente α.

Para explicar el significado preciso de tales desarrollos, considérese el polinomio de orden n formado por la suma de los $n+1$ primeros términos de la serie, esto es, la *n-ésima "suma parcial"*,

$$S_n = \sum_{\nu=0}^{n} a_\nu x^\nu.$$

La fórmula

$$f(x) = \sum_{\nu=0}^{\infty} a_\nu x^\nu, \quad \text{para } |x| < a$$

significa entonces: Para $n \to \infty$ la sucesión S_n tiende al valor de la función $f(x)$ en cada punto del intervalo $|x| < a$. Se dice entonces que la serie infinita *converge* a $f(x)$ en el intervalo $|x| < a$. La diferencia

$$R_n(x) = f(x) - S_n(x),$$

el *"residuo"* de la serie, mide la precisión con la cual el polinomio $S_n(x)$ aproxima a la función $f(x)$ en x. Por ejemplo,

$$(1b) \qquad \frac{1}{1-x} = 1 + x + x^2 + \cdots + x^n + R_n(x),$$

donde el residuo $R_n(x) = x^{n+1}/(1-x)$ tiende a cero para $|x| < 1$ conforme n crece; y resulta así la serie geométrica infinita $\sum_{\nu=0}^{\infty} x^\nu = 1/(1-x)$. Encontrar en casos específicos estimaciones sencillas para R_n que sean fácilmente utilizables es una tarea de importancia tanto teórica como práctica.

En este capítulo estamos interesados en tales desarrollos para una amplia clase de funciones, incluyendo todas las funciones trascendentales "elementales". Es notable el hecho de que en estos desarrollos de funciones trascendentes los coeficientes sean expresiones simples en términos de enteros. El tratamiento de estos desarrollos se hará mediante el *teorema de Taylor*. Más adelante, en el capítulo 7, se discutirá un enfoque diferente por medio del estudio directo de las series de potencias.

Debe hacerse notar que a menudo, como ocurre para la serie geométrica de la Ec. (1*a*), el desarrollo infinito no es válido fuera de algún intervalo de x en el caso de la serie geométrica, el intervalo $x^2 < 1$), aunque la función representada por la serie esté bien definida fuera de este intervalo.

5.2 Desarrollo del logaritmo y de la tangente inversa

a. El logaritmo

Como ejemplos sencillos se deducen primero los desarrollos para las funciones logaritmo y tangente inversa usando integración y a partir de la serie geométrica

$$\frac{1}{1-t} = 1 + t + t^2 + \cdots + t^{n-1} + r_n(t)$$

con $r_n(t) = t^n/(1-t)$.

El integrando en la fórmula

$$-\log(1-x) = \int_0^x \frac{dt}{1-t}$$

se reemplaza por la suma anterior y se realiza la integración término a término, obteniéndose para $x < 1$

$$-\log(1-x) = x + \frac{x^2}{2} + \frac{x^3}{3} + \frac{x^4}{4} + \cdots + \frac{x^n}{n} + R_n(x),$$

donde el residuo es

$$R_n(x) = \int_0^x r_n\,dt = \int_0^x \frac{t^n}{1-t}\,dt.$$

Por tanto, para cualquier entero positivo n la función $-\log(1-x)$ es aproximada por el polinomio de grado n

$$x + \frac{x^2}{2} + \frac{x^3}{3} + \cdots + \frac{x^n}{n},$$

donde el residuo R_n indica el "error" de esta aproximación.

Para evaluar la exactitud de dicha aproximación se estima el residuo R_n. Si primero se supone que $-1 \le x \le 0$, entonces en todo el intervalo de integración el valor absoluto del integrando $t^n/(1-t)$ nunca excede a $|t^n| = (-1)^n t^n$. Así,

$$|R_n| \le \left|\int_0^x t^n\,dt\right| = \frac{|x|^{n+1}}{n+1}.$$

por tanto, para todo valor de x en el intervalo *cerrado* $-1 \le x \le 0$, incluyendo $x = -1$, este residuo puede hacerse tan pequeño como se quiera con sólo escoger n suficientemente grande (ver p. 84). Cuando $x > 0$, el punto extremo $x = 1$ debe omitirse y x tiene que restringirse al intervalo $0 \le x < 1$. Aquí el integrando no cambia de signo y su valor absoluto no supera a $t^n/(1-x)$. Así se obtiene para $0 \le x < 1$ la estimación

$$|R_n| \le \frac{1}{1-x}\int t^n\,dt = \frac{x^{n+1}}{(1-x)(n+1)}.$$

Otra vez, por tanto, si x es fijo el residuo es arbitrariamente pequeño para n suficientemente grande. Por supuesto, la estimación anterior no tiene significado para $x = 1$.

En resumen,

$$(2) \qquad \log(1-x) = -x - \frac{x^2}{2} - \frac{x^3}{3} - \cdots - \frac{x^n}{n} - R_n$$

donde el residuo R_n tiende a cero conforme n crece, suponiendo que x está en el intervalo semiabierto $-1 \le x < 1$.

De hecho, este razonamiento establece una estimación "uniforme" para el residuo, independiente de x y válida para todo x en el intervalo $-1 \le x \le 1 - h$, donde h es cualquier número tal que $0 < h \le 1$, a saber, $|R_n| \le 1/[(n+1)h]$.

El hecho de que el residuo R_n tienda a cero en el intervalo semiabierto $-1 \le x < 1$ queda expresado diciendo que en este intervalo la función logaritmo está dada por la serie infinita [1]

$$(3) \qquad \log(1-x) = -x - \frac{x^2}{2} - \frac{x^3}{3} - \frac{x^4}{4} - \cdots$$

Si en esta fórmula se inserta el valor particular $x = -1$, se obtiene una fórmula notable:

$$(4) \qquad \log 2 = 1 - \tfrac{1}{2} + \tfrac{1}{3} - \tfrac{1}{4} + - \cdots.$$

Esta es una de las relaciones cuyo descubrimiento produjo una honda impresión en los iniciadores del cálculo.

Para el intervalo abierto $-1 < x < 1$ sólo se necesita reemplazar x por $-x$ en (2) para obtener

$$(2a) \qquad \log(1+x) = x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} + - \cdots + (-1)^{n-1}\frac{x^n}{n} - R_n',$$

[1] Se deja como ejercicio al lector verificar que para todos los valores de x tales que $|x| > 1$ el residuo no solamente deja de tender a cero sino que, de hecho, R_n crece más allá de toda cota al crecer n; de manera que para tales valores de x el polinomio ya no es una buena aproximación del logaritmo, sino que empeora al aumentar n.

donde

$$R_n'(x) = \int_0^x \frac{t^n\,dt}{1-t} = (-1)^{n+1}\int_0^x \frac{t^n\,dt}{1+t}.$$

Tomando n par y restando (2) de (2a) se tiene

$$\frac{1}{2}\log\left(\frac{1+x}{1-x}\right) = \operatorname{ar\,tanh} x = x + \frac{x^3}{3} + \frac{x^5}{5} + \cdots + \frac{x^{n-1}}{n-1} + \bar{R}_n,$$

donde el residuo $\bar{R}_n$ está dado por

$$\bar{R}_n = \frac{1}{2}(R_n - R_n') = \int^x \frac{t^n}{1-t^2}\,dt,$$

y ar tanh x se define como en la p. 254.

Si se observa que $1/(1-t^2) \leq 1/(1-x^2)$, se encuentra, mediante una estimación elemental de la integral, que

$$|\bar{R}_n| \leq \frac{|x^{n-1}|}{n+1}\cdot\frac{1}{1-x^2};$$

de esta manera el residuo $\bar{R}_n$ tiende a cero conforme n crece, lo cual se expresa escribiendo el desarrollo como una serie infinita:

$$(5) \qquad \frac{1}{2}\log\frac{1+x}{1-x} = \operatorname{ar\,tanh} x = x + \frac{x^3}{3} + \frac{x^5}{5} + \frac{x^7}{7} + \cdots,$$

para todos los valores de x con $|x| < 1$. Es de notar, este resultado también podría derivarse directamente integrando la serie geométrica para $1/(1-x^2)$. Esta fórmula presenta la ventaja de que al recorrer x el intervalo de -1 a 1, la expresión $(1+x)/(1-x)$ corre sobre todos los números positivos. Así, si el valor de x se escoge en forma adecuada, la serie nos permite calcular el valor del logaritmo de cualquier número positivo con un error no mayor que el valor estimado arriba como cota para $\bar{R}_n$.

b. La tangente inversa

La tangente inversa puede tratarse de manera similar a la usada para el logaritmo, partiendo con la fórmula

$$\frac{1}{1+t^2} = 1 - t^2 + t^4 - + \cdots + (-1)^{n-1}t^{2n-2} + r_n$$

donde es ahora $r_n = (-1)^n\dfrac{t^{2n}}{1+t^2}$.

Por integración [ver Ec. (14), p. 284] se obtiene

$$\operatorname{arc\,tan} x = x - \frac{x^3}{3} + \frac{x^5}{5} - + \cdots + (-1)^{n-1}\frac{x^{2n-1}}{2n-1} + R_n,$$

$$R_n = (-1)^n\int_0^x \frac{t^{2n}}{1+t^2}\,dt;$$

y se ve de inmediato que en el intervalo cerrado $-1 \leq x \leq 1$ el residuo R_n tiende a cero conforme n crece, pues

$$|R_n| \leq \int_0^{|x|} t^{2n}\,dt = \frac{|x|^{2n+1}}{2n+1}.$$

De la fórmula para el residuo puede también fácilmente demostrarse que para $|x| > 1$ el valor absoluto del residuo crece más allá de toda cota conforme n crece.

De esta manera se ha deducido la serie infinita

$$(6) \quad \operatorname{arc\,tan} x = x - \frac{x^3}{3} + \frac{x^5}{5} - + \cdots + (-1)^{n-1}\frac{x^{2n-1}}{2n-1} + - \cdots,$$

válida para el intervalo cerrado $|x| \leq 1$. Puesto que arc tan $1 = \pi/4$, se obtiene para $x = 1$ la serie de *Leibnitz-Gregory*

$$(7) \qquad \frac{\pi}{4} = 1 - \frac{1}{3} + \frac{1}{5} - + \cdots,$$

una expresión tan notable como la encontrada antes para log 2.

5.3 Teorema de Taylor

Taylor, alumno de Newton, observó que el desarrollo elemental de polinomios se prestaba a una amplia generalización para funciones no polinomiales suponiendo que estas funciones fuesen suficientemente derivables y con dominios restringidos adecuadamente.

a. Representación de Taylor para polinomios

Es ésta una fórmula algebraica completamente elemental relativa a un polinomio en x de orden n, por ejemplo

$$f(x) = a_0 + a_1x + a_2x^2 + \cdots + a_nx^n.$$

Si se reemplaza x por $a + h = b$ y se desarrolla cada término en potencias de h resulta inmediatamente una representación en la forma

$$(8) \qquad f(a+h) = c_0 + c_1h + c_2h^2 + \cdots + c_nh^n.$$

La fórmula de Taylor está dada por la relación

$$(8a) \qquad c_\nu = \frac{1}{\nu!} f^{(\nu)}(a),$$

para el coeficiente c_ν en términos de f y sus derivadas en $x = a$. Para demostrar esto, consideremos la cantidad $h = b - a$ como la variable independiente y apliquemos la regla de la cadena, según la cual derivar con respecto a h es lo mismo que derivar con respecto a $b = a + h$. De esta manera, derivando una y otra vez la fórmula (8) con respecto a h y substituyendo $h = 0$ se obtienen sucesivamente los resultados

$$c_0 = f(a),\ c_1 = f'(a),\ \ldots,\ \nu!\, c_\nu = f^{(\nu)}(a),$$

y, por consiguiente, la fórmula de Taylor para polinomios:

$$(9) \qquad f(a + h) = f(a) + hf'(a) + \frac{h^2}{2} f''(a) + \cdots + \frac{h^n}{n!} f^{(n)}(a).$$

La derivada $(n + 1)$-ésima se anula para un polinomio de grado n, de manera que nuestra fórmula (9) termina naturalmente.

La fórmula (9), tal como se la establece, no es más que un elemental reordenamiento algebraico de un polinomio en potencias de $a + h$ para dar un polinomio en potencias de h.

b. Fórmula de Taylor para funciones no polinomiales

Newton y sus discípulos inmediatos aplicaron audazmente la fórmula (9) a funciones no polinomiales, para las cuales el desarrollo no termina automáticamente en el n-ésimo término. En vez de cortar el desarrollo permitieron que n tendiera a infinito, procedimiento que para muchas funciones especiales importantes será justificado posteriormente.

Suponiendo que la función f es derivable por lo menos n veces en un intervalo que contiene los puntos a y $a + h$, ya no podemos escribir $f(a + h)$ como una expresión de la fórmula de (9) consistente de un número finito de potencias de h, sino que debe tomarse en cuenta la discrepancia introduciendo un "residuo" adicional R_n, lo cual se hace escribiendo formalmente

$$(10) \qquad f(b) = f(a + h) = f(a) + hf'(a) + \cdots + \frac{h^n}{n!} f^n(a) + R_n.$$

De hecho (10), no es sino la definición del *término residual* de corrección R_n e indica la esperanza de que R_n se vuelva pequeño y, más aún, tienda a cero para $n \to \infty$. Si, efectivamente, el residuo tiende a cero, entonces la fórmula (10) en el límite $n \to \infty$ conduce al desarrollo

$$(11) \qquad f(a + h) = f(a) + hf'(a) + \cdots + \frac{h^n}{n!} f^{(n)}(a) + \cdots$$

para $f(x)$ como una *serie infinita de potencias* de h.

El problema crucial, que mucho supera en dificultad al de las manipulaciones algebraicas de la sección 5.3*a*, es entonces el de encontrar estimaciones para el residuo R_n de manera que la exactitud de la representación mediante el *polinomio de Taylor* de orden n en h,

$$(12) \qquad T_n(h) = \sum_0^n \frac{f^{(\nu)}(a)}{\nu!} h^\nu,$$

y el paso al límite para $n \to \infty$, puedan explorarse rigurosamente. El polinomio de Taylor $T_n(h)$ es una aproximación a $f(a + h)$ en el sentido de que en $h = 0$ las funciones T_n y f, así como sus derivadas hasta el orden n, coinciden, de manera que la diferencia $R_n = f - T_n$ se anula en $x = a$ junto con las n primeras derivadas.

5.4 Expresiones y estimaciones para el residuo

a. Expresiones de Cauchy y de Lagrange

Una representación directa del residuo R_n que permite hacer estimaciones de su valor absoluto, $|R_n|$, es lo que constituye el núcleo del teorema de Taylor. Los resultados se obtienen fácilmente sobre la base del teorema del valor medio del Cálculo. Están, además, relacionados con la *aproximación lineal de funciones mediante diferenciales* (ver p. 200).

Primeramente examinemos de nuevo esta aproximación.

La definición de derivada en un punto a establece simplemente que $f(a + h) = f(a) + hf'(a) + h\varepsilon$, donde $\varepsilon \to 0$ para $h \to 0$. Podemos lograr una aproximación algo más fina averiguando si ε es, en efecto, de orden por lo menos igual al de h, siempre que tanto f' como f'' existan y sean continuas en nuestro intervalo J. La estimación se obtiene si escribimos nuevamente $a + h = b$, introducimos un residuo R mediante

$$(13) \qquad f(b) = f(a) + (b - a)f'(a) + R,$$

y consideramos ahora b fijo y el punto inicial a variable. Esta ecuación define R como una función de a en el intervalo J. Derivando entonces con respecto a la variable a se obtiene cero para el miembro izquierdo, pues $f(b)$ es constante, y la regla para derivar un producto muestra que

$$0 = f'(a) - f'(a) + (b - a)f''(a) + R'(a)$$

y, por tanto,

$$-R'(a) = (b - a)f''(a). \tag{14}$$

Ahora bien, para $a = b$ tenemos obviamente $R(b) = 0$. Por el teorema del valor medio del Cálculo se tiene $[R(a) - R(b)]/(b - a) = -R'(\xi)$, donde ξ es algún valor entre a y b; y, debido a que $R(b) = 0$, concluimos por consiguiente que $R(a) = -(b - a)R'(\xi) = -hR'(\xi)$. Entonces, por (14) se tiene $R'(\xi) = -(b - \xi)f''(\xi)$, y, por tanto $|R'(\xi)| < h\,|f''(\xi)|$ puesto que $|b - \xi| \leq h$. Como $|f''(\xi)|$ es acotado en un intervalo en torno a a, obtenemos finalmente una estimación que muestra que el residuo o "error" R_n es pequeño al menos hasta el segundo orden en h:

$$|R(a)| < h^2\,|f''(\xi)|. \tag{15}$$

Del caso especial $n = 1$ pasamos ahora al de cualquier orden n. La caracterización directa del residuo R_n se logra mediante el mismo artificio que para $n = 1$. Suponemos que a y $b = a + h$ son puntos de un intervalo J en el cual $f(x)$ está definida y tiene derivadas continuas hasta de orden $n + 1$. Consideramos a como la variable independiente y mantenemos fijo el punto extremo b. En la fórmula (10), p. 465, que define $R_n(a)$, escribimos $b - a$ en vez de h. Al derivar y tomar en cuenta que $f(b)$ es constante, se encuentra, por la regla del producto, que casi todos los términos se cancelan, quedando la fórmula

$$0 = \frac{(b - a)^n}{n!} f^{(n+1)}(a) + R_n'(a) \tag{16}$$

para todo valor a en el intervalo. Puesto que para $a = b$ el residuo R_n es cero, esta expresión directa para su derivada como función de a caracteriza completamente a R_n como una integral $\int_b^a R_n'(t)\,dt = -\int_a^b R_n'(t)\,dt$; o sea:

$$R_n(a) = \int_a^b \frac{(b - t)^n}{n!} f^{n+1}(t)\,dt. \tag{17}$$

Esta es una *representación integral exacta para el residuo.*

Una estimación para R_n análoga a la obtenida antes para $n = 1$ se desprende directamente del teorema del valor medio del Cálculo, aplicado a (16):

$$\frac{R_n(a) - R_n(b)}{b - a} = \frac{R_n(a)}{b - a} = -R_n'(\xi) = \frac{(b - \xi)^n}{n!} f^{n+1}(\xi);$$

o sea

$$R_n(a) = \frac{(b - a)(b - \xi)^n}{n!} f^{n+1}(\xi), \tag{18}$$

donde ξ es un valor no especificado que se elige adecuadamente entre a y b. La misma estimación puede también obtenerse aplicando a la expresión (17) el teorema del valor medio del cálculo integral (capítulo 2, p. 163).

Forma de Cauchy para el residuo. Si se define $\xi = a + \theta h = a + \theta(b - a)$, se obtiene la *fórmula de Cauchy* para el residuo en la fórmula de Taylor (10):

$$R_n(a) = \frac{h^{n+1}}{n!}(1 - \theta)^n f^{(n+1)}(a + \theta h), \tag{19}$$

donde θ es una cantidad entre 0 y 1 no especificada.

Puede también aplicarse a la integral (17) para R_n el *teorema generalizado del valor medio* del cálculo integral (ver p. 164), tomando para la "función peso" $p(t)$ la expresión $p(t) = (b - t)^n$, que no cambia de signo en el intervalo de integración.[1] Entonces,

$$R_n = \frac{1}{n!} f^{(n+1)}(\xi) \int_a^b (b - t)^n\,dt = \frac{(b - a)^{n+1}}{(n + 1)!} f^{(n+1)}(\xi). \tag{20}$$

Forma de Lagrange para el residuo. Haciendo de nuevo $\xi = a + \theta h$ se obtiene la forma de *Lagrange* para el residuo,

$$R_n(a) = \frac{h^{n+1}}{(n + 1)!} f^{(n+1)}(a + \theta h), \tag{21}$$

con una cantidad adecuada θ tal que $0 \leq \theta \leq 1$. La forma de Lagrange es particularmente sugestiva, y, por lo tanto, más comúnmente aplicada, puesto que hace que el residuo en la fórmula

$$f(a + h) = f(a) + \frac{h}{1!} f'(a) + \frac{h^2}{2!} f''(a) + \cdots + \frac{h^n}{n!} f^{(n)}(a) + R_n = P_n(h) + R_n \tag{22}$$

se parezca al término $h^{n+1}f^{(n+1)}(a)/(n + 1)!$ que aparecería en el desarrollo (22) para un orden mayor, con sólo reemplazar el argumento a por el valor intermedio $a + \theta h$.

Para una función f cuya derivada $f^{(n+1)}$ es continua en un intervalo cerrado conteniendo al punto a, la cantidad $|f^{(n+1)}(\xi)|$ tiene una cota fija M. Puesto que entonces

$$|R_n| \leq \frac{h^{n+1}}{(n + 1)!} M,$$

[1] El teorema generalizado del valor medio fue demostrado para el caso de una $p(t)$ positiva; pero vale igualmente cuando $p(t)$ es negativa en todo el intervalo de integración.

el polinomio de Taylor $P_n(h)$ da para un n fijo una aproximación a la función $f(a + h)$ con un error de orden al menos $n + 1$ en h.

Nuestro interés se enfocará principalmente sobre la cuestión de si el residuo R_n tiende a cero conforme n crece; y si éste es el caso, decimos que la función se ha desarrollado en una serie de Taylor infinita

$$(23) \quad f(a + h) = f(a) + \frac{h}{1!}f'(a) + \frac{h^2}{2!}f''(a) + \frac{h^3}{3!}f'''(a) + \cdots;$$

en particular, si primero se hace $a = 0$ y después se escribe x en lugar de h, se obtiene la "serie de potencias"

$$f(x) = f(0) + \frac{x}{1!}f'(0) + \frac{x^2}{2!}f''(0) + \cdots$$

En la sección 5.5 se discutirán algunos ejemplos.

Para las aplicaciones, igualmente importante es el desarrollo finito de Taylor para un n fijo, con el término del residuo. Si se hace h tender a cero en esta fórmula, según la terminología del capítulo 3, p. 273, los diferentes términos de la serie tienden a cero con diferentes órdenes de magnitud en h. La expresión $f(a)$ representa el término de orden cero en la serie de Taylor; la expresión $hf'(a)$, el término de primer orden; la expresión $h^2f''(a)/2!$ el término de segundo orden, etc. De la forma para el residuo se observa que al desarrollar una función hasta el término de orden n se comete un error que tiende a cero con orden $(n + 1)$ conforme h tiende a cero. Mientras más próximo a a esté $a + h$, mejor será la representación de la función $f(a + h)$ mediante el polinomio de aproximación $P_n(h)$. En los casos de mayor interés la aproximación en la inmediata vecindad del punto x puede mejorarse aumentando el valor de n.

b. Una derivación alternativa de la fórmula de Taylor

La representación integral (17) para el residuo R_n en el teorema de Taylor se basó en la fórmula (16) para $R_n'(a)$. Debido a la importancia del teorema, damos aquí una versión diferente de su derivación, que lleva directamente a expresar R_n mediante repetidas *integraciones por partes* comenzando con la fórmula

$$(24) \quad f(b) - f(a) = \int_a^b f'(t)\,dt.$$

Para transformar (24) usando sucesivamente integración por partes, se introducen las funciones

$$\phi_1(t), \phi_2(t), \ldots, \phi_\nu(t), \ldots,$$

por medio de las relaciones

$$(25) \quad \phi_0(t) = 1, \qquad \phi_\nu'(t) = \phi_{\nu-1}(t),$$

y de las condiciones

$$(26) \quad \phi_\nu(b) = 0, \qquad \text{para} \quad \nu \geq 1,$$

donde b se considera como un parámetro fijo. Claramente, las condiciones (25) y (26) determinan en forma sucesiva todas las $\phi_\nu(t)$. Como puede verificarse de inmediato, las $\phi_\nu(t)$ son precisamente los polinomios

$$\phi_\nu(t) = \frac{(t - b)^\nu}{\nu!}.$$

De paso observamos que las funciones ϕ_ν se obtienen una de otra mediante integraciones sucesivas, dejando indeterminadas las constantes de integración; por consiguiente, las condiciones de definición (25) podrían satisfacerse también por otras funciones que cumplieran condiciones colaterales diferentes a las (26) (ver p. 210).

Puesto que $\phi_\nu(a) = (-1)^\nu(b - a)^\nu/\nu!$ y $\phi_\nu(b) = 0$, se obtiene

$$f(b) - f(a) = \int_a^b \phi_0 f'\,dt = \int_a^b \phi_1' f'\,dt = \phi_1 f'\Big|_a^b - \int_a^b \phi_1 f''\,dt;$$

e integrando por partes otra vez el último término, resulta

$$f(b) - f(a) = (b - a)f'(a) - \int_a^b \phi_2' f''\,dt$$

$$= (b - a)f'(a) + \frac{(b - a)^2}{2!}f''(a) + \int_a^b \phi_2 f'''\,dt,$$

y repitiendo n veces el proceso,

$$f(b) - f(a) = (b - a)f'(a) + \frac{(b - a)^2}{2!}f''(a) + \cdots + \frac{(b - a)^n}{n!}f^{(n)}(a)$$

$$+ \int_a^b (-1)^n \phi_n(t) f^{(n+1)}(t)\,dt$$

$$= (b - a)f'(a) + \frac{(b - a)^2}{2!}f''(a) + \cdots + \frac{(b - a)^n}{n!}f^{(n)}(a) + R_n,$$

donde, por definición de ϕ_n,

$$R_n = \int_a^b f^{(n+1)}(t)\,\frac{(b - t)^n}{n!}\,dt.$$

Hemos así demostrado nuevamente el

TEOREMA DE TAYLOR. *Si una función $f(t)$ tiene derivadas continuas hasta de orden $n + 1$ en un intervalo cerrado que contiene los puntos a y b, entonces:*

$$f(b) = f(a) + (b-a)f'(a) + \cdots + \frac{(b-a)^n}{n!}f^{(n)}(a) + R_n,$$

donde el residuo R_n, que depende de n, a y b, está dado por

$$R_n = \frac{1}{n!}\int_a^b (b-t)^n f^{(n+1)}(t)\,dt. \tag{27}$$

Mediante cambios de notación se obtienen expresiones ligeramente diferentes para la fórmula de Taylor. Así, reemplazando a por x y b por $x+h$ se tiene

$$f(x+h) = f(x) + hf'(x) + \cdots + \frac{h^n}{n!}f^{(n)}(x) + R_n, \tag{27a}$$

con

$$R_n = \frac{1}{n!}\int_x^{x+h} (x+h-t)^n f^{(n+1)}(t)\,dt;$$

y si $t = x + \tau$,

$$R_n = \frac{1}{n!}\int_0^h (h-\tau)^n f^{(n+1)}(x+\tau)\,d\tau. \tag{27b}$$

Haciendo $x = 0$ y escribiendo x en lugar de h, se obtiene [1]

$$f(x) = f(0) + \frac{x}{1!}f'(0) + \frac{x^2}{2!}f''(0) + \cdots + \frac{x^n}{n!}f^{(n)}(0) + R_n \tag{27c}$$

con el residuo

$$R_n = \frac{1}{n!}\int_0^x (x-t)^n f^{(n+1)}(t)\,dt.$$

Aplicando el teorema del valor medio del cálculo integral, o su forma generalizada, a la integral anterior, se llega, respectivamente, a la fórmula de Cauchy

$$R_n = \frac{(1-\theta)^n}{n!}x^{n+1}f^{(n+1)}(\theta x)$$

y a la fórmula de Lagrange

$$R_n = \frac{x^{n+1}}{(n+1)!}f^{(n+1)}(\theta x)$$

[1] Este caso especial del teorema, algunas veces se lo llama, sin justificación histórica, teorema de Maclaurin. El teorema general de Taylor fue publicado en 1715; el resultado especial de Maclaurin, en 1742.

para el residuo, tal como se vio antes (p. 467). Aquí θ es un número no especificado pero elegido adecuadamente y tal que $0 \leq \theta \leq 1$ (no necesariamente el mismo en ambas fórmulas).

Como un ejercicio, el lector podría construir funciones ϕ_ν que satisfagan (25) pero para las cuales las condiciones (26) se reemplazan por las relaciones

$$\int_0^1 \phi_\nu(t)\,dt = 0$$

para $\nu \geq 1$ (ver capítulo 8, Apéndice A).

5.5 Desarrollos de funciones elementales

Los resultados generales anteriores nos permiten desarrollar en serie de Taylor las funciones elementales más sencillas. En el capítulo 7 se discutirán los desarrollos para otras funciones.

a. La función exponencial

En primer lugar desarrollamos la función exponencial, $f(x) = e^x$. En este caso todas las derivadas son idénticas a $f(x)$ y tienen el valor 1 para $x = 0$. La forma de Lagrange para el residuo (p. 468, ecuación (21) conduce de inmediato a la fórmula:

$$e^x = 1 + \frac{x}{1!} + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots + \frac{x^n}{n!} + \frac{x^{n+1}}{(n+1)!}e^{\theta x}; \qquad 0 < \theta < 1.$$

Si ahora dejamos que n crezca más allá de toda cota, el residuo R_n tiende a cero para cualquier valor fijo de x. Para demostrar esto, nótese primero que $e^{\theta x} \leq e^{|x|}$, puesto que e^x es una función monótona creciente. Sea m cualquier entero mayor que $2|x|$. Entonces para todo $k > m$ se cumple $|x|/k < \frac{1}{2}$ y

$$\left|\frac{x^{n+1}}{(n+1)!}\right| = \frac{|x^m|}{m!}\cdot\frac{|x|}{m+1}\cdots\frac{|x|}{n+1} \leq \frac{|x^m|}{m!}\cdot\frac{1}{2^{n+1-m}} \leq \frac{|2x|^m}{m!}\cdot\frac{1}{2^n},$$

de manera que

$$|R_n| \leq \frac{|2x|^m}{m!}\cdot e^{|x|}\cdot\frac{1}{2^n}.$$

Puesto que los dos primeros factores de la derecha son independientes de n, en tanto que $1/2^n \to 0$ para $n \to \infty$, nuestra afirmación queda demostrada.

La función e^x se representa, por consiguiente, mediante la serie infinita

$$e^x = 1 + \frac{x}{1!} + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots$$

$$= \sum_{\nu=0}^{\infty} \frac{x^\nu}{\nu!}.$$

Este desarrollo es *válido para todos los valores de* x. En particular, para $x = 1$ se obtiene de nuevo la serie infinita que sirvió para definir el número e en el capítulo 1 (ver p. 100).

Por supuesto, para cálculos numéricos debe usarse la forma del teorema de Taylor con residuo. Por ejemplo, para $x = 1$ (compárese con el cálculo similar de la p. 101) se tiene

$$e = 1 + 1 + \frac{1}{2!} + \frac{1}{3!} + \cdots + \frac{1}{n!} + \frac{e^\theta}{(n+1)!}.$$

Si se desea calcular e con un error máximo de 1/10,000, lo único que se necesita es tomar n tan grande que el residuo sea menor que 1/10,000 y, puesto que este residuo es ciertamente menor[1] que $3/(n+1)!$, basta escoger $n = 7$, ya que $8! > 30{,}000$. De esta manera se obtiene el valor aproximado $e = 2.71825$, cuyo error es menor que 0.0001.

b. Desarrollos de sen x, cos x, senh x, cosh x

Para las funciones sen x, cos x, senh x, cosh x se encuentran las siguientes fórmulas·

$f(x) =$	sen x	cos x	senh x	cosh x
$f'(x) =$	cos x	$-$sen x	cosh x	senh x
$f''(x) =$	$-$sen x	$-$cos x	senh x	cosh x
$f'''(x) =$	$-$cos x	sen x	cosh x	senh x
$f^{(4)}(x) =$	sen x	cos x	senh x	cosh x

Así, en los polinomios de aproximación para sen x y senh x los coeficientes de las potencias pares de x se anularán, en tanto que para cos x y cosh x los coeficientes que se anulan son los de las potencias impares.

[1] Aquí se ha usado el hecho de que $e < 3$. Esto surge (ver p. 101) de nuestra serie para e; pues siempre es cierto que $1/n! \leq 1/2^{n-1}$, y por consiguiente

$$e < 1 + 1 + \tfrac{1}{2} + \tfrac{1}{4} + \cdots = \frac{1}{1 - \frac{1}{2}} = 3.$$

Cuando se usa la forma de Lagrange para el residuo, (21), p. 468, las series de Taylor para nuestras funciones resultan ser:

$$\operatorname{sen} x = x - \frac{x^3}{3!} + \frac{x^5}{5!} + \cdots + \frac{(-1)^n x^{2n+1}}{(2n+1)!} + \frac{(-1)^{n+1} x^{2n+3} \cos(\theta x)}{(2n+3)!};$$

$$\cos x = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - + \cdots + \frac{(-1)^n x^{2n}}{(2n)!} + \frac{(-1)^{n+1} x^{2n+2} \cos(\theta x)}{(2n+2)!},$$

$$\operatorname{senh} x = x + \frac{x^3}{3!} + \frac{x^5}{5!} + \cdots + \frac{x^{2n+1}}{(2n+1)!} + \frac{x^{2n+3} \cosh(\theta x)}{(2n+3)!},$$

$$\cosh x = 1 + \frac{x^2}{2!} + \frac{x^4}{4!} + \cdots + \frac{x^{2n}}{(2n)!} + \frac{x^{2n+2} \cosh(\theta x)}{(2n+2)!}$$

donde, por supuesto, en cada una de nuestras fórmulas θ denota un número diferente en el intervalo $0 \leq \theta \leq 1$, número que, además, depende de n y de x. Puesto que en cada una de estas fórmulas el residuo tiende a cero conforme n crece, lo cual puede verse por exactamente el mismo argumento que el usado para e^x, pueden hacerse las aproximaciones tan precisas como se desee. Se obtienen así las cuatro series infinitas, válidas para cualquier valor de x,

$$\operatorname{sen} x = x - \frac{x^3}{3!} + \frac{x^5}{5!} - + \cdots = \sum_{\nu=0}^{\infty} \frac{(-1)^\nu x^{2\nu+1}}{(2\nu+1)!},$$

$$\cos x = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - + \cdots = \sum_{\nu=0}^{\infty} \frac{(-1)^\nu x^{2\nu}}{(2\nu)!},$$

$$\operatorname{senh} x = x + \frac{x^3}{3!} + \frac{x^5}{5!} + \cdots = \sum_{\nu=0}^{\infty} \frac{x^{2\nu+1}}{(2\nu+1)!},$$

$$\cosh x = 1 + \frac{x^2}{2!} + \frac{x^4}{4!} + \cdots = \sum_{\nu=0}^{\infty} \frac{x^{2\nu}}{(2\nu)!}.$$

Las últimas dos pueden también obtenerse formalmente a partir de la serie para e^x de acuerdo con las definiciones de las funciones hiperbólicas (ver p. 250).

c. La serie binomial

Volvemos a las series de Taylor para las funciones $\log(1+x)$ y $\arctan x$, ya tratadas directamente en la sección 5.2. Sin embargo, trataremos la generalización del teorema binomial para exponentes arbitrarios, lo cual constituyó uno de los descubrimientos matemáticos de Newton más espectaculares. Deseamos desarrollar la función $f(x) = (1+x)^\alpha$ en una serie de Taylor para el caso en que $x > -1$ y α es un número arbitrario, positivo o negativo, racional o irracional. Se escoge la función $(1+x)^\alpha$ en vez de x^α porque para esta última no es cierto que todas las derivadas sean continuas para $x = 0$, excepto en el caso trivial de valores de α enteros no negativos. Primero calculamos las derivadas de $f(x)$, obteniendo

$$\begin{aligned} f'(x) &= \alpha(1+x)^{\alpha-1}, \\ f''(x) &= \alpha(\alpha-1)(1+x)^{\alpha-2}, \ldots, \\ &\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ f^{(\nu)}(x) &= \alpha(\alpha-1)\cdots(\alpha-\nu+1)(1+x)^{\alpha-\nu} \end{aligned}$$

En particular, para $x = 0$ resulta

$$\begin{aligned} f'(0) &= \alpha, \\ f''(0) &= \alpha(\alpha-1), \ldots, \\ f^{(\nu)}(0) &= \alpha(\alpha-1)\cdots(\alpha-\nu+1). \end{aligned}$$

El teorema de Taylor establece entonces que

$$(1+x)^\alpha = 1 + \alpha x + \frac{\alpha(\alpha-1)}{2!}x^2 + \cdots + \frac{\alpha(\alpha-1)(\alpha-2)\cdots(\alpha-n+1)}{n!}x^n + R_n.$$

Convergencia

Debemos todavía discutir el residuo. Este problema no es muy difícil pero, sin embargo, no es tan simple como los casos tratados anteriormente. Obtendremos una estimación para el residuo directamente y también como un caso especial del resultado general de la sección A.4. Esto nos permitirá concluir que, siempre que $|x| < 1$, el residuo R_n para el desarrollo binomial tiende a cero. Así, la expresión $(1+x)^\alpha$ podrá desarrollarse en una serie binomial infinita

$$\begin{aligned} (1+x)^\alpha &= 1 + \frac{\alpha}{1!}x + \frac{\alpha(\alpha-1)}{2!}x^2 + \cdots \\ &= \sum_{\nu=0}^{\infty}\binom{\alpha}{\nu}x^\nu, \end{aligned}$$

donde, por brevedad, se han introducido los coeficientes binomiales generales

$$\binom{\alpha}{\nu} = \frac{\alpha(\alpha-1)\cdots(\alpha-\nu+1)}{\nu!} \qquad (\text{para } \nu > 0),$$

$$\binom{\alpha}{0} = 1.$$

*Para ver directamente que $R_n \to 0$ para $n \to \infty$ en el caso en que $-1 < x < 1$, se hace uso de la forma de Cauchy para el residuo, (19), p. 468:

$$\begin{aligned} R_n &= \frac{(1-\theta)^n}{n!}x^{n+1}f^{(n+1)}(\theta x) \\ &= \frac{(1-\theta)^n}{n!}\alpha(\alpha-1)(\alpha-2)\cdots(\alpha-n)x^{n+1}(1+\theta x)^{\alpha-n-1} \end{aligned}$$

$(0 \le \theta \le 1)$. Puesto que $|x| < 1$, se tiene $0 \le (1-\theta)/(1+\theta x) \le 1$, de manera que

$$|R_n| \le (1+\theta x)^{\alpha-1}|\alpha x|\left|\left(1-\frac{\alpha}{1}\right)x\right|\left|\left(1-\frac{\alpha}{2}\right)x\right|\cdots\left|\left(1-\frac{\alpha}{n}\right)x\right|$$

Existe un número q tal que $|x| < q < 1$. Entonces es obvio que también

$$\left|\left(1-\frac{\alpha}{m}\right)x\right| < q$$

para todo m suficientemente grande (digamos, $m > N$). Así, para $n > N$ se cumple

$$|R_n| \le (1+\theta x)^{\alpha-1}|\alpha|\,(1+|\alpha|)^N q^{n-N}.$$

El factor $(1+x\theta)^{\alpha-1}$ es acotado (por $2^{\alpha-1}$ si $\alpha \ge 1$, y por $(1-q)^{\alpha-1}$ si $(\alpha < 1)$, de manera que, claramente, $R_n \to 0$.

Una fórmula un poco más general da una expresión para $(a+b)^\alpha$. Sólo tiene que factorizarse a^α y aplicarse el desarrollo binomial con $x = b/a$ para obtener, con $a > 0$ y $|b| < a$,

$$\begin{aligned} (a+b)^\alpha &= a^\alpha\left(1+\frac{b}{a}\right)^\alpha = a^\alpha\left(1+\alpha\frac{b}{a}+\frac{\alpha(\alpha-1)}{1\cdot 2}\left(\frac{b}{a}\right)^2+\cdots\right) \\ &= a^\alpha + \frac{\alpha}{1}a^{\alpha-1}b + \frac{\alpha(\alpha-1)}{1\cdot 2}a^{\alpha-2}b^2 + \cdots. \end{aligned}$$

5.6 Aplicaciones geométricas

El comportamiento de una función $f(x)$ en una vecindad del punto $x = a$, o el comportamiento de una cierta curva en la vecindad de uno de

sus puntos, puede describirse en detalles mediante el teorema de Taylor, ya que éste nos permite descomponer el incremento de la función, conforme se recorre una vecindad del punto $x = a + h$, en una suma de cantidades de primer orden, segundo orden, etc., en h.

a. Contactos entre curvas

Contacto de orden superior

Si en un punto $x = a$ dos curvas, $y = f(x)$ e $y = g(x)$, se intersectan y tienen una tangente común, se dice que las curvas *se* tocan una con otra o que tienen *contacto de primer orden.* En este caso los desarrollos de Taylor para las funciones $f(a + h)$ y $g(a + h)$ tienen los mismos términos de orden cero y de primer orden en h. Si, además, en el punto $x = a$ las segundas derivadas de $f(x)$ y $g(x)$ son iguales entre sí, se dice que las curvas tienen *contacto de segundo orden.* Entonces, los términos de segundo orden en los desarrollos de Taylor de f y de g también serán idénticos. Si se supone que ambas funciones tienen derivadas continuas por lo menos hasta de tercer orden, entonces la diferencia

$$D(x) = f(x) - g(x)$$

puede expresarse en la forma

$$\begin{aligned} D(a + h) &= f(a + h) - g(a + h) \\ &= \frac{h^3}{3!} D'''(a + \theta h) = \frac{h^3}{3!} F(h), \end{aligned}$$

donde la expresión $F(h)$ tiende a $f'''(a) - g'''(a)$ conforme h tiende a cero. Por lo tanto, la diferencia $D(a + h)$ se hace nula al menos de tercer orden con h.

Puede proseguirse de esta manera y considerarse el caso general en que las series de Taylor para $f(x)$ y $g(x)$ coincidan hasta los términos de n-ésimo orden: esto es,

$$f(a) = g(a),\ f'(a) = g'(a), \ldots, f^{(n)}(a) = g^{(n)}(a).$$

Se supone que las $(n + 1)$-ésimas derivadas son continuas. Bajo estas condiciones se dice que las curvas definidas por nuestras dos funciones tienen *contacto* de *n-ésimo orden* en el punto $x = a$. La diferencia de las dos funciones es entonces de la forma

$$\begin{aligned} D(a + h) &= f(a + h) - g(a + h) = \frac{h^{n+1}}{(n + 1)!} D^{(n+1)}(a + \theta h) \\ &= \frac{h^{n+1}}{(n + 1)!} F(h), \end{aligned}$$

donde la cantidad $F(h) = D^{(n+1)}(a + \theta h)$ tiende a $f^{(n+1)}(a) - g^{(n+1)}(a)$ conforme h tiende a cero, puesto que $0 \leq \theta \leq 1$. De esta fórmula se ve que en el punto de contacto. la diferencia $f(x) - g(x)$ se anula con, por lo menos, $(n + 1)$-ésimo orden.

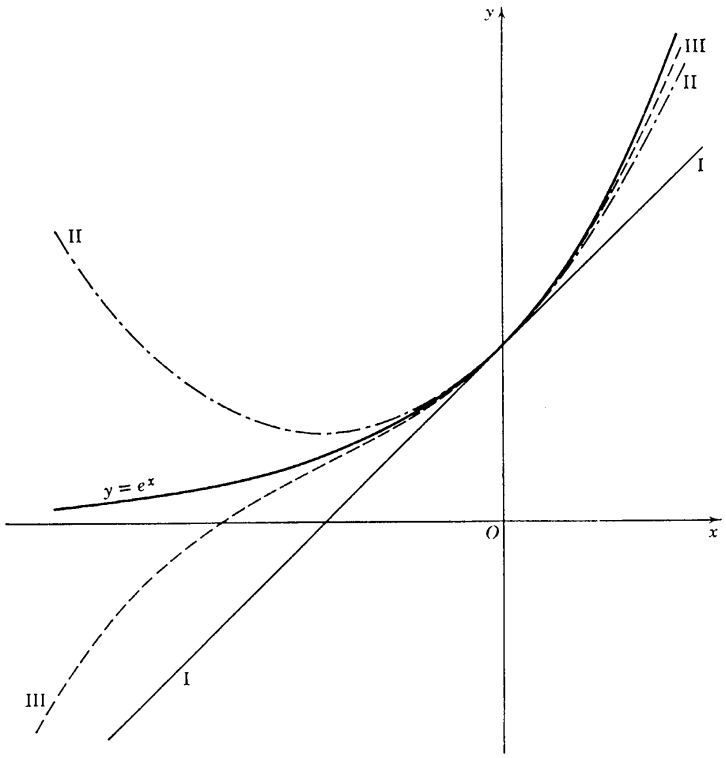


Figura 5.1 Parábolas osculadoras de e^x.

Los polinomios de Taylor definidos por

$$P_n(x) = f(a) + \frac{x - a}{1!} f'(a) + \cdots + \frac{(x - a)^n}{n!} f^{(n)}(a)$$

se caracterizan geométricamente como las "parábolas" de orden n que tienen contacto del mayor orden posible con la gráfica de la función en el punto dado. Por tanto, estas parábolas se llaman a veces *parábolas osculadoras.* (Solamente para $n = 2$ son estas curvas "parábolas" en el sentido ordinario.)

Para la función $y = e^x$ la Fig. 5.1 muestra las tres primeras parábolas osculadoras en el punto $x = 0$.

Dos curvas, $y = f(x)$ e $y = g(x)$, que tienen contacto de orden n en un punto $x = a$, podrían posiblemente tener contacto de orden aún mayor; esto es, la ecuación $f^{(n+1)}(a) = g^{(n+1)}(a)$ podría también ser cierta. Si no es este el caso, es decir, si $f^{(n+1)}(a) \neq g^{(n+1)}(a)$, se dice que el orden del contacto es exactamente n.[1]

Contacto de orden par o impar

Usando nuestras fórmulas y la intuición, podemos establecer un hecho notable a menudo no observado por los principiantes. Supóngase que el contacto entre dos curvas es exactamente de orden par; esto es, un número par n de derivadas de las dos funciones tienen el mismo valor en el punto en cuestión, en tanto que las $(n + 1)$-ésimas derivadas difieren entre sí. Las fórmulas anteriores muestran entonces que la diferencia $f(a + h) - g(a + h)$ tienen signos diferentes para pequeños valores positivos de h y también para valores negativos numéricamente pequeños de h. Por tanto, las dos curvas *se cruzan* en el punto de contacto. Esto ocurre, por ejemplo, para el contacto de segundo orden si las terceras derivadas tienen valores diferentes. Por el contrario, contactos de orden precisamente impar (por ejemplo, un contacto ordinario de primer orden) implica que la diferencia $f(a + h) - g(a + h)$ tenga el mismo signo para todos los valores de h positivos o negativos numéricamente pequeños; por consiguiente, las dos curvas no se cruzan en una vecindad del punto de contacto. El ejemplo más simple es el contacto de una curva con su tangente. La tangente puede cruzar la curva sólo en puntos en que el contacto es por lo menos de segundo orden; cruzará efectivamente la curva en puntos en que el contacto es par, por ejemplo, un punto de inflexión ordinario, en el cual $f''(x) = 0$ pero $f'''(x) \neq 0$. En los puntos en que el contacto es de orden impar la tangente no cruza la curva; esto ocurre, por ejemplo, en un punto ordinario de la curva en que la segunda derivada no es cero, tal como para la curva $y = x^4$ en el origen.

Del capítulo 4, p. 378, se sabe que para la circunferencia de curvatura en el punto $x = a$, dado por la función $y = g(x)$ en una vecindad de ese punto, no sólo se tiene $g(a) = f(a)$ y $g'(a) = f'(a)$, sino también $g''(a) = f''(a)$. Por tanto, la circunferencia de curvatura es al mismo tiempo la circunferencia osculadora de la curva en el punto considerado, es decir, es la que en ese punto tiene contacto de segundo orden con la curva. En el caso límite de un punto de inflexión o, en general, de un punto en el cual la curvatura sea cero y el radio de curvatura infinito, la circunferencia de curvatura degenera en la tangente. En casos ordinarios, cuando el contacto en el punto que se considera no es de orden superior al segundo, la circunferencia de curvatura no sólo toca la curva sino que también la cruza (ver Fig. 4.23, p. 377).

[1] Que el orden del contacto entre dos curvas es una genuina relación geométrica no afectada por cambios de ejes, es un hecho que puede confirmarse fácilmente por medio de las fórmulas para cambio de ejes (ver capítulo 4, p. 378).

Como conclusión, se debe mencionar que a veces un contacto de orden exactamente igual a m se describe diciendo: las curvas tienen $m + 1$ puntos en común infinitamente próximos. Por supuesto, el significado preciso de tal afirmación se refiere obviamente a un proceso de límite. Si las curvas tienen en efecto $m + 1$ puntos distintos $P, P_1, \ldots, P_m$ en común y si se deja que todos los puntos P_i tiendan a P, modificando una de las curvas si es necesario, entonces podría esperarse que la posición límite fuese la de dos curvas con un contacto de orden m. Por ejemplo, si se dibuja una circunferencia que pase por tres puntos P, P_1 P_2 de una curva C y luego se hace que P_1 y P_2 tiendan a P, puede verse que la circunferencia tiende a la circunferencia de curvatura de C en P. (Ver problema 4, p. 455.)

b. Sobre la teoría de máximos y mínimos relativos

Como ya se ha visto en el capítulo 3, pp. 264-265, una función $f(x)$ cuya primera derivada se anula en $x = a$ tiene un máximo relativo en ese punto si $f''(a)$ es negativo y un mínimo si $f''(a)$ es positivo. Estas condiciones son, por tanto, suficientes para la existencia de un máximo o un mínimo. Pero de ninguna manera son necesarias, pues en el caso en que $f''(a) = 0$ existen tres posibilidades: en ese punto la función puede tener un máximo, un mínimo o ninguna de las dos cosas. Ejemplos de las tres posibilidades se tienen en las funciones $y = -x^4$, $y = x^4$ e $y = x^3$ en el punto $x = 0$. El teorema de Taylor nos permite de inmediato establecer un teorema general sobre condiciones suficientes para máximo o mínimo. Solamente se necesita desarrollar la función $f(a + h)$ en potencias de h. El punto esencial es entonces determinar si el primer término que no se anula contiene una potencia par o impar de h. En el primer caso se tiene un máximo o un mínimo según que el coeficiente de h sea negativo o positivo; en el segundo caso se tiene una tangente de inflexión horizontal y no se tiene ni máximo ni mínimo. El lector puede por sí mismo completar el argumento usando la fórmula para el residuo.[1]

[1] La condición necesaria y suficiente dada antes (p. 263) es, sin embargo, más general y más conveniente para las aplicaciones: suponiendo que la primera derivada $f'(x)$ se anule solamente en un número finito de puntos, una condición necesaria y suficiente para la existencia de un máximo o de un mínimo en esos puntos es que la primera derivada $f'(x)$ cambie de signo conforme la curva pasa a través del punto.

Apéndice I

A.I.1 Ejemplo de una función que no se puede desarrollar en una serie de Taylor

La posibilidad de expresar una función mediante una serie de Taylor con residuo de orden $(n + 1)$ depende esencialmente de la continuidad y la derivabilidad de la función en el punto considerado. Por esta razón $\log x$ no puede representarse por medio de una serie de Taylor en potencias de x; y lo mismo es cierto para la función $x^{1/3}$, cuya derivada es infinita en $x = 0$.

Para que una función pueda ser desarrollada en una serie de Taylor infinita todas sus derivadas deben necesariamente existir en el punto en cuestión. Esta condición, sin embargo, de ninguna manera es suficiente. Una función tal que todas sus derivadas existan y sean continuas en todo un intervalo no necesariamente es desarrollable en serie de Taylor; esto es, el residuo R_n en el teorema de Taylor puede no tender a cero al crecer n, no importa cuán pequeño sea el intervalo en el cual se desee desarrollar la función.

Un ejemplo importante y sencillo de este fenómeno lo constituye la función

$$y = f(x) = e^{-1/x^2} \quad \text{para} \quad x \neq 0,\ f(0) = 0,$$

que ya ha sido considerada en el apéndice del capítulo 3, p. 276. Esta función y sus derivadas son continuas en cualquier intervalo, aún en $x = 0$, como ya se ha visto, y en este punto todas las derivadas se anulan, es decir, $f^{(n)}(0) = 0$ para todo valor de n. (Geométricamente esto significa que la recta $y = 0$ tiene *contacto de orden infinito* con la curva en el punto $x = 0$). Por tanto, en el desarrollo de Taylor

$$f(0) + \frac{x}{1!}f'(0) + \frac{x^2}{2!}f''(0) + \cdots$$

todos los coeficientes del polinomio de aproximación $P_n(x)$ se anulan sin importar qué valor se escoja para n. Así, el residuo permanece igual a la función misma y, por tanto, excepto para $x = 0$, no puede tender a cero conforme n crece, ya que la función es positiva para todo otro valor de x.

De paso se observa que esta función es útil para la construcción de funciones que exhiben fenómenos no esperados por la intuición. Por ejemplo, la función

$$g(x) = e^{-1/x^2} \operatorname{sen}(1/x)$$

complementada con $g(0) = 0$ es una función con derivadas de todo orden, las cuales se anulan en $x = 0$. La gráfica de $y = g(x)$ cerca de $x = 0$ intersecta el eje x una infinidad de veces oscilando otras tantas.

A.I.2 Ceros e infinitos de funciones

a. Ceros de orden n

El desarrollo de Taylor de una función $f(x)$ permite caracterizar el orden con el cual se anula la función en un punto $x = a$. Se dice que una función $f(x)$ tiene un cero de multiplicidad exactamente n en $x = a$, o que se anula ahí exactamente con orden n, si $f(a) = 0, f'(a) = 0, f''(a) = 0, \ldots, f^{(n-1)}(a) = 0$ y $f^{(n)}(a) \neq 0$. Se supone expresamente que en la vecindad del punto la función tiene derivadas continuas al menos hasta de orden n. De la definición se sigue que la serie de Taylor para la función en la vecindad de dicho punto puede escribirse en la forma

$$f(a + h) = \frac{h^n}{n!}F(h) = \frac{h^n}{n!}f^{(n)}(a + \theta h), \quad 0 < 1, \tag{28}$$

expresión en la cual, conforme h tiende a cero, el factor $F(h) = n!f(a+h)/h^n$ tiende a un límite diferente de cero, a saber el valor $f^{(n)}(a)$. Por consiguiente, $f(a + h)$ tiene el mismo orden que h^n para $h \to 0$, o se anula con orden n (en el sentido de la definición dada en el capítulo 3, p. 273).

Análogamente, desarrollando las derivadas $f'(x), f''(x), \ldots, f^{(\nu)}(x)$ mediante el teorema de Taylor con la forma de Lagrange para el residuo, se obtiene una serie de expresiones

$$\begin{aligned} f'(a + h) &= \frac{h^{n-1}}{(n-1)!}F_1(h) = \frac{h^{n-1}}{(n-1)!}f^{(n)}(a + \theta h), \\ &\vdots \\ f^{\nu}(a + h) &= \frac{h^{n-\nu}}{(n-\nu)!}F_\nu(h) = \frac{h^{n-\nu}}{(n-\nu)!}f^{(n)}(a + \theta h), \end{aligned} \tag{29}$$

en todas las cuales los factores θ *pueden ser diferentes* en tanto que los factores $F_1, F_2, \ldots, F_\nu$ tienden *continuamente* a $f^{(n)}(a)$ conforme $h \to 0$. Por tanto, f' se anula con orden $n - 1$, f'' con orden $n - 2$, etc.

En estas fórmulas, por supuesto, se hace la hipótesis de que $f(x)$ se anula con orden $n \geqslant \nu$.

b. Infinito de orden ν

Si una función $\phi(x)$ está definida en una vecindad del punto $x = a$, excepto quizá en ese punto, y si puede escribirse $\phi(x) = f(x)/g(x)$ de manera que en $x = a$ el numerador no se anule pero el denominador tenga un cero de multiplicidad ν, se dice que $\phi(x)$ se vuelve infinita de orden ν en el punto $x = a$. Si en $x = a$ el numerador tiene un cero de multiplicidad μ

y si $\mu > \nu$, la función tiene en ese punto un cero de multiplicidad $\mu - \nu$; si $\mu < \nu$, la función tiene una infinidad de orden $\nu - \mu$ en el punto en cuestión.

Estas definiciones están de acuerdo con las convenciones adoptadas antes (ver sección 3.7) relativas al comportamiento de una función.

A.I.3 Expresiones indeterminadas

Discutiremos ahora de manera más precisa las *"expresiones indeterminadas"* de la forma $\phi(x) = f(x)/g(x)$ tal que $f(x)$ y $g(x)$ se anulan ambas en el mismo punto $x = a$, como ocurre para la función $(\operatorname{sen} x)/x$ en $x = 0$. A tales funciones les asignaremos siempre el valor

$$\phi(a) = \lim_{h \to 0} \phi(a + h) \tag{30}$$

suponiendo que este límite exista.

Estos valores límite pueden ser caracterizados mediante una sencilla regla, conocida como *regla de L'Hospital,* para lo cual supondremos la continuidad, en un intervalo que contenga al punto a, de todas las derivadas de f y g que aparezcan. Supondremos además que el denominador $g(x)$ se anula en $x = a$ con orden ν no mayor que el orden correspondiente para el numerador, de manera que la función $\phi(x)$ no se vuelva infinita en $x = a$. La regla establece entonces que

$$\phi(a) = \frac{f^{(\nu)}(a)}{g^{(\nu)}(a)}. \tag{31}$$

Por la definición de continuidad, la función $\phi(x)$ es entonces continua en $x = a$ y, puesto que es continua en todas partes, siempre que $g(x) \neq 0$, es continua en un intervalo en torno de a.

La prueba se sigue inmediatamente de los resultados de A.2. Aplicando las Ecs. (28) tanto a f como a g, se encuentra que la función ϕ está dada. en una vecindad de a, por la relación

$$\phi(a + h) = \frac{f(a + h)}{g(a + h)} = \frac{f^{(\nu)}(a + \theta h)}{g^{(\nu)}(a + \theta_1 h)},$$

de la cual se sigue (31) al considerar la continuidad del numerador y del denominador y el hecho de que $g^{(\nu)}(a)$ no se anula. El significado de las últimas ecuaciones puede expresarse de la siguiente manera: si el numerador y el denominador de una función $\phi(x) = f(x)/g(x)$ se anulan en $x = a$, puede determinarse el valor límite de la función para $x \to a$ derivando numerador y denominador el mismo número de veces hasta que al menos una de las derivadas no se anule en ese punto. Si se encuentra en el denominador una derivada no nula antes que en el numerador, la fracción tiende a cero. Por el contrario, si la derivada no nula aparece primero en el numerador, el valor absoluto de la fracción crece más allá de toda cota.

Se tiene así un método para evaluar la llamada "expresión indeterminada" 0/0, esto es, para determinar el límite de un cociente en el que tanto el numerador como el denominador tienden a cero.

Estos resultados pueden obtenerse de manera un tanto diferente si se basa la deducción en el teorema generalizado del valor medio en vez del teorema de Taylor (ver p. 243). Según esto, si $g'(x) \neq 0$ en una vecindad del punto a, se tiene

$$\frac{f(a + h) - f(a)}{g(a + h) - g(a)} = \frac{f'(a + \theta h)}{g'(a + \theta h)},$$

donde θ es el mismo en numerador y denominador. Por tanto, en el caso particular en que $f(a) = 0 = g(a)$,

$$\frac{f(a + h)}{g(a + h)} = \frac{f'(a + \theta h)}{g'(a + \theta h)}.$$

Aquí θ es un valor en el intervalo $0 < \theta < 1$, y si se hace $k = \theta h$ se obtiene

$$\lim_{h \to 0} \frac{f(a + h)}{g(a + h)} = \lim_{k \to 0} \frac{f'(a + k)}{g'(a + k)},$$

suponiendo que el límite de la derecha existe.

Si $f'(a) = 0 = g'(a)$, se procede de la misma manera hasta llegar a un primer índice μ para el cual ya no se sigue teniendo $f^{(\mu)}(a) = 0 = g^{(\mu)}(a)$. Entonces

$$\lim_{h \to 0} \frac{f(a + h)}{g(a + h)} = \lim_{l \to 0} \frac{f^{(\mu)}(a + l)}{g^{(\mu)}(a + l)} = \frac{f^{(\mu)}(a)}{g^{(\mu)}(a)},$$

expresión que incluye el caso en que ambos miembros son infinitos.

Ejemplos. Los siguientes ejemplos, importantes por sí mismos, ilustran la aplicación de la regla de L'Hospital.

$$\lim_{x \to 0} \frac{\operatorname{sen} x}{x} = \frac{\cos 0}{1} = 1;$$

$$\lim_{x \to 0} \frac{1 - \cos x}{x} = \frac{\operatorname{sen} 0}{1} = 0;$$

$$\lim_{x \to 0} \frac{e^{2x} - 1}{\log (1 + x)} = \lim_{x \to 0} \frac{2e^{2x}}{1/(1 + x)} = 2;$$

$$\lim_{x \to 0} \frac{1 - \cos x}{x^2} = \lim_{x \to 0} \frac{\operatorname{sen} x}{2x} = \lim_{x \to 0} \frac{\cos x}{2} = \frac{1}{2}.$$

Otras formas indeterminadas. Como siguiente paso se observa que otras de las llamadas formas indeterminadas pueden también reducirse al caso ya considerado; por ejemplo, el límite de

$$\frac{1}{\operatorname{sen} x} - \frac{1}{x}$$

cuando x tiende a cero es el límite de la diferencia de dos expresiones que se hacen infinitas, o bien, es una "forma indeterminada" $\infty - \infty$. Mediante la transformación

$$\frac{1}{\operatorname{sen} x} - \frac{1}{x} = \frac{x - \operatorname{sen} x}{x \operatorname{sen} x}$$

se llega de inmediato a una expresión cuyo límite, cuando x tiende a cero, puede encontrarse aplicando la regla. Así resulta

$$\lim_{x\to 0} \frac{1 - \cos x}{x \cos x + \operatorname{sen} x} = \lim_{x\to 0} \frac{\operatorname{sen} x}{2 \cos x - x \operatorname{sen} x} = 0.$$

Derivadas de formas indeterminadas

Las expresiones $\phi(x) = f(x)/g(x)$ definidas por la regla en $x = a$ no solamente son continuas sino que tienen derivadas también continuas, suponiendo que f y g tienen derivadas continuas de orden suficientemente alto.

Bastará aquí establecer este hecho en el caso en que g se anula en a con primer orden, o sea, $g(a) = 0$, $g'(a) \neq 0$. Para $x \neq a$,

$$\phi'(x) = \frac{g(x)f'(x) - f(x)g'(x)}{(g(x))^2} = \frac{z(x)}{N(x)},$$

donde tanto el numerador como el denominador se anulan en $x = a$, pues $f(a) = g(a) = 0$. Por tanto, el valor límite puede determinarse aplicando la regla

$$\lim_{x\to a} \phi'(x) = \lim_{x\to 0} \frac{z'(x)}{N'(x)}.$$

Claramente, $d(N(x))/dx = 2g(x)g'(x)$, $d(z(x))/dx = g(x)f''(x) - f(x)g''(x)$, las cuales se anulan de nuevo para $x = a$. Aplicando la regla de L'Hospital una vez más,

$$\lim_{x\to a} \phi'(x) = \frac{z''(x)}{N''(x)},$$

y observando que $N''(x) = 2g(x)g''(x) + 2(g'(x))^2$, que no se anula en $x = a$, se encuentra que

$$\lim_{x\to a} \phi'(x) = \frac{g'(a)f''(a) - f'(a)g''(a)}{2(g'(a))^2},$$

y este límite es, sin duda, la derivada de $\phi'(x)$ en $x = a$ (ver capítulo 3, p. 282).

Reglas análogas para formas indefinidas valen para $x \to \infty$. Así, por ejemplo, sean $f(x)$ y $g(x)$ funciones tales que $\lim_{x\to\infty} f(x) = \lim_{x\to\infty} g(x) = 0$, en tanto que $\lim_{x\to\infty} f'(x)$ y $\lim_{x\to\infty} g'(x)$ existen y son diferentes de cero. Entonces

$$\lim_{x\to\infty} \frac{f(x)}{g(x)} = \frac{\lim_{x\to\infty} f'(x)}{\lim_{x\to\infty} g'(x)}$$

La demostración surge otra vez del teorema del valor medio del cálculo diferencial.

*A.I.4 La convergencia de la serie de Taylor para una función con derivadas no negativas de todos los órdenes

Se incluye a continuación un teorema general sobre la convergencia del desarrollo de Taylor para funciones cuyas derivadas son todas no negativas.

Considérese la clase de funciones $f(x)$ derivables hasta cualquier orden en el intervalo cerrado $a \leq x \leq b$ y tales que todas sus derivadas son no negativas en este intervalo:

$$f^{(\nu)}(x) \geq 0, \qquad \nu = 1, 2, \ldots.$$

Demostraremos que: Para cualquiera de tales funciones el correspondiente desarrollo de Taylor de $f(x + h)$ en potencias de h converge, y la serie representa el valor de $f(x + h)$ cuando x y $\xi = x + h$ están en el intervalo abierto (a, b) y $|h| < b - x$.

La demostración se inicia observando que $f'(x) \geq 0$, por hipótesis, y que, en consecuencia,

$$0 \leq f(x) - f(a) = \int_a^x f'(\xi)\, d\xi$$

$$\leq \int_a^b f'(\xi)\, d\xi = f(b) - f(a) = M.$$

Además, para x y $\xi = x + h$ en el intervalo (a, b), puede escribirse

$$f(x + h) - f(x) = hf'(x) + \cdots + \frac{f^{(n)}(x)}{n!} h^n + R_n.$$

Supóngase primero que $h > 0$, o sea, $x < \xi < b$. Entonces todos los términos del miembro derecho son no negativos,[1] de manera que ninguno de ellos

[1] Esto se deduce para R_n partiendo de las fórmulas de Cauchy o de Lagrange y con la hipótesis $f^{(n+i)} \geq 0$.

es mayor que el valor del miembro izquierdo o que M. Así resulta que

$$0 \leq \frac{f^{(n)}(x)}{n!} \leq \frac{M}{h^n} = \frac{M}{(\xi - x)^n}.$$

Para $\xi \to b$ surge que

$$\frac{f^{(n)}(x)}{n!} \leq \frac{M}{(b-x)^n}. \tag{32}$$

Ahora bien, por la fórmula de auchy ((19), p. 468) para el residuo, se sabe que existe algún θ en el intervalo $0 \leq \theta \leq 1$ tal que

$$0 \leq R_n = \frac{(1-\theta)^n}{n!} h^{(n+1)} f^{(n+1)}(x + \theta h)$$

$$\leq \frac{h^{n+1}(n+1)(1-\theta)^n M}{(b - x - \theta h)^{n+1}}.$$

Puesto que $\xi = x + h < b$, puede escogerse un número positivo p de manera que

$$0 \leq h \leq \frac{b-x}{1+p}, \quad \text{o sea,} \quad b - x - \theta h \geq h(1 + p - \theta).$$

Se tiene entonces

$$0 \leq R_n \leq \frac{M h^{n+1}(n+1)(1-\theta)^n}{h^{n+1}(1+p-\theta)^{n+1}},$$

o sea,

$$0 \leq R_n \leq \frac{M(n+1)}{(1+p-\theta)} \left(\frac{1-\theta}{1-\theta+p}\right)^n \leq \frac{M(n+1)}{p} \frac{1}{(1+p)^n},$$

ya que

$$\frac{1-\theta}{1-\theta+p} = \frac{1}{1 + p/(1-\theta)} \leq \frac{1}{1+p} < 1.$$

Se sabe (capítulo 1, p. 93) que $(n+1)/(1+p)^n$ tiende a cero al crecer n, de manera que cuando $0 \leq h < b - x$ y n crece, R_n tiende a cero; así, la serie de Taylor tiende a la función f para $h \geq 0$.

Para h negativo, el que R_n tienda a cero al crecer n se deduce aplicando la fórmula de Lagrange ((21), p. 468) para R_n:

$$|R_n| = \frac{1}{(n+1)!} |h^{n+1}| \, |f^{(n+1)}(x - \theta |h|)|.$$

Ahora bien, $f^{(n+2)}$ es no negativa y, por tanto, $f^{(n+1)}$ es monótona no decreciente. Se sigue entonces de la estimación (32) usada antes que

$$\frac{f^{(n+1)}(x - \theta |h|)}{(n+1)!} \leq \frac{f^{(n+1)}(x)}{(n+1)!} \leq \frac{M}{(b-x)^{n+1}}.$$

Por consiguiente,

$$|R_n| \leq \left(\frac{|h|}{b-x}\right)^{n+1} M,$$

y así resulta que R_n tiende a cero conforme n crece, cuando

$$0 < -h < b - x.$$

Entonces, para cualquier punto x tal que $a \leq x < b$, el residuo R_n en la serie de Taylor para $f(x+h)$ en potencias de h tenderá a cero siempre que $|h| < b - x$ y $h > -(x-a)$.

Obsérvese que el resultado anterior es válido aun en el caso en que se suponga la desigualdad $f^{(\nu)}(x) \geq 0$ válida sólo para valores de ν suficientemente grandes (digamos, para $\nu > N$, siendo N algún entero, pudiendo tener $f^{(\nu)}(x)$ cualquier signo para $\nu \leq N$). Para demostrar esto solamente es necesario reemplazar en la demostración anterior la función f por la función

$$g(x) = f(x) + M(x - a + 1)^N,$$

donde M es alguna constante positiva. Entonces, $g^{(\nu)}(x) = f^{(\nu)}(x) \geq 0$ para $\nu > N$, y $g^{(\nu)}(x) = f^{(\nu)}(x) + MN(N-1)\cdots(N-\nu+1)(x-a+1)^{N-\nu} \geq f^{(\nu)}(x) + M$ para $\nu \leq N$. Así, $g^{(\nu)}(x) \geq 0$ para todo ν si M se escoge suficientemente grande. Esto prueba que $g(x)$ puede desarrollarse en potencias de x; y el mismo resultado se deduce entonces para la función f, la cual difiere de g sólo por un polinomio.

El teorema de la serie binomial (p. 475) es una consecuencia inmediata de este resultado: Cámbiese ligeramente la notación y considérese primero la función $\phi(x) = (1-x)^\alpha$ en vez de $(1+x)^\alpha$. Las derivadas de ϕ están dadas entonces por

$$\phi^{(\nu)}(x) = (-1)^\nu \binom{\alpha}{\nu} (1-x)^{\alpha-\nu} \nu!$$

Puesto que los coeficientes binomiales,

$$\binom{\alpha}{\nu} = \frac{\alpha(\alpha-1)\cdots(\alpha-\nu+1)}{\nu!},$$

tienen signos alternantes cuando $\alpha - \nu$ es negativo, una de las dos funciones, $\phi(x)$ o $-\phi(x)$, pertenece a la clase de funciones con derivadas no negativas a partir de cierto orden, para x limitada a valores $x < 1$. Así, para $a = -1$, $b = 1$, $x = 0$ y $|h| < b - x = 1$ nuestro teorema general afirma que

$$(1-h)^\alpha = \sum_{\nu=0}^{\infty} (-1)^\nu \binom{\alpha}{\nu} h^\nu.$$

Si aquí se escribe x en vez de $-h$, se obtiene el desarrollo binomial

$$(1+x)^\alpha = \sum_{\nu=0}^{\infty}\binom{\alpha}{\nu}x^\nu = 1 + \alpha x + \frac{\alpha(\alpha-1)}{1\cdot 2}x^2 + \frac{\alpha(\alpha-1)(\alpha-2)}{1\cdot 2\cdot 3}x^3 + \cdots$$

para cualquier exponente α y cualquier x tal que $-1 < x < 1$.

Apéndice II Interpolación

*A.II.1 El problema de la interpolación. Unicidad

El polinomio de Taylor $P_n(x)$ aproxima a la función $f(x)$ de manera tal que las gráficas de $f(x)$ y $P_n(x)$ tienen contacto de orden n en un punto a, o sea, de manera que $f(x)$ y $P_n(x)$ coinciden en $n+1$ puntos "infinitamente próximos" a a. Podría "descomponerse" el punto con abscisa a en $n+1$ puntos distintos con abscisas $x_0, x_1, \ldots, x_n$ y buscarse una aproximación para $f(x)$ mediante un polinomio $\phi(x)$ de grado n que coincida con $f(x)$ en estos puntos. Este polinomio, como es fácil ver, está determinado unívocamente por un sistema de ecuaciones lineales. Mediante el paso al límite, $x_i \to x$ para todo i, se obtienen de nuevo los polinomios de Taylor. La "interpolación", esto es, la aproximación mediante polinomios coincidentes con $f(x)$ en puntos distintos, es de gran importancia en muchas aplicaciones. La siguiente discusión proporciona una idea sumaria de la teoría de la interpolación.

Considérese el siguiente problema: Determinar un polinomio $\phi(x)$ de grado n tal que en $n+1$ puntos distintos, $x_0, x_1, \ldots, x_n$, tome los $n+1$ valores dados $f_0, f_1, \ldots, f_n$, o sea,

$$\phi(x_0) = f_0, \quad \phi(x_1) = f_1, \ldots, \quad \phi(x_n) = f_n.$$

Si los números f_i son los valores $f_i = f(x_i)$ que toma una función dada, $f(x)$ (posiblemente menos elemental), en los puntos x_i, entonces el polinomio $\phi(x)$ se llamará el *polinomio de interpolación* de n-ésimo grado de la función $f(x)$ para los puntos $x_0, x_1, \ldots, x_n$.

Existe a lo más solamente uno de tales polinomios de grado n, pues si existiesen dos de ellos, $\phi(x)$ y $\psi(x)$ distintos, su diferencia $D(x) = \phi(x) - \psi(x)$ sería un polinomio de grado m, con $0 \le m \le n$ y con $n+1$ raíces distintas, lo cual no es posible según el álgebra elemental.[1]

Puede probarse la *unicidad* del polinomio de interpolación aún por otro método, basado en el

[1] Pues se tiene

$$D(x) = c_0(x-x_1)(x-x_2)\cdots(x-x_m), \quad c_0 \neq 0,$$

ya que $x_1, \ldots, x_m$ son ceros de $D(x)$; por tanto, como $D(x_0) = 0$,

$$c_0(x_0-x_1)(x_0-x_2)\cdots(x_0-x_m) = 0,$$

contrario al hecho de que $x_0, x_1, \ldots, x_m$ son todos distintos.

TEOREMA GENERAL DE ROLLE. *Si una función $F(x)$ tiene derivadas continuas hasta el orden n en un intervalo y se anula por lo menos en $n+1$ puntos distintos, $x_0, x_1, \ldots, x_n$, del mismo, entonces existe un punto en el interior de dicho intervalo para el cual $F^{(n)}(\xi) = 0$.*

DEMOSTRACIÓN. El teorema general se deduce fácilmente del caso especial en que $n = 1$, que es el teorema de Rolle demostrado en la p. 196. Supónganse los números $x_0, x_1, \ldots, x_n$ dispuestos en orden creciente. Entonces, por el teorema del valor medio (o por el teorema de Rolle), la primera derivada $F'(x)$ debe anularse por lo menos una vez dentro de cada uno de los n subintervalos (x_i, x_{i+1}). Esta misma consideración aplicada a $F'(x)$ y a los intervalos entre sus ceros nos dice que $F''(x)$ se anula en $n-1$ puntos. Aplicando repetidamente este argumento se demuestra el teorema.

Apliquemos ahora este teorema a la diferencia

$$F(x) = D(x) = \phi(x) - \psi(x)$$
$$= d_0x^n + d_1x^{n-1} + \cdots + d_n,$$

que, por hipótesis, se anula en $n+1$ puntos. Obtenemos un punto ξ en el cual la n-ésima derivada se anula. $D^{(n)}(\xi) = 0$. Esta es, precisamente, $n!d_0$, de manera que $d_0 = 0$, y la diferencia dada es un polinomio de grado a lo más igual a $n-1$ y que se anula en $n+1$ puntos. Aplicando nuevamente el teorema de Rolle se obtiene $d_1 = 0$; etc. O sea $D(x)$ es idénticamente cero, como se afirmó.

Estas consideraciones pueden extenderse al caso en que los x_i no son todos distintos entre sí, sino que, posiblemente r valores x_i coinciden, es decir, $x_0 = x_1 = \cdots = x_{r-1}$. En el problema de la interpolación para estos casos se requiere entonces que $\phi(x)$ y las derivadas $\phi'(x), \ldots, \phi^{(r-1)}(x)$ tomen valores prefijados para $x = x_0$, y, correspondientemente, también para los otros puntos x_ν. El polinomio $D(x)$ es entonces de la forma $c(x-x_0)^r(x-x_r)\cdots$. El teorema general de Rolle y el teorema de unicidad, así como también las demostraciones, permanecen inalterados en este caso.

A.II.2 Construcción de la solución. Fórmula de interpolación de Newton

Construiremos ahora un polinomio de interpolación $\phi(x)$ de grado n, tal que $\phi(x_0) = f_0, \ldots, \phi(x_n) = f_n$. Para construirlo paso a paso, comenzamos con la constante f_0, que es un polinomio $\phi_0(x)$ de orden 0 que toma el valor $A_0 = f_0$ para todo x y en particular para $x = x_0$. Sumamos a éste

un polinomio de primer orden que se anula para $x = x_0$ y es, por consiguiente, de la forma $A_1(x - x_0)$. A continuación determinamos A_1 de manera que la suma tenga para $x = x_1$ el valor correcto f_1. Al polinomio de primer grado que resulta le llamamos $\phi_1(x)$. Sumamos ahora a $\phi_1(x)$ un polinomio de segundo grado que se anula para $x = x_0$ y $x = x_1$, debiendo por consiguiente ser de la forma $A_2(x - x_0)(x - x_1)$, cuya adición no cambia el comportamiento en estos dos puntos. En seguida se determina el factor A_2 de manera que el polinomio de segundo grado resultante, $\phi_2(x)$, tome también el valor fijado, que es f_2 para $x = x_2$ en este caso. Este procedimiento se continúa hasta cubrir todos los puntos, obteniéndose así el polinomio

$$\phi(x) = \phi_n(x) = A_0 + A_1(x - x_0) + A_2(x - x_0)(x - x_1) + \cdots + A_n(x - x_0) \cdots (x - x_{n-1}). \tag{33}$$

El método para la obtención de los coeficientes A_i en la expresión de ϕ se aclara substituyendo $x = x_0$, $x = x_1$, ..., $x = x_n$ en este mismo orden; y así se obtiene el sistema de $n + 1$ ecuaciones

$$\begin{aligned} f_0 &= A_0 \\ f_1 &= A_0 + A_1(x_1 - x_0) \\ f_2 &= A_0 + A_1(x_2 - x_0) + A_2(x_2 - x_0)(x_2 - x_1) \\ &\cdots\cdots\cdots\cdots\cdots\cdots \\ f_n &= A_0 + A_1(x_n - x_0) + \cdots + A_n(x_n - x_0)(x_n - x_1) \cdots (x_n - x_{n-1}). \end{aligned} \tag{34}$$

Obviamente, los coeficientes $A_0, A_1, \ldots, A_n$ pueden determinarse en forma sucesiva de manera tal que satisfagan estas ecuaciones, pudiendo construirse así el polinomio de interpolación.

Cuando los valores x_ν son *equidistantes*, o sea $x_\nu = x_{\nu-1} + h$, el resultado puede escribirse explícitamente en una forma más elegante. Las ecuaciones para los A_i se convierten ahora en

$$\begin{aligned} f_0 &= A_0 \\ f_1 &= A_0 + hA_1, \\ f_2 &= A_0 + 2hA_1 + 2!\,h^2A_2 \\ f_3 &= A_0 + 3hA_1 + 3 \cdot 2h^2A_2 + 3!\,h^3A_3 \\ &\cdots\cdots\cdots\cdots\cdots \\ f_n &= A_0 + nhA_1 + \cdots + \frac{n!}{(n-i)!}h^iA_i + \cdots + n!\,h^nA_n. \end{aligned} \tag{35}$$

Las soluciones pueden fácilmente ser expresadas como *diferencias sucesivas* de f: Dada cualquier sucesión (finita o infinita) de términos $f_0, f_1, f_2, \ldots$, llamamos a las expresiones

$$\Delta f_0 = f_1 - f_0, \quad \Delta f_1 = f_2 - f_1, \quad \Delta f_2 = f_3 - f_2, \ldots$$

las *primeras diferencias* de las f_k. Aplicando nuevamente el proceso de diferencias a la sucesión de Δf_k, se obtienen las expresiones

$$\Delta^2 f_0 = \Delta f_1 - \Delta f_0, \quad \Delta^2 f_1 = \Delta f_2 - \Delta f_1, \quad \Delta^2 f_2 = \Delta f_3 - \Delta f_2, \ldots,$$

o sea,

$$\Delta^2 f_0 = f_2 - 2f_1 + f_0, \quad \Delta^2 f_1 = f_3 - 2f_2 + f_1, \ldots,$$

que son las *segundas diferencias* de las f_k. La n-ésima diferencia se define recursivamente como $\Delta^{n-1}f_{k+1} - \Delta^{n-1}f_k$. Y cuando se la expresa directamente en términos de las f_k, está dada por la fórmula

$$\Delta^n f_k = f_{k+n} - \binom{n}{1} f_{k+n-1} + \binom{n}{2} f_{k+n-2} - \cdots + (-1)^n f_k, \tag{36}$$

obtenida mediante un simple argumento inductivo que se deja al lector como ejercicio. Con esta terminología, los coeficientes A_ν pueden escribirse en la forma

$$A_\nu = \frac{1}{\nu!}h^{-\nu}\Delta^\nu f_0, \tag{37}$$

como puede verificarse por inducción.[1]

Fórmula de interpolación de Newton. Haciendo $\xi = (x - x_0)/h$ se tiene $x - x_r = h(\xi - r)$. Las expresiones $(x - x_0)(x - x_1) \cdots (x - x_\nu)$ adoptan entonces la forma $\xi(\xi - 1) \cdots (\xi - \nu)h^{\nu+1}$. Así se obtiene para los polinomios $\phi(x)$, a partir de (33), (37), la *fórmula de interpolación de Newton*:[2]

$$\phi(x) = \phi(x_0 + \xi h) = f_0 + \binom{\xi}{1}\Delta f_0 + \binom{\xi}{2}\Delta^2 f_0 + \cdots + \binom{\xi}{n}\Delta^n f_0.$$

Si $f_0, f_1, f_2, \ldots$ son los valores de una función $f(x)$ en los puntos $x_0, x_1, x_2, \ldots$, donde f tiene derivadas continuas hasta la de orden n, entonces $\Delta^\nu f_0/h^\nu$ es una aproximación a la derivada $f^{(\nu)}(x_0)$. En la p. 495 se demostrará que

$$\lim_{h\to 0} \frac{1}{h^\nu}\Delta^\nu f_0 = f^{(\nu)}(x_0).$$

---

[1] Debe verificarse que los valores A_ν dados por (37) satisfacen las ecuaciones (35); o sea, que para cualquier sucesión $f_0, f_1, f_2, \ldots$, la identidad

$$f_k = f_0 + \binom{k}{1}\Delta f_0 + \binom{k}{2}\Delta^2 f_0 + \cdots + \binom{k}{k}\Delta^k f_0$$

se satisface. Suponiendo que esto se cumple para un cierto k, debe demostrarse que

$$\begin{aligned} f_{k+1} &= f_1 + \binom{k}{1}\Delta f_1 + \binom{k}{2}\Delta^2 f_1 + \cdots \\ &= (f_0 + \Delta f_0) + \binom{k}{1}(\Delta f_0 + \Delta^2 f_0) + \binom{k}{2}(\Delta^2 f_0 + \Delta^3 f_0) + \cdots \\ &= f_0 + \binom{k+1}{1}\Delta f_0 + \binom{k+1}{2}\Delta^2 f_0 + \cdots \end{aligned}$$

que es la identidad para el caso de $k + 1$.

[2] Como en la p. 476, aquí se definen los coeficientes binomiales $\binom{\xi}{k}$ para un ξ general y enteros de posición k, por medio de $\binom{\xi}{k} = \xi(\xi - 1) \cdots (\xi - k + 1)/k!$

Puesto que también

$$\lim_{h\to 0} h^k \binom{\xi}{k} = \frac{(x-x_0)^k}{k!},$$

se ve que en este caso $\phi(x)$ tiende al polinomio de Taylor $P_n(x)$ cuando h tiende a cero.

Obsérvese que de la misma manera es posible la construcción del polinomio de interpolación en el caso en que, por ejemplo, los primeros r valores, $x_0, \ldots, x_{r-1}$, coinciden y se prefijan los correspondientes valores f_0, $f_0', \ldots, f_0^{(r-1)}$ para $\phi(x_0)$, $\phi'(x_0), \ldots, \phi^{(r-1)}(x_0)$, los cuales coinciden con los valores

$$f(x_0), f'(x_0), \ldots, f^{(r-1)}(x_0),$$

siendo f una función dada. Para $\phi(x)$ se escribe la forma

$$\phi(x) = A_0 + A_1(x-x_0) + A_2(x-x_0)^2 + \cdots + A_r(x-x_0)^r + A_{r+1}(x-x_0)^r(x-x_r) + \cdots;$$

y entonces se determinan los A_ν en forma sucesiva a partir de las ecuaciones

$$f_0 = A_0 \quad f_0' = A_1 \quad f_0'' = 2A_2$$

$$\cdots\cdots\cdots\cdots$$

$$f_0^{(r-1)} = (r-1)!\,A_{r-1}$$

$$f_r = A_0 + A_1(x_r - x_0) + \cdots + A_r(x_r - x_0)^r$$

$$f_{r+1} = A_0 + A_1(x_{r+1} - x_0) + \cdots + A_r(x_{r+1} - x_0)^r + A_{r+1}(x_{r+1} - x_0)^r(x_{r+1} - x_r)$$

$$\cdots\cdots\cdots\cdots$$

A.II.3 La estimación del residuo

En vista de las consideraciones anteriores, no importa cómo estén dados originalmente los valores $f_0, f_1, \ldots, f_n$. Por ejemplo, si estos valores fuesen obtenidos mediante observaciones físicas, el problema de construir el polinomio de interpolación estaría resuelto completamente, resultando $\phi(x)$ como una función "suave" o "lisa" definida para todo x, que toma los valores observados para los puntos en cuestión, los cuales podrían usarse para "predecir" valores aproximados de $f(x)$ en otros x. Sin embargo, si la función $f(x)$, que toma en los puntos x_k dados los $n+1$ valores f_k, está definida también para valores intermedios x, se tiene que encarar el problema de estimar la *diferencia* $R(x) = f(x) - \phi(x)$, llamada *error de interpolación*. Al principio sólo se sabe que $R(x_0) = R(x_1) = \cdots R(x_n) = 0$. Para disponer de mayor información deberán hacerse hipótesis adicionales relativas al comportamiento de la función $f(x)$ y que afecten al residuo $R(x)$. Se

supondrá, por consiguiente, que en el intervalo en consideración $f(x)$ tiene derivadas continuas, al menos hasta la de orden $n+1$.

Se observa en primer lugar que para cualquier elección de la constante c, la función

$$K(x) = R(x) - c(x-x_0)(x-x_1)\cdots(x-x_n)$$

se anula en los $n+1$ puntos $x_0, \ldots, x_n$. Escójase ahora cualquier valor y distinto de $x_0, x_1, \ldots, x_n$. Se puede entonces determinar c de manera que $K(y) = 0$, o sea,

$$c = \frac{R(y)}{(y-x_0)(y-x_1)\cdots(y-x_n)}.$$

Existen, por tanto, $n+2$ puntos en los que $K(x)$ se anula. Aplicando a $K(x)$ el teorema de Rolle generalizado, se ve que existe un valor $x = \xi$ entre el mayor y el menor de los valores $x_0, x_1, \ldots, x_n$, y, tal que $K^{(n+1)}(\xi) = 0$. Puesto que $R(x) = f(x) - \phi(x)$, y en vista de que $\phi(x)$, por ser un polinomio de orden n, tiene derivada $(n+1)$-ésima idénticamente nula, resulta que

$$f^{(n+1)}(\xi) - c(n+1)! = 0,$$

[habiéndose usado el hecho de que $(n+1)!$ es la $(n+1)$-ésima derivada de $(x-x_0)\cdots(x-x_n)$]. Así se ha obtenido para c una segunda expresión, $c = f^{(n+1)}(\xi)/(n+1)!$, que contiene a ξ y que de alguna manera depende de y. Usando ahora la ecuación $K(y) = 0$, en la cual y es completamente arbitraria y puede entonces ser reemplazada por x, se obtiene la representación

$$R(x) = \frac{(x-x_0)(x-x_1)\cdots(x-x_n)}{(n+1)!} f^{(n+1)}(\xi), \tag{38}$$

donde ξ es algún valor comprendido entre el menor y el mayor de los puntos $x, x_0, x_1, \ldots, x_n$.

En esta forma el problema general de la interpolación para una función $f(x)$ dada queda resuelto completamente. Se tiene para $f(x)$ la representación

$$f(x) = A_0 + A_1(x-x_0) + A_2(x-x_0)(x-x_1) + \cdots + A_n(x-x_0)(x-x_1)\cdots(x-x_{n-1}) + R_n, \tag{39}$$

donde los coeficientes $A_0, A_1, \ldots, A_n$ pueden ser encontrados sucesivamente a partir de los valores de f en los puntos $x_0, x_1, \ldots, x_n$, usando las fórmulas de recurrencia (34) de la p. 491, y donde el residuo R_n es de la forma

$$R_n = \frac{(x-x_0)(x-x_1)\cdots(x-x_n)}{(n+1)!} f^{(n+1)}(\xi), \tag{40}$$

siendo ξ un número escogido adecuadamente entre el mayor y el menor de los valores $x, x_0, x_1, \ldots, x_n$.

Si la fórmula (39) que da $f(x)$ se escribe para $n-1$ en vez de n y el resultado se resta de la original, se obtiene

$$A_n(x-x_0)(x-x_1)\cdots(x-x_{n-1}) + R_n - R_{n-1} = 0.$$

Para $x = x_n$ resulta $R_n = 0$ y, por tanto, el coeficiente A_n (usando (40) con $n-1$ en vez de n) tiene la representación

$$A_n = \frac{f^{(n)}(\xi)}{n!}$$

donde ξ es entre el menor y el mayor de los valores $x_0, x_1, \ldots, x_n$. Representaciones análogas existen para $A_{n-1}, A_{n-2}, \ldots, A_0$. En esta forma se ve que si los $x_0, x_1, \ldots, x_n$ se aproximan entre sí tendiendo hacia uno y el mismo punto, posiblemente el origen, la fórmula de interpolación (39) se convierte, término a término, en la fórmula de Taylor ((27*a*), p. 471), con la forma de Lagrande ((21), p. 468) para el residuo. La fórmula de Taylor puede en esta forma considerarse como un caso límite de la fórmula de interpolación de Newton.

Esta fórmula nos permite dar un significado preciso a una expresión usada muy a menudo en geometría. La parábola osculante, que tiene contacto de orden n en un punto con una curva dada, se dice también que posee "$(n+1)$ puntos consecutivos en común" con dicha curva en ese punto. De hecho, esta parábola osculante se obtiene encontrando una parábola que tenga $n+1$ puntos en común con la curva y haciendo después que esos puntos se aproximen entre sí. En forma analítica esto corresponde justamente a la transición del polinomio de interpolación al polinomio de Taylor. De manera semejante puede caracterizarse la condición de osculación de dos curvas arbitrarias. Por ejemplo, la circunferencia de curvatura es aquélla que tiene tres puntos consecutivos en común con la curva dada.

Es de esperarse que la fórmula de interpolación aplicada al caso de una función cuyos valores en puntos definidos se conocen dé los valores de ésta con un alto grado de precisión *entre* dichos puntos (tanto $|f^{(n+1)}(\xi)|$ como $|x - x_i|$ son entonces acotadas). Si el valor x cae *fuera* de los intervalos de los puntos $x_0, x_1, \ldots, x_n$, se habla de *extrapolación*. Mediante tal extrapolación puede obtenerse una precisión bastante buena siempre que el punto x esté lo suficientemente próximo a los puntos dados. La fórmula de Taylor, en cierto sentido, corresponde a *extrapolación completa*. En general, sólo es adecuada en la vecindad de un punto.

A.II.4 La fórmula de interpolación de Lagrange

Para terminar, se resuelve el problema de la interpolación usando una fórmula, debida a Lagrange, ligeramente distinta y que difiere de la fórmula de interpolación de Newton por contener cada término individual sólo

uno de los valores dados de la función. Además, la fórmula da $\phi(x)$ en forma completamente explícita, sin necesidad de fórmulas de recurrencia. Se introduce, por brevedad, el polinomio de grado $(n+1)$,

$$\psi(x) = (x-x_0)(x-x_1)\cdots(x-x_n),$$

que corresponde a los puntos x_ν dados. Derivando con la regla del producto y substituyendo luego sucesivamente x por los valores $x_0, \ldots, x_n$, se obtienen las relaciones

$$\psi'(x_0) = (x_0-x_1)(x_0-x_2)\cdots(x_0-x_n)$$

$$\psi'(x_\nu) = (x_\nu-x_0)\cdots(x_\nu-x_{\nu-1})(x_\nu-x_{\nu+1}),\cdots(x_\nu-x_n),$$

$$\psi'(x_n) = (x_n-x_0)(x_n-x_1)\cdots(x_n-x_{n-1}).$$

Obsérvese que

$$\frac{\psi(x)}{(x-x_\nu)\psi'(x_\nu)} = \frac{(x-x_0)\cdots(x-x_{\nu-1})(x-x_{\nu+1})\cdots(x-x_n)}{(x_\nu-x_0)\cdots(x_\nu-x_{\nu-1})(x_\nu-x_{\nu+1})\cdots(x_\nu-x_n)}$$

es un polinomio de grado n cuyo valor en el punto $x = x$ es 1 y en los puntos x_i restantes es 0. Se ve entonces de inmediato que la expresión

$$\phi(x) = \psi(x)\left[\frac{f_0}{(x-x_0)\psi'(x_0)} + \frac{f_1}{(x-x_1)\psi'(x_1)} + \cdots + \frac{f_n}{(x-x_n)\psi'(x_n)}\right] \tag{41}$$

es el polinomio de interpolación deseado. Esta es la *fórmula de interpolación de Lagrange*.

PROBLEMAS

SECCION 5.4*b*, página 469

1. Dar la derivación formal completa de la fórmula del residuo (27), de p. 471, usando inducción matemática.

2. *(Una variante de la demostración del teorema de Taylor)*

(*a*) Si $g(h)$ tiene sus primeras $(n+1)$ derivadas continuas en el intervalo $0 \le h \le A$ y si $g(0) = g'(0) = \cdots = g^{(n)}(0) = 0$, en tanto que $|g^{(n+1)}(h)| \le M$ en $[0, A]$, donde M es una constante, demostrar que $|g^{(n)}(h)| \le Mh$, $|g^{(n+1)}(h)| \le Mh^2/2!, \ldots, |g^{(n-i)}(h)| \le Mh^i/i!, \ldots, |g(h)| \le Mh^n/n!$, para todo h en el intervalo.

(*b*) Sean $f(x)$ una función suficientemente derivable en $a \le x \le b$, y $T_n(h)$ el polinomio de Taylor para $f(x)$ en $x = a$. Aplíquese el resultado de (*a*) a la función $g(h) = R_n = f(a+h) - T_n(h)$ para obtener la fórmula de Taylor con una estimación grosera para el residuo.

3. Supóngase que $f(x)$ tiene una derivada continua en el intervalo $a \le x \le b$, y sea $f''(x) \ge 0$ para todo valor de x. Entonces, si ξ es cualquier punto en el inter-

valo, la curva en ninguna parte queda por debajo de su tangente en el punto $x = \xi$, $y = f(\xi)$.

(Usar el desarrollo de Taylor hasta el tercer término.)

4. Deducir la fórmula integral para el residuo R_n integrando por partes

$$f(x + h) - f(x) = \int_0^h f'(x + \tau)\, d\tau.$$

5. Integrar por partes la fórmula

$$R_n = \frac{1}{n!}\int_0^h (h - \tau)^n f^{(n+1)}(x + \tau)\, d\tau,$$

y así obtener

$$R_n = f(x + h) - f(x) - hf'(x) - \cdots - \frac{h^n}{n!} f^{(n)}(x).$$

***6.** Supóngase que de alguna manera se ha obtenido una serie para la función $f(x)$, a saber

$$f(x) = a_0 + a_1 x + a_2 x^2 + \cdots + a_n x^n + R_n(x),$$

donde $a_0, a_1, \ldots, a_n$ son constantes, $R_n(x)$ tiene n derivadas continuas y $R_n(x)/x^n \to 0$ cuando $x \to 0$. Demostrar que $a_k = (f^k(0)/k!)$ $(k = 0, \ldots, n)$, esto es, que la serie es una serie de Taylor.

SECCION 5.5, página 472

1. Encontrar los primeros cuatro términos no nulos en la serie de Taylor para las siguientes funciones en la vecindad de $x = 0$:

(a) $x \cot x$ (d) $e^{\operatorname{sen} x}$

(b) $\dfrac{\sqrt{\operatorname{sen} x}}{\sqrt{x}}$ (e) e^{e^x}

(c) $\sec x$ (f) $\log \operatorname{sen} x - \log x$.

2. Encontrar la serie de Taylor para arc sen x en la vecindad de $x = 0$,, usando

$$\operatorname{arc\,sen} x = \int_0^x \frac{dt}{\sqrt{1 - t^2}}.$$

Comparar con el problema 2, sección 3.2.

***3.** Encontrar los tres primeros términos no nulos de la serie de Taylor para $\operatorname{sen}^2 x$ en la vecindad de $x = 0$, mediante la multiplicación de la serie de Taylor para sen x por sí misma. Justificar este procedimiento.

***4.** Encontrar los tres primeros términos no nulos de la serie de Taylor para tan x en la vecindad de $x = 0$, usando la relación $\tan x = \operatorname{sen} x/\cos x$. Justificar el procedimiento.

***5.** Encontrar los tres primeros términos no nulos de la serie de Taylor para $\sqrt{\cos x}$ en la vecindad de $x = 0$, aplicando el teorema del binomio a la serie de Taylor para cos x. Justificar el procedimiento.

***6.** Encontrar la serie de Taylor para $(\operatorname{arc\,sen} x)^2$. Comparar con el problema 2, sección 3.2.

***7.** Encontrar las series de Taylor para las siguientes funciones en la vecindad de x = 0:

(a) $\operatorname{senh}^{-1} x$. (b) $\displaystyle\int_0^x e^{-t^2}\, dt$. (c) $\displaystyle\int_0^x \frac{\operatorname{sen} t}{t}\, dt$.

***8.** Estimar el error que se comete al usar sólo los n primeros términos en el problema 7.

***9.** La función elíptica $s(u)$ ha sido definida (sección 3.14*a*) como la inversa de la integral elíptica

$$u(s) = \int_0^s \frac{dx}{\sqrt{(1 - x^2)(1 - k^2x^2)}}.$$

Encontrar el desarrollo de Taylor para $s(u)$ hasta el término de grado 5.

***10.** Evaluar los siguientes límites:

(a) $\displaystyle\lim_{x\to\infty} x\left[\left(1 + \frac{1}{x}\right)^x - e\right]$,

(b) $\displaystyle\lim_{x\to\infty} \left\{\frac{e}{2}x + x^2\left[\left(1 + \frac{1}{x}\right)^x - e\right]\right\}$,

*(c) $\displaystyle\lim_{x\to\infty} x\left[\left(1 + \frac{1}{x}\right)^x - e\log\left(1 + \frac{1}{x}\right)^x\right]$.

(d) $\displaystyle\lim_{x\to 0}\left(\frac{\operatorname{sen} x}{x}\right)^{1/x^2}$,

(e) $\displaystyle\lim_{x\to\infty}\left(\frac{\operatorname{sen} x}{x}\right)^{1/x^2}$

***11.** Encontrar los tres primeros términos de la serie de Taylor para $[1 + (1/x)]^x$ en potencias de $1/x$.

***12.** Dos partículas con cargas de signo opuesto, $+e$, $-e$, situadas a una pequeña distancia d una de la otra, forman un *dipolo* eléctrico con momento $M = ed$. Demostrar que la energía potencial:

(*a*) En un punto situado sobre el eje del dipolo a una distancia r del centro del mismo, es $(M/r^2)(1 + \varepsilon)$, donde ε es aproximadamente igual a $d^2/4r^2$.

(*b*) En un punto situado sobre la bisectriz perpendicular al dipolo, es 0.

(*c*) En un punto con coordenadas polares r, θ, relativas al centro y al eje del dipolo, es $[M\cos(\theta/r^2)](1 + \varepsilon)$, donde ε es aproximadamente igual a

$$(d^2/8r^2)(5\cos^2\theta - 3).$$

(La energía potencial de una carga puntual única q en un punto situado a una distancia r de ella es q/r; y la energía potencial de varias cargas es la suma de las energías potenciales de las cargas por separado.)

SECCION 5.6, página 476

1. Demostrar que si $f(a) = 0$ y $f(x)$ tiene derivadas en $x = a$ hasta de un orden suficientemente alto, entonces $f(x)^n$ tiene al menos un contacto de orden $(n - 1)$ con el eje x.

2. La curva $y = f(x)$ pasa por el origen O y toda al eje x en ese punto. Demostrar que el radio de curvatura de la curva en O está dado por

$$\rho = \lim_{x\to 0} \frac{x^2}{2y}.$$

***3.** K es una circunferencia que toca en un punto P a una curva dada y pasa por un punto cercano Q situado sobre la curva. Demostrar que el límite de circunferencia K para $Q \to P$ es la circunferencia de curvatura de la curva en P.

***4.** Demostrar que el orden de contacto entre una curva y su circunferencia osculante es al menos igual a tres en puntos en que el radio de curvatura es un máximo o un mínimo.

***5.** Demostrar que la circunferencia osculante en un punto en que el radio de curvatura es un máximo o un mínimo no cruza a la curva a menos que el contacto sea de orden superior al tercero.

***6.** Encontrar los máximos y mínimos de las siguientes funciones:

$$(a)\ \cos x \cosh x \qquad (b)\ x + \cos x$$

***7.** Determinar los máximos y mínimos de la función $y = e^{-1/x^2}$ (ver p. 263).

SECCION A.I.3, página 483

1. Demostrar que si f es continua en el intervalo $[0, 1]$, entonces

$$\lim_{x \to 0} x \int_x^1 \frac{f(z)}{z^2}\, dz = f(0).$$

2. Demostrar que la función $y = (x^2)^x$, $y(0) = 1$ es continua en $x = 0$.

6

Métodos numéricos

La tarea de resolver un problema analítico queda siempre inconclusa. Demostrar la existencia y algunas de las propiedades básicas de la solución se considera por lo general satisfactorio, pero quedan siempre importantes preguntas por responder. Así, cuando la solución es definida mediante un proceso de límite, usando una integral, por ejemplo, surge el problema de encontrar realmente aproximaciones numéricas a este límite y de estimar la exactitud de estas aproximaciones. Tales cuestiones no sólo son de importancia básica desde el punto de vista teórico, sino que son también inevitables cuando se desea aplicar el análisis a la descripción y al control de los fenómenos naturales, los cuales pueden describirse en principio sólo de manera aproximada.

Se plantea así el difícil problema de llevar la solución al punto en que puedan obtenerse respuestas numéricas y estimaciones de su exactitud.

Recientemente, con el advenimiento de las máquinas calculadoras automáticas de alta velocidad, los aspectos teóricos y prácticos del "análisis numérico" han recibido un gran impulso. Estos aspectos son presentados en una gran variedad de libros de texto.[1] Durante siglos, sin embargo, muchos de los matemáticos más notables, como Newton, Euler y, en particular, Gauss, han aportado grandes contribuciones a los métodos numéricos.

En este volumen no es posible presentar una visión completa del análisis numérico, pero, al menos, se discutirán algunos de los resultados clásicos más simples.

[1] Véanse, por ejemplo: Hildebrand, *Introduction to Numerical Analysis,* McGraw-Hill Book Co., 1956; Householder, *Principles of Numerical Analysis,* McGraw-Hill Book Co., 1953; Whittaker and Robinson, *The Calculus of Observations,* Blackie and Sons, Ltd., 1929.

6.1 Cálculo de integrales

Aunque la existencia de la integral de una función (continua) es asegurada por la teoría que se dio en el capítulo 2, la evaluación o "cuadratura" [1] de dicha integral no puede realizarse mediante funciones elementales, salvo en casos relativamente raros. Deberán desarrollarse, por consiguiente, métodos de integración numérica y también de estimación de la exactitud de tales aproximaciones numéricas.

Para calcular en forma aproximada la integral

$$J = \int_a^b f(x)\,dx \tag{1}$$

con $a < b$, se subdivide el intervalo $a \leq x \leq b$ en n partes iguales, cada una de longitud $h = (b - a)/n$, mediante los $n + 1$ puntos

$$x_\nu = a + \nu h, \quad nh = b - a, \quad \nu = 0, 1, \ldots, n. \tag{2}$$

Entonces,

$$J = \sum_{\nu=1}^{n} J_\nu,$$

donde

$$J_\nu = \int_{x_{\nu-1}}^{x_\nu} f(x)\,dx. \tag{3}$$

El problema de calcular la integral J queda reducido al de obtener buenas aproximaciones para las áreas J_ν de anchura h en las cuales se ha dividido el área total representada por J.

a. Aproximación mediante rectángulos

La aproximación más directa, sugerida por la definición original de la integral, conduce a la relación

$$J = \sum_{\nu=1}^{n} J_\nu,$$

$$\approx h(f_1 + f_2 + \cdots + f_n),$$

donde, por brevedad, se ha hecho

$$f_\nu = f(x_\nu).$$

Aquí (y en todo este capítulo) el símbolo $\approx$ significa "es aproximadamente igual a".

[1] La palabra "cuadratura" indica el proceso de "cuadrar", esto es, de medir el área dentro de una curva encontrando un cuadrado de área igual (como en el problema de "cuadrar el círculo").

Para estimar la exactitud o el "error" de esta aproximación, supóngase que $f(x)$ es continua, con una derivada acotada uniformemente en el intervalo $a \leq x \leq b$: $|f'(x)| \leq M_1$. Puede entonces demostrarse fácilmente (ver problema 4, p. 527, 6.1), que

$$|J_\nu - hf_\nu| \leq \frac{M_1 h^2}{2}, \tag{4}$$

o sea,

$$\left| J - h \sum_{\nu=1}^{n} f_\nu \right| \leq n \frac{M_1 h^2}{2}$$

$$= \tfrac{1}{2} M_1 (b - a) h. \tag{5}$$

Así, el error de la aproximación para la integral mediante la suma finita es del orden de h, la "anchura de malla", en la terminología del capítulo 3, p. 273.

b. Aproximaciones refinadas — Regla de Simpson

Una mejor aproximación se obtiene casi sin esfuerzo adicional si las áreas J_ν se aproximan, no mediante franjas rectangulares, sino mediante trapezoides angostos, como en la Fig. 6.1*a*. La fórmula de aproximación (fórmula del trapezoide) es entonces

$$J \approx \tfrac{1}{2}h(f_0 + f_1) + \tfrac{1}{2}h(f_1 + f_2) + \cdots + \tfrac{1}{2}h(f_{n-1} + f_n)$$

$$= h(f_1 + f_2 + \cdots + f_{n-1}) + \frac{h}{2}(f_0 + f_n), \tag{6}$$

pues cada valor de la función, excepto el primero y el último, aparece dos veces.

Una aproximación que resulta por lo general un poco más precisa que la fórmula del trapezoide es aquélla en que la ν-ésima franja se aproxima mediante un tropezoide limitado superiormente por la tangente a la curva en el punto medio $x_{\nu-1} + h/2$ del intervalo $x_{\nu-1} \leq x \leq x_\nu$. El área de este trapezoide es simplemente

$$hf_{\nu-1/2} = hf\left(x_{\nu-1} + \frac{h}{2}\right),$$

y por adición se obtiene la *fórmula de la tangente,*

$$J \approx h(f_{1/2} + f_{3/2} + \cdots + f_{(2n-1)/2}). \tag{7}$$

Como se verá en la p. 506, el error de esta aproximación es del orden de h^2 cuando la segunda derivada de f es continua en el intervalo $a \leq x \leq b$ y $|f''(x)| < M_2$, siendo M_2 alguna cota constante.

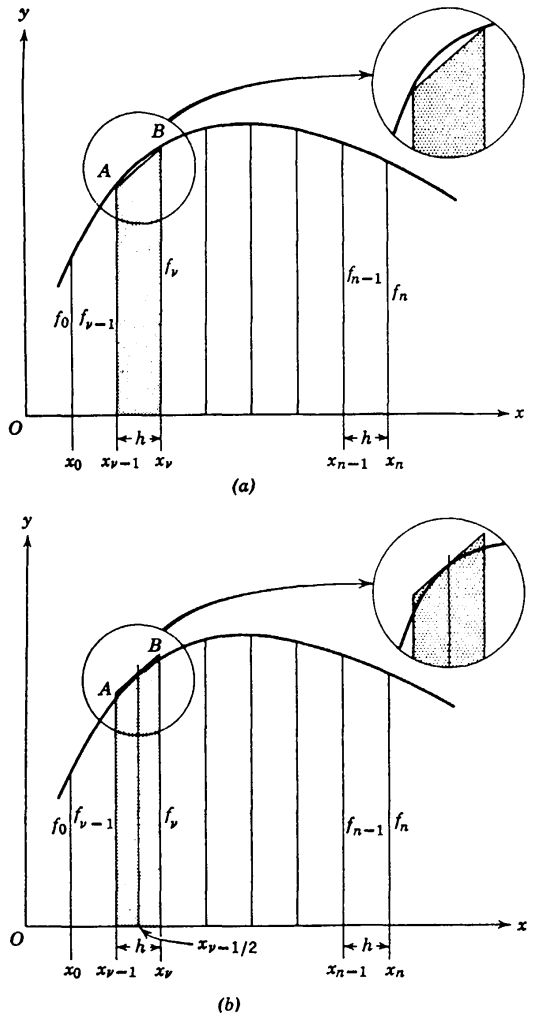


Figura 6.1 (*a*) La fórmula del trapezoide. (*b*) La fórmula de la tangente.

Finalmente, se menciona la famosa aproximación de *Simpson,* la cual con poco mayor trabajo conduce a una aproximación mucho más precisa si la cuarta derivada de f existe y es uniformemente acotada en el intervalo dado.

$$|f^{(4)}(x)| \leq M_4,$$

donde M_4 es una constante. La fórmula de Simpson para $n = 2m$ es

$$(8)\quad J \approx \frac{4h}{3}(f_1 + f_3 + f_5 + \cdots + f_{2m-1}) + \frac{2h}{3}(f_2 + f_4 + f_6 + \cdots + f_{2m-2}) + \frac{h}{3}(f_0 + f_{2m}).$$

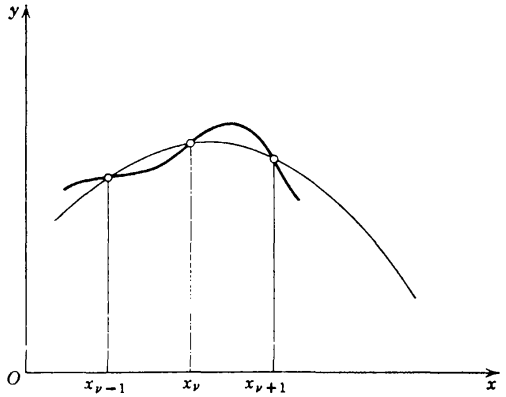


Figura 6.2 Regla de Simpson.

La fórmula es obtenida fácilmente si la región formada por las franjas ν y $(\nu + 1)$ se aproxima mediante una franja de anchura $2h$ limitada superiormente por la parábola que coincide con f para las tres abscisas $x_{\nu-1}$, $x_\nu = x_{\nu-1} + h$ y $x_{\nu+1} = x_{\nu-1} + 2h$ (ver Fig. 6.2). La fórmula de interpolación de Newton (p. 492) conduce a la ecuación de esta parábola:

$$y = f_{\nu-1} + (x - x_{\nu-1}) \cdot \frac{f_\nu - f_{\nu-1}}{h} + \frac{(x - x_{\nu-1})(x - x_{\nu-1} - h)}{2} \frac{f_{\nu-1} - 2f_\nu + f_{\nu-1}}{h^2}.$$

teniéndose, por tanto, la aproximación

$$J_\nu + J_{\nu+1} \approx \int_{x_{\nu-1}}^{x_{\nu+1}} y\,dx = \int_{x_{\nu-1}}^{x_{\nu-1}+2h} y\,dx$$
$$= 2hf_{\nu-1} + 2h(f_\nu - f_{\nu-1}) + \frac{\frac{8}{3}h - 2h}{2}(f_{\nu+1} - 2f_\nu + f_{\nu-1})$$
$$= \frac{h}{3}(f_{\nu-1} + 4f_\nu + f_{\nu+1}).$$

La fórmula se obtiene ahora para n par $(n = 2m)$ sumando todos estos valores aproximados para $\nu = 1, 3, 4, \ldots, 2m - 1$, o sea, todas las áreas de los pares de franjas.

**Exactitud*

No es difícil estimar la exactitud de las anteriores aproximaciones. Toda cuadratura consiste en aproximar la función $f(x)$ en un intervalo por medio de una función $\phi(x)$ fácilmente integrable (un polinomio). Entonces, una estimación del error de la fórmula de integración puede obtenerse estimando $|f(x) - \phi(x)|$.

En la fórmula de la tangente (p. 503) se reemplaza $f(x)$ en el intervalo $[x_{\nu-1}, x_\nu]$ por su tangente en el punto medio $x_\nu - (h/2)$, esto es, por

$$\phi(x) = f\left(x_\nu - \frac{h}{2}\right) + \left(x - x_\nu + \frac{h}{2}\right) f'\left(x_\nu - \frac{h}{2}\right).$$

Usando el teorema de Taylor con la forma de Lagrange para el residuo:

$$f(x) = \phi(x) + \frac{1}{2}\left(x - x_\nu + \frac{h}{2}\right)^2 f''(\xi),$$

donde ξ está entre x y $x_\nu - h/2$. Por tanto, el error correspondiente a una franja se estima en

$$|J_\nu - hf_{\nu-\frac{1}{2}}| = \left| \int_{x_{\nu-1}}^{x_\nu} [f(x) - \phi(x)]\,dx \right|$$
$$\leq \int_{x_{\nu-1}}^{x_\nu} |f(x) - \phi(x)|\,dx \leq M_2 \int_{x_\nu - h}^{x_\nu} \frac{1}{2}\left(x - x_\nu + \frac{h}{2}\right)^2 dx$$
$$= \frac{h^3}{24} M_2.$$

Para el error total en la fórmula de la tangente al considerar todos los intervalos,[1] se encuentra entonces la cota superior

$$n\frac{h^3}{24}M_2 = \frac{h^2}{24}M_2(b - a).$$

Esta derivación se usará como modelo para estimar el error en las otras fórmulas de cuadraturas. Así, en la regla del trapezoide (6) $f(x)$ se aproxima en el intervalo $[x_{\nu-1}, x_\nu]$ por el polinomio de interpolación lineal

$$\phi(x) = f_{\nu-1} + (x - x_{\nu-1})\frac{f_\nu - f_{\nu-1}}{h}.$$

Usando la estimación para el error del residuo en la fórmula de interpolación [ver p. 494, Ec. (40)] con $n = 1$, se encuentra

$$f(x) - \phi(x) = \tfrac{1}{2}(x - x_{\nu-1})(x - x_\nu) f''(\xi),$$

[1] Este es el error total inherente al uso de la fórmula de aproximación, el así llamado error "por truncamiento". En la práctica aparecen otros errores debido al redondeo que se hace en los cálculos. El efecto total de los errores por redondeo aumenta generalmente con el número de pasos efectuados, o sea, al disminuir h, en tanto que el error por truncamiento disminuye.

donde ξ está entre $x_{\nu-1}$ y x_ν. Por tanto, el valor absoluto del error en el cálculo de J_ν es a lo más

$$M_2 \int_{x_{\nu-1}}^{x_\nu} |\tfrac{1}{2}(x - x_{\nu-1})(x - x_\nu)|\, dx = \frac{h^3}{12} M_2,$$

y el error total es a lo más n veces la cantidad anterior:

$$\frac{h^2}{12} M_2 (b - a).$$

La misma técnica puede aplicarse a la regla de Simpson, (8), tomando como $\phi(x)$ el polinomio cuadrático que coincide con f en los puntos $x_{\nu-1}$, x_ν, $x_{\nu+1}$. Así resulta un error en $J_\nu + J_{\nu+1}$ del orden de h^4. Realmente, sin embargo, puede mejorarse en un orden de magnitud la estimación del error usando un polinomio cúbico $\phi(x)$, el cual da una mejor aproximación que el cuadrático para f en el intervalo $[x_{\nu-1}, x_{\nu+1}]$ y tiene aún la misma integral, conduciendo, por tanto, a la misma fórmula de aproximación (9) para J. Simplemente se usa el polinomio de interpolación que coincide con $f(x)$ en los puntos $x_{\nu-1}$, x_ν, $x_{\nu+1}$ y para el cual $\phi'(x_\nu) = f'(x_\nu)$. Tiene la forma

$$\phi(x) = A_0 + A_1(x - x_{\nu-1}) + A_2(x - x_{\nu-1})(x - x_\nu) + A_3(x - x_{\nu-1})(x - x_\nu)(x - x_{\nu+1}).$$

Aquí los tres primeros términos representan el polinomio de interpolación cuadrático que coincide con f en los tres puntos $x_{\nu-1}$, x_ν, $x_{\nu+1}$. La constante A_3 tiene que determinarse a partir de la condición $\phi'(x_\nu) = f'(x_\nu)$.

El último término,

$$A_3(x - x_\nu + h)(x - x_\nu)(x - x_\nu - h) = A_3[(x - x_\nu)^2 - h^2] \cdot [x - x_\nu],$$

es obviamente una función impar de $x - x_\nu$, y, por consiguiente, no contribuye a la integral entre los límites $x_\nu - h$ y $x_\nu + h$. Para el error en la aproximación de f se tiene entonces la estimación siguiente [ver (40), p. 494, con $n = 3$ y con dos de los puntos de interpolación coincidiendo en x_ν]:

$$f - \phi = \frac{1}{4!}(x - x_{\nu-1})(x - x_\nu)^2(x - x_{\nu+1}) f^{(4)}(\xi).$$

Esto conduce a la siguiente estimación para el error total en el cálculo de $J_\nu + J_{\nu+1}$:

$$\frac{h^5}{90} M_4,$$

y, por tanto, para el error total se tiene la estimación

$$\frac{n}{2}\frac{h^5}{90} M_4 = \frac{h^4}{180}(b - a) M_4.$$

Naturalmente, puede lograrse una mayor precisión aproximando la función $f(x)$ en una franja por un polinomio de orden aún más alto.

Ejemplos. Apliquemos estos métodos al cálculo de

$$\log_e 2 = \int_1^2 \frac{dx}{x}.$$

Dividiendo el intervalo $1 \leq x \leq 2$ en diez partes de longitud $h = 1/10$ y usando la regla del trapezoide (6) se obtiene

$x_1 = 1.1$	$f_1 = 0.90909$
$x_2 = 1.2$	$f_2 = 0.83333$
$x_3 = 1.3$	$f_3 = 0.76923$
$x_4 = 1.4$	$f_4 = 0.71429$
$x_5 = 1.5$	$f_5 = 0.66667$
$x_6 = 1.6$	$f_6 = 0.62500$
$x_7 = 1.7$	$f_7 = 0.58824$
$x_8 = 1.8$	$f_8 = 0.55556$
$x_9 = 1.9$	$f_9 = 0.52632$
	Suma 6.18773
$x_0 = 1.0$	$\frac{1}{2}f_0 = 0.5$
$x_{10} = 2.0$	$\frac{1}{2}f_{10} = 0.25$
	$6.93773 \cdot 1/10$

$$\log_e 2 \approx 0.69377.$$

Puesto que la gráfica del integrando tiene su lado convexo hacia el eje x, este valor es demasiado grande.

Usando la regla de la tangente (7) se tiene

$x_0 + \frac{1}{2}h = 1.05$	$f_{1/2} = 0.95238$
$x_1 + \frac{1}{2}h = 1.15$	$f_{3/2} = 0.86957$
$x_2 + \frac{1}{2}h = 1.25$	$f_{5/2} = 0.80000$
$x_3 + \frac{1}{2}h = 1.35$	$f_{7/2} = 0.74074$
$x_4 + \frac{1}{2}h = 1.45$	$f_{9/2} = 0.68966$
$x_5 + \frac{1}{2}h = 1.55$	$f_{11/2} = 0.64516$
$x_6 + \frac{1}{2}h = 1.65$	$f_{13/2} = 0.60606$
$x_7 + \frac{1}{2}h = 1.75$	$f_{15/2} = 0.57143$
$x_8 + \frac{1}{2}h = 1.85$	$f_{17/2} = 0.54054$
$x_9 + \frac{1}{2}h = 1.95$	$f_{19/2} = 0.51282$
	$6.92836 \cdot 1/10$

$$\log_e 2 \approx 0.69284,$$

valor que, debido a la convexidad de la curva, es demasiado pequeño.

Con la misma subdivisión se obtiene un resultado más preciso usando la regla de Simpson (8). Resulta

$$\begin{array}{ll} x_1 = 1.1 & f_1 = 0.90909 \\ x_3 = 1.3 & f_3 = 0.76923 \\ x_5 = 1.5 & f_5 = 0.66667 \\ x_7 = 1.7 & f_7 = 0.58824 \\ x_9 = 1.9 & f_9 = 0.52632 \end{array}$$

$$\text{Suma } 3.45955 \cdot 4$$

$$13.83820$$

$$\begin{array}{ll} x_2 = 1.2 & f_2 = 0.83333 \\ x_4 = 1.4 & f_4 = 0.71429 \\ x_6 = 1.6 & f_6 = 0.62500 \\ x_8 = 1.8 & f_8 = 0.55556 \end{array}$$

$$\text{Suma } 2.72818 \cdot 2$$

$$5.45636$$

$$13.83820$$

$$\begin{array}{ll} x_0 = 1.0 & f_0 = 1.0 \\ x_{10} = 2.0 & f_{10} = 0.5 \end{array}$$

$$20.79456 \cdot \tfrac{1}{30}$$

$$\log_e \approx 0.69315.$$

En realidad,

$$\log_e 2 = 0.693147\ldots.$$

6.2 Otros ejemplos de métodos numéricos

a. El "cálculo de errores"

El "cálculo de errores" es simplemente una aplicación numérica de un teorema básico del cálculo diferencial: una función $f(x)$, derivable un número suficiente de veces, puede ser representada en la vecindad de un punto por una función lineal con un error de orden superior al primero, por una función cuadrática con un error de orden superior al segundo, y así sucesivamente.

Considérese la aproximación lineal para la función $y = f(x)$. Si $y + \Delta y = f(x + \Delta x) = f(x + h)$, por el teorema de Taylor se tiene

$$\Delta y = hf'(x) + \frac{h^2}{2} f''(\xi),$$

donde $\xi = x + \theta h$ $(0 < \theta < 1)$ es un valor intermedio no necesariamente conocido en forma precisa. Si $h = \Delta x$ es pequeño, se obtiene la siguiente aproximación de interés práctico

$$\Delta y \approx hf'(x).$$

Así, el cociente incremental se reemplaza por la derivada, que lo aproxima, y el incremento de y se reemplaza por la aproximación dada por la expresión lineal en h.

Este sencillo resultado se usa para propósitos numéricos de la siguiente manera. Sean x e y dos cantidades físicas relacionadas por $y = f(x)$. Se pregunta entonces qué efecto tiene sobre la determinación de y una imprecisión en la medición de x. Si en vez del valor "verdadero" x se usa el valor aproximado $x + h$, el correspondiente valor de y difiere del valor verdadero $y = f(x)$ en la cantidad $\Delta y = f(x + h) - f(x)$. Por tanto, el error está dado aproximadamente por la relación anterior.

Mediante ejemplos se ilustra la utilidad de tales aproximaciones lineales.

Ejemplos. (*a*) Supóngase que en un triángulo ABC (ver Fig. 6.3) los lados b y c se determinan con precisión en tanto que el ángulo $\alpha = x$ se determina con un error $|\Delta x| < \delta$. ¿Cuál es el error correspondiente en el valor del tercer lado $y = a = \sqrt{b^2 + c^2 - 2bc\cos\alpha}$?

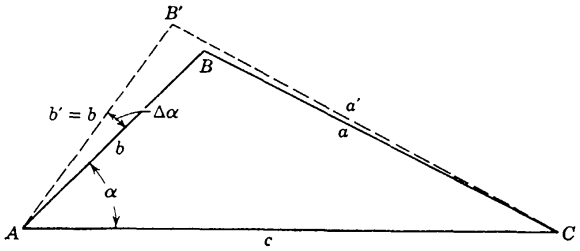


Figura 6.3

Se tiene $\Delta a \approx (bc \operatorname{sen}\alpha\, \Delta\alpha)/a$; y el porcentaje de error es por consiguiente

$$\frac{100\,\Delta a}{a} \approx \frac{100\,bc}{a^2} \operatorname{sen}\alpha\, \Delta\alpha.$$

En el caso especial en que $b = 400$ metros, $c = 500$ metros y $\alpha = 60°$, resulta $y = a = 458.2576$ metros, de manera que

$$\Delta a \approx \frac{200000}{458.2576} \times \tfrac{1}{2}\sqrt{3}\,\Delta\alpha.$$

Si $\Delta\alpha$ pudiera medirse con un error máximo de 10 segundos de arco, es decir, si

$$\Delta\alpha = 10'' = 4846 \times 10^{-8} \text{ radianes},$$

se encontraría que en el caso más desfavorable

$$\Delta a \approx 1.83 \text{ cm}.$$

Así, el error sería a lo más del 0.004% aproximadamente.

(*b*) El ejemplo que sigue ilustra la utilidad de la "linealización" en problemas físicos.

Experimentalmente se sabe que si una varilla de metal tiene una longitud l_0 a la temperatura t_0, entonces a la temperatura t su longitud será $l = l_0(1 + \alpha(t - t_0))$, donde α depende sólo de t_0 y del material que compone la varilla. Ahora, si un reloj de péndulo conserva la hora exacta a la temperatura t_0, ¿cuántos segundos se atrasará diariamente si la temperatura se eleva a t_1?

Para el período de oscilación $T(l)$ se tiene (ver pp. 429-430)

$$T(l) = 2\pi\sqrt{\frac{l}{g}};$$

por tanto,

$$\frac{dT}{dl} = \frac{\pi}{\sqrt{lg}}.$$

Si el cambio en longitud es Δl, el cambio correspondiente en el período de oscilación es

$$\Delta T \approx \frac{\pi\,\Delta l}{\sqrt{l_0 g}},$$

donde $l_1 = l_0(1 + \alpha(t_1 - t_0))$ y $\Delta l = \alpha l_0(t_1 - t_0)$. Este es el tiempo de retraso por oscilación. El tiempo de retraso en un segundo es $\Delta T/T \approx \Delta l/2l_0$; por tanto, en un día el reloj se atrasa $43{,}200\,\Delta l/l_0 = 43{,}200\,\alpha(t_1 - t_0)$ segundos.

En éste y muchos otros casos en que la función considerada es un producto de varios factores, pueden simplificarse los cálculos tomando logaritmos en ambos miembros antes de derivar. Para el presente ejemplo se tiene:

$$\log T = \log 2\pi - \tfrac{1}{2}\log g + \tfrac{1}{2}\log l;$$

y al derivar,

$$\frac{1}{T}\frac{dT}{dl} = \frac{1}{2l}.$$

Reemplazando dT/dl por $\Delta T/\Delta l$, resulta

$$\frac{\Delta T}{T} = \frac{\Delta l}{2l},$$

lo cual coincide con el resultado anterior.

*b. Cálculo de π

Un ejemplo diferente, en el que se emplean recursos artificiales, es el siguiente. Es un ejemplo clásico, aunque quizá las máquinas calculadoras modernas lo han hecho obsoleto.

La serie de Leibnitz, $\pi/4 = 1 - \frac{1}{3} + \frac{1}{5} - \frac{1}{7} + \cdots$ [Ec. (7), sección 5.2, p. 464], obtenida a partir de la serie para la tangente inversa, no es adecuada para el cálculo de π debido a la extrema lentitud de su convergencia. Sin embargo, se puede calcular π con relativa facilidad mediante el artificio siguiente. Si en el teorema de adición para la tangente,

$$\tan(\alpha + \beta) = \frac{\tan\alpha \tan\beta}{1 - \tan\alpha \tan\beta},$$

se introducen las funciones inversas $\alpha = \arctan u$, $\beta = \arctan v$, se obtiene la fórmula

$$\arctan u + \arctan v = \arctan\left(\frac{u + v}{1 - uv}\right).$$

Ahora bien, escogiendo u y v de manera que $(u + v)/(1 - uv) = 1$, se obtiene el valor $\pi/4$ en el miembro derecho y, si u y v son números pequeños, puede fácilmente calcularse el miembro izquierdo por medio de series conocidas. Por ejemplo, si se hacen $u = \frac{1}{2}$, $v = \frac{1}{3}$, como hizo Euler, se obtiene

$$\frac{\pi}{4} = \arctan\tfrac{1}{2} + \arctan\tfrac{1}{3}. \tag{9}$$

Si, además, se observa que $\dfrac{\frac{1}{3} + \frac{1}{7}}{1 - \frac{1}{21}} = \dfrac{1}{2}$,

se tiene $\arctan\frac{1}{2} = \arctan\frac{1}{3} + \arctan\frac{1}{7}$, de manera que, por (9),

$$\frac{\pi}{4} = \arctan\tfrac{1}{3} + \arctan\tfrac{1}{7}.$$

Usando esta fórmula Vega calculó el número π con 140 cifras decimales.

Por medio de la ecuación $(\frac{1}{5} + \frac{1}{8})/(1 - \frac{1}{40}) = \frac{1}{3}$, se obtiene además

$$\arctan\tfrac{1}{3} = \arctan\tfrac{1}{5} + \arctan\tfrac{1}{8}$$

en donde:

$$\frac{\pi}{4} = 2\arctan\tfrac{1}{5} + \arctan\tfrac{1}{7} + 2\arctan\tfrac{1}{8}.$$

Este desarrollo es sumamente útil para el cálculo de π por medio de la serie $\arctan x = x - x^3/3 + x^5/5 - \cdots$; pues si se substituye x por uno de los valores $\frac{1}{5}$, $\frac{1}{7}$ ó $\frac{1}{8}$, se obtiene con pocos términos un alto grado de precisión ya que éstos disminuyen rápidamente.

El lector que no tenga especial interés en estas manipulaciones, que requieren habilidad aunque sean artificiosas, puede conformarse con entender el principio.

*c. Cálculo de logaritmos

Para el cálculo numérico de logaritmos la serie logarítmica [Ec. (5), p. 463]

$$\tfrac{1}{2}\log\frac{1 + x}{1 - x} = x + \frac{x^3}{3} + \frac{x^5}{5} + \cdots,$$

donde $0 < x < 1$, mediante la substitución

$$\frac{1+x}{1-x} = \frac{p^2}{p^2-1}, \qquad x = \frac{1}{2p^2-1},$$

se transforma en la serie

$$\log p = \tfrac{1}{2}\log(p-1) + \tfrac{1}{2}\log(p+1) + \frac{1}{2p^2-1} + \frac{1}{3(2p^2-1)^3} + \cdots$$

donde $2p^2 - 1 > 1$ o $p^2 > 1$. Si p es un entero y $p + 1$ puede descomponerse en factores enteros más pequeños (por ejemplo, si $p + 1$ es par), esta última serie expresa el logaritmo de p mediante los logaritmos de enteros menores, más una serie cuyos términos disminuyen muy rápidamente y cuya suma puede por tanto calcularse con suficiente precisión usando sólo unos cuantos términos. A partir de esta serie pueden entonces calcularse sucesivamente los logaritmos de cualesquier números primos y, en consecuencia, de cualquier número, suponiendo que ya se haya calculado el valor log 2 (por ejemplo, mediante su representación integral, como en las pp. 508-509).

La exactitud de esta determinación de log p puede estimarse más fácilmente por medio de la serie geométrica que usando la fórmula general del residuo. Para el residuo R_n de la serie, esto es, la suma de todos los términos que siguen al término $1/n(2p^2-1)^n$, se tiene

$$R_n < \frac{1}{(n+2)(2p^2-1)^{n+2}}\left(1 + \frac{1}{(2p^2-1)^2} + \frac{1}{(2p^2-1)^4} + \cdots\right)$$
$$= \frac{1}{(n+2)(2p^2-1)^n} \cdot \frac{1}{(2p^2-1)^2 - 1},$$

y esta fórmula da inmediatamente la estimación requerida para el error.

Calculemos, por ejemplo, $\log_e 7$ (con la hipótesis de que log 2 y log 3 ya han sido determinados numéricamente), usando los cuatro primeros términos de la serie Tenemos

$$p = 7, \quad 2p^2 - 1 = 97,$$

$$\log 7 = 2\log 2 + \tfrac{1}{2}\log 3 + \frac{1}{97} + \frac{1}{3\cdot 97^3} + \cdots;$$

$$\tfrac{1}{97} \approx 0.01030928, \qquad \frac{1}{3\cdot 97^3} \approx 0.00000037,$$

$$2\log 2 \approx 1.38629436, \qquad \tfrac{1}{2}\log 3 \approx 0.54930614;$$

y, por tanto,

$$\log_e 7 \approx 1.94591015.$$

La estimación del error es

$$R_n < \frac{1}{5\cdot 97^3} \times \frac{1}{97^2-1} < \frac{1}{36\times 10^9}.$$

Sin embargo, obsérvese que cada uno de los cuatro números que se han sumado fue dado con un error de 5×10^{-9}, de manera que la última cifra decimal en el valor calculado de log 7 podría estar afectada por un error de ± 2. De hecho, sin embargo, esta última cifra es también correcta.

6.3 Solución numérica de ecuaciones

Se agregan aquí algunas observaciones sobre la solución numérica de la ecuación $f(x) = 0$, donde $f(x)$ no es necesariamente un polinomio.[1] Se comienza con un primer valor de prueba x_0 para una de las raíces y luego se afina esta aproximación. Queda abierta la cuestión de cómo se escoge el primer valor de aproximación para la raíz y cuán buena es esta aproximación. Puede, por ejemplo, hacerse una conjetura o, mejor aún, obtenerse una primera aproximación a partir de la gráfica de la función $y = f(x)$, cuya intersección con el eje x indica la raíz deseada.

Luego se intenta afinar la aproximación mediante un proceso o transformación que lleve el valor x_0 a una "segunda aproximación", y se repite este proceso. Resolver numéricamente la ecuación $f(x) = 0$ consiste en llevar a cabo esas aproximaciones sucesivas en forma repetida (o, como se dice, "iterar" el proceso) con la esperanza de que los valores iterados $x_1, x_2, \ldots, x_n$ converjan satisfactoriamente a la raíz ξ. A continuación se consideran dichos procedimientos y se discute su precisión.

a. Método de Newton

Descripción del método. El procedimiento iterativo de Newton se basa en un principio fundamental del cálculo diferencial: el reemplazo de una curva por una tangente en la inmediata vecindad del punto de contacto. Partiendo con un primer valor aproximado, x_0, para una raíz ξ de la ecuación $f(x) = 0$, se considera el punto de la gráfica de la función $y = f(x)$ cuyas coordenadas son $x = x_0$, $y = f(x_0)$. Con el fin de encontrar una mejor aproximación para la intersección ξ de la curva con el eje x, se determina el punto x_1 en el cual la tangente en $x = x_0$, $y = f(x_0)$ intersecta al eje x. La abscisa x_1 de esta intersección representa una nueva y, en ciertas circunstancias, mejor aproximación que x_0 para la raíz deseada, ξ, de la ecuación.

[1] Estamos, por supuesto, interesados solamente en la determinación de raíces reales de $f(x) = 0$.

De la figura 6.4 se obtiene de inmediato

$$\frac{f(x_0)}{x_0 - x_1} = f'(x_0);$$

y, por tanto, la nueva aproximación

$$x_1 = x_0 - \frac{f(x_0)}{f'(x_0)}. \tag{10}$$

Partiendo con x_1 como aproximación, se repite el proceso para encontrar $x_2 = x_1 - f(x_1)/f'(x_1)$, y así sucesivamente.

La utilidad de este procedimiento depende esencialmente de la naturaleza de la curva. En la situación que se indica en la Fig. 6.4, las aproximaciones sucesivas x_n convergen con precisión creciente a la raíz deseada ξ.

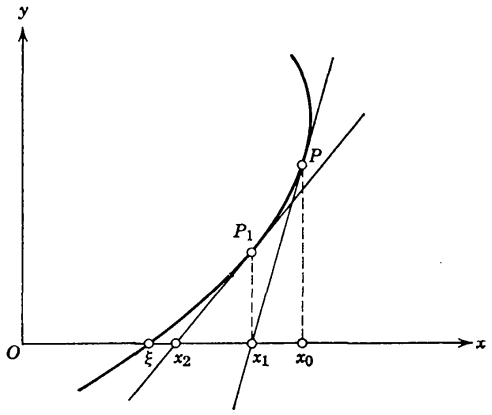


Figura 6.4 Método de aproximación de Newton.

Sin embargo, la Fig. 6.5 muestra que, con una elección plausible del valor original x_0, nuestra construcción no necesariamente converge a la raíz requerida. Es necesario, por consiguiente, examinar en general las circunstancias bajo las cuales el método de Newton proporciona aproximaciones útiles a la solución de la ecuación.

Convergencia cuadrática del método de Newton

Suponiendo que en un intervalo suficientemente amplio en torno a la raíz ξ la segunda derivada $f''(x)$ no es "demasiado grande" y la primera derivada $f'(x)$ no "demasiado pequeña, el hecho principal relativo a la aproximación de Newton es que los "errores" sucesivos

$$h_1 = \xi - x_1, \quad h_2 = \xi - x_2, \ldots, h_n = \xi - x_n, \ldots$$

convergen cuadráticamente a cero en el sentido de que $|h_{n+1}| \leq \mu h_n^2$ para una constante fija μ. Lo anterior indica una razón de convergencia extremadamente rápida. La desigualdad, escrita en la forma $|h_{n+1}\mu| \leq |h_n\mu|^2$, implica, por ejemplo, que cuando $|h_n\mu| < 10^{-m}$ se tiene $|h_{n+1}\mu| < 10^{-2m}$, o sea, el número de "dígitos significativos" en μx_n se duplica en cada etapa.

La demostración de la convergencia cuadrática es inmediata. De las relaciones $x_{n+1} = x_n - f(x_n)/f'(x_n)$ y $f(\xi) = 0$, se encuentra que

$$h_{n+1} = \xi - x_{+1} = \xi - x_n - \frac{f(\xi) - f(x_n)}{f'(x_n)}.$$

Por la fórmula de Taylor,

$$f(\xi) - f(x_n) = (\xi - x_n)f'(x_n) + \tfrac{1}{2}(\xi - x_n)^2 f''(\eta),$$

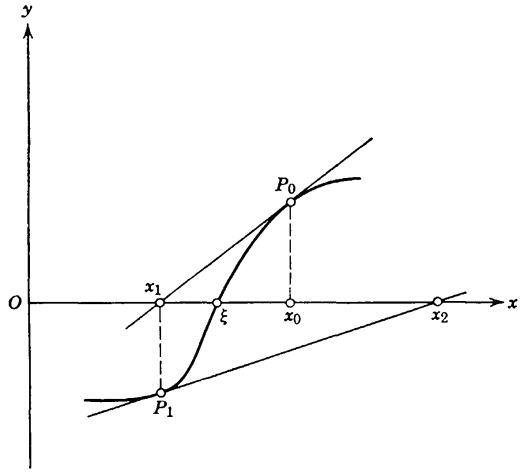


Figura 6.5

con η comprendido entre ξ y x_n. Por tanto,

$$h_{n+1} = -\frac{f''(\eta)}{2f'(x_n)} h_n^2 . \tag{11}$$

Para establecer la convergencia, supóngase que x_n pertenece ya a un intervalo fijo $\xi - \delta < x < \xi + \delta$ en el cual $|f''|$ tiene el valor máximo M_2, $|f'|$ el valor mínimo m_1 (positivo), y δ es tan pequeño que $\frac{1}{2}\delta M_2/m_1 < 1$. Haciendo $\mu = \frac{1}{2}M_2/m_1$, se tiene $\mu\delta < 1$ y

$$|h_{n+1}| \leq \mu\,|h_n|^2 \leq \mu\delta\,|h_n| < |h_n|.$$

Esta desigualdad muestra, en primer lugar, que x_{n+1} pertenece nuevamente a la misma vecindad δ de ξ, de manera que el argumento puede repetirse.

De esta manera, si x_0 está en la vecindad δ de ξ, lo mismo ocurrirá para todos los x_n subsiguientes. De $|h_{n+1}| \leq \mu\delta\,|h_n|$ se deduce entonces que $|h_{n+1}| \leq (\mu\delta)^{n+1}\,|h_0|$, lo cual implica que $h_n \to 0$, o sea, que $x_n \to \xi$. Además, la ley del decrecimiento cuadrático, $|h_{n+1}| \leq \mu\,|h_n|^2$, se cumplirá para los errores. Resulta entonces evidente que el método de Newton proporciona una sucesión x_n que efectivamente converge a la solución ξ siempre que f' y f'' existan y sean continuas cerca de ξ, que $f'(\xi) \neq 0$ y que x_0 esté suficientemente cerca de la solución. El carácter cuadrático de la aproximación es a menudo una ventaja decisiva del método de Newton respecto a otros (ver p. 522).

*b. La Regla de la posición falsa

El método de Newton es el caso límite de un método más antiguo, el de la "regla de la posición falsa", en el cual aparece la secante en lugar de la tangente. Supongamos que se conocen dos puntos (x_0, y_0) y (x_1, y_1) en la vecindad de la intersección con el eje x requerida. Si reemplazamos la curva por la secante que une estos dos puntos, la intersección de ésta con el eje x puede ser una mejor aproximación a la raíz[1] requerida de la ecuación. Para la abscisa ξ del punto de intersección se tiene (Fig. 6.6)

$$\frac{\xi - x_0}{f(x_0)} = \frac{\xi - x_1}{f(x_1)}, \tag{12}$$

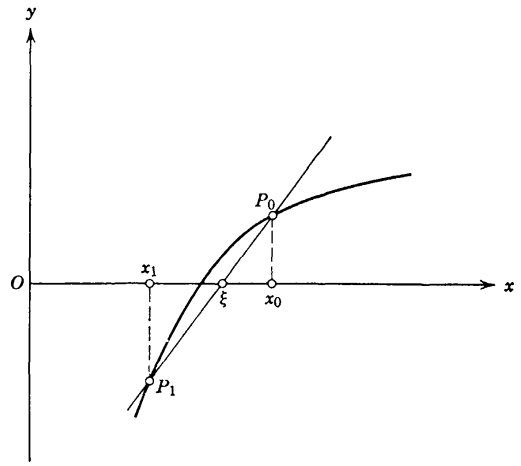


Figura 6.6 La regla de la posición falsa.

[1] Esto equivale esencialmente a aplicar la interpolación lineal a la función inversa.

lo cual conduce a

$$\xi = \frac{x_0 f(x_1) - x_1 f(x_0)}{f(x_1) - f(x_0)} = \frac{x_0 f(x_1) - x_0 f(x_0) + x_0 f(x_0) - x_1 f(x_0)}{f(x_1) - f(x_0)};$$

o sea,

$$\xi = x_0 - \frac{f(x_0)}{\dfrac{f(x_1) - f(x_0)}{x_1 - x_0}}. \tag{13}$$

Esta fórmula, que determina la aproximación ξ con respecto a x_0 y x_1, constituye la regla de la 'posición falsa". Es útil cuando uno de los valores de la función es positivo y el otro es negativo, como en la Fig. 6.6, por ejemplo, en donde $y_0 > 0$ y $y_1 < 0$.

La fórmula de aproximación de Newton resulta como un caso límite cuando $x_1 \to x_0$, pues el denominador del segundo término del miembro derecho de la fórmula (13) tiende a $f'(x_0)$ conforme x_1 tiende a x_0.

Aunque la regla de la posición falsa puede considerarse como más elemental que el método de Newton, este último tiene la gran ventaja de requerir como aproximación inicial un solo valor de x en vez de dos.

c. El método iterativo

El esquema de iteración. Consideremos ahora un esquema de grandes posibilidades en la solución de ecuaciones escritas en la forma

$$x = \phi(x),$$

donde ϕ es una función continua con una derivada también continua. La solución de ecuaciones de la forma $f(x) = 0$ puede reducirse a la de $x = \phi(x)$ haciendo $\phi(x) = x - c(x)f(x)$, donde $c(x)$ es cualquier función diferente de cero.

En el *método iterativo*,[1] particularmente sugerente, se empieza también con un valor de aproximación inicial x_0 escogido adecuadamente y luego se determina una sucesión de valores $x_1, x_2, x_3 \ldots$, por medio de las condiciones

$$x_{n+1} = \phi(x_n), \qquad n = 0, 1, 2, \ldots.$$

Si esta sucesión "de iteración", x_n, converge a un límite ξ, entonces $\xi = \phi(\xi)$ es una solución de nuestra ecuación, pues en ese caso $\lim_{n\to\infty} x_{n+1} = \xi$ y $\lim_{n\to\infty} \phi(x_n) = \phi(\xi)$ debido a la continuidad de la función ϕ.

[1] Llamado algunas veces "método de aproximaciones sucesivas". El método es usado en muchos contextos matemáticos diferentes para la solución de ecuaciones de una clase u otra.

Convergencia. La sucesión de valores x_n en el proceso iterativo converge, con una hipótesis muy general, a una solución: Si la primera aproximación x_n está en un intervalo [2] J en torno a ξ, en el cual

$$|\phi'(x)| < q$$

con una constante $q < 1$, entonces x_n converge a ξ.

Puesto que, por hipótesis, x_0 está en J, se tiene

$$x_1 - \xi = \phi(x_0) - \phi(\xi).$$

El miembro derecho de esta ecuación, es por el teorema del valor medio, igual a $(x_0 - \xi)\phi'(\bar{x})$, siendo $\bar{x}$ un punto de J. Así, por nuestras hipótesis es

$$|x_1 - \xi| \leq q\,|x_0 - \xi|,$$

de manera que x_1 pertenece a J y, por tanto, también es

$$|x_2 - \xi| \leq q\,|x_1 - \xi| \leq q^2\,|x_0 - \xi|.$$

En general se obtiene

$$|x_n - \xi| \leq q^n\,|x_0 - \xi|;$$

y, como $q^n \to 0$ conforme $n \to \infty$, nuestro aserto queda demostrado.

De lo anterior se ve, además, que la sucesión de iteración x_n no converge cuando $\phi'(x) > 1$ en un intervalo que contiene a ξ; y si $|\phi'(\xi)| = 1$ no es posible hacer una afirmación de validez general.

Puntos fijos atractivos y repulsivos

Es útil considerar el proceso de iteración en términos de transformaciones. La función $y = \phi(x)$ representa una transformación que proyecta un punto x del eje numérico sobre un punto imagen y de dicho eje (ver p. 44). La solución ξ es entonces un punto que no cambia bajo la transforción ϕ, denominado *punto fijo,* y el problema en este caso consiste en encontrar un punto fijo de la transformación. Este problema puede resolverse mediante la iteración siempre que $|\phi'(\xi)| \leq q < 1$, como ya se ha visto.

La transformación $y = \phi(x)$ en la vecindad de la raíz o punto fijo ξ tiene, para $|\phi'(x)| < q < 1$, la propiedad de ser una *contracción,* esto es, de disminuir la distancia del punto original al punto fijo. Los puntos fijos de transformaciones de contracción se llaman *puntos fijos atractivos.* Su construcción mediante iteraciones converge como los términos de una serie geométrica de cociente q.

Si la raíz ξ o el correspondiente punto fijo de nuestra transformación está en un intervalo en el cual $|\phi'(x)| > r$, donde r es una constante mayor que 1, la *transformación es de expansión,* el proceso de iteración diverge y el *punto fijo* se denomina *repulsivo.*

[2] Aunque se desconoce ξ, a menudo es posible determinar *a priori* tal intervalo.

Si en el punto fijo se tiene $|\phi'(\xi)| = 1$, no es posible hacer afirmaciones generales relativas a la convergencia de las iteraciones. Tales puntos fijos se denominan a veces *indiferentes.*

Debe recalcarse lo siguiente: un punto fijo ξ de una transformación ϕ es automáticamente también un punto fijo para ψ, la transformación inversa: $\xi = \psi(\xi)$. Si $|\phi'(\xi)| > 1$ en la vecindad de una raíz ξ y $x = \psi(y)$ es la función inversa de ϕ, entonces $|\psi'(\xi)| < 1$. Así, ξ es un punto fijo atractivo para esta transformación inversa, y es posible reemplazar el esquema de iteración originalmente divergente por uno convergente para la transformación inversa. Como un ejemplo, considérese la ecuación

$$x = \tan x.$$

De las gráficas de las funciones $y = x$ e $y = \tan x$ resulta claro que éstas se intersectan en algún punto del intervalo $\pi < x < \frac{3}{2}\pi$ y que la ecuación tendrá un raíz ξ en ese intervalo (Fig. 6.7). Puesto que

$$\frac{d \tan x}{dx} = \frac{1}{\cos^2 x} > 1,$$

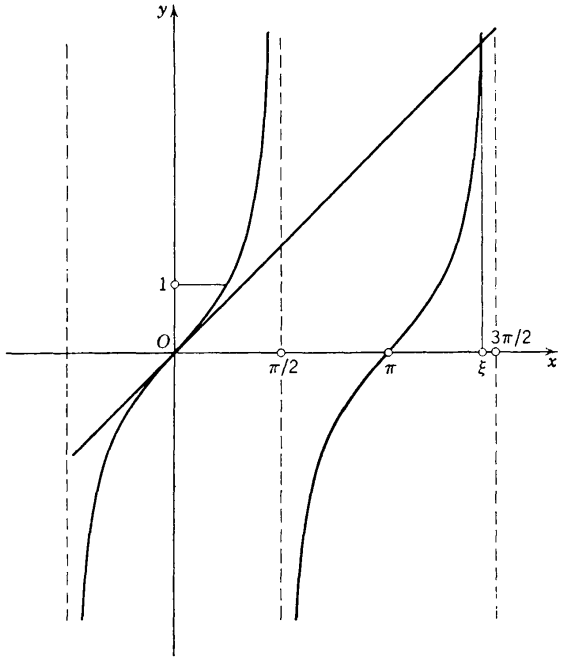


Figura 6.7 Intersección (ξ, ξ) de las curvas $y = \tan x$ y $y = x$.

el procedimiento de iteración con un punto cualquiera x_0 del intervalo no converge. Sin embargo, se obtiene una sucesión de iteración convergente si se escribe la ecuación en la forma inversa (usando la notación arc tan x para la rama principal),

$$x = \arctan x + \pi.$$

Puesto que aquí

$$\frac{d}{dx}\arctan x = \frac{1}{1+x^2} < 1,$$

la sucesión definida por $x_{n+1} = \arctan x_n + \pi$ y, digamos, $x_0 = \pi$, converge a ξ.

d. Iteraciones y procedimiento de Newton

Como ya se ha mencionado, el problema de resolver una ecuación de la forma $f(x) = 0$ puede reducirse al de resolver la ecuación $x = \phi(x)$ si se escoge para ϕ cualquier expresión de la forma

$$\phi(x) = x - c(x)f(x),$$

donde $c(x)$ es una función que no se anula. Si se desea resolver la ecuación resultante $x = \phi(x)$ por iteración es necesario asegurar, escogiendo $c(x)$ adecuadamente, que el punto fijo ξ de la transformación ϕ es "atractivo", esto es, que $|\phi'(\xi)| < 1$. En estas condiciones, se tiene para la solución ξ de $f(\xi) = 0$:

$$\phi'(\xi) = 1 - c'(\xi)f(\xi) - c(\xi)f'(\xi) = 1 - c(\xi)f'(\xi).$$

Lo más sencillo es tomar para $c(x)$ la expresión $1/f'(x)$. Entonces, en efecto, $|\phi'(\xi)| = 0 < 1$. Esta elección de $c(x)$ conduce a la sucesión de iteración

$$x_{n+1} = \phi(x_n) = x_n - \frac{f(x_n)}{f'(x_n)},$$

que es precisamente la sucesión de aproximaciones (10) (pp. 514-515) del método de Newton. Para el error $x_n - \xi = h_n$ se tiene la estimación

$$|h_{n+1}| = |\phi(x_n) - \phi(\xi)| \leq q h_n,$$

donde q es el máximo de $|\phi'(x)|$ en el intervalo cuyos puntos extremos son ξ y x_n. Puesto que en este caso

$$\phi'(x) = \frac{f(x)f''(x)}{f'^2(x)}$$

y $f(x) = f(x) - f(\xi) = f'(\eta)(x - \xi)$, se ve que q mismo es del orden de h_n, confirmándose así nuevamente el carácter cuadrático de la aproximación en el método de Newton.

Otra elección sencilla para $c(x)$ la constituye el valor constante $1/f'(x_0)$, lo cual conduce a la fórmula de recurrencia

$$x_{n+1} = \phi(x_n) = x_n - \frac{f(x_0)}{f'(x_0)}.$$

Aquí es $\phi'(\xi) = 1 - f'(\xi)/f'(x_0)$. Si f' es continua y diferente de cero se tendrá un punto fijo atractivo ξ cuando la aproximación inicial x_0 esté tan cerca de la solución ξ que sea

$$|\phi'(\xi)| = \frac{|f'(x_0) - f'(\xi)|}{|f'(x_0)|} < 1.$$

Esta sucesión de iteración es algo más simple que la usada en el método de Newton; sin embargo, su convergencia será más lenta (parecida a la de una serie geométrica), como ocurre con la mayor parte de los esquemas de iteración.

Ejemplos. Como un ejemplo, considérese la ecuación cúbica

$$f(x) = x^3 - 2x - 5 = 0.$$

Puesto que $f(2) = -1 < 0$, $f(3) = 16 > 0$, es claro que existe una raíz en el intervalo $2 < x < 3$. Además, como $f'(x) = 3x^2 - 2 > 3(2)^2 - 2 > 0$, ese intervalo contiene sólo una raíz. Usando el método de Newton con la aproximación inicial $x_0 = 2$ se encuentra sucesivamente

$$x_1 = x_0 - \frac{f(x_0)}{f'(x_0)} = 2 - \frac{-1}{3(2)^2 - 2} = 2.1, \quad f(x_1) = 0.061,$$

$$x_2 = x_1 - \frac{f(x_1)}{f'(x_1)} = 2.1 - \frac{0.061}{3(2.1)^2 - 2} = 2.094568.$$

Análogamente, puesto que $f(2.1) > 0$, $f(2) < 0$, la raíz ξ está entre 2 y 2.1. En el intervalo $1.9 < x < 2.2$ y, por consiguiente, también en el intervalo $\xi - 0.1 < x < \xi + 0.1$, se tienen las estimaciones

$$|f''(x)| = |6x| < 6(2.2) = 13.2,$$

$$f'(x) = 3x^2 - 2 > 3(1.9)^2 - 2 = 8.83.$$

Se deduce entonces [ver (11), p. 516] que

$$|\xi - x_{n+1}| \leq \frac{13.2}{2(10.83)}|x_n - \xi|^2 < 0.75\,|x_n - \xi|^2;$$

en el supuesto de que $|x_n - \xi| < 0.1$. Como $|x_0 - \xi| = |\xi - 2| < 0.1$, se obtiene sucesivamente

$$|x_1 - \xi| < (0.75)(0.1)^2 = 0.0061,$$

$$|x_2 - \xi| < (0.61)(0.0061)^2 < 0.000042.$$

Si este grado de aproximación no es suficiente, puede obtenerse una aproximación posterior, x_3, con un error $< (0.75)(0.000042)^2 < 0.000\,000\,001\,3$.

Todo x_n posterior a x_0 debe ser mayor que ξ, como se desprende del hecho de que f' y f'' son positivas, lo cual implica que

$$h_{n+1} = -f''(\eta)h_n^2/2f'(x_n) < 0.$$

Por otra parte, la aplicación de la regla de la posición falsa [(13, p. 518] a los valores x_0, x_1 da para la intersección del eje x con la secante que une los puntos $(x_0, f(x_0))$ y $(x_1, f(x_1))$ el resultado

$$\xi = x_0 - \frac{f(x_0)(x_1 - x_0)}{f(x_1) - f(x_0)} = 2.09425\cdots.$$

Puesto que la curva es convexa en el intervalo considerado, la secante queda por encima de la curva y la aproximación de ξ debe ser menor que la raíz ξ.

Como un segundo ejemplo, resolvamos la ecuación

$$f(x) = x\log_{10} x - 2 = 0.$$

Tenemos $f(3) = -0.6$ y $f(4) = +0.4$, y por consiguiente usamos $x_0 = 3.5$ como primera aproximación. Usando tablas logarítmicas de diez cifras decimales obtenemos las siguientes aproximaciones sucesivas

$$x_0 = 3.5, \qquad x_1 = 3.598,$$
$$x_2 = 3.5972849, \qquad x_3 = 3.5972850235.$$

Apéndice

*A.1 Fórmula de Stirling

En muchas aplicaciones, particularmente en Estadística y en la Teoría de Probabilidades, se presenta la necesidad de disponer de una aproximación sencilla para $n!$ en la forma de una función elemental de n. Tal expresión está dada por el siguiente teorema, que lleva el nombre de su descubridor, Stirling (ver también capítulo 8, p. 630).

Conforme $n \to \infty$,

$$\frac{n!}{\sqrt{2\pi}\, n^{n+1/2}e^{-n}} \to 1; \tag{14}$$

o, en forma más precisa,

$$\sqrt{2\pi}\, n^{n+1/2}e^{-n} < n! < \sqrt{2\pi n}^{n+1/2}e^{-n}\left(1 + \frac{1}{4n}\right) \tag{14a}$$

En otras palabras, las expresiones $n!$ y $\sqrt{\pi}\, n^{n+1/2}e^{-n}$ difieren solamente por un porcentaje pequeño cuando el valor de n es grande —o, como usualmente se dice, las dos expresiones son *asintóticamente iguales*— y, al mismo tiempo, el factor $1 + 1/4n$ proporciona una estimación del grado de exactitud de la aproximación.

Esta notable fórmula aparece cuando se trata de evaluar el área bajo la curva $y = \log x$.[1] Por integración (p. 297) se encuentra que A_n, el área exacta bajo esta curva entre las abscisas $x = 1$ y $x = n$, está dada por

$$A_n = \int_1^n \log x\, dx = x\log x - x\Big]_1^n = n\log n - n + 1. \tag{15}$$

Sin embargo, si se estima el área usando la regla del trapezoide, trazando verticales por los puntos $x = 1, x = 2, \ldots, x = n$ como en la Fig. 6.8, se obtiene un valor aproximado T_n para el área [ver (6), p. 503]:

$$\begin{aligned} T_n &= \log 2 + \log 3 + \cdots + \log(n-1) + \tfrac{1}{2}\log n \\ &= \log n! - \tfrac{1}{2}\log n. \end{aligned} \tag{16}$$

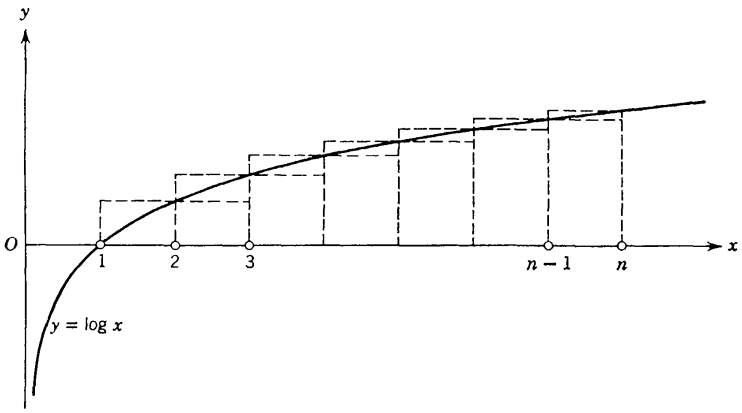


Figura 6.8

Haciendo la hipótesis razonable de que A_n y T_n son del mismo orden de magnitud, se encuentra que $n!$ es del mismo orden de magnitud que $n^{n+1/2}e^{-n}$, lo cual constituye la esencia de lo expresado por la fórmula de Stirling.

Para precisar este argumento demostraremos primero que la diferencia $a_n = A_n - T_n$ es acotada, de lo cual se desprenderá inmediatamente que $T_n = A_n(1 - a_n/A_n)$ es del mismo orden de magnitud que A_n. La can-

[1] El método empleado aquí es un caso especial de la fórmula de Euler MacLaurin que será discutida en el capítulo 8, p. 642.

tidad $a_{k+1} - a_k$ es la diferencia entre el área bajo la curva y el área bajo la secante, en la franja $k < x < k + 1$. Puesto que la curva es cóncava y queda por encima de la secante, $a_{k+1} - a_k$ es positivo y $a_n = (a_n - a_{n-1}) + (a_{n-1} - a_{n-2}) + \cdots + (a_2 - a_1) + a_1$, monótona creciente. Además, la diferencia $a_{k+1} - a_k$ es obviamente menor (ver Fig. 6.9) que la diferencia

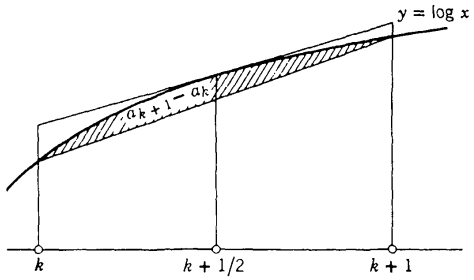


Figura 6.9

entre el área bajo la tangente en $x = k + \frac{1}{2}$ y el área bajo la secante. Por tanto, se tiene la desigualdad

$$\begin{aligned} a_{k+1} - a_k &< \log(k + \tfrac{1}{2}) - \tfrac{1}{2}\log k - \tfrac{1}{2}\log(k+1) \\ &= \tfrac{1}{2}\log\left(1 + \frac{1}{2k}\right) - \tfrac{1}{2}\log\left[1 + \frac{1}{2(k+\frac{1}{2})}\right] \\ &< \tfrac{1}{2}\log\left(1 + \frac{1}{2k}\right) - \tfrac{1}{2}\log\left[1 + \frac{1}{2(k+1)}\right]. \end{aligned}$$

Sumando estas desigualdades para $k = 1,2, \ldots, n-1$, se encuentra que todos los términos del lado derecho, excepto dos, se cancelan y (puesto que $a_1 = 0$) resulta

$$a_n < \tfrac{1}{2}\log \tfrac{3}{2} - \tfrac{1}{2}\log\left(1 + \frac{1}{2n}\right) < \tfrac{1}{2}\log \tfrac{3}{2}.$$

Como a_n, además de ser acotada, es monótona creciente, tiende a un límite a conforme $n \to \infty$. La desigualdad para $a_{k+1} - a_k$ da ahora

$$a - a_n = \sum_{k=n}^{\infty} (a_{k+1} - a_k) < \tfrac{1}{2}\log\left(1 + \frac{1}{2n}\right).$$

Ya que, por definición, $A_n - T_n = a_n$, por (15) y (16) se tiene

$$\log n! = 1 - a_n + (n + \tfrac{1}{2})\log n - n,$$

o bien, escribiendo $\alpha_n = e^{1-a_n}$,

$$n! = \alpha_n n^{n+1/2} e^{-n}.$$

La sucesión α_n es monótona *decreciente* y tiende al límite $\alpha = e^{1-a}$; y, por tanto,

$$\begin{aligned} 1 < \frac{\alpha_n}{\alpha} &= e^{a-a_n} < e^{(1/2)\log(1+1/2n)} \\ &= \sqrt{1 + \frac{1}{2n}} < 1 + \frac{1}{4n}. \end{aligned}$$

Se tiene, por consiguiente,

$$\alpha n^{n+1/2} e^{-n} < n! < \alpha n^{n+1/2} e^{-n}\left(1 + \frac{1}{4n}\right).$$

Sólo nos resta encontrar el valor verdadero del límite a. Para ello usaremos la fórmula (80) del capítulo 3, p. 302:

$$\sqrt{\pi} = \lim_{n\to\infty} \frac{(n!)^2 2^{2n}}{(2n)!\sqrt{n}}.$$

Reemplazando $n!$ por $\alpha_n n^{n+1/2} e^{-n}$ y $(2n)!$ por $\alpha_{2n} 2^{2n+1/2} n^{2n+1/2} e^{-2n}$ se obtiene inmediatamente

$$\begin{aligned} \sqrt{\pi} &= \lim_{n\to\infty} \frac{\alpha_n^2}{\alpha_{2n}\sqrt{2}} \\ &= \frac{\alpha^2}{\alpha\sqrt{2}}, \end{aligned}$$

por lo que $\alpha = \sqrt{2\pi}$. La demostración de la fórmula de Stirling queda así concluida.

Además de su interés teórico, la fórmula de Stirling es una herramienta de gran utilidad para el cálculo numérico de $n!$ cuando n es grande. En vez de multiplicar un gran número de enteros, sólo tiene que calcularse la expresión de Stirling por medio de logaritmos, lo que supone un número de operaciones considerablemente menor. Así, para $n = 10$ se obtiene el valor 3598696 mediante la expresión de Stirling (usando tablas de siete cifras), en tanto que el valor exacto de 10! es 3628800. El porcentaje de error es apenas de 5%%.

PROBLEMAS

SECCION 6.1, página 502

1. Si $f''(x) \geq 0$, demostrar que la regla del trapezoide conduce a un valor mayor y la regla de la tangente a uno menor que el valor exacto de la integral de f.

2. Estimar el valor $h = (b - a)/n$ necesario para calcular hasta la p-ésima cifra decimal, usando la regla de Simpson, el valor de

(*a*) $\log 2 = \displaystyle\int_1^2 \frac{1}{x}\,dx,$ (*b*) $\pi = 4\displaystyle\int_0^1 \frac{1}{1+x^2}\,dx.$

3. Estimar en términos de k y s ($k < 1$ y $s < 1$) el número de puntos necesarios para calcular con un error menor de ε el valor de la integral elíptica

$$u(s) = \int_0^s \frac{dx}{\sqrt{(1 - x^2)(1 - k^2x^2)}}.$$

4. Sea $f(x)$ una función continua en el intervalo $\alpha \leq x \leq \alpha + h$, con una derivada acotada uniformemente: $|f'(x)| \leq M_1$ para una constante M_1. Demostrar que para cualquier punto fijo ξ ($\alpha \leq \xi \leq \alpha + h$) se tiene la estimación

$$\left| \int_\alpha^{\alpha+h} f(x)\, dx - hf(\xi) \right| \leq \frac{M_1 h^2}{2}$$

5. Calcular numéricamente $\int_\alpha^\infty e^{-x^2}\, dx$ con un error menor que 1/100.

SECCION 6.2, página 509

1. El período de un péndulo está dado por

$$T = 2\pi \sqrt{\frac{l}{g}}$$

donde l es la longitud del péndulo. Si dicho péndulo acciona un reloj que se adelanta un minuto por día, determínese la corrección necesaria en l.

2. Para medir la altura de una colina, se observa desde la planicie una torre de 100 metros situada en la cima. El ángulo de elevación de la base de la torre es de 42° y la torre misma subtiende un ángulo de 6°. ¿Cuáles son los límites del error en la determinación de la altura si el ángulo de 42° está sujeto a un error de 1°?

SECCION 6.3, página 514

1. (*a*) Para resolver la ecuación $x = f(x)$ muéstrese la mejor forma de escoger la constante a de manera que el esquema de iteración

$$x_{k+1} = x_k + a[x_k - f(x_k)]$$

converja lo más rápidamente posible en la vecindad de la solución.

(*b*) Aplíquese este método con el fin de resolver la ecuación $\sqrt{A}$,

$$x = \frac{A}{x}.$$

(*c*) Demostrar que si $A \geq 1$, el número de cifras decimales exactas por lo menos se duplica en cada paso del esquema de iteración obtenido en (*b*).

2. (*a*) Muéstrese la mejor forma de elegir un polinomio

$$g(x) = a + bx^2$$

de manera tal que el esquema de iteración para $\sqrt{A}$,

$$x_{k+1} = x_k + g(x_k)\left(x_k - \frac{A}{x_k}\right),$$

converja lo más rápidamente posible en la vecindad de la solución.

(*b*) Estimar la rapidez de la convergencia.

(*c*) Mostrar cómo mejorar posteriormente la convergencia mediante la elección adecuada de polinomios $g(x)$ de grado superior.

3. Investigar la posibilidad de esquemas adecuados del tipo tratado en los problemas 1 y 2, para el cálculo de $\sqrt[n]{A}$.

SECCION A.1, página 523

1. Probar que $\lim_{n\to\infty} \frac{\sqrt[n]{n!}}{n} = \frac{1}{e}$.

***2.** Considerando $\int_{1/2}^{n+1/2} \log(\alpha + x)\, dx$, $\alpha > 0$, demuéstrese que

$$\alpha(\alpha + 1) \cdots (\alpha + n) = a_n n!\, n^\alpha,$$

donde a_n tiene como cota inferior un número positivo. Demostrar que a_n es monótona decreciente para valores de n suficientemente grandes. [El límite de a_n conforme $n \to \infty$ es $1/\Gamma(\alpha)$.]

3. Encontrar una expresión aproximada para $\frac{n_1!\, n_2! \cdots n_l!}{n!}$, donde $n_1 + n_2 + \cdots + n_l = n$.

4. Demostrar que el coeficiente de x^n en el desarrollo binomial de $\frac{1}{\sqrt{1 - x}}$ está dado asintóticamente por $\frac{1}{\sqrt{\pi n}}$.

7

Sumas y productos infinitos

La serie geométrica, la serie de Taylor y cierto número de ejemplos previamente discutidos en este libro, sugieren la conveniencia de estudiar aquellos procesos de límite pertenecientes al análisis en los cuales la suma de una *serie infinita* aparece desde un punto de vista más general. En principio, cualquier valor límite

$$S = \lim_{n\to\infty} s_n$$

puede expresarse como una serie infinita; sólo se necesita hacer $a_n = s_n - s_{n-1}$ para $n > 1$ y $a_1 = s_1$, obteniéndose

$$s_n = a_1 + a_2 + \cdots + a_n,$$

y así el valor S resulta ser el límite de s_n, la suma de n términos, cuando n crece. Este hecho se expresa diciendo que S es la "suma de la serie infinita"

$$a_1 + a_2 + a_3 + \cdots.$$

Una "suma infinita" tal es simplemente una manera de representar un límite cuando cada aproximación sucesiva se encuentra a partir de la anterior sumando un término más. Así, la expresión decimal de un número a es, en principio, simplemente la representación del número en la forma de una serie infinita $a = a_1 + a_2 + a_3 + \cdots$, en la que, si $0 \leq a \leq 1$, el término a_n se reemplaza por $\alpha_n \times 10^{-n}$, siendo α_n uno de los enteros 0, 1, . . . , 9.

Puesto que todo valor límite puede ser escrito en la forma de una serie infinita, un estudio especial de las series puede parecer superfluo. Sin embargo, tales valores límite aparecen con frecuencia, en forma natural, como series infinitas que exhiben leyes de formación particularmente simples. *No toda* serie tiene una ley de formación fácilmente reconocible. Por ejemplo, el número π puede ciertamente ser representado como un decimal

(que es una serie $\Sigma c_\nu 10^{-\nu}$), aunque no conocemos una ley simple que nos permita establecer el valor de una cifra arbitraria, digamos, la número 7000, de este decimal. Si, no obstante, en vez de ello consideramos la serie de Leibnitz-Gregory para $\pi/4$, tendremos una expresión con una ley general de formación perfectamente clara [ver (7), p. 464].

Análogos a las series infinitas, en las cuales las aproximaciones al límite se forman por la adición repetida de nuevos términos, los llamados *productos infinitos* proporcionan aproximaciones al límite mediante la multiplicación repetida por nuevos factores. Sin embargo, no profundizaremos mucho en la teoría general de los productos infinitos. El tema principal de éste y del capítulo 8 serán las series infinitas.

7.1 Los conceptos de convergencia y divergencia

a. Conceptos básicos

Criterios de convergencia de Cauchy. Considérese una serie infinita con "término general" a_n. La serie [1] es entonces de la forma

$$a_1 + a_2 + \cdots = \sum_{\nu=1}^{n} a_\nu.$$

El símbolo en el lado derecho, en el que aparece el signo de suma, es simplemente una manera abreviada de escribir la expresión de la izquierda.

Si, conforme n crece, la *n-ésima suma parcial*

$$s_n = a_1 + a_2 + \cdots + a_n = \sum_{\nu=1}^{n} a_\nu$$

se aproxima a un límite

$$S = \lim_{n\to\infty} s_n,$$

se dice que la serie es *convergente;* en caso contrario, se dice que es *divergente.* En el primer caso, S se llama la *suma de la serie.*

Hemos encontrado ya muchos casos de series convergentes; como ejemplos podemos citar la serie geométrica $1 + q + q^2 + \cdots$, que converge a la suma $1/(1 - q)$ cuando $|q| < 1$, la serie para log 2, la serie para e, y otras.

En el lenguaje de las series infinitas, el criterio de convergencia de Cauchy (ver capítulo 1, pp. 97-98) se expresa como sigue:

[1] Para completar el formalismo incluimos la posibilidad de que ciertos a_n se anulen. Si *todos* los términos a partir de cierto índice N (es decir, cuando $n > N$) se anulan, hablamos de una *serie que termina.*

Una condición necesaria y suficiente para la convergencia de una serie es que el número

$$|s_m - s_n| = |a_{n+1} + a_{n+2} + \cdots + a_m| \tag{1}$$

$(m > n)$, *llegue a ser arbitrariamente pequeño al escogerse para m y n valores suficientemente grandes.* En otras palabras: *Una serie es convergente si, y sólo si, se satisface la siguiente condición: dado un número positivo* ε, *es posible encontrar siempre un índice* $N = N(\varepsilon)$ *tal que la expresión anterior,* $|s_m - s_n|$, *sea menor que* ε *para* $m > N$ *y* $n > N$.

El criterio de convergencia puede ilustrarse con la serie geométrica para $q = \frac{1}{2}$. Si se toma $\varepsilon = \frac{1}{10}$, basta hacer $N = 4$. En efecto,

$$|s_m - s_n| = \frac{1}{2^n} + \cdots + \frac{1}{2^{m-1}}$$

$$= \frac{1}{2^{n-1}}\left(\frac{1}{2} + \frac{1}{2^2} + \cdots + \frac{1}{2^{m-n}}\right) < \frac{1}{2^{n-1}}$$

$$\frac{1}{2^{n-1}} < \frac{1}{10} \quad \text{si } n > 4.$$

Cuando ε se escoge igual a $\frac{1}{100}$, es suficiente con tomar para el correspondiente valor de N el número 7, como puede verificarse fácilmente.

Es obvio que la condición

$$\lim_{n\to\infty} a_n = 0$$

es *necesaria* para la convergencia de una serie. Pues, de lo contrario, el criterio de convergencia daría un resultado negativo para $m = n + 1$. Sin embargo, esta condición necesaria de ninguna manera es *suficiente*. Por el contrario, es fácil encontrar series infinitas cuyos términos generales a_n tienden a 0 al crecer n pero cuyas sumas no existen, debido a que las sumas parciales s_n crecen sin límite al hacerlo n.

Ejemplos. Un ejemplo lo tenemos en la serie

$$1 + \frac{1}{\sqrt{2}} + \frac{1}{\sqrt{3}} + \cdots + \frac{1}{\sqrt{n}} + \cdots,$$

para la cual el término general es $1/\sqrt{n}$. Se ve de inmediato que

$$s_n > \frac{1}{\sqrt{n}} + \cdots + \frac{1}{\sqrt{n}} = \frac{n}{\sqrt{n}} = \sqrt{n}.$$

La n-ésima suma parcial crece más allá de toda cota al crecer n, por lo que la serie diverge.

Lo mismo puede decirse en el caso, ya clásico, de la *serie armónica*

$$1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \cdots.$$

Aquí

$$a_{n+1} + \cdots + a_{2n} = \frac{1}{n+1} + \cdots + \frac{1}{2n} > \frac{1}{2n} + \cdots + \frac{1}{2n} = \frac{1}{2}.$$

Puesto que n y $m = 2n$ pueden escogerse tan grandes como se desee, la serie diverge, ya que el criterio de Cauchy no es satisfecho. En efecto, por ser positivos todos los términos de la n-ésima suma parcial, ésta tiende a infinito. Por otra parte, la serie formada con los mismos números, pero alternando los signos,

$$1 - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} + \frac{1}{5} - + \cdots + \frac{(-1)^{n-1}}{n} + \cdots,$$

converge [ver (4) capítulo 5, p. 462] y tiene por suma $\log 2$.

No hay razón para afirmar que en toda serie divergente S_n tiende a $+\infty$ o a $-\infty$. Así, en la serie infinita

$$1 - 1 + 1 - 1 + 1 + - \cdots,$$

se observa que la suma parcial s_n toma los valores 1 y 0 alternadamente y, como consecuencia de esta oscilación hacia atrás y hacia adelante, ni tiende a un límite definido ni aumenta numéricamente más allá de toda cota.

El hecho siguiente, aunque evidente por sí mismo, es importante y debe hacerse notar. *La convergencia o divergencia de una serie no es alterada insertando o eliminando un número finito de términos.* Por lo que respecta a la convergencia o a la divergencia, carece completamente de importancia el que la serie comience con el término a_0 o con el a_1 o el a_5, o bien cualquier otro término elegido arbitrariamente.

b. Convergencia absoluta y convergencia condicional

La serie armónica $1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} \cdots$, como es sabido, diverge; pero, cuando se cambia en ella el signo de los términos cuyo denominador es par, resulta la serie para log 2, la cual es convergente. Por otra parte, la serie geométrica $1 - q + q^2 - q^3 + - \cdots$, cuando $0 \leq q < 1$, converge y tiene por suma $1/(1 + q)$. Haciendo positivos todos los signos que en ella aparecen resulta la serie

$$1 + q + q^2 + q^3 + \cdots,$$

que es también convergente y cuya suma es $1/(1 - q)$.

Surge aquí una diferencia que debe ser examinada. Para una serie cuyos términos son todos positivos, existen sólo dos casos posibles: o es convergente o la suma parcial n-ésima crece más allá de toda cota al crecer n. Esto es así porque, constituyendo las sumas parciales una sucesión monótona creciente, convergirían si permaneciesen acotadas. Hay convergencia cuando los términos individuales tienden a cero con rapidez suficiente al crecer n. Se tiene, por otra parte, divergencia si los términos no tienden a cero o bien se acercan a cero con extrema lentitud. Sin embargo, para series con términos positivos y negativos puede ocurrir que los cambios de signo produzcan la convergencia en los casos en que un incremento demasiado grande en las sumas parciales, debido a los términos positivos, sea compensado por los términos negativos, de manera que, como resultado final, dichas sumas tiendan a un límite definido.

Para entender mejor las diferentes posibilidades, considérese una serie $\sum_{\nu=1}^{\infty} a_\nu$ con términos positivos y negativos, y fórmese, para fines de comparación, la serie que tiene los mismos términos pero todos con signos positivos, esto es,

$$|a_1| + |a_2| + \cdots = \sum_{\nu=1}^{\infty} |a_\nu|.$$

Si esta serie converge, entonces, para valores de n y $m > n$ suficientemente grandes, la expresión

$$|a_{n+1}| + |a_{n+2}| + \cdots + |a_m|$$

será, en efecto, tan pequeña como se desee. Teniendo en cuenta que

$$|a_{n+1} + \cdots + a_m| \leq |a_{n+1}| + \cdots + |a_m|$$

la expresión de la izquierda es también arbitrariamente pequeña y, por el criterio de Cauchy, la serie $\sum_{\nu=1}^{\infty} a_\nu$ original converge. En este caso se dice que la serie original es *absolutamente convergente*. Su convergencia se debe a la pequeñez absoluta de sus términos y no depende de cambios en los signos.

Si, por otra parte, la serie con términos a_n diverge en tanto que la serie original converge, se dice que esta última es *condicionalmente convergente*. La convergencia condicional es el resultado de que los términos de signo opuesto se compensan mutuamente.

Criterio de Leibnitz. El *criterio de convergencia de Leibnitz,* enunciado a continuación, es con frecuencia útil para investigar la convergencia condicional:

Si los términos de una serie tienen signos alternados y si, además, sus valores absolutos, $|a_n|$, *tienden monótonamente a* 0 *(de manera que* $|a_{n+1}| \leqslant |a_n|$*), la serie* $\sum_{\nu=1}^{\infty} a_\nu$ *converge.* [Ejemplo: la serie de Leibnitz, (7), p. 464.]

Para la demostración, se supondrá que $a_1 > 0$, lo cual no resta generalidad al razonamiento, y se escribirá la serie en la forma

$$b_1 - b_2 + b_3 - + \cdots,$$

donde todos los términos b_n son ahora positivos, tienden a cero y satisfacen la condición $b_{n+1} \leq b_n$. Agrupando los términos en dos formas diferentes,

$$b_1 - (b_2 - b_3) - (b_4 - b_5) - \cdots$$

y

$$(b_1 - b_2) + (b_3 - b_4) + (b_5 - b_6) + \cdots,$$

se ve de inmediato que las sumas parciales $s_n = \sum_{1}^{n} a_\nu$ satisfacen las dos relaciones siguientes

$$s_1 \geq s_3 \geq s_5 \geq \cdots \geq s_{2n+1} \geq \cdots,$$

$$s_2 \leq s_4 \leq s_6 \leq \cdots \leq s_{2n} \leq \cdots.$$

Por otra parte, $s_{2n} \leq s_{2n+1} \leq s_1$ y $s_{2n+1} \geq s_{2n} \geq s_2$. Las sumas parciales impares $s_1, s_3, \ldots$ forman por consiguiente, una sucesión monótona decreciente

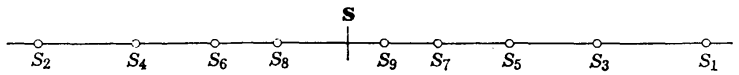


Figura 7.1 Convergencia de una serie alternante.

que nunca cae por debajo de s_2; por tanto, esta sucesión posee un límite, L (p. 96). De la misma manera, las sumas parciales pares $s_2, s_4, \ldots$ forman una sucesión monótona creciente cuyos términos nunca exceden el número fijo s_1, razón por la cual esta sucesión debe tener un valor límite, L'. Puesto que los números s_{2n} y s_{2n+1} difieren entre sí sólo por el número b_{2n+1}, que tiende a cero cuando n crece, los valores límite L y L' son iguales. O sea, las sumas parciales pares e impares tienden al mismo límite, el cual se denotará por S. Esto, sin embargo, implica que la serie es convergente, tal como se afirmó; y su suma es S.

**Criterio de Abel*

Un criterio para la convergencia, que incluye el de Leibnitz como un caso especial, es el *criterio de convergencia de Abel.* Sea $a_1 + a_2 + \cdots$ una serie infinita cuyas sumas parciales $s_n = a_1 + \cdots + a_n$ son acotadas para cualquier valor de n. Sea, además, $p_1, p_2, \ldots$ una sucesión de números positivos que tienden monótonamente a cero. Entonces la serie infinita

$$(2) \qquad p_1 a_1 + p_2 a_2 + \cdots$$

converge. (Para el caso especial de la serie $a_1 + a_2 + \cdots = +1 - 1 + 1 - 1 + - \cdots$ se encuentra que $p_1 - p_2 + p_3 - \cdots$ converge, lo que no es sino el criterio de

Leibnitz.) La demostración se realiza aplicando el criterio de Cauchy y usando una "suma por partes" para estimar

$$\begin{aligned}
&|p_{n+1}a_{n+1} + p_{n+2}a_{n+2} + \cdots + p_m a_m| \\
&\quad = |p_{n+1}(s_{n+1} - s_n) + p_{n+2}(s_{n+2} - s_{n+1}) + \cdots + p_m(s_m - s_{m-1})| \\
&\quad = |-p_{n+1}s_n + p_m s_m + (p_{n+1} - p_{n+2})s_{n+1} + (p_{n+2} - p_{n+3})s_{n+2} + \cdots \\
&\qquad + (p_{m-1} - p_m)s_{m-1}| \\
&\quad \leq p_{n+1}M + p_m M + (p_{n+1} - p_{n+2} + p_{n+2} - p_{n+3} + - \cdots + p_{m-1} - p_m)M \\
&\quad = 2p_{n+1}M,
\end{aligned}$$

donde M es una cota para los valores $|s_i|$. Puesto que $p_{n+1} \to 0$, la convergencia de la serie (2) se desprende del criterio de Cauchy.

*Resulta, en conclusión, otra observación general relativa a la fundamental diferencia que existe entre la convergencia absoluta y la convergencia condicional. Considérese una serie convergente $\sum_{\nu=1}^{\infty} a_\nu$. Denótense sus términos positivos por $p_1, p_2, p_3, \ldots$, y sus términos negativos por $-q_1, -q_2, -q_3, \ldots$. Al formar la n-ésima suma parcial, $s_n = \sum_{\nu=1}^{\infty} a_\nu$, de la serie dada debe aparecer un cierto número (llamémosle n') de términos positivos y otro cierto número (digamos, n'') de términos negativos, debiendo tenerse $n' + n'' = n$. Además, si el número de términos positivos de la serie, así como también el número de sus términos negativos, es infinito, entonces los números n' y n'' crecerán más allá de toda cota al crecer n. Se ve de inmediato que la suma parcial s_n simplemente es igual a la suma parcial $\sum_{\nu=1}^{n'} p_\nu$, de los términos positivos más la suma parcial, $-\sum_{\nu=1}^{n''} q_\nu$, de los términos negativos. Si la serie dada es absolutamente convergente, la serie de términos positivos, $\sum_{\nu=1}^{\infty} p_\nu$, y la serie de los valores absolutos de los términos negativos, $\sum_{\nu=1}^{\infty} q_\nu$, convergen. Esto es así porque cuando m crece las sumas parciales $\sum_{\nu=1}^{m} p_\nu$ y $\sum_{\nu=1}^{m} q_\nu$ constituyen sucesiones monótonas no decrecientes acotadas superiormente por $\sum_{\nu=1}^{\infty} |a_\nu|$.

La suma de una serie absolutamente convergente es, por lo tanto, igual a la suma de la serie formada por los términos positivos exclusivamente más la suma de la serie que consiste sólo de los términos negativos; en otras palabras, *es igual a la diferencia entre las dos series con términos positivos.*

En efecto, $\sum_{\nu=1}^{n} a_\nu = \sum_{\nu=1}^{n'} p_\nu - \sum_{\nu=1}^{n''} q_\nu$; al crecer n, n' y n'' crecen también

más allá de toda cota, y el límite de la izquierda debe, por tanto, ser igual a la diferencia entre las dos sumas de la derecha. Si la serie contiene solamente un número finito de términos de un signo particular, las cosas, en consecuencia, se simplifican. Si por, otra parte, la serie no converge absolutamente pero lo hace en forma condicional, las serie $\sum_{\nu=1}^{\infty} p_\nu$ y $\sum_{\nu=1}^{\infty} q_\nu$ deben ambas divergir; pues si las dos convergiesen, la serie original debería ser absolutamente convergente, lo que es contrario a la hipótesis. Si sólo una de ellas diverge, digamos, $\sum_{\nu=1}^{\infty} p_\nu$, en tanto que la otra converge, la separación en partes positiva y negativa, $s_n = \sum_{\nu=1}^{n'} p_\nu - \sum_{\nu=1}^{n''} q_\nu$, muestra que la serie no podría ser convergente. En efecto, al crecer n, n' y $\sum_{\nu=1}^{n'} p_\nu$ crecen sin límite, en tanto que el término $\sum_{\nu=1}^{n''} q_\nu$ se aproxima a un límite definido, de manera que la suma parcial s_n crece más allá de toda cota.

Se ve, por tanto, que *una serie condicionalmente convergente no puede considerarse como la diferencia de dos series convergentes, una de ellas formada por sus términos positivos y la otra por los valores absolutos de sus términos negativos.*

En estrecha relación con este hecho, existe otra diferencia entre las series absolutamente convergentes y las series condicionalmente convergentes, de la cual se hace en seguida una breve mención.

*c. Reordenamiento de términos

Las sumas finitas tienen la propiedad de que en ellas es posible cambiar el orden de los términos o, como se dice, reordenar a voluntad los términos, sin que por elo cambie e vaor de lla suma. Surgen en seguida las preguntas: ¿cuál es el significado exacto de un cambio en el orden de los términos de una serie infinita? ¿deja tal reordenamiento inalterado el valor de la suma? Aunque para sumas finitas no existe dificultad en sumar, por ejemplo, los términos en orden inverso, en el caso de una serie infinita tal posibilidad no existe; no hay un último término por el cual empezar. Ahora bien, un cambio de orden en una serie infinita puede sólo significar lo siguiente: se dice que una serie $a_1 + a_2 + a_3 + \cdots$ se transforma mediante un reordenamiento en una serie $b_1 + b_2 + b_3 + \cdots$, si todo término a_n de la primera serie aparece exactamente una vez en la segunda, y recíprocamente. Por ejemplo, a_n puede desplazarse al reordenar, según una cantidad que crezca más allá de toda cota al crecer n. Lo único que debe cumplirse es que a_n aparezca en *alguna parte* de la nueva serie. Si algunos de los términos son desplazados hacia lugares posteriores en la serie, otros términos deberán, por supuesto, ser desplazados hacia sitios anteriores Por ejemplo, la serie

$$1 + q + q^2 + q^4 + q^3 + q^8 + q^7 + q^6 + q^5 + q^{16} + \cdots$$

es un reordenamiento [1] de la serie geométrica $1 + q + q^2 + \cdots$.

[1] Para todo $n > 0$, los términos q^k con $2^n < k \leq 2^{n+1}$ se escriben en orden invertido.

Con respecto al cambio de orden, entre las series absolutamente convergentes y las series condicionalmente convergentes existe una diferencia fundamental:

En las series absolutamente convergentes, un reordenamiento de sus términos no afecta la convergencia y el valor de la suma de la serie queda inalterado, exactamente como en las sumas finitas.

En las series condicionalmente convergentes, por otra parte, el valor de la suma de la serie puede cambiarse a voluntad reordenando adecuadamente sus términos; más aún, por reordenamiento puede hacerse que la serie diverja.

El primero de estos hechos, que se refiere a series absolutamente convergentes, es fácil de establecer. Supóngase inicialmente que la serie en cuestión tiene solamente términos positivos, y considérese la n-ésima suma parcial, $s_n = \sum_{\nu=1}^{n} a_\nu$. Todos los términos de esta suma parcial aparecen en la m-ésima suma parcial $t_m = \sum_{\nu=1}^{m} b_\nu$ de la serie reordenada, con sólo escoger m suficientemente grande. Por tanto $t_m \geq s_n$. Por otra parte, es posible determinar un índice n' tan grande que la suma parcial $s_{n'} = \sum_{\nu=1}^{n'} a_\nu$ de la primera serie contenga todos los términos $b_1, b_2, \ldots, b_m$. Se deduce entonces que $t_m \leq s_{n'} \leq A$, donde A es la suma de la primera serie. Así, para todos los valores de m suficientemente grandes se tiene $s_n \leq t_m \leq A$, y puesto que s_n puede hacerse que difiera de A en una cantidad arbitrariamente pequeña, se concluye que la serie reordenada también converge y, precisamente, lo hace al mismo límite A que la serie original.

Si la serie absolutamente convergente tiene términos tanto positivos como negativos, puede considerarse, en realidad, como la diferencia de dos series formadas exclusivamente por términos positivos. Puesto que al reordenar los términos de la serie original cada una de esas dos series sufre simplemente un reordenamiento y, por consiguiente, converge al mismo valor que antes, la serie original, cuando se reordena, converge también al mismo valor. Pues, por el caso considerado antes, la nueva serie es absolutamente convergente y, por tanto, es la diferencia de las dos series reordenadas consistentes de términos positivos.

Al principiante puede parecer trivial el aserto recién demostrado. Sin embargo, puede mostrarse mediante un ejemplo acerca del comportamiento opuesto de las series condicionalmente convergentes que la demostración es realmente necesaria y que en ella es esencial la convergencia absoluta. Consideremos la ya conocida serie para log 2 y escribamos a continuación el resultado de multiplicarla por el factor $\frac{1}{2}$:

$$1 - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} + \frac{1}{5} - \frac{1}{6} + \frac{1}{7} - \frac{1}{8} + - \cdots = \log 2,$$

$$\frac{1}{2} \qquad - \frac{1}{4} \qquad + \frac{1}{6} \qquad - \frac{1}{8} + - \cdots = \frac{1}{2} \log 2\,;$$

y ahora sumemos, combinando los términos situados en las columnas verticales.[1] Resulta así

$$1 + \frac{1}{3} - \frac{1}{2} + \frac{1}{5} + \frac{1}{7} - \frac{1}{4} + \frac{1}{9} + \frac{1}{11} - \frac{1}{6} + + - \cdots = \frac{3}{2} \log 2.$$

Es claro que esta última serie puede obtenerse reordenando la serie original. Sin embargo, el valor de la suma de la serie ha sido multiplicado por el factor $\frac{3}{2}$. Es fácil imaginar la impresión que debió causar el descubrimiento de esta aparente paradoja a los matemáticos del siglo XVIII, quienes estaban acostumbrados a operar con series infinitas sin considerar cuestiones de convergencia.

[1] Para la adición de series véase sección 7.1*d*.

*Damos la demostración del teorema anterior que se refiere al cambio en la suma de una serie condicionalmente convergente, Σa_n, debido a un cambio de orden de sus términos, aunque no tendremos oportunidad de usar tal resultado. Sean $p_1, p_2, \ldots$ los términos positivos y $-q_1, -q_2, \ldots$ los términos negativos de la serie. Como el valor absoluto $|a_n|$ tiende a cero cuando n crece, los números p_n y q_n deben también tender a cero al crecer n. Además, como ya se ha visto, la suma $\sum_1^\infty p_\nu$ debe divergir, así como también $\sum_1^\infty q_\nu$.

Se puede encontrar ahora fácilmente un reordenamiento de la serie original que tenga como suma un número arbitrario a. Supóngase, por conveniencia, que a es positivo. Entonces, súmense los n_1 primeros términos necesarios para justamente lograr que la suma $\sum_1^{n_1} p_\nu$ sea mayor que a. Puesto que la suma $\sum_1^{n_1} p_\nu$ crece con n_1 más allá de toda cota, siempre es posible hacer que la suma parcial sea mayor que a tomando para ello suficientes términos. En estas condiciones, la suma diferirá del valor exacto a en a lo más p_{n_1}. Súmense ahora suficientes términos negativos $-\sum_1^{m_1} q_\nu$ para asegurar que la suma $\sum_1^{n_1} p_\nu - \sum_1^{m_1} q_\nu$ sea menor que a. Esto también es posible debido a la divergencia de la serie $\sum_1^\infty q_\nu$. La diferencia entre esta nueva suma y a es q_{m_1}, a lo sumo. A continuación agréguense otros términos positivos $\sum_{n_1+1}^{n_2} p_\nu$ para hacer que la suma parcial sea de nuevo ligeramente mayor que a, lo cual es posible ya que la serie de términos positivos diverge. La diferencia entre la suma parcial y a es a lo más p_{n_2}. Otra vez, súmense términos negativos $-\sum_{m_1+1}^{m_2} q_\nu$, comenzando con el término siguiente al último usado en el paso anterior, para hacer una vez más que la suma sea menor que a, y prosígase de esta manera. Los valores de las sumas así obtenidas oscilarán en torno al número a; y si el proceso se repite un número suficiente de veces, dichas oscilaciones ocurrirán entre cotas arbitrariamente próximas una de otra. En efecto, la longitud del intervalo en que tales oscilaciones tienen lugar tiende a O ya que los términos p_ν y q_ν mismos tienden a 0 cuando ν crece. En esta forma queda demostrado el teorema.

De manera análoga es posible reordenar la serie para hacerla divergente: sólo se necesita escoger de entre los términos positivos números tan grandes, comparados con los negativos, que la compensación ya no sea posible.

d. Operaciones con series infinitas

Es claro que dos series infinitas que convergen, $a_1 + a_2 + \cdots = S$ y $b_1 + b_2 + \cdots = T$, pueden *sumarse término a término,* esto es, que la serie formada por los términos $c_n = a_n + b_n$ converge y su suma tiene el valor $S + T$.[1] En efecto,

$$\sum_{\nu=1}^{n} c_\nu = \sum_{\nu=1}^{n} a_\nu + \sum_{\nu=1}^{n} b_\nu \to S + T.$$

[1] Este teorema, en realidad, no es sino otra expresión del hecho (ver capítulo 1, p. 95) de que el límite de la suma de dos términos es la suma de sus límites.

También es claro que si se multiplica cada término de una serie infinita, convergente, por el mismo factor, la serie sigue siendo convergente y su suma queda multiplicada por dicho factor.

En estas operaciones no importa si la convergencia es absoluta o condicional. Por otra parte, un estudio más a fondo muestra que la multiplicación de dos series infinitas mediante el método usado para multiplicar sumas finitas no necesariamente da como resultado una serie convergente, a no ser que por lo menos una de las dos series originales sea absolutamente convergente (ver Apéndice, p. 555).

7.2 Criterios de convergencia absoluta y de divergencia

En la sección 7.1*b* se encontró ya el útil criterio de Leibnitz para investigar la convergencia condicional de las series. En las páginas siguientes se considerarán exclusivamente criterios relativos a la convergencia *absoluta*.

a. El criterio de comparación. Mayorantes

Todas las consideraciones de este tipo relativas a la convergencia dependen de la comparación de la serie dada con una segunda serie. Esta se escoge de manera tal que su convergencia pueda investigarse rápidamente. El *criterio general de comparación* puede establecerse como sigue:

Si los números $b_1, b_2, \ldots$ *son todos positivos y la serie* $\sum_{\nu=1}^{\infty} b_\nu$ *converge, y si*

$$|a_n| \leq b_n$$

para todos los valores de n, entonces la serie $\sum_{n=1}^{\infty} a_n$ *es absolutamente convergente.*

Usando el criterio de Cauchy, la demostración se vuelve casi trivial. Pues si $m \geq n$, se tiene

$$|a_n + \cdots + a_m| \leq |a_n| + \cdots + |a_m| \leq b_n + \cdots + b_m.$$

Puesto que la serie $\sum_{n=1}^{\infty} b_n$ converge, el miembro de la derecha es arbitrariamente pequeño siempre que n y m sean suficientemente grandes. Como resultado de ello, se tiene que para esos valores de n y m el miembro de la izquierda es también arbitrariamente pequeño, y así, por el criterio de Cauchy, la serie dada converge. La convergencia es absoluta pues nuestro argumento es aplicable también a la convergencia de la serie de los valores absolutos $|a_n|$.

La demostración (análoga) del teorema siguiente puede dejarse al lector. *Si*

$$|a_n| \geq b_n > 0,$$

y la serie $\sum_{n=1}^{\infty} b_n$ *diverge, entonces la serie* $\sum_{n=1}^{\infty} a_n$ *no es absolutamente convergente.*

Algunas veces las series anteriores formadas por los términos positivos b_n se llaman, respectivamente, serie *mayorante* y serie *minorante* de la formada por los términos a_n.

b. Investigación de la convergencia mediante la comparación con la serie geométrica

En las aplicaciones del criterio anterior la serie de comparación más frecuentemente usada como mayorante es la serie geométrica. De inmediato se obtiene el siguiente teorema.

TEOREMA. *La serie* $\sum_{n=1}^{\infty} a_n$ *es absolutamente convergente si a partir de cierto término se cumple una relación de la forma*

$$|a_n| < cq^n, \tag{3}$$

donde c es un número positivo independiente de n y q es cualquier número positivo fijo menor que 1.

Criterios del cociente y de la raíz. Estos criterios se expresan generalmente en una de las formas siguientes: la serie $\sum_{n=1}^{\infty} a_n$ converge absolutamente si a partir de cierto término se cumple una relación de la forma

$$\left|\frac{a_{n+1}}{a_n}\right| < q, \tag{4a}$$

donde q es un número positivo menor que 1 e independiente de n, o sea, si a partir de cierto término se cumple una relación de la forma

$$\sqrt[n]{|a_n|} < q \tag{4b}$$

donde q es un número positivo menor que 1. En particular, las condiciones anteriores se satisfacen si una de las relaciones

$$\lim_{n \to \infty} \left|\frac{a_{n+1}}{a_n}\right| = k < 1 \tag{5a}$$

o

$$(5b) \qquad \lim_{n\to\infty} \sqrt[n]{|a_n|} = k < 1$$

es cierta.

Estas afirmaciones se fundamentan fácilmente de la siguiente manera.

Supóngase que (4*a*), el *criterio del cociente,* se satisface a partir de cierto subíndice n_0, esto es, para $n > n_0$. Por brevedad, hágase $a_{n_0+m+1} = b_m$. Se encuentra entonces que

$$|b_1| < q\,|b_0|, \qquad |b_2| < q\,|b_1| < q^2\,|b_0|, \qquad |b_3| < q\,|b_2| < q^3\,|b_0|,$$

y así sucesivamente. Por tanto,

$$|b_m| < q^m\,|b_0|,$$

y si $n > n_0$ y $c = q^{-n_0-1}\,|b_0|$,

$$|a_n| = |b_{n-n_0-1}| < q^{n-n_0-1}\,|b_0| = cq^n,$$

lo que prueba nuestra afirmación. Para el criterio (4*b*), el *criterio de la raíz,* se tiene de inmediato que $|a_n| < q^n$, de lo cual se sigue su veracidad.

Finalmente, para establecer el criterio (5) considérese un número arbitrario, q, tal que $k < q < 1$. Entonces a partir de cierto n_0, esto es, cuando $n > n_0$, las Ecs. 4*a, b* implican que $\left|\frac{a_{n+1}}{a_n}\right| < q$ y $\sqrt[n]{|a_n|} < q$, respectivamente, pues a partir de cierto término los valores de $\left|\frac{a_{n+1}}{a_n}\right|$ o de $\sqrt[n]{|a_n|}$ difieren de k en menos de $(q - k)$. El criterio queda entonces establecido sobre la base de los resultados demostrados previamente.

Debe hacerse notar que los cuatro criterios, 4*a, b,* 5*a, b,* obtenidos a partir del criterio original $|a_n| < cq^n$, *no* son equivalentes entre sí, ni tampoco equivalentes al original, es decir, no pueden ser derivados unos de otros en ambos sentidos. Pronto veremos, mediante ejemplos, que si una serie satisface una de las condiciones, no necesariamente satisface todas las demás.

Para completar lo anterior, puede señalarse que una serie será, en efecto, divergente si a partir de cierto término se tiene

$$|a_n| > c$$

para algún número positivo c, o bien, si a partir de cierto término

$$\sqrt[n]{|a_n|} > 1,$$

o también si $\lim_{n\to\infty} \left|\frac{a_{n+1}}{a_n}\right| = k$, o $\lim_{n\to\infty} \sqrt[n]{|a_n|} = k$,

donde k es un número mayor que 1. En efecto, como es fácil ver, en tal serie los términos no pueden tender a cero al crecer n. La serie, por tanto, debe divergir. (En estas condiciones, la serie ni siquiera puede ser condicionalmente convergente.)

Nuestros criterios proporcionan condiciones *suficientes* para la convergencia absoluta de una serie; es decir, cuando se satisfacen puede concluirse que la serie converge absolutamente. Definitivamente, sin embargo, dichas condiciones *no* son *necesarias;* esto es, pueden construirse series absolutamente convergentes que no satisfagan esas condiciones.

Por lo anterior, el saber que

$$\lim_{n\to\infty} \left|\frac{a_{n+1}}{a_n}\right| = 1 \qquad \text{o} \qquad \lim_{n\to\infty} \sqrt[n]{|a_n|} = 1$$

nada implica acerca de la convergencia de la serie; ella puede ser convergente o divergente. Por ejemplo, la serie

$$\sum_{n\to\infty}^{\infty} \frac{1}{n},$$

para la cual $\lim_{n\to\infty} \sqrt[n]{|a_n|} = 1$ y $\lim_{n\to\infty} \left|\frac{a_{n+1}}{a_n}\right| = 1$, es divergente, como se vió en la p. 532. Por otra parte, más adelante se verá que la serie $\sum_{n=1}^{\infty} \frac{1}{n^2}$; que satisface las mismas condiciones y es convergente.

Como un ejemplo de la aplicación de nuestros criterios, consideremos la serie

$$q + 2q^2 + 3q^3 + \cdots + nq^n + \cdots.$$

Para ella se tiene

$$\lim_{n\to\infty} \sqrt[n]{|a_n|} = |q| \cdot \lim_{n\to\infty} \sqrt[n]{n} = |q|,$$

$$\lim_{n\to\infty} \left|\frac{a_{n+1}}{a_n}\right| = |q| \cdot \lim_{n\to\infty} \frac{n+1}{n} = |q|.$$

La convergencia de la serie cuando $|q| < 1$ se deduce del criterio del cociente y también del criterio de la raíz, aún en la forma más débil dada por (5).

Si, por otra parte, consideramos la serie

$$1 + 2q + q^2 + 2q^3 + \cdots + q^{2n} + 2q^{2n+1} + \cdots,$$

no podremos investigar la convergencia mediante el criterio del cociente cuando $\frac{1}{2} \le |q| < 1$, pues en ese caso $\left|\frac{2q^{2n+1}}{q^{2n}}\right| = 2\,|q| \ge 1$. Sin embargo, el criterio de la raíz inmediatamente dará $\lim_{n\to\infty} \sqrt[n]{|a_n|} = |q|$, mostrando así que la serie converge cuando $|q| < 1$, cosa que, por supuesto, podría también observarse directamente.

c. Comparación con una integral [1]

Procedemos ahora a discutir un método completamente diferente para el estudio de la convergencia. Lo explicaremos para el caso típico, particularmente simple e importante, de la serie

$$\sum_{n=1}^{\infty} \frac{1}{n^\alpha} = 1 + \frac{1}{2^\alpha} + \frac{1}{3^\alpha} + \cdots,$$

en la que el término general, a_n, es $1/n$, siendo α un número positivo. Para investigar la convergencia o divergencia de esta serie, considérese la gráfica de la función $y = 1/x^\alpha$ y márquense sobre el eje x las abscisas enteras $x = 1$, $x = 2$, ... Para empezar, constrúyase un rectángulo de altura $1/n^\alpha$ sobre el intervalo $n - 1 \leq x \leq n$ del eje x (siendo $n > 1$), y compárese su área con el área de la región limitada por el mismo intervalo, las ordenadas en los extremos y la curva $y = 1/x^\alpha$ (esta región aparece sombreada en la Fig. 7.2). A continuación constrúyase el rectángulo de altura $1/n^\alpha$

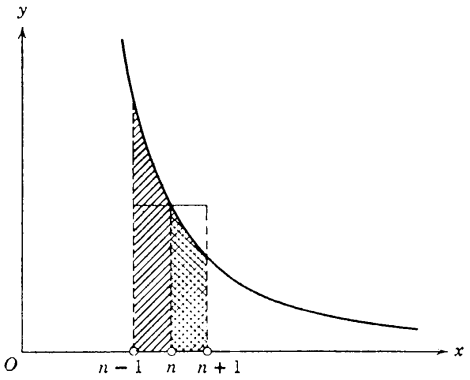


Figura 7.2 Comparación de una serie con una integral.

sobre el intervalo $n \leq x \leq n + 1$, y, de igual forma, compárese con el área de la región que está por encima del intervalo y por debajo de la curva (esta región está indicada en la Fig. 7.2 por el sombreado en cruz). En el primer caso el área bajo la curva es, evidentemente, mayor que el área del rectángulo; en el segundo, es menor. En otras palabras,

$$\int_n^{n+1} \frac{dx}{x^\alpha} < \frac{1}{n^\alpha} < \int_{n-1}^{n} \frac{dx}{x^\alpha}.$$

[1] En relación con esto, véase también el apéndice al capítulo 5, p. 523.

Escribiendo en seguida estas desigualdades para $n = 1, 2, 3, \ldots, m$, y $n = 2, 3, \ldots, m$, respectivamente, y sumando, se obtiene la siguiente estimación para la m-ésima suma parcial $s_m = \sum_{n=1}^{m} \frac{1}{n^\alpha}$:

$$\int_1^{m+1} \frac{dx}{x^\alpha} < s_m < 1 + \int_1^m \frac{dx}{x^\alpha}. \tag{6}$$

Ahora bien, cuando n crece la integral $\int_1^m \frac{1}{x^\alpha}\,dx$ tiende a un límite finito o bien crece sin límites según que $\alpha > 1$ o $\alpha \leq 1$. En consecuencia, la sucesión monótona de números s_m es acotada o crece más allá de toda cota según que $\alpha > 1$ o $\alpha \leq 1$. En esta forma se tiene el siguiente teorema.

TEOREMA. *La serie de potencias recíprocas,*

$$\sum_{n=1}^{\infty} \frac{1}{n^\alpha} = \frac{1}{1^\alpha} + \frac{1}{2^\alpha} + \frac{1}{3^\alpha} + \cdots,$$

es convergente si, y sólo si, $\alpha > 1$.

Para $\alpha = 1$ resulta, como una consecuencia inmediata, la divergencia de la serie armónica, hecho establecido antes de manera distinta. De manera análoga, las series

$$\frac{1}{1^2} + \frac{1}{2^2} + \frac{1}{3^2} + \cdots,$$

$$\frac{1}{1^3} + \frac{1}{2^3} + \frac{1}{3^3} + \cdots,$$

convergen, en tanto que la serie $\frac{1}{\sqrt{1}} + \frac{1}{\sqrt{2}} + \frac{1}{\sqrt{3}} + \cdots$ diverge.

La serie convergente $\sum_{\nu=1}^{\infty} \frac{1}{\nu^\alpha}$, con $\alpha > 1$, sirve a menudo como una serie de comparación en la investigación de la convergencia. Por ejemplo, se ve de inmediato que para $\alpha > 1$ la serie $\sum_{\nu=1}^{\infty} \frac{c_\nu}{\nu^\alpha}$ converge absolutamente si los valores absolutos, $|c_\nu|$, de los coeficientes permanecen por debajo de una cota fija independiente de ν.

Constante de Euler. De la estimación (6) con $\alpha = 1$ se deduce en seguida que la sucesión de números $C_n = 1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{n} - \log n = s_n - \log n > \log (n + 1) - \log n > 0$ está acotada inferiormente. Puesto que, por la desigualdad $\frac{1}{n+1} < \int_n^{n+1} \frac{dx}{x} = \log (n + 1) - \log n = \frac{1}{n+1} +$

$C_n - C_{n+1}$, la sucesión es monótona decreciente, debe tender a un límite:

$$\lim_{n\to\infty} C_n = \lim_{n\to\infty} \left(1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{n} - \log n \right) = C.$$

El número C, cuyo valor es 0.5772..., se llama *constante de Euler*. En contraste con otros números especiales e importantes del Análisis, tales como π y e, no se ha encontrado para la constante de Euler otra expresión con una ley de formación sencilla. Hasta la fecha no se sabe si C es racional o irracional.

7.3 Sucesiones de funciones

Como ya se ha hecho resaltar muchas veces, el proceso de límite sirve no sólo para representar números conocidos en términos de otros más simples, sino también para extender el conjunto de números conocidos a otro más amplio. De importancia decisiva en el Análisis es el estudio de los límites, no solamente para sucesiones —o series infinitas— de números constantes, sino, de manera similar, para sucesiones de funciones o series cuyos términos son funciones de una variable x. Este es el caso, por ejemplo, de la serie de Taylor y de las series de potencias en general. Tales procesos de límite se requieren lo mismo para aproximar funciones dadas por otras más simples que para definir y describir analíticamente nuevas funciones, lo que a menudo se realiza sobre la base del concepto de límite para sucesiones de funciones: $f(x) = \lim f_n(x)$ para $n \to \infty$. En forma equivalente, es posible considerar $f(x)$ como la suma y las $f_n(x)$ como las sumas parciales de una serie infinita $f(x) = \sum_{r=1}^{\infty} g_r(x)$ de funciones $g_n(x)$, donde $g_n(x) = f_n(x) - f_{n-1}(x)$ para $n > 1$ y $g_1(x) = f_1(x)$.

Discutiremos a continuación las definiciones precisas y las interpretaciones geométricas.

a. Procesos de límite con funciones y curvas

Definición. La sucesión $f_1(x), f_2(x), \ldots$ en el intervalo $a \leq x \leq b$ converge a la función límite $f(x)$ si en cada punto x de ese intervalo los valores $f_n(x)$ convergen, en el sentido usual, al valor $f(x)$. En este caso se escribe $\lim_{n\to\infty} f_n(x) = f(x)$. De acuerdo con el criterio de Cauchy (ver p. 98), la convergencia de la sucesión puede expresarse sin hacer referencia a la función límite: la sucesión de funciones converge a una función límite si, y sólo si, en cada punto x del intervalo en cuestión y para cada número positivo ε, la cantidad $|f_n(x) - f_m(x)|$ es menor que ε, siempre que n y m se tomen suficientemente grandes, es decir, mayores que cierto número. Este número, $N = N(\varepsilon, x)$, depende por lo general de ε y de x y crece más allá de toda cota cuando ε tiende a cero.

Hemos encontrado ya algunos casos de límites de sucesiones de funciones. Mencionamos aquí solamente la definición de la potencia x^α para valores irracionales α mediante la ecuación

$$x^\alpha = \lim_{n\to\infty} x^{r_n},$$

donde $r_1, r_2, \ldots, r_n, \ldots$ es una sucesión de números racionales que tiende a α; y la ecuación

$$e^x = \lim_{n\to\infty} \left(1 + \frac{x}{n} \right)^n,$$

en la que las funciones de aproximación, $f_n(x)$, en el miembro derecho son polinomios de grado n.

La representación gráfica de funciones por medio de curvas sugiere la posibilidad de hablar de límites de sucesiones de curvas, diciendo, por ejemplo, que las gráficas de las anteriores funciones límite, x^α y e^x, pueden considerarse como las curvas límite de las gráficas de las funciones x^{r_n} y $\left(1 + \frac{x}{n} \right)^n$ respectivamente.

Existe, sin embargo, una diferencia sutil entre el paso al límite con funciones y el paso correspondiente cuando se trata de curvas, la cual no fue vista con claridad sino hasta mediados del siglo XIX. Ilustraremos este punto con un ejemplo y después, en la siguiente sección, lo discutiremos en forma sistemática.

Considérense las funciones

$$f_n(x) = x^n, \qquad n = 1, 2, \ldots,$$

en el intervalo $0 \leq x \leq 1$. Todas estas funciones son continuas y la función límite $\lim_{n\to\infty} f_n(x) = f(x)$ existe. Sin embargo, esta función límite no es continua. Por el contrario, como $f_n(1) = 1$ para todos los valores de n, se tiene el límite

$$f(1) = 1;$$

y, por otra parte, para $0 \leq x < 1$ el límite es $f(x) = \lim_{n\to\infty} f_n(x) = 0$, como se vio en el capítulo 1, p. 88. Por consiguiente, $f(x)$ es una función discontinua que toma el valor 1 en $x = 1$, en tanto que es igual a cero para todo otro valor de x en el intervalo.

Esta discontinuidad queda ilustrada geométricamente por las gráficas C_n de las funciones $y = f_n(x)$. Son éstas (ver Fig. 1.44, p. 89) curvas continuas que pasan por el origen y por el punto $x = 1$, $y = 1$ y que se aproxi-

man más y más al eje x cuando n crece. Las *curvas* tienen una *curva límite* C que no es discontinua sino que consiste (ver Fig. 7.3) de la porción del eje x comprendida entre $x = 0$ y $x = 1$ y la porción de la recta $x = 1$ comprendida entre $y = 0$ e $y = 1$. Las *curvas* convergen, por consiguiente, a una curva límite *continua* que tiene una porción vertical, en tanto que las *funciones* convergen a una función límite *discontinua*. Se ve

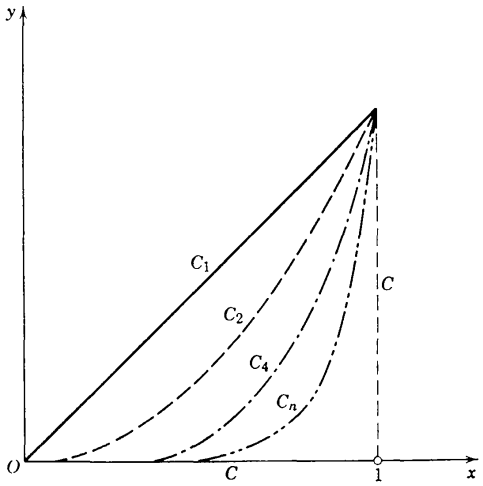


Figura 7.3 Curva límite y función límite.

entonces que esta discontinuidad de la función límite queda expresada por la aparición en la curva límite de una porción perpendicular al eje x. Esta curva límite *no* es la gráfica de la función límite; pues al valor de x para el cual se tiene la porción vertical, la curva hace corresponder un número infinito de valores de y, en tanto que la función le asocia sólo uno. Por consiguiente, el límite de las gráficas de las funciones $f_n(x)$ no es el mismo que la gráfica del límite, $f(x)$, de estas funciones.

7.4 Convergencia uniforme y convergencia no uniforme

a. Comentarios generales y definiciones

La distinción entre los conceptos de convergencia de funciones y de convergencia de curvas es un fenómeno que el estudiante debe comprender con claridad. Involucra la llamada *convergencia no uniforme* de sucesiones o series infinitas de funciones, que será discutida con algún detalle.

Que una función $f(x)$ es el límite de una sucesión $f_1(x)$, $f_2(x)$, ... en un intervalo $a \leq x \leq b$ significa, simplemente por definición, que la relación de límite usual, $f(x) = \lim_{n\to\infty} f_n(x)$, se satisface *en cada punto* del intervalo.

Tal convergencia es una *propiedad local*, en el punto x, de la sucesión. Es natural, sin embargo, pedir algo más que la sola convergencia local de dichas aproximaciones. Se requiere también que si se asigna una medida arbitraria, ε, para la aproximación, entonces, a partir de cierto índice N *todas* las funciones $f_n(x)$ estén comprendidas entre $f(x) - \varepsilon$ y $f(x) + \varepsilon$, para *todos* los valores de x, de manera que sus gráficas $y = f_n(x)$ queden completamente dentro de la franja mostrada en la Fig. 7.4. Se dice que la

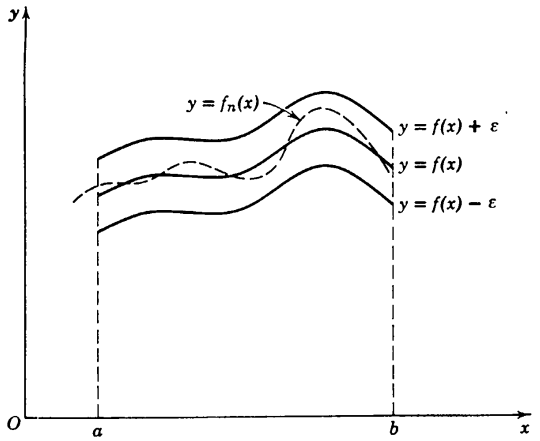


Figura 7.4 Para ilustrar la convergencia uniforme.

aproximación es *uniforme* [1] si su exactitud puede hacerse por lo menos igual a un número positivo prefijado ε, simultáneamente en todos los puntos del intervalo, es decir, escogiendo en todas partes el mismo número $N(\varepsilon)$ independiente de x. Si $\lim_{n\to\infty} f_n(x) = f(x)$ *uniformemente* para $a \leq x \leq b$, existe para todo $\varepsilon > 0$ un correspondiente $N = N(\varepsilon)$ tal que $|f(x) - f_n(x)| < \varepsilon$ para todo $n > N$ y todo x en el intervalo. Mucha gente quedó sorprendida cuando, a mediados del siglo XIX, Seidel y otros observaron que la convergencia de funciones no siempre es uniforme, como ingenuamente se había supuesto.

[1] Compárese con la definición análoga que se dio cuando vimos *continuidad uniforme*, p. 65, según la cual es posible escoger el mismo número $\delta(\varepsilon)$ independientemente de x.

Ejemplos de convergencia no uniforme. El concepto de convergencia uniforme resulta aclarado con algunos ejemplos sobre convergencia no uniforme.

(*a*) El primer ejemplo lo constituye la sucesión de funciones que se acaba de considerar, $f_n(x) = x^n$. En el intervalo $0 \leq x \leq 1$ esta sucesión converge a la función límite $f(x) = 0$ para $0 \leq x < 1$, $f(1) = 1$. La convergencia se presenta en cada punto del intervalo, es decir, si ε es cualquier número positivo y si se escoge cualquier valor definido y fijo $x = \xi$, entonces la desigualdad $|\xi^n - f(\xi)| < \varepsilon$ se cumple, en efecto, si n es suficientemente grande. Sin embargo, esta aproximación no es uniforme, pues si se toma $\varepsilon = \frac{1}{2}$ puede encontrarse un punto $x = \eta \neq 1$ para el cual $|\eta^n - f(\eta)| = \eta^n > \frac{1}{2}$, sin importar cuán grande se escoja el número n. En efecto, esto es cierto para todos los puntos $x = \eta$ con $1 > \eta > \sqrt[n]{\frac{1}{2}}$. Por consiguiente, es imposible escoger el número n tan grande que la diferencia entre $f(x)$ y $f_n(x)$ sea menor que $\frac{1}{2}$ *en todo el intervalo.*

Este comportamiento se hace comprensible si se observan las gráficas de estas funciones (Fig. 7.3). Por ellas es claro que, no importa cuán grande se escoja el valor de n, para valores de ξ ligeramente menores que 1 el valor de la función $f_n(\xi)$ estará muy próximo a 1 y, por consiguiente, no podrá ser una buena aproximación para $f(\xi)$, que es 0.

Un comportamiento análogo es exhibido por las funciones

$$f_n(x) = \frac{1}{1 + x^{2n}}$$

en la vecindad de los puntos $x = 1$ y $x = -1$. Esto puede ser establecido fácilmente. Aquí $f(x) = 1$ para $|x| < 1$, $f(x) = \frac{1}{2}$ para $x = 1$ y $f(x) = 0$ para $|x| > 1$.

(*b*) En los dos ejemplos anteriores la no uniformidad de la convergencia tiene relación con el hecho de que *la función límite es discontinua.* Sin embargo, es también fácil construir una sucesión de funciones continuas convergentes a una función límite continua pero sin que la convergencia sea uniforme. Enfoquemos nuestra atención al intervalo $0 \leq x \leq 1$ y hagamos las siguientes definiciones para $n > 2$:

$$f_n(x) = xn^\alpha \quad \text{para} \quad 0 \leq x \leq \frac{1}{n},$$

$$f_n(x) = \left(\frac{2}{n} - x\right) n^\alpha \quad \text{para} \quad \frac{1}{n} \leq x \leq \frac{2}{n},$$

$$f_n(x) = 0 \quad \text{para} \quad \frac{2}{n} \leq x \leq 1,$$

en donde, para comenzar, puede darse a α cualquier valor; pero, una vez hecho esto, debe mantenerse fijo tal valor de α para todos los términos de la sucesión. Gráficamente, nuestras funciones están representadas por una figura en forma de V invertida, consistente en dos segmentos de recta definidos en el intervalo $0 \leq x \leq 2/n$ del eje x, en tanto que de $x = 2/n$ en adelante la gráfica es el eje x mismo (ver Fig. 7.5).

Si $\alpha < 1$, la ordenada del punto más alto de la gráfica, que en general tiene el valor $n^{\alpha-1}$, tenderá a cero al crecer n; y la curva tenderá entonces a confundirse con el eje x, y las funciones $f_n(x)$ convergirán uniformemente a la función límite $f(x) = 0$.

Si $\alpha = 1$, el pico de la gráfica tendrá altura 1 para todo valor de n. Si $\alpha > 1$, la altura del pico crecerá más allá de toda cota al crecer n.

No obstante, sin importar cómo se escoja α, la sucesión $f_1(x), f_2(x), \ldots$ siempre tiende a la función límite $f(x) = 0$. En efecto, si x es positivo se tiene $2/n < x$ para todos los valores de n suficientemente grandes, de manera que x no está debajo de la parte en forma de V invertida de la gráfica, y $f_n(x) = 0$. Para $x = 0$ todos los valores funcionales $f_n(x)$ son iguales a 0, de manera que en ambos casos

$$\lim_{n\to\infty} f_n(x) = 0.$$

Sin embargo, si $\alpha \geq 1$ la convergencia ciertamente es no uniforme, pues resulta entonces imposible escoger n tan grande que la expresión $|f(x) - f_n(x)| = f_n(x)$ sea menor que $\frac{1}{2}$ en *todo* el intervalo.

(*c*) Un comportamiento exactamente análogo es exhibido por la sucesión de funciones

$$f_n(x) = xn^\alpha e^{-nx},$$

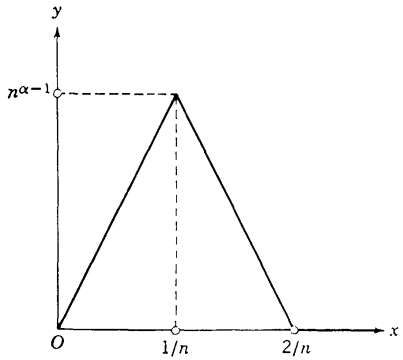


Figura 7.5 Para ilustrar la convergencia no uniforme.

donde, en contraste con el caso anterior, cada función de la sucesión está representada por una única expresión analítica. Aquí, otra vez, la ecuación $\lim_{n\to\infty} f_n(x) = 0$ se cumple para todo valor positivo de x, pues cuando n crece la función e^{-nx} tiende a cero con un orden mayor que el de cualquier potencia de $1/n$ (ver sección 3.7*b*, p. 271). Para $x = 0$ se tiene siempre $f_n(x) = 0$, y así

$$f(x) = \lim_{n\to\infty} f_n(x) = 0$$

para todo valor de x en el intervalo $0 \leq x \leq a$, donde a es un número positivo arbitrario. Sin embargo, aquí la convergencia a la función límite es no uniforme. Pues en el punto $x = 1/n$, en el que $f_n(x)$ alcanza su máximo, se tiene

$$f_n(x) = f_n\left(\frac{1}{n}\right) = \frac{n^{\alpha-1}}{e},$$

de lo cual se sigue que si $\alpha \geq 1$ la convergencia es no uniforme, pues toda curva $y = f_n(x)$, sin importar cuán grande se escoja n, contendrá puntos (a saber, el punto $x = 1/n$, que varía con n) en los cuales $f_n(x) - f(x) = f_n(x) > 1/2e$ (ver Fig. 7.6).

(*d*) Los conceptos de convergencia uniforme y no uniforme pueden, por supuesto, extenderse a una serie infinita. Se dice que una serie

$$g_1(x) + g_2(x) + \cdots$$

es o no uniformemente convergente según sea el comportamiento de sus sumas parciales $f_n(x)$. Un ejemplo muy sencillo de una serie cuya convergencia es no uniforme está dado por

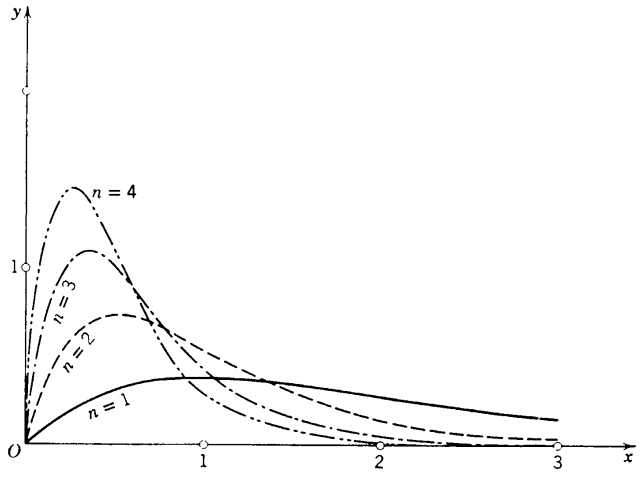


Figura 7.6 Convergencia no uniforme de la sucesión $f_n(x) = n^2xe^{-nx}$.

$$f(x) = x^2 + \frac{x^2}{1+x^2} + \frac{x^2}{(1+x^2)^2} + \frac{x^2}{(1+x^2)^3} + \cdots.$$

Para $x = 0$ toda suma parcial $f_n(x) = x^2 + \cdots + x^2/(1+x^2)^{n-1}$ tiene el valor 0; y, por tanto, $f(0) = 0$. Para $x \neq 0$ la serie es simplemente una serie geométrica de cociente positivo $1/(1+x^2) < 1$; y puede, por consiguiente, sumarse usando las reglas elementales, y así se obtiene para todo $x \neq 0$ la suma

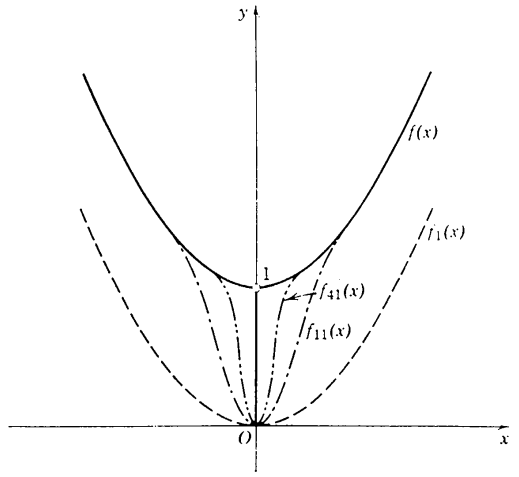


Figura 7.7 Convergencia hacia una función que tiene una discontinuidad por salto eliminable.

$$\frac{x^2}{1 - 1/(1+x^2)} = 1 + x^2.$$

La función límite $f(x)$ está dada entonces en todas partes excepto en $x = 0$ por la expresión $f(x) = 1 + x^2$, en tanto que $f(0) = 0$. Ella tiene, por consiguiente, una discontinuidad eliminable en el origen.

Una vez más, se tiene aquí convergencia no uniforme en todo intervalo que contiene al origen. Pues la diferencia $f(x) - f_n(x) = r_n(x)$ es siempre igual a 0 para $x = 0$, en tanto que está dada por la expresión $r_n(x) = 1/(1+x^2)^{n-1}$ para todos los demás valores de x, como puede verificar fácilmente el lector. Si se pide que esta expresión sea menor que, digamos, $\frac{1}{2}$, entonces para cada valor fijo de x esto puede conseguirse escogiendo n suficientemente grande. Pero no puede encontrarse un valor de n lo suficientemente grande para asegurar que $r_n(x)$ sea en todas partes menor que $\frac{1}{2}$; pues si se escoge cualquier valor de n, no importa cuán grande, puede hacerse $r_n(x)$ mayor que $\frac{1}{2}$ con sólo tomar x suficientemente próximo a 0. Por tanto, es imposible una aproximación uniforme con diferencias menores que $\frac{1}{2}$. El asunto se aclara si se consideran las curvas de aproximación (ver Fig. 7.7). Estas curvas, excepto cerca de $x = 0$, quedan cada vez más próximas a la parábola $y = 1 + x^2$ cuando n crece. Cerca de $x = 0$, sin embargo, de las curvas se deriva una protuberancia cada vez más estrecha que se extiende hacia el origen y que, cuando n crece, se aproxima más y más a un cierto segmento de recta, una porción del eje y, de manera que como curva límite se tiene la parábola más una extensión lineal que baja verticalmente hasta el origen.

Como un ejemplo adicional de convergencia no uniforme mencionamos la serie $\sum_{\nu=0}^{\infty} g_\nu(x)$, donde $g_\nu(x) = x^\nu - x^{\nu-1}$ para $\nu \geq 1$, $g_0(x) = 1$, definida en el intervalo $0 \leq x \leq 1$. Las sumas parciales de esta serie son las funciones x^ν ya consideradas en el ejemplo (*a*), p. 549.

b. Un criterio de convergencia uniforme

Las consideraciones anteriores nos muestran que la convergencia uniforme de una sucesión o de una serie es una propiedad especial, que no todas las sucesiones o todas las series tienen. Repetimos ahora la definición de convergencia uniforme según se aplica a las series infinitas: la serie

$$g_1(x) + g_2(x) + \cdots$$

converge uniformemente a una función $f(x)$ en un intervalo si $f(x)$ puede ser aproximada dentro de un margen de aproximación ε (donde ε es un número positivo arbitrariamente pequeño) por la suma de un número suficientemente grande, pero fijo, de términos, $g_1(x) + \cdots + g_N(x) = f_N(x)$, independientemente del punto x del intervalo.

Ahora tenemos un criterio (criterio de Cauchy) para la convergencia uniforme, en el cual no es necesario conocer la función límite $f(x)$: la serie converge uniformemente (o, lo que es lo mismo, la sucesión de funciones $f_n(x)$ converge uniformemente) si, y sólo si, la diferencia $|f_n(x) - f_m(x)|$ puede hacerse, en todo el intervalo, menor que una cantidad positiva arbitraria ε escogiendo n y m mayores que un número N independiente de x. Pues, en primer lugar, si la convergencia es uniforme podemos hacer tanto

$|f_n(x) - f(x)|$ como $|f_m(x) - f(x)|$ menores que $\varepsilon/2$ con sólo escoger n y m mayores que un cierto número N independiente de x; de lo cual se sigue que $|f_n(x) - f_m(x)| < \varepsilon$. Y, en segundo lugar, si $|f_n(x) - f_m(x)| < \varepsilon$ para todo x, y siempre que n y m sean mayores que N, entonces escogiendo cualquier valor fijo de $n > N$ y dejando que m crezca más allá de toda cota se obtiene la relación

$$|f_n(x) - f(x)| = \lim_{m\to\infty} |f_n(x) - f_m(x)| \leq \varepsilon,$$

para todo valor de x, de manera que la convergencia es uniforme.

Como veremos, es precisamente esta condición de convergencia uniforme la que hace que las series infinitas y otros procesos de límite sean herramientas útiles y convenientes en el Análisis. Afortunadamente, en los procesos de límite que, por lo general, se encuentran en el Análisis y sus aplicaciones, la convergencia no uniforme aparece sólo en puntos aislados excepcionales y, por ahora, apenas nos ocasionará algunas dificultades.

Por lo general, la uniformidad de la convergencia de una serie queda establecida mediante el siguiente criterio (comparación de la serie con una serie mayorante de términos constantes):

Si los términos de la serie $\sum_{\nu=1}^{\infty} g_\nu(x)$ *satisfacen la condición* $|g_\nu(x)| \leq a_\nu$, *donde los números* a_ν *son constantes positivas que forman una serie convergente* $\sum_{\nu=1}^{\infty} a_\nu$, entonces la serie $\sum_{\nu=1}^{\infty} g_\nu(x)$ *converge uniformemente (y absolutamente).*

En efecto, se tiene entonces que

$$\left| \sum_{\nu=n}^{m} g_\nu(x) \right| \leq \sum_{\nu=n}^{m} |g_\nu(x)| \leq \sum_{\nu=n}^{m} a_\nu,$$

y puesto que, por el criterio de Cauchy, la suma $\sum_{\nu=n}^{m} a_\nu$ puede hacerse arbitrariamente pequeña con sólo escoger n y $m > n$ suficientemente grandes, esto expresa exactamente la condición necesaria y suficiente para la convergencia uniforme.

Un primer ejemplo lo ofrece la serie geométrica $1 + x + x^2 + \cdots$, con x restringido al intervalo $|x| \leq q$, siendo q cualquier número positivo menor que 1. Los términos de la serie son entonces numéricamente menores que o iguales a los términos de la serie geométrica convergente Σq^ν.

Un segundo ejemplo lo constituye la "serie trigonométrica"

$$\frac{c_1 \operatorname{sen}(x - \delta_1)}{1^2} + \frac{c_2 \operatorname{sen}(x - \delta_2)}{2^2} + \frac{c_3 \operatorname{sen}(x - \delta_3)}{3^2} + \cdots,$$

siempre que $|c_n| < c$, donde c es una constante positiva independiente de n. En efecto, en tal caso se tiene

$$g_n(x) = \frac{c_n \operatorname{sen}(x - \delta_n)}{n^2}, \quad \text{de manera que} \quad |g_n(x)| < \frac{c}{n^2}.$$

Por tanto, la convergencia uniforme y absoluta de la serie trigonométrica se deduce de la convergencia de la serie $\sum_{\nu=1}^{\infty} \frac{c}{n^2}$.

c. Continuidad de la suma de una serie uniformemente convergente de funciones continuas

La importancia de la convergencia uniforme radica en el hecho de que una serie uniformemente convergente se comporta, en muchos aspectos, exactamente como la suma de un número finito de funciones. Así, por ejemplo, la suma de un número finito de funciones continuas es a su vez continua. Conforme a esto, se tiene el siguiente teorema.

TEOREMA. *Si una serie de términos continuos converge uniformemente en un intervalo, su suma es también una función continua.*

DEMOSTRACIÓN. La demostración es muy sencilla. La serie

$$f(x) = g_1(x) + g_2(x) + \cdots$$

se subdivide en la n-ésima suma parcial $f_n(x)$ más el residuo $R_n(x)$. Como ya es costumbre, $f_n(x) = g_1(x) + \cdots + g_n(x)$. Si ahora se asigna cualquier número positivo ε, en virtud de la convergencia uniforme es posible escoger el número n tan grande que el residuo sea menor que $\varepsilon/4$ en todo el intervalo, y, por tanto,

$$|R_n(x + h) - R_n(x)| < \frac{\varepsilon}{2}$$

para cualquier par de números x y $x + h$ del intervalo. La suma parcial $f_n(x)$ consiste en la suma de un número finito de funciones continuas y es, por consiguiente, continua. Por tanto, para cada punto x del intervalo es posible escoger un δ positivo tan pequeño que

$$|f_n(x + h) - f_n(x)| < \frac{\varepsilon}{2}$$

siempre que $|h| < \delta$ y los puntos x y $x + h$ estén en el intervalo. Se deduce entonces que

$$\begin{aligned} |f(x + h) - f(x)| &= |f_n(x + h) - f_n(x) + R_n(x + h) - R_n(x)| \\ &\leq |f_n(x + h) - f_n(x)| + |R_n(x + h) - R_n(x)| < \varepsilon, \end{aligned}$$

lo cual expresa la continuidad de la función.

La importancia de este teorema se pone de relieve cuando recordamos, por nuestros ejemplos anteriores, que las sumas de series convergentes por lo

general no son uniformemente discontinuas. Del teorema anterior podemos concluir que: si la suma de una serie convergente de funciones continuas tiene un punto de discontinuidad, entonces en toda vecindad de este punto la convergencia es no uniforme. Por tanto, toda representación de funciones discontinuas mediante series de funciones continuas debe basarse en el uso de procesos de límite no uniformemente convergentes.

d. Integración de series uniformemente convergentes

La suma de un número finito de funciones continuas puede integrarse "término a término", o sea, la integral de la suma de obtiene integrando cada término separadamente y sumando luego las integrales. En una serie infinita convergente formada por funciones continuas es permisible el mismo procedimiento siempre y cuando la serie converja uniformemente en el intervalo de integración.

Una serie $\sum_{\nu=1}^{\infty} g_\nu(x) = f(x)$ *que converge uniformemente en un intervalo puede integrarse término a término en él.* O bien, en forma más precisa, *si a y x son dos números en el intervalo de convergencia uniforme, la serie* $\sum_{\nu=1}^{\infty} \int_a^x g_\nu(t)\,dt$ *converge y, de hecho, converge uniformemente con respecto a x, siendo su suma igual a* $\int_a^x f(t)\,dt$.[1]

Para demostrar esto, escríbase como antes

$$f(x) = \sum_{\nu=1}^{\infty} g_\nu(x) = f_n(x) + R_n(x).$$

Se ha supuesto que los términos separados de la serie son continuos; y, en consecuencia, por la sección 7.4*c* la suma es también continua y, por tanto, integrable. Ahora bien, si ε es cualquier número positivo, es posible encontrar un número N tan grande que para todo $n > N$ la desigualdad $|R_n(x)| < \varepsilon$ se cumpla para todo valor de x en el intervalo. Por el teorema del valor medio del cálculo integral se tiene

$$\left| \int_a^x [f(t) - f_n(t)]\,dt \right| \leq \varepsilon l,$$

[1] Obsérvese que en este teorema deben tomarse integrales *definidas*. Así, por ejemplo, la serie $\sum_{\nu=1}^{\infty} g_\nu(x)$, con $g_\nu(x) = 0$, converge uniformemente. Sin embargo, tomando la integral indefinida $\int g_\nu(x)\,dx$ = constante = c de cada término, se llega a la serie $\sum_{\nu=1}^{\infty} c$, que, en general, es divergente.

donde l es la longitud del intervalo de integración. Puesto que la integración de la suma finita $f_n(x)$ puede realizarse término a término, resulta

$$\left| \int_a^n f(t)\,dt - \sum_1^x \int_a^x g_\nu(t)\,dt \right| < \varepsilon l.$$

Pero εl puede hacerse tan pequeño como se quiera; por tanto,

$$\sum_{\nu=1}^{\infty} \int_n^x g_\nu(t)\,dt = \lim_{n\to\infty} \sum_{\nu=1}^{a} \int_a^x g_\nu(t)\,dt = \int_a^x f(t)\,dt,$$

que es lo que debía demostrarse.

Si, en vez de series infinitas, se trata de sucesiones de funciones, el resultado puede expresarse de la manera siguiente:

Si en un intervalo la sucesión de funciones $f_1(x), f_2(x), \ldots$ *tiende uniformemente a la función límite* $f(x)$, *entonces*

$$\int_a^b f(x)\,dx = \lim_{n\to\infty} \int_a^b f_n(x)\,dx \tag{7}$$

para todo par de números a y b pertenecientes al intervalo. En otras palabras, *pueden intercambiarse las operaciones de integración y de paso al límite.*

Este hecho no es trivial. Desde un punto de vista ingenuo, como el que prevalecía en el siglo XVIII, es cierto que la intercambiabilidad de los dos procesos difícilmente puede ponerse en duda. Sin embargo, un vistazo a los ejemplos de 7.4*a* muestra que en el caso de la convergencia no uniforme la ecuación anterior podría no cumplirse. Basta considerar el ejemplo *b* de p. 549, en el cual la integral de la función límite es 0, en tanto que la integral de la función $f_n(x)$ sobre el intervalo $0 \leq x \leq 1$, es decir, el área del triángulo en la Fig. 7.5, tiene el valor

$$\int_0^1 f_n(x)\,dx = n^{\alpha-2},$$

que no tiende a cero cuando $\alpha \geq 2$. En este caso se ve inmediatamente por la figura que la razón de la diferencia entre $\int_0^1 f(x)\,dx$ y $\lim_{n\to\infty} \int_0^1 f_n(x)\,dx$ radica en la no uniformidad de la convergencia.

Por otra parte, considerando valores de α tales que $1 \leq \alpha < 2$, vemos que la ecuación $\lim_{n\to\infty} \int_0^1 f_n(x)\,dx = \int_0^1 f(x)\,dx$ se satisface aun cuando la convergencia es no uniforme. Como un ejemplo adicional, considérese la serie $\sum_0^{\infty} g_n(x)$, con $g_n(x) = x^n - x^{n-1}$ para $n \geq 1$ y $g_0(x) = 1$, la que puede integrarse término a término entre los límites 0 y 1 aun cuando no converge uniformemente. Así, aunque la uniformidad de la convergencia es una condición *suficiente* para la integrabilidad término a término, de ninguna manera es una condición *necesaria*.

e. Derivación de series infinitas

El comportamiento de las series o de las sucesiones uniformemente convergentes con respecto a la derivación es completamente diferente de su comportamiento respecto a la integración. Por ejemplo, la sucesión de funciones $f_n(x) = \dfrac{\operatorname{sen} n^2x}{n}$ converge uniformemente a la función límite $f(x) = 0$, pero la derivada $f_n'(x) = n \cos n^2x$ no converge en todas partes a la derivada de la función límite $f'(x) = 0$, como puede verse considerando el punto $x = 0$. Por consiguiente, a pesar de la uniformidad de la convergencia, no es posible intercambiar los procesos de derivación y de paso al límite.

Por supuesto, los correspondientes asertos para series infinitas son válidos. Por ejemplo, la serie

$$\operatorname{sen} x + \frac{\operatorname{sen} 2^4x}{2^2} + \frac{\operatorname{sen} 3^4x}{3^2} + \cdots$$

es absoluta y uniformemente convergente, pues sus términos son numéricamente no mayores que los términos de la serie convergente $\dfrac{1}{1^2} + \dfrac{1}{2^2} + \dfrac{1}{3^2} + \cdots$. Si, no obstante, se deriva la serie término a término, se obtiene la serie

$$\cos x + 2^2 \cos 2^4x + 3^2 \cos 3^4x + \cdots,$$

que es divergente en $x = 0$.

El único criterio útil que asegura en casos especiales que la derivación término a término es permisible queda expresado en el siguiente teorema.

Si al derivar término a término una serie infinita convergente, $\sum_{\nu=0}^{\infty} G_\nu(x) = F(x)$, *se obtiene una serie uniformemente convergente de términos continuos,* $\sum_{\nu=0}^{\infty} g_\nu(x) = f(x)$, *entonces la suma de esta última serie es igual a la derivada de la suma de la primera serie.*

Este teorema, por tanto, requiere expresamente que después de derivar término a término la serie se investigue si el resultado de tal derivación es o no una serie uniformemente convergente.

La demostración del teorema es casi trivial. En efecto, por el teorema de la sección 7.4*d*, la serie obtenida por derivación puede integrarse término a término. Teniendo en cuenta que $g_\nu(t) = G_\nu'(t)$, se obtiene

$$\int_a^x f(t)\,dt = \int_a^x \left(\sum_{\nu=0}^{\infty} g_\nu(t)\right) dt = \sum_{\nu=0}^{\infty} \int_a^x g_\nu(t)\,dt = \sum_{\nu=0}^{\infty} (G_\nu(x) - G_\nu(a))$$
$$= F(x) - F(a).$$

Lo cual, siendo cierto para todo valor de x en el intervalo de convergencia uniforme, implica que

$$f(x) = F'(x),$$

como se quería demostrar.

7.5 Series de potencias

Las *series de potencias* ocupan un lugar preponderante entre las series infinitas. Una serie de potencias es una serie del tipo

$$(8) \qquad P(x) = c_0 + c_1x + c_2x^2 + \cdots = \sum_{\nu=0}^{\infty} c_\nu x^\nu$$

("serie de potencias en x"), o, más generalmente,

$$(8a) \quad P(x) = c_0 + c_1(x - x_0) + c_2(x - x_0)^2 + \cdots = \sum_{\nu=0}^{\infty} c_\nu (x - x_0)^\nu$$

("serie de potencias en $(x - x_0)$"), donde x_0 es un número fijo. Si en esta última serie introducimos $\xi = x - x_0$ como una nueva variable, obtenemos una serie de potencias $\sum_{\nu=0}^{\infty} c_\nu \xi^\nu$ en la nueva variable ξ, por lo que podemos enfocar nuestra atención a las series de potencias de la forma especial $\sum_{\nu=0}^{\infty} c_\nu x^\nu$ sin pérdida de generalidad.

En el capítulo 5 (p. 465) consideramos la representación aproximada de funciones mediante polinomios, y en esa forma llegamos al desarrollo de funciones en serie de Taylor, las cuales son, de hecho, series de potencias. En esta sección estudiaremos con más detalle las series de potencias y obtendremos, en una forma más conveniente que antes, los desarrollos en series de algunas de las funciones más notables.

a. Propiedades de convergencia de las series de potencias—Intervalo de convergencia

Existen series de potencias que para ningún valor de x convergen, excepto, claro está, para $x = 0$. Como ejemplo tenemos la serie

$$x + 2^2x^2 + 3^3x^3 + \cdots + n^nx^n + \cdots.$$

En efecto, si $x \neq 0$ podemos encontrar un entero N tal que $|x| > 1/N$. Entonces, todos los términos n^nx^n para los cuales $n > N$ serán mayores que 1 en valor absoluto y, de hecho, cuando n crece n^nx^n crece más allá de toda cota, por lo que la serie no puede ser convergente.

Por otra parte, existen series que convergen para *todo* valor de x; por ejemplo, la serie de potencias para la función exponencial

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots,$$

cuya convergencia para todo valor de x se deduce de inmediato a partir del criterio del cociente (criterio 5*a*, p. 540). En $(n+1)$-ésimo término dividido por el n-ésimo término da x/n, y, cualquiera que sea el número x escogido, este cociente tiende a cero al crecer n.

El comportamiento de las series de potencias respecto a la convergencia es expresado en el siguiente teorema fundamental.

Si una serie de potencias en x converge para un valor $x = \xi$, converge absolutamente para todo valor de x tal que $|x| < |\xi|$, y la convergencia es uniforme en todo intervalo $|x| \leq \eta$ siendo η cualquier número positivo menor que $|\xi|$. Aquí η puede estar tan próximo a $|\xi|$ como se quiera.

La demostración es sencilla. Si la serie $\sum_{\nu=0}^{\infty} c_\nu \xi^\nu$ converge, sus términos tienden a cero cuando n crece. De aquí se desprende el aserto, más débil, de que todos los términos están por debajo de una cota M independiente de ν, es decir, $|c_\nu \xi^\nu| < M$. Si en estas condiciones representamos por q un número cualquiera que satisfaga la condición $0 < q < 1$, y si x se restringe al intervalo $|x| \leq q\,|\xi|$, entonces $|c_\nu x^\nu| \leq |c_\nu \xi^\nu|\, q^\nu < M q^\nu$. En este intervalo, por consiguiente, los términos de la serie $\sum_{0}^{\infty} c_\nu x^\nu$ son en valor absoluto más pequeños que los términos de la serie geométrica convergente $\Sigma M q^\nu$. Por tanto, del teorema de la p. 553 se deduce la convergencia absoluta y uniforme de la serie en el intervalo $-q\,|\xi| \leq x \leq q\,|\xi|$.

Si una serie de potencias no converge en todas partes, es decir, si existe un valor $x = \xi$ para el cual la serie diverge, debe divergir para todo valor de x tal que $|x| > |\xi|$. Pues si fuese convergente para tal valor de x, por el teorema anterior tendría que convergir para el valor ξ numéricamente menor.

De lo anterior se desprende que una serie de potencias que converge al menos para un valor de x diferente de 0 y que diverge al menos para un valor de x, tiene un *intervalo de convergencia;* esto es, existe un número positivo determinado ρ tal que para $|x| > \rho$ la serie diverge y para $|x| < \rho$ la serie converge. Para $|x| = \rho$ no es posible hacer afirmación general alguna. Aquí ρ es precisamente el *extremo superior* de los valores de x para los cuales la serie converge (y la existencia de tal extremo superior queda garantizada por el teorema de la p. 121, pues los valores de x para los cuales la serie converge forman un conjunto acotado). Los casos límite,

aquéllos en los que la serie converge solamente para $x = 0$ y aquéllos en que converge en todas partes, son expresados simbólicamente escribiendo $\rho = 0$ y $\rho = \infty$ respectivamente.[1]

Por ejemplo, para la serie geométrica $1 + x + x^2 + \cdots$ tenemos $\rho = 1$. En los puntos extremos del intervalo de convergencia la serie diverge. De manera análoga, para la serie de la tangente inversa (p. 463),

$$\arctan x = x - \frac{x^3}{3} + \frac{x^5}{5} - + \cdots,$$

tenemos $\rho = 1$, y en ambos puntos extremos ($x = \pm 1$) del intervalo de convergencia la serie converge, lo cual puede verse usando el criterio de Leibnitz (p. 533).

De la convergencia uniforme se deriva el importante resultado de que dentro de su intervalo de convergencia (si tal intervalo existe) una serie de potencias representa una función continua.

b. Integración y derivación de series de potencias

Por la uniformidad de la convergencia, *siempre es válido integrar término a término una serie de potencias*

$$f(x) = \sum_{\nu=0}^{\infty} c_\nu x^\nu$$

sobre cualquier intervalo cerrado contenido completamente por el intervalo de convergencia. Obtenemos así la función

$$F(x) = c + \sum_{\nu=0}^{\infty} \frac{c_\nu}{\nu+1} x^{\nu+1}, \tag{9}$$

para la cual

$$F'(x) = f(x) \quad \text{y} \quad F(0) = c.$$

También es posible derivar término a término una serie de potencias dentro de su intervalo de convergencia, obteniéndose así la ecuación

$$f'(x) = \sum_{\nu=1}^{\infty} \nu c_\nu x^{\nu-1}. \tag{10}$$

Para demostrar esta afirmación basta probar que la serie de la derecha converge uniformemente si x se restringe a un intervalo contenido com-

[1] Es posible encontrar este intervalo de convergencia directamente a partir de los coeficientes c_ν de la serie. Si el límite $\lim_{n\to\infty} \sqrt[n]{|c_n|}$ existe, entonces

$$\rho = \frac{1}{\lim_{n\to\infty} \sqrt[n]{|c_n|}}.$$

Para el caso general véase problema 8, p. 586.

pletamente por el intervalo de convergencia. Por tanto, supóngase que ξ es un número, tan próximo a ρ como se quiera, para el cual $\sum_{\nu=1}^{\infty} c_\nu \xi^\nu$ converge. Entonces, como ya se ha visto, los números $|c_\nu \xi^\nu|$ quedan todos por debajo de una cota M independiente de ν, de manera que $|c_\nu \xi^{\nu-1}| < \frac{M}{|\xi|} = N$. Ahora bien, sea q cualquier número tal que $0 < q < 1$. Si x se restringe al intervalo $|x| \leq q\,|\xi|$, los términos de la serie infinita (10) no son mayores que los de la serie $\sum_{\nu=1}^{\infty} |\nu c_\nu\, q^{\nu-1}\, \xi^{\nu-1}|$, y, por consiguiente, son menores que los términos de la serie $\sum_{\nu=0}^{\infty} N\nu q^{\nu-1}$. Sin embargo, en esta última serie la razón del $(n+1)$-ésimo término sobre el n-ésimo es $q(n+1)/n$, que tiende a q cuando n crece. Como $0 < q < 1$, se concluye [criterio (5*a*)] que esta serie converge. Por tanto, la serie obtenida por derivación converge uniformemente y, por el teorema de la p. 557, representa la derivada $f'(x)$ de la función $f(x)$, lo que prueba nuestra afirmación.

Si este resultado se aplica nuevamente a la serie de potencias

$$f'(x) = \sum_{\nu=1}^{\infty} \nu c_\nu x^{\nu-1},$$

al derivar término a término se encuentra que

$$f''(x) = \sum_{\nu=2}^{\infty} \nu(\nu-1) c_\nu x^{\nu-2}.$$

y, continuando el proceso, se llega al teorema: *Toda función representada por una serie de potencias puede derivarse tantas veces como se desee dentro del intervalo de convergencia, y esta derivación puede realizarse término a término.*[1]

c. Operaciones con series de potencias

Los teoremas anteriores relativos al comportamiento de las series de potencias son la justificación para operar con series de potencias de la misma manera que con polinomios. Es obvio que dos series de potencias pueden

[1] Como una expresión explícita para la k-ésima derivada obtenemos

$$f^{(k)}(x) = \sum_{\nu=k}^{\infty} \nu(\nu-1)\cdots(\nu-k+1) c\, x^{\nu-k},$$

o, en forma ligeramente diferente,

$$\frac{f^{(k)}(x)}{k!} = \sum_{\nu=k}^{\infty} \binom{\nu}{k} c_\nu x^{\nu-k} = \sum_{\nu=0}^{\infty} \binom{k+\nu}{k} c_{k+\nu} x^\nu.$$

Estas dos fórmulas son a menudo útiles.

sumarse o restarse sumando o restando los coeficientes correspondientes (ver p. 538). También es claro que una serie de potencias, lo mismo que cualquier otra serie convergente, puede multiplicarse por un factor constante multiplicando cada término por ese factor. Por otra parte, la multiplicación y la división de dos series requieren un estudio algo más detallado, para el cual remitimos al lector al Apéndice (p. 572). Aquí mencionamos solamente, sin demostrarlo, que dos series de potencias,

$$f(x) = \sum_{\nu=0}^{\infty} a_\nu x^\nu$$

y

$$g(x) = \sum_{\nu=0}^{\infty} b_\nu x^\nu,$$

pueden multiplicarse entre sí, como los polinomios. Concretamente, tenemos los siguientes teoremas: *En toda la parte común de sus intervalos de convergencia el producto de esas dos series está dado por la serie de potencias convergente* $\sum_{\nu=0}^{\infty} c_\nu x^\nu$, *en la que los coeficientes están dados por las fórmulas*

$$\begin{aligned}
c_0 &= a_0 b_0, \\
c_1 &= a_0 b_1 + a_1 b_0, \\
c_2 &= a_0 b_2 + a_1 b_1 + a_2 b_0, \\
&\cdots\cdots\cdots\cdots \\
c_n &= a_0 b_n + a_1 b_{n-1} + \cdots + a_n b_0, \\
&\cdots\cdots\cdots\cdots
\end{aligned}$$

d. Unicidad del desarrollo

En la teoría de las series de potencias es importante el hecho siguiente: si dos series de potencias, $\sum_{\nu=0}^{\infty} a_\nu x^\nu$ y $\sum_{\nu=0}^{\infty} b_\nu x^\nu$, convergen en un intervalo que contiene al punto $x = 0$ en su interior, y si en ese intervalo las dos series representan la misma función $f(x)$, entonces son idénticas entre sí, es decir, la ecuación $a_n = b_n$ es válida para todo valor de n. En otras palabras:

Una función $f(x)$ puede representarse mediante una serie de potencias en x de una única manera.

Brevemente: la representación de una función mediante una serie de potencias es "única".

Para la demostración sólo se necesita observar que la diferencia de dos series de potencias, esto es, la serie de potencias $\phi(x) = \sum_{\nu=0}^{\infty} c_\nu x^\nu$ con coeficientes $c_\nu = a_\nu - b_\nu$, representa la función

$$\phi(x) = f(x) - f(x) = 0$$

en el intervalo; es decir, esta última serie de potencias converge al límite 0 en todo el intervalo. En particular, para $x = 0$ la suma de la serie debe ser 0; o sea, $c_0 = 0$, de manera que $a_0 = b_0$. Ahora derivamos la serie en el interior del intervalo; y así resulta $\phi'(x) = \sum_{\nu=1}^{\infty} \nu c_\nu x^{\nu-1}$. Sin embargo, $\phi'(x)$ es también 0 en todo el intervalo; por tanto, para el caso particular de $x = 0$ se tiene $c_1 = 0$, o sea, $a_1 = b_1$. Continuando con este proceso consistente en derivar y luego hacer $x = 0$, se encuentra sucesivamente que todos los coeficientes c_ν son iguales a cero, lo que prueba nuestro teorema.

Además, de esta discusión puede extraerse la conclusión siguiente: si se toma la ν-ésima derivada de una serie $f(x) = \Sigma a_\nu x^\nu$ y luego se hace $x = 0$, se obtiene de inmediato

$$a_\nu = \frac{1}{\nu!} f^{(\nu)}(0),$$

esto es:

Toda serie de potencias que converge en puntos distintos de $x = 0$ es la serie de Taylor de la función que representa.

La unicidad del desarrollo corresponde al hecho de que los coeficientes pueden ser expresados en términos de la función misma.

*e. Funciones analíticas

El nombre "funciones analíticas" ha sido usado para designar las funciones $f(x)$ que pueden ser expresadas mediante series de potencias, desde que la importancia de tales funciones fue reconocida por Lagrange. Específicamente, $f(x)$ se llama *analítica* en la vecindad de $x = a$ si en esta vecindad es posible desarrollar $f(x)$ como una serie de potencias en $x - a$ que sea convergente.

En tanto que las funciones que no son totalmente o en todas partes analíticas desempeñan un papel central en el Análisis y sus aplicaciones, las funciones analíticas son particularmente importantes porque comparten con los polinomios muchas características simples.

Por ejemplo, una función analítica que no es idénticamente nula tendrá alguna derivada no nula en $x = a$. Sea r el número más pequeño para el cual $f^{(r)}(a) \neq 0$. Entonces, teniendo f un cero de orden r en el punto $x = a$, puede ser representada como un producto $f(x) = (x - a)^r g(x)$, donde $g(x)$ es una función analítica para la cual $g(a) = \frac{1}{r!} f^{(r)}(a)$ es diferente de cero. (Compárese con capítulo 5, p. 482.) Sin duda, la posibilidad de sacar como factor común la potencia $(x - a)^r$ se deduce de inmediato de la convergencia de la serie de potencias respectiva.

También, como se desprende de la continuidad de la serie de potencias convergente que se tiene para $g(x)$, el factor $g(x)$ no se anula en una vecindad adecuada de $x = a$, o sea: *los ceros de $f(x)$ están aislados,* a menos, claro está, que f se anule idénticamente.

Como lo mismo es cierto para la función $f'(x)$, se concluye que en un intervalo finito una función analítica es monótona por pedazos, es decir, no puede cambiar su carácter de monotonía un número infinito de veces. Así, la gráfica de $y = f(x)$ no puede intersectar en una infinidad de puntos a una recta $y =$ constante (o cualquier línea recta) dentro de un intervalo finito.

Puede observarse que estas últimas afirmaciones no son ciertas necesariamente para funciones no analíticas, como ocurre para $y = \text{sen}\,(1/x)\, e^{-1/x^2}$ en la vecindad de $x = 0$ (ver p. 481).

7.6 Desarrollos en series de potencias de funciones dadas. El método de los coeficientes indeterminados. Ejemplos

Toda serie de potencias representa dentro de su intervalo de convergencia una función continua con derivadas de todos los órdenes. Discutiremos ahora el problema inverso consistente en desarrollar en una serie de potencias una función dada. Teóricamente esto siempre puede hacerse por medio del teorema de Taylor. En la práctica surgen a menudo dificultades en el cálculo concreto de la n-ésima derivada y en la estimación del residuo. Sin embargo, en muchos casos es posible alcanzar dicho objetivo haciendo uso del siguiente artificio. Como primer paso se escribe $f(x) = \sum_{\nu=0}^{\infty} c_\nu x^\nu$, donde los coeficientes c_ν son inicialmente desconocidos. A continuación, usando alguna propiedad conocida de la función $f(x)$ se determinan los coeficientes, y luego se demuestra que la serie converge. Esta serie representa una función, y lo único que falta es demostrar que tal función es idéntica a $f(x)$. Por la unicidad del desarrollo en serie de potencias, se sabe que el desarrollo deseado lo constituye la serie obtenida, y ninguna otra. En realidad, las series para $\arctan x$ y $\log(1 + x)$ fueron obtenidas ya por un método relacionado con la idea que se expone en este capítulo; pues simplemente se integraron término a término las series para las derivadas de estas funciones, las cuales se sabía eran series geométricas. Consideraremos ahora algunos ejemplos de este método.

a. La función exponencial

Como se vio en el capítulo 3, sección 4*a*, p. 244, la función $y = e^x$ está completamente caracterizada por la ecuación diferencial $y' = y$ y la condición inicial $y = 1$ para $x = 0$. Estas propiedades pueden usarse directamente para encontrar la serie de potencias de la función exponencial. El problema consiste en hallar una función $f(x)$ para la cual $f'(x) = f(x)$ y $f(0) = 1$. Si se escribe formalmente la serie con sus coeficientes indeterminados,

$$f(x) = c_0 + c_1 x + c_2 x^2 + \cdots.$$

y se la deriva, se obtiene

$$f'(x) = c_1 + 2c_2 x + 3c_3 x^2 + \cdots.$$

Puesto que, por hipótesis, estas dos series de potencias deben ser idénticas, resulta la ecuación

$$nc_n = c_{n-1},$$

válida para todos los valores de $n \geq 1$. Si se observa que, por la relación $f(0) = 1$, el coeficiente c_0 debe tener el valor 1, es posible calcular entonces todos los coeficientes en forma sucesiva y de esta manera obtener la serie de potencias

$$f(x) = 1 + \frac{x}{1!} + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots.$$

Como se deduce fácilmente del criterio del cociente, esta serie converge para cualquier valor de x y, por consiguiente, representa una función para la cual las relaciones $f'(x) = f(x)$ y $f(0) = 1$ efectivamente se cumplen. (Intencionalmente se ha evitado aquí cualquier uso de lo previamente aprendido acerca del desarrollo de la función exponencial.)

Como únicamente la función e^x posee estas propiedades, se concluye en seguida que la función $f(x)$ es idéntica a e^x.

b. La serie del binomio

Podemos ahora volver a la serie del binomio (sección 5.5*c*, p. 475), haciendo uso esta vez del método de los coeficientes indeterminados. Deseamos desarrollar la función $f(x) = (1 + x)$ en una serie de potencias, por lo cual escribimos

$$f(x) = (1 + x)^\alpha = c_0 + c_1x + c_2x^2 + \cdots,$$

estando indeterminados los coeficientes c_ν. Observamos ahora que la función satisface obviamente la relación

$$(1 + x)f'(x) = \alpha f(x) = \sum_{\nu=0}^{\infty} \alpha c_\nu x^\nu:$$

Por otra parte, si derivamos término a término la serie para $f(x)$ y el resultado lo multiplicamos por $(1 + x)$, obtenemos

$$(1 + x)f'(x) = c_1 + (2c_2 + c_1)x + (3c_3 + 2c_2)x^2 + \cdots.$$

Ahora bien, puesto que estas dos series de potencias para $(1 + x)f'(x)$ deben ser idénticas entre sí, debe ser

$$\alpha c_0 = c_1, \quad \alpha c_1 = 2c_2 + c_1, \quad \alpha c_2 = 3c_3 + 2c_2, \ldots.$$

Además, como nuestra serie debe tener el valor 1 para $x = 0$, se encuentra que $c_0 = 1$, y así resultan sucesivamente las expresiones

$$c_1 = \alpha, \quad c_2 = \frac{(\alpha - 1)\alpha}{2}, \quad c_3 = \frac{(\alpha - 2)(\alpha - 1)\alpha}{3 \cdot 2}, \ldots,$$

para los coeficientes. En general se tiene

$$c_\nu = \frac{(\alpha - \nu + 1)(\alpha - \nu + 2)\cdots(\alpha - 1)\alpha}{\nu(\nu - 1)\cdots 2 \cdot 1} = \binom{\alpha}{\nu}.$$

como puede establecerse fácilmente. Al substituir estos valores para los coeficientes, resulta la serie $\sum_{\nu=0}^{\infty}\binom{\alpha}{\nu}x^\nu$. Hasta aquí no se ha investigado la convergencia de la serie ni demostrado que ella representa realmente la función $(1 + x)$.

Por el criterio del cociente, se encuentra que cuando α es un entero no positivo la serie converge si $|x| < 1$ y diverge si $|x| > 1$; pues en ese caso la relación entre los términos $(n + 1)$-ésimo y n-ésimo es $\dfrac{\alpha - n + 1}{n}x$, y el valor absoluto de esta expresión tiende a x cuando n crece más allá de toda cota.[1] Por tanto, si $|x| < 1$ la serie representa una función $f(x)$ que satisface la condición $(1 + x)f'(x) = \alpha f(x)$, según

[1] Mencionamos aquí, sin demostración, las condiciones exactas bajo las cuales esta serie converge. Si el índice α es un entero ≥ 0, la serie termina, y es válida, por consiguiente, para todos los valores de x (reduciéndose al teorema del binomio ordinario). Para todo otro valor de α la serie es absolutamente convergente si $|x| < 1$ y divergente si $|x| > 1$. Para $x = +1$ la serie converge absolutamente si $\alpha > 0$, converge condicionalmente si $-1 < \alpha < 0$ y diverge si $\alpha \leq -1$. Finalmente, para $x = -1$ la serie es absolutamente convergente si $\alpha > 0$, y es divergente si $\alpha < 0$.

se deduce del método usado para determinar los coeficientes. Además, $f(0) = 1$. Estas dos condiciones juntas garantizan que la función $f(x)$ es idéntica a $(1 + x)^\alpha$. En efecto, haciendo

$$\phi(x) = \frac{f(x)}{(1 + x)^\alpha}$$

se encuentra que

$$\phi'(x) = \frac{(1 + x)^\alpha f'(x) - \alpha(1 + x)^{\alpha-1}f(x)}{(1 + x)^{2\alpha}} = 0;$$

por tanto, $\phi(x)$ es una constante y, de hecho, es siempre igual a 1 ya que $\phi(0) = 1$. Hemos, pues, probado que para $|x| < 1$ se tiene

$$(1 + x)^\alpha = \sum_{\nu=0}^{\infty}\binom{\alpha}{\nu}x^\nu,$$

que es la serie del binomio.

Mencionamos aquí los siguientes casos especiales de la serie del binobio: la serie geométrica

$$\frac{1}{1 + x} = (1 + x)^{-1} = 1 - x + x^2 - x^3 + x^4 - + \cdots$$

$$= \sum_{\nu=0}^{\infty}(-1)^\nu x^\nu$$

la serie

$$\frac{1}{(1 + x)^2} = (1 + x)^{-2} = 1 - 2x + 3x^2 - 4x^3 + - \cdots$$

$$= \sum_{\nu=0}^{\infty}(-1)^\nu(\nu + 1)x^\nu;$$

que también puede obtenerse por derivación a partir de la serie geométrica; y la serie

$$\sqrt{(1 + x)} = (1 + x)^{1/2} = 1 + \frac{1}{2}x - \frac{1}{2 \cdot 4}x^2 + \frac{1 \cdot 3}{2 \cdot 4 \cdot 6}x^3 - \frac{1 \cdot 3 \cdot 5}{2 \cdot 4 \cdot 6 \cdot 8}x^4 + - \cdots,$$

$$\frac{1}{\sqrt{(1 + x)}} = (1 + x)^{-1/2} = 1 - \frac{1}{2}x + \frac{1 \cdot 3}{2 \cdot 4}x^2 - \frac{1 \cdot 3 \cdot 5}{2 \cdot 4 \cdot 6}x^3 + \frac{1 \cdot 3 \cdot 5 \cdot 7}{2 \cdot 4 \cdot 6 \cdot 8}x^4 - + \cdots,$$

para la cual constituyen buenas aproximaciones los dos o tres primeros términos.

c. La serie para arc sen x

Tal serie puede obtenerse de manera muy sencilla desarrollando la expresión $1/\sqrt{(1 - t^2)}$ de acuerdo con la serie del binomio:

$$(1 - t^2)^{-1/2} = 1 + \frac{1}{2}t^2 + \frac{1 \cdot 3}{2 \cdot 4}t^4 + \cdots.$$

Esta serie converge si $|t| \leq 1$, de manera que converge uniformemente si $|t| \leq q < 1$. Al integrar término a término entre 0 y x, resulta

$$\operatorname{arc\,sen} x = x + \frac{1}{2}\frac{x^3}{3} + \frac{1 \cdot 3}{2 \cdot 4}\frac{x^5}{5} + \cdots,$$

la cual, por el criterio del cociente, converge si $|x| < 1$ y diverge si $|x| > 1$.

La obtención de esta serie a partir del teorema de Taylor sería menos conveniente, debido a la dificultad en estimar el residuo.

d. La serie para arc senh $x = \log[x + \sqrt{(1 + x^2)}]$

Este desarrollo se obtiene de manera análoga. Usando el anterior teorema de la serie del binomio, escribimos la serie para la derivada de arc senh x.

$$\frac{1}{\sqrt{1 + x^2}} = 1 - \frac{1}{2}x^2 + \frac{1 \cdot 3}{2 \cdot 4}x^4 - \frac{1 \cdot 3 \cdot 5}{2 \cdot 4 \cdot 6}x^6 + - \cdots,$$

y luego integramos término a término. Así se obtiene el desarrollo

$$\operatorname{ar\,senh} x = x - \frac{1}{2}\frac{x^3}{3} + \frac{1 \cdot 3}{2 \cdot 4}\frac{x^5}{5} - + \cdots,$$

cuyo intervalo de convergencia es $-1 < x < 1$.

e. Ejemplo de multiplicación de series

El desarrollo de la función

$$\frac{\log(1 + x)}{1 + x}$$

es un ejemplo sencillo de la aplicación de la regla para multiplicar dos series de potencias. Sólo tiene que multiplicarse la serie logarítmica

$$\log(1 + x) = x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} + - \cdots$$

por la serie geométrica

$$\frac{1}{1 + x} = 1 - x + x^2 - x^3 + x^4 - + \cdots.$$

Como el lector puede fácilmente verificar, resulta el notable desarrollo

$$\frac{\log(1 + x)}{1 + x} = x - (1 + \tfrac{1}{2})x^2 + (1 + \tfrac{1}{2} + \tfrac{1}{3})x^3 - (1 + \tfrac{1}{2} + \tfrac{1}{3} + \tfrac{1}{4})x^4 + - \cdots$$

válido para $|x| < 1$.

f. Ejemplo de integración término a término (integral elíptica)

En aplicaciones anteriores (pp. 320, 430) hemos encontrado la integral elíptica

$$K = \int_0^{\pi/2} \frac{d\phi}{\sqrt{(1 - k^2 \operatorname{sen}^2 \phi)}}; \quad \text{para } (k^2 < 1)$$

que da el período de oscilación de un péndulo. Para evaluar la integral, puede primero desarrollarse el integrando mediante el teorema de la serie del binomio. Tal desarrollo da

$$\frac{1}{\sqrt{(1 - k^2 \operatorname{sen}^2 \phi)}} = 1 + \tfrac{1}{2}k^2 \operatorname{sen}^2 \phi + \frac{1 \cdot 3}{2 \cdot 4}k^4 \operatorname{sen}^4 \phi + \frac{1 \cdot 3 \cdot 5}{2 \cdot 4 \cdot 6}k^6 \operatorname{sen}^6 \phi + \cdots.$$

Puesto que $k^2 \operatorname{sen}^2 \phi$ nunca es mayor que k^2, esta serie converge uniformemente para todos los valores de ϕ, y de esta manera es posible integrar término a término:

$$K = \int_0^{\pi/2} \frac{d\phi}{\sqrt{(1 - k^2 \operatorname{sen}^2 \phi)}} = \int_0^{\pi/2} d\phi + \frac{1}{2}k^2 \int_0^{\pi/2} \operatorname{sen}^2 \phi\, d\phi + \frac{1 \cdot 3}{2 \cdot 4}k^4 \int_0^{\pi/2} \operatorname{sen}^4 \phi\, d\phi + \cdots.$$

Las integrales que aquí aparecen ya han sido calculadas [ver Ec. (76), p. 300]. Substituyendo sus valores se obtiene

$$K = \int_0^{\pi/2} \frac{d\phi}{\sqrt{(1 - k^2 \operatorname{sen}^2 \phi)}} = \frac{\pi}{2}\left[1 + \left(\frac{1}{2}\right)^2 k^2 + \left(\frac{1 \cdot 3}{2 \cdot 4}\right)^2 k^4 + \left(\frac{1 \cdot 3 \cdot 5}{2 \cdot 4 \cdot 6}\right)^2 k^6 + \cdots\right].$$

7.7 Series de potencias con términos complejos

a. Introducción de términos complejos en una serie de potencias. Representaciones complejas de funciones trigonométricas

La semejanza entre ciertas series de potencias representativas de funciones sin relación aparente entre sí condujo a Euler a establecer una conexión puramente formal entre dichas series, asignando valores complejos y, en particular, valores imaginarios puros, a la variable x. Primero describiremos el descubrimiento de Euler, formal pero por demás fructífero y notable, dejando de lado las cuestiones del rigor lógico. Después indicaremos una justificación más rigurosa.

La primer relación de esta clase se obtiene al reemplazar la cantidad x en la serie para e^x por una cantidad imaginaria pura, $i\phi$, donde ϕ es un número real. Teniendo en cuenta la ecuación fundamental para la unidad imaginaria i, esto es, $i^2 = -1$, de la cual se obtienen $i^3 = -1$, $i^4 = -1$, $i^5 = i$, ..., se ve que al separar los términos reales e imaginarios de la serie resulta

$$e^{i\phi} = \left(1 - \frac{\phi^2}{2!} + \frac{\phi^4}{4!} - \frac{\phi^6}{6!} + - \cdots\right) + i\left(\phi - \frac{\phi^3}{3!} + \frac{\phi^5}{5!} - \frac{\phi^7}{7!} + - \cdots\right),$$

o, en otra forma

$$e^{i\phi} = \cos\phi + i \operatorname{sen}\phi. \tag{11}$$

Esta es la bien conocida e importante "fórmula de Euler", fundamental en el Análisis, la que, por ahora, no es sino una relación puramente formal.[1] Está de acuerdo con el teorema de De Moivre (p. 128), expresado por la ecuación

$$(\cos\phi + i \operatorname{sen}\phi)(\cos\psi + i \operatorname{sen}\psi) = \cos(\phi + \psi) + i \operatorname{sen}(\phi + \psi).$$

A la luz de la fórmula de Euler, esta ecuación simplemente establece que la relación

$$e^x \cdot e^y = e^{x+y}$$

sigue siendo válida para valores imaginarios puros, $x = i\phi$, $y = i\psi$.

Debe dejarse establecido que esta fórmula de Euler y el teorema de adición $e^{i\phi}e^{i\psi} = e^{i(\phi+\psi)}$ pueden ser usados rigurosamente, sin necesidad de justificación posterior, *definiendo* simplemente $e^{i\phi}$ como el número complejo $\cos\phi + i \operatorname{sen}\phi$. Esta definición es compatible con las reglas ordinarias para operar con exponenciales. En particular, la regla ordinaria para multiplicar potencias de e proporciona justamente expresiones concisas y simples de los teoremas de adición de la trigonometría expresados por la fórmula de De Moivre, la que es de un carácter enteramente elemental. Por tanto, podremos usar las relaciones de Euler con la confianza de estar pisando terreno firme, sin necesidad de recurrir a un análisis más general de las funciones de una variable compleja, cosa que haremos en la siguiente sección.

En forma más general, la función exponencial puede ser definida para un exponente complejo arbitrario $x + iy$ (donde x e y son reales) mediante la fórmula

$$e^{x+iy} = e^x e^{iy} = e^x(\cos y + i \operatorname{sen} y).$$

Al reemplazar en la serie de potencias para $\cos x$ la variable x por la cantidad imaginaria pura ix, se obtiene de inmediato la serie para $\cosh x$. Esta relación queda expresada por la ecuación

$$\cosh x = \cos ix. \tag{12}$$

De manera similar se obtiene

$$\operatorname{senh} x = \frac{1}{i} \operatorname{sen} ix. \tag{13}$$

Puesto que la fórmula de Euler también da $e^{-i\phi} = \cos\phi - i \operatorname{sen}\phi$, llegamos a las *expresiones exponenciales para las funciones trigonométricas:*

$$\operatorname{sen} x = \frac{e^{ix} - e^{-ix}}{2i}, \qquad \cos x = \frac{e^{ix} + e^{-ix}}{2}. \tag{14}$$

[1] Una consecuencia que resulta para $\phi = \pi$ es la fórmula $e^{\pi i} = -1$, notable relación entre tres de los números más importantes, e, π e i.

Estas expresiones son análogas a las expresiones exponenciales de las funciones hiperbólicas y, en efecto, se transforman en éstas mediante las relaciones $x = \cos ix$, $\operatorname{sen} x = \frac{1}{i} \operatorname{sen} ix$.

Pueden también obtenerse, por supuesto, las correspondientes relaciones formales para las funciones $\tan x$, $\tanh x$, $\cot x$, $\coth x$, relacionadas entre sí por las ecuaciones $x = \frac{1}{i}\tan ix$, $\coth x = i \cot ix$.

Finalmente, diremos que es posible encontrar también relaciones análogas para las funciones trigonométricas inversas y las hiperbólicas. Por ejemplo, a partir de

$$y = \tan x = \frac{e^{ix} - e^{-ix}}{i(e^{ix} + e^{-ix})} = \frac{e^{2ix} - 1}{i(e^{2ix} + 1)}$$

se encuentra inmediatamente que

$$e^{2ix} = \frac{1 + iy}{1 - iy}.$$

Tomando logaritmos en ambos lados de esta ecuación y luego escribiendo x en vez de y y $\operatorname{arc\,tan} x$ en vez de x, se obtiene la ecuación

$$\operatorname{arc\,tan} x = \frac{1}{2i} \log \frac{1 + ix}{1 - ix}, \tag{15}$$

la cual expresa una relación notable entre la tangente inversa y el logaritmo. Si en la serie de potencias para $\frac{1}{2}\log\frac{1+x}{1-x}$ (p. 444), ya conocida, se reemplaza x por ix, se obtiene la serie de potencias para $\operatorname{arc\,tan} x$,

$$\operatorname{arc\,tan} x = \frac{1}{i}\left(ix + \frac{(ix)^3}{3} + \frac{(ix)^5}{5} + \cdots\right)$$

$$= x - \frac{x^3}{3} + \frac{x^5}{5} - + \cdots.$$

Estas relaciones son, hasta este punto, de carácter puramente formal, y, naturalmente, es necesario establecer de manera más precisa el contenido de las mismas. Sin embargo, como se ha visto antes, mediante el uso de definiciones adecuadas es posible dar a esas relaciones significado satisfactoriamente riguroso.

*b. Un vistazo a la teoría general de funciones de una variable compleja

Aunque el punto de vista puramente formal indicado en la última sección está libre de objeciones, es deseable, no obstante, reconocer en las fórmulas anteriores algo de carácter más fundamental que simples relaciones formales. Esta idea conduce a la *teoría general de funciones complejas,* como se denomina (por brevedad) a la teoría general de las llamadas *funciones analíticas de una variable compleja.* Como punto de partida podemos usar una discusión general de la teoría de series de potencias con variables complejas y coeficientes también complejos. La construcción de una teoría tal de series de potencias no presenta dificultades una vez que se han definido los conceptos de límite en el dominio de los números complejos. De hecho, esta teoría se desarrolla en forma casi paralela a la teoría de series de potencias reales. Sin embargo, puesto que no haremos uso alguno de estas cuestiones en lo que sigue, nos concretaremos aquí a establecer ciertos hechos, omitiendo las demostraciones. Uno de los resultados que se encuentran es la siguiente generalización del teorema de la sección 7.5*a,* válida para series de potencias complejas:

Si una serie de potencias converge para cualquier valor complejo $x = \xi$, *entonces converge absolutamente para todo valor* x *tal que* $|x| < |\xi|$; *y si dicha serie diverge para un valor* $x = \xi$, *entonces diverge también para todo otro valor* x *tal que* $|x| > |\xi|$. *Una serie de potencias que no converge en todas partes pero que converge para algunos puntos, además de hacerlo para* $x = 0$, *posee un círculo de convergencia, es decir, existe un número* $\rho > 0$ *tal que la serie converge absolutamente para* $|x| < \rho$ *y diverge para* $|x| > \rho$.

Una vez que se ha establecido el concepto de "funciones de una variable compleja representadas por series de potencias", y habiendo desarrollado las reglas para operar con tales funciones, podemos considerar las funciones e^x, sen x, cos x, arc tan x, etc., de la variable *compleja* x simplemente como *definidas* por las series de potencias que las representan para valores reales de x.

Ilustraremos mediante dos ejemplos la forma en que la introducción de variables complejas arroja luz sobre el comportamiento de las funciones elementales. La serie geométrica para $1/(1 + x^2)$ deja de ser convergente cuando x cae fuera del intervalo $-1 \leq x \leq 1$, y lo mismo ocurre para la serie que representa a arc tan x, aunque el comportamiento de estas funciones no exhibe peculiaridades en los extremos del intervalo de convergencia. Estas funciones y sus derivadas son continuas para todos los valores reales de x. Por otra parte, es fácil comprender por qué las series para

$1/(1 - x^2)$ y $\log(1 - x)$ dejan de ser convergentes cuando x pasa por el valor uno, ya que en este punto las funciones se vuelven infinitas.

Sin embargo, la divergencia de la serie que representa a la tangente inversa y la de la serie $\sum_{\nu=0}^{\infty} (-1)^\nu x^{2\nu}$, para $|x| > 1$, se aclara de inmediato si consideramos valores complejos de x, como antes se hizo. En efecto, en este caso encontramos que cuando $x = i$ las funciones se vuelven infinitas y por esta razón no pueden ser representadas mediante series convergentes. Por tanto, como una consecuencia del teorema sobre el círculo de convergencia, la serie debe ser divergente para todos los valores de x tales que $|x| > |i| = 1$. En particular, para valores reales de x la serie diverge fuera del intervalo $-1 \leq x \leq 1$.

Otro ejemplo lo constituye la función $f(x) = e^{-1/x^2}$ para $x \neq 0$, con $f(0) = 0$ (ver p. 481), la cual, a pesar de su comportamiento completamente suave,[1] no puede ser desarrollada en serie de Taylor. De hecho, esta función deja de ser continua si consideramos valores imaginarios puros $x = i\xi$. La función toma entonces la forma e^{1/ξ^2} y crece más allá de toda cota cuando $\xi \to 0$. Es claro, por consiguiente, que esta función no puede ser representada por ninguna serie de potencias en x para todos los valores complejos de x en una vecindad del origen, sin importar cuán pequeña se escoja dicha vecindad.

Estas observaciones sobre la teoría de funciones y series de potencias de una variable compleja bastarán por ahora.

Apéndice

*A.1 Multiplicación y división de series

a. Multiplicación de series absolutamente convergentes

Sean

$$A = \sum_{\nu=0}^{\infty} a_\nu \quad \text{y} \quad B = \sum_{\nu=0}^{\infty} b_\nu$$

dos series absolutamente convergentes. Junto con éstas consideremos las correspondientes series convergentes de valores absolutos

$$\bar{A} = \sum_{\nu=0}^{\infty} |a_\nu| \quad \text{y} \quad \bar{B} = \sum_{\nu=0}^{\infty} |b_\nu|.$$

A continuación hagamos

$$A_n = \sum_{\nu=0}^{n-1} a_\nu, \quad B_n = \sum_{\nu=0}^{n-1} b_\nu, \quad \bar{A}_n = \sum_{\nu=0}^{n-1} |a_\nu|, \quad \bar{B}_n = \sum_{\nu=0}^{n-1} |b_\nu|$$

y

[1] Es una función "lisa", como se vio, en todo el campo real. *(N. del T.)*

$$c_n = a_0 b_n + a_1 b_{n-1} + \cdots + a_n b_0$$

Afirmamos que la serie $\sum_{\nu=0}^{\infty} c_n$ es absolutamente convergente y que su suma es igual a AB.

Para demostrarlo escribimos la serie

$$\begin{aligned} &a_0 b_0 + a_1 b_0 + a_1 b_1 + a_0 b_1 + a_2 b_0 + a_2 b_1 \\ &+ a_2 b_2 + a_1 b_2 + a_0 b_2 + \cdots + a_n b_0 + a_n b_1 \\ &+ \cdots + a_n b_n + \cdots + a_1 b_n + a_0 b_n + \cdots, \end{aligned}$$

en la cual la n^2-ésima suma parcial es $A_n B_n$. Lo que se afirma es que tal serie converge absolutamente. En efecto, las sumas parciales de las correspondientes series de valores absolutos crecen monótonamente; y la n^2-ésima suma parcial es igual a $\bar{A}_n \bar{B}_n$, que es menor que $\bar{A}\bar{B}$ (y que tiende a $\bar{A}\bar{B}$). Por consiguiente, la serie de valores absolutos converge y la serie en cuestión converge absolutamente. La suma de la serie es evidentemente AB, puesto que la n^2-ésima suma parcial es $A_n B_n$, que tiende a AB cuando $n \to \infty$. Intercambiemos ahora los términos, lo cual es válido para series absolutamente convergentes, y agrupemos en paréntesis, en forma sucesiva, los términos. En una serie convergente es posible agrupar los términos sucesivos en la forma que se quiera sin alterar la convergencia o el valor de la suma de la serie, pues si, por ejemplo, se agrupan todos los términos $(a_{n+1} + a_{n+2} + \cdots + a_m)$, entonces al formar las sumas parciales omitiremos aquéllas sumas que originalmente caían entre s_n y s_m, lo cual no afecta la convergencia ni cambia el valor del límite. También, si antes de insertar los paréntesis la serie es absolutamente convergente, lo sigue siendo después. Puesto que la serie

$$\sum_{\nu=0}^{\infty} c_\nu = (a_0 b_0) + (a_0 b_1 + a_1 b_0) + (a_0 b_2 + a_1 b_1 + a_2 b_0) + \cdots$$

está formada de esta manera por las series consideradas antes, la demostración requerida queda completa.

*b. Multiplicación y división de series de potencias

La aplicación principal del teorema anterior tiene lugar en la teoría de las series de potencias. El siguiente aserto es una consecuencia inmediata de él: El producto de dos series de potencias,

$$\sum_{\nu=0}^{\infty} a_\nu x^\nu \quad \text{y} \quad \sum_{\nu=0}^{\infty} b_\nu x^\nu;$$

está representado en el intervalo de convergencia común a ambas series de potencias por una tercera serie, $\sum_{\nu=0}^{\infty} c_\nu x^\nu$, cuyos coeficientes están dados por

$$c_\nu = a_0 b_\nu + a_1 b_{\nu-1} + \cdots + a_\nu b_0.$$

*Como en el caso de la multiplicación de series de potencias, el cociente de las dos series anteriores puede representarse por una serie de potencias $\sum_{\nu=0}^{\infty} q_\nu x^\nu$, siempre que b_0, el término constante en el denominador, no se anule. (En este último caso tal representación es por lo general imposible, pues podría no haber convergencia para $x = 0$ debido a la anulación del denominador, en tanto que, por otra parte, toda serie de potencias debe converger para $x = 0$.) Los coeficientes de la serie de potencias

$$\sum_{\nu=0}^{\infty} q_\nu x^\nu$$

pueden calcularse recordando que $\sum_{\nu=0}^{\infty} q_\nu x^\nu \cdot \sum_{\nu=0}^{\infty} b_\nu x^\nu = \sum_{\nu=0}^{\infty} a_\nu x^\nu$, y que, por consiguiente, son válidas las siguientes ecuaciones:

$$\begin{aligned} a_0 &= q_0 b_0, \\ a_1 &= q_0 b_1 + q_1 b_0, \\ a_2 &= q_0 b_2 + q_1 b_1 + q_2 b_0, \\ &\cdots\cdots\cdots\cdots\cdots \\ a_\nu &= q_0 b_\nu + q_1 b_{\nu-1} + \cdots + q_\nu b_0. \end{aligned}$$

De la primera de estas ecuaciones fácilmente se encuentra q_0, de la segunda se obtiene el valor de q_1, de la tercera (usando los valores de q_0 y q_1) se encuentra el valor de q_2, etc. Para dar una justificación estricta de la expresión del cociente de dos series de potencias como una tercera serie, hemos de investigar la convergencia de la serie de potencias $\sum_{\nu=0}^{\infty} q_\nu x^\nu$, que se ha calculado formalmente. Sin embargo, no haremos uso posterior del resultado y nos conformaremos con mencionar que la serie para el cociente realmente converge en algún intervalo en torno al origen. Se omite la demostración.

A.2 Series infinitas e integrales impropias

Las series infinitas y los conceptos desarrollados en relación con ellas tienen aplicaciones y analogías simples en la teoría de las integrales impropias (ver capítulo 3, p. 320). Nos concretamos aquí al caso de una integral convergente con un intervalo de integración infinito; digamos, una integral de la forma $\int_0^\infty f(x)\,dx$. Si dividimos el intervalo de integración mediante la sucesión de números $x_0 = 0, x_1, \ldots$ que tiende monótonamente a $+\infty$, podemos escribir la integral impropia en la forma

$$\int_0^\infty f(x)\,dx = a_1 + a_2 + \cdots,$$

donde cada término de la serie infinita es una integral:

$$a_1 = \int_0^{x_1} f(x)\,dx, \qquad a_2 = \int_{x_1}^{x_2} f(x)\,dx, \ldots,$$

y así sucesivamente. Esto es válido independientemente de la forma en que se escojan los puntos x_ν. Podemos, por consiguiente, relacionar la idea de una integral impropia convergente con la de una serie infinita, de muchas maneras.

Es especialmente conveniente escoger los puntos x_ν de manera tal que el integrando no cambie de signo dentro de ninguno de los subintervalos individuales. La serie $\sum_{\nu=1}^{\infty} |a_\nu|$ corresponde entonces a la integral del valor absoluto de nuestra función,

$$\int_0^\infty |f(x)|\,dx.$$

Hemos así llegado, de manera natural, al siguiente concepto: *se dice que una integral impropia,* $\int_0^\infty f(x)\,dx$, *es absolutamente convergente si la integral* $\int_0^\infty |f(x)|\,dx$ *converge.* De lo contrario, si la primera integral existe y la segunda no, decimos que aquélla es *condicionalmente convergente.*

Algunas de las integrales antes consideradas (pp. 326 a 328), tales como

$$\int_0^\infty \frac{1}{1+x^2}\,dx, \quad \int_0^\infty e^{-x^2}\,dx, \quad \Gamma(x) = \int_0^\infty e^{-t}t^{x-1}\,dt,$$

son absolutamente convergentes.

Por otra parte, la importante integral de Dirichlet

$$J = \int_0^\infty \frac{\operatorname{sen} x}{1}\,dx = \lim_{A\to\infty}\int_0^A \frac{\operatorname{sen} x}{x}\,dx,$$

estudiada en la p. 328, es el ejemplo típico de una integral condicionalmente convergente. La demostración más sencilla posible de su convergencia se basa en la reducción a una integral absolutamente convergente: Escribimos $\operatorname{sen} x = (1-\cos x)' = 2(\operatorname{sen}^2 x/2)'$ y aplicamos integración por partes, con lo cual J se reduce a la forma absolutamente convergente

$$J = 2\int_0^\infty \left(\operatorname{sen}^2\frac{x}{2}\right)\frac{1}{x^2}\,dx$$

(Obsérvese que el nuevo integrando tiende continuamente al límite $\frac{1}{2}$ cuando $x\to 0$ y se anula con el orden de x^{-2} cuando $x\to\infty$.)

*Una demostración diferente de la convergencia se obtiene subdividiendo el intervalo de 0 a A mediante los puntos $x_\nu = \nu\pi (\nu = 0, 1, 2, \ldots, \mu_A)$, donde μ_A es el mayor entero posible que satisface $\mu_A\pi \le A$. Por consiguiente, la integral resulta dividida en términos de la forma

$$a_\nu = \int_{(\nu-1)\pi}^{\nu\pi} \frac{\operatorname{sen} x}{x}\,dx, \quad \text{para} \quad \nu = 1, 2, \ldots,$$

y un residuo R_A de la forma

$$\int_{\mu_A\pi}^{A} \frac{\operatorname{sen} x}{x}\,dx \qquad (0 \le A - \mu_A\pi < \pi).$$

Es obvio que las cantidades a_ν tienen signos alternantes, pues $\operatorname{sen} x$ es alternadamente positivo y negativo en intervalos consecutivos. Además, $|a_{\nu+1}| < |a_\nu|$, pues al aplicar la transformación $x = \xi - \pi$, se tiene

$$|a_\nu| = \int_{(\nu-1)\pi}^{\nu\pi} \frac{|\operatorname{sen} x|}{x}\,dx = \int_{\nu\pi}^{(\nu+1)\pi} \frac{|\operatorname{sen}(\xi-\pi)|}{\xi-\pi}\,d\xi = \int_{\nu\pi}^{(\nu+1)\pi} \frac{|\operatorname{sen}\xi|}{\xi-\pi}\,d\xi > \int_{\nu\pi}^{(\nu+1)\pi} \frac{|\operatorname{sen}\xi|}{\xi}\,d\xi = |a_{\nu+1}|.$$

En consecuencia, por el criterio de Leibnitz se ve que Σa_ν converge. Además, el residuo R_A tiene el valor absoluto

$$|R_A| = \left|\int_{\mu_A\pi}^{A} \frac{\operatorname{sen} x}{x}\,dx\right| \le \int_{\mu_A\pi}^{(\mu_A+1)\pi} \frac{|\operatorname{sen} x|}{x}\,dx \le \frac{1}{\mu_A\pi}\int_{\mu_A\pi}^{(\mu_A+1)\pi} |\operatorname{sen} x|\,dx = \frac{2}{\mu_A\pi},$$

que tiende a 0 cuando A crece. Así, si se hace que A tienda a infinito en la ecuación

$$\int_0^A \frac{\operatorname{sen} x}{x}\,dx = a_1 + a_2 + a_3 + \cdots + a_{\mu_A} + R_A,$$

el lado derecho tiende a Σa_ν como límite, y así la integral es convergente. Sin embargo, la convergencia no es absoluta para

$$|a_\nu| > \int_{(\nu-1)\pi}^{\nu\pi} \frac{|\operatorname{sen} x|}{\nu\pi}\,dx = \frac{2}{\nu\pi}, \qquad \text{de manera que } \sum |a_\nu| \text{ diverge.}$$

*A.3 Productos infinitos

En la introducción a este capítulo (p. 530) se dijo que las series infinitas son sólo *una* forma, aunque particularmente importante, de representar números o funciones mediante procesos infinitos. Como un ejemplo de otro de tales procesos consideraremos aquí los productos infinitos. Pero no daremos demostraciones.

En la p. 301 encontramos el producto de Wallis,

$$\frac{\pi}{2} = \frac{2}{1}\cdot\frac{2}{3}\cdot\frac{4}{3}\cdot\frac{4}{5}\cdot\frac{6}{5}\cdot\frac{6}{7}\cdots,$$

en el cual el número $\pi/2$ se expresa como un "producto infinito". Generalmente hablando, por *valor* del producto infinito

$$\prod_{\nu=1}^{\infty} a_\nu = a_1\cdot a_2\cdot a_3\cdot a_4\cdots$$

se entiende el *límite* de la sucesión de "productos parciales"

$$a_1, \quad a_1\cdot a_2, \quad a_1\cdot a_2\cdot a_3, \quad a_1\cdot a_2\cdot a_3\cdot a_4, \ldots,$$

en el supuesto de que exista.

Es claro que los factores $a_1, a_2, a_3, \ldots$ pueden ser también funciones de una variable x. Un ejemplo especialmente interesante es el del "producto infinito" para la función sen x,

$$\text{(16)} \qquad \operatorname{sen} \pi x = \pi x \left(1 - \frac{x^2}{1^2}\right)\left(1 - \frac{x^2}{2^2}\right)\left(1 - \frac{x^2}{3^2}\right)\cdots,$$

que obtendremos en la sección 8.5, p. 620.

El producto infinito para la *función zeta* juega un papel muy importante en la teoría de los números. Para conservar la notación usual de la teoría de los números, denotaremos aquí por s la variable independiente y definiremos la función zeta para $s > 1$, siguiendo a Riemann, por la expresión

$$\zeta(s) = \sum_{n=1}^{\infty} \frac{1}{n^s}.$$

Sabemos (sección 7.2*c*, p. 543) que la serie de la derecha converge cuando $s > 1$. Si p es cualquier número mayor que 1, se obtiene la ecuación

$$\frac{1}{1 - \dfrac{1}{p^s}} = 1 + \frac{1}{p^s} + \frac{1}{p^{2s}} + \frac{1}{p^{3s}} + \cdots$$

al desarrollar el lado de la izquierda en una serie geométrica de cociente p^{-s}. Imaginemos que esta serie se escribe para todos los números primos $p_1, p_2, p_3, \ldots$ considerados en orden de magnitud creciente, y que todas las ecuaciones que así resultan se multiplican entre sí. Obtenemos en el lado izquierdo un producto de la forma

$$\frac{1}{1 - p_1^{-s}} \cdot \frac{1}{1 - p_2^{-s}} \cdots.$$

Sin detenernos a justificar este procedimiento multipliquemos entre sí las series del lado derecho de nuestras ecuaciones. Como resultado se obtiene una suma de términos

$$p_1^{-k_1 s} p_2^{-k_2 s} p_3^{-k_3 s} \cdots = (p_1^{k_1} p_2^{k_2} p_3^{k_3} \cdots)^{-s},$$

donde $k_1, k_2, k_3, \ldots$ son enteros no negativos. Recordamos también que, por un teorema elemental, todo entero $n > 1$ puede ser expresado de manera única como un producto de potencias de números primos diferentes: $n = p_1^{k_1} p_2^{k_2} \cdots$. En esta forma encontramos que el producto del lado derecho es la función $\zeta(s)$, de manera que obtenemos la notable "fórmula del producto" de Euler:

$$\text{(17)} \qquad \zeta(s) = \frac{1}{1 - p_1^{-s}} \cdot \frac{1}{1 - p_2^{-s}} \cdot \frac{1}{1 - p_3^{-s}} \cdots.$$

Esta "fórmula del producto", cuya derivación sólo se ha bosquejado aquí, es en realidad una expresión de la función zeta como un producto infinito, puesto que el número de números primos es infinito.

En la teoría general de los productos infinitos se excluye, por lo general, el caso en que el producto $a_1 a_2 \cdots a_n$ tiene como límite el cero. Por consiguiente, es especialmente importante que ninguno de los factores a_n se anule. Para que el producto pueda ser convergente los factores a_n deben, según esto, tender a 1 cuando n crece; y, puesto que, si es necesario, pode-

mos omitir un número finito de factores (lo cual no influye en las cuestiones de convergencia), es válido suponer que $a_n > 0$. El siguiente teorema, casi trivial, es aplicable a este caso:

Una condición necesaria y suficiente para la convergencia del producto $\prod_{\nu=1}^{\infty} a_\nu$, *donde* $a_\nu > 0$, *es que la serie* $\sum_{\nu=1}^{\infty} \log a_\nu$ *sea convergente.* En efecto, como una consecuencia de la continuidad del logaritmo, las sumas parciales $\sum_{\nu=1}^{n} \log a_\nu = \log (a_1 a_2 \cdots a_n)$ de esta serie tenderán a un límite definido si, y sólo si, los productos parciales $a_1 a_2 \cdots a_n$ tienen un límite positivo.

En el estudio de la convergencia es de gran utilidad el siguiente *criterio que da una condición suficiente,* en la que $a_\nu = 1 + \alpha_\nu$. *El producto*

$$\prod_{\nu=1}^{\infty} (1 + \alpha_\nu)$$

converge si la serie

$$\sum_{\nu=1}^{\infty} |\alpha_\nu|$$

converge y ninguno de los factores $(1 + \alpha_\nu)$ *es nulo.* Para la demostración podemos suponer, omitiendo un número finito de factores si es necesario, que $|\alpha_\nu| < \frac{1}{2}$ para todo ν. Entonces tenemos $1 - |\alpha_\nu| > \frac{1}{2}$ y, por el teorema del valor medio, $\log (1 + h) = \log (1 + h) - \log 1 = h/(1 + \theta h)$ con $0 < \theta < 1$.

Por consiguiente

$$|\log (1 + \alpha_\nu)| = \left| \frac{\alpha_\nu}{1 + \theta \alpha_\nu} \right| \leq \frac{|\alpha_\nu|}{1 - |\alpha_\nu|} \leq 2\,|\alpha_\nu|,$$

y en esta forma la convergencia de la serie $\sum_{\nu=1}^{\infty} \log (1 + \alpha_\nu)$ se deduce de la convergencia de $\sum_{\nu=1}^{\infty} |\alpha_\nu|$.

De este criterio se deduce que el producto infinito considerado antes para sen πx, (16), converge para todos los valores de x excepto para los valores $x = 0, \pm 1, \pm 2, \ldots$, en los cuales los factores del producto se anulan. Lo mismo que para la función ζ de Riemann, cuando $p \geq 2$ y $s > 1$ se encuentra de inmediato que

$$\frac{1}{1 - p^{-s}} = 1 + \frac{1}{p^s - 1}, \qquad 0 < \frac{1}{p^s - 1} < \frac{2}{p^s}.$$

Ahora bien, si p toma todos los valores primos, la serie $\Sigma \dfrac{1}{p^s}$ debe converger, puesto que sus términos son sólo una parte de la serie convergente $\sum_{\nu=1}^{\infty} \dfrac{1}{\nu^s}$. Queda así demostrada la convergencia para $s > 1$ del producto de la Ec. (17). Puesto que la serie

para $\zeta(s)$ con $s = 1$ (esto es, la serie armónica) diverge, puede concluirse que la serie de inversos de números primos, es decir, la serie

$$\sum_{k=1}^{\infty} \frac{1}{p_k} = \frac{1}{2} + \frac{1}{3} + \frac{1}{5} + \frac{1}{7} + \frac{1}{11} + \frac{1}{13} + \frac{1}{17} + \frac{1}{19} + \cdots$$

diverge también. (De paso esto demuestra que el número de números primos es infinito.) En efecto, si la serie de números primos recíprocos fuese convergente, entonces también lo sería la serie con términos

$$\alpha_k = \frac{1}{1 - p_k{}^{-1}} - 1 = \frac{p_k{}^{-1}}{1 - p_k{}^{-1}}$$

puesto que $p_k \geq 2$ y

$$0 < \alpha_k \leq 2p_k{}^{-1}.$$

En consecuencia, como indica el criterio anterior, el producto infinito

$$\prod_{k=1}^{\infty} (1 + \alpha_k) = \prod_{k=1}^{\infty} \frac{1}{1 - p_k{}^{-1}} = \prod_{k=1}^{\infty} \left(1 + \frac{1}{p_k} + \frac{1}{p_k{}^2} + \cdots\right)$$

también sería convergente; pero entonces la serie armónica debería converger, lo cual es imposible.

*A.4 Series en que aparecen números de Bernoulli

Hasta ahora no hemos dado los desarrollos en series de potencias para ciertas funciones elementales; por ejemplo, para tan x. La razón de ello es que los coeficientes numéricos que aparecen en esos casos no tienen ninguna forma sencilla. Estos coeficientes, y los que aparecen en las series de ciertas funciones, pueden expresarse en términos de los llamados *números de Bernoulli*. Son éstos curiosos números racionales, con una ley de formación algo obscura, que aparecen en muchas partes del Análisis. La manera más simple de obtenerlos consiste en desarrollar la función

$$\frac{x}{e^x - 1} = \frac{1}{1 + \dfrac{x}{2!} + \dfrac{x^2}{3!} + \cdots}$$

en una serie de potencias formal del tipo

$$\frac{x}{e^x - 1} = \sum_{\nu=0}^{\infty} \frac{B_\nu{}^*}{\nu!} x^\nu.$$

Escribiendo esta ecuación en la forma

$$x = (e^x - 1) \sum_{\nu=0}^{\infty} \frac{B_\nu{}^*}{\nu!} x^\nu$$

y substituyendo en el lado derecho $e^x - 1$ por su serie de potencias, se obtiene una relación de recurrencia para los $B_n{}^*$:

$$\binom{n+1}{1} B_n{}^* + \binom{n+1}{2} B_{n-1}{}^* + \binom{n+1}{3} B_{n-2}{}^* + \cdots + \binom{n+1}{n+1} B_0{}^* = 0,$$

donde $n > 0$ y $B_0{}^* = 1$. A partir de ella es posible calcular con facilidad los $B_i{}^*$ en forma sucesiva. Estos números racionales se llaman números de Bernoulli.[1] Son números racionales ya que en su formación sólo intervienen operaciones racionales. Como es fácil ver, se anulan para todos los índices impares diferentes de $\nu = 1$. Algunos de ellos se dan a continuación

$$B_0{}^* = 1, \quad B_1{}^* = -\frac{1}{2}, \quad B_2{}^* = \frac{1}{6}, \quad B_4{}^* = -\frac{1}{30}, \quad B_6{}^* = \frac{1}{42},$$

$$B_8{}^* = -\frac{1}{30}, \quad B_{10}{}^* = \frac{5}{66}, \cdots$$

Nos conformaremos aquí con dar una idea breve de cómo estos números aparecen en las series de potencias que estamos considerando. Primero, haciendo uso de la transformación

$$1 + \frac{B_2{}^*}{2!} x^2 + \cdots = \frac{x}{e^x - 1} + \frac{x}{2} = \frac{x}{2} \cdot \frac{e^x + 1}{e^x - 1} = \frac{x}{2} \cdot \frac{e^{\frac{1}{2}x} + e^{-\frac{1}{2}x}}{e^{\frac{1}{2}x} - e^{-\frac{1}{2}x}},$$

obtenemos

$$\frac{x}{2} \coth \frac{x}{2} = \sum_{\nu=0}^{\infty} \frac{B_{2\nu}{}^*}{(2\nu)!} x^{2\nu}.$$

(Esta fórmula prueba que $B_{2\nu+1}{}^* = 0$ para $\nu > 0$, ya que $(x/2) \coth (x/2)$ es una función par de x.)

Si reemplazamos x por $2x$ obtenemos la serie

$$x \coth x = \sum_{\nu=0}^{\infty} \frac{2^{2\nu} B_{2\nu}{}^*}{(2\nu)!} x^{2\nu},$$

válida, como puede demostrarse, para $|x| < \pi$, y de la cual se obtiene

$$x \cot x = \sum_{\nu=0}^{\infty} (-1)^\nu \frac{2^{2\nu} B_{2\nu}{}^*}{(2\nu)!} x^{2\nu}, \qquad |x| < \pi,$$

al reemplazar x por $-ix$ (ver p. 567).

Por medio de la ecuación $2 \cot 2x = \cot x - \tan x$ obtenemos ahora la serie

$$\tan x = \sum_{\nu=1}^{\infty} (-1)^{\nu-1} \frac{2^{2\nu}(2^{2\nu} - 1)}{(2\nu)!} B_{2\nu}{}^* x^{2\nu-1},$$

que se cumple para $|x| < \dfrac{\pi}{2}$.

Para una información adicional remitimos al lector al capítulo 8 y a textos más detallados.[2]

---

[1] Con una notación ligeramente diferente (pp. 638-639), la fórmula básica se escribirá

$$\frac{x}{e^x - 1} = 1 - \tfrac{1}{2}x + \sum_{\nu=1}^{\infty} (-1)^{\nu+1} \frac{B_\nu}{(2\nu)!} x^{2\nu}.$$

[2] Véase, por ejemplo, K. Knopp, *Theory and Applications of Infinite Series*, página 183, Blackie & Son, Ltd., 1928, y K Knopp, *Infinite Sequences and Series*, Dover Publications, 1956.

PROBLEMAS

SECCION 7.1, página 528

1. Demostrar que

$$\sum_{\nu=1}^{\infty} \frac{1}{\nu(\nu+1)} = \frac{1}{1\cdot 2} + \frac{1}{2\cdot 3} + \cdots + = 1$$

[ver problema 1.6, 12(*a*)] y úsese el resultado para demostrar que $\sum_{\nu=1}^{\infty} \frac{1}{\nu^2}$ converge.

2. Usar el resultado del problema 1 para obtener cotas superior e inferior para

$$\sum_{\nu=1}^{\infty} \frac{1}{\nu^2}.$$

3. Demostrar que $\sum_{\nu=0}^{\infty} (-1)^{\nu} \frac{2\nu+3}{(\nu+1)(\nu+2)} = 1.$

4. ¿Para qué valores de α converge la serie $1 - \frac{1}{2^{\alpha}} + \frac{1}{3^{\alpha}} - \frac{1}{4^{\alpha}} + \cdots$?

5. Demostrar que si $\sum_{\nu=1}^{\infty} a_{\nu}$ converge y $s_n = a_1 + a_2 + \cdots + a_n$, entonces la sucesión

$$\frac{s_1 + s_2 + \cdots + s_N}{N}$$

también converge y tiene por límite $\sum_{\nu=1}^{\infty} a_{\nu}$.

6. ¿Es convergente la serie $\sum_{n=1}^{\infty} \left(\frac{2n}{2n+1} - \frac{2n-1}{2n} \right)$?

7. ¿Es convergente la serie $\sum_{\nu=1}^{\infty} (-1)^{\nu} \frac{\nu}{\nu+1}$?

8. Demostrar que si $\sum_{\nu=1}^{\infty} a_{\nu}^2$ converge, lo mismo es cierto para $\sum_{\nu=1}^{\infty} \frac{a_{\nu}}{\nu}$.

9. *a*) Si a_n es una sucesión monótona creciente con términos positivos, ¿en qué casos converge la serie $\frac{1}{a_1} + \frac{1}{a_1 a_2} + \frac{1}{a_1 a_2 a_3} + \cdots$?

b) Dar un ejemplo de una sucesión monótona decreciente tal que $\lim_{n\to\infty} a_n = 1$ y para la cual la serie asociada diverja.

c) Demostrar que si se admiten sucesiones decrecientes, es posible entonces obtener sumas convergentes aun cuando $\lim_{n\to\infty} a_n = 1$.

10. Si la serie $\sum_{\nu=1}^{\infty} a_{\nu}$ con términos positivos decrecientes converge, entonces $\lim_{n\to\infty} n a_n = 0$.

11. Demostrar que la serie $\sum_{\nu=1}^{\infty} \operatorname{sen} \frac{\pi}{\nu}$ diverge.

12. Demostrar que si Σa_{ν} converge y $b_1, b_2, b_3, \ldots$ es una sucesión monótona acotada de números, entonces $\Sigma a_{\nu} b_{\nu}$ converge. Además, demostrar que si $S = \Sigma a_{\nu} b_{\nu}$ y $\Sigma a_{\nu} \leq M$, entonces $|S| \leq M b_1$.

13. Se dice que una sucesión $\{a_n\}$ es de *variación acotada* si la serie

$$\sum_{i=1}^{\infty} |a_{i+1} - a_i|$$

converge

a) Demostrar que si la sucesión $\{a_n\}$ es de variación acotada, entonces la sucesión $\{a_n\}$ converge.

b) Encontrar una serie infinita divergente, Σa_i cuyos elementos a_i constituyan una sucesión de variación acotada.

c) Demostrar la siguiente generalización del criterio de convergencia de Abel (ver página 532), debido a Dedekind:

La serie $\Sigma a_i p_i$ es convergente si Σa_i oscila entre cotas finitas y $\{p_i\}$ es una sucesión nula de variación acotada.

d) Establecer la convergencia de las siguientes series infinitas:

a) $\sum_{n=2}^{\infty} \frac{\operatorname{sen} nx}{\log n} (-1)^n$;

b) $\sum_{n=2}^{\infty} \frac{\cos nx}{\log n} (-1)^n$,

siendo x un número real arbitrario pero fijo.

14. Discutir la convergencia o divergencia de las siguientes series:

a) $\Sigma \frac{(-1)^{\nu}}{\nu}$ *d*) $\Sigma \frac{\operatorname{sen} \nu\theta}{\nu}$

b) $\Sigma \frac{(-1)^{\nu} \cos(\theta/\nu)}{\nu}$ *e*) $\Sigma \frac{(-1)^{\nu} \cos \nu\theta}{\nu}$

c) $\Sigma \frac{\cos \nu\theta}{\nu}$ *f*) $\Sigma \frac{(-1)^{\nu} \operatorname{sen} \nu\theta}{\nu}$

15. Encontrar las sumas de los siguientes reordenamientos de la serie para log 2,

$$1 - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} + \frac{1}{5} - \frac{1}{6} + \cdots:$$

a) $1 - \frac{1}{2} - \frac{1}{4} + \frac{1}{3} - \frac{1}{6} - \frac{1}{8} + \frac{1}{5} - \frac{1}{10} - \frac{1}{12} + - - \cdots.$

b) $1 + \frac{1}{3} + \frac{1}{5} - \frac{1}{2} - \frac{1}{4} - \frac{1}{6} + + + \cdots.$

16. Determinar si las siguientes series convergen o divergen:

a) $1 + \frac{1}{2} - \frac{1}{3} + \frac{1}{4} + \frac{1}{5} - \frac{1}{6} + \frac{1}{7} + \frac{1}{8} - \frac{1}{9} + + - \cdots.$

b) $1 + \frac{1}{2} - \frac{2}{3} + \frac{1}{4} + \frac{1}{5} - \frac{2}{6} + \frac{1}{7} + \frac{1}{8} - \frac{2}{9} + + - \cdots.$

SECCION 7.2, página 539

1. Demostrar que $\sum_{\nu=2}^{\infty} \frac{1}{\nu(\log \nu)^{\alpha}}$ converge cuando $\alpha > 1$ y diverge cuando $\alpha \leq 1$.

2. Demostrar que $\sum_{\nu=3}^{\infty} \frac{1}{\nu \log \nu (\log\log \nu)^{\alpha}}$ converge cuando $\alpha > 1$ y diverge cuando $\alpha \leq 1$

3. Demostrar que si n es un entero arbitrario mayor que 1,

$$\sum_{\nu=1}^{\infty} \frac{a_\nu{}^n}{\nu} = \log n,$$

donde $a_\nu{}^n$ se define como sigue:

$$a_\nu{}^n = \begin{cases} 1 \text{ si } n \text{ no es un factor de } \nu \\ -(n-1) \text{ si } n \text{ es un factor de } \nu \end{cases}$$

4. Demostrar que $\sum_{\nu=2}^{\infty} \frac{\log(\nu+1) - \log \nu}{(\log \nu)^2}$ converge.

5. Demostrar que $\sum_{\nu=1}^{\infty} \frac{1 \cdot 2 \cdot 3 \cdots \nu}{(\alpha+1)(\alpha+2)\cdots(\alpha+\nu)}$ converge si $\alpha > 1$ y diverge si $\alpha \leq 1$.

***6.** Mediante comparación con la serie $\sum_{\nu=1}^{\infty} \frac{1}{\nu^{\alpha}}$, establecer el criterio siguiente: Si $\frac{\log(1/|a_n|)}{\log n} > 1 + \varepsilon$ para algún número fijo $\varepsilon > 0$ independiente de n, y para todo n suficientemente grande, entonces la serie Σa_ν converge absolutamente. Si $\frac{\log(1/|a_n|)}{\log n} < 1 - \varepsilon$ para todo n suficientemente grande y algún número $\varepsilon > 0$ independiente de n, entonces la serie Σa_ν no converge absolutamente.

7. Demostrar que la serie $\sum_{\nu=1}^{\infty} \left(1 - \frac{1}{\sqrt{\nu}}\right)^{\nu}$ converge.

8. ¿Para qué valores de α convergen las siguientes series

a) $1 - \frac{1}{2^\alpha} + \frac{1}{3} - \frac{1}{4^\alpha} + \frac{1}{5} - \frac{1}{6^\alpha} + - \cdots$,

b) $1 + \frac{1}{3^\alpha} - \frac{1}{2^\alpha} + \frac{1}{5^\alpha} + \frac{1}{7^\alpha} - \frac{1}{4^\alpha} + - \cdots$?

9. Por comparación con la serie $\Sigma \frac{1}{\nu(\log \nu)^{\alpha}}$ establecer el criterio siguiente: La serie $\Sigma |a_\nu|$ converge o diverge según que

$$\frac{\log(1/n\,|a_n|)}{\log\log n}$$

sea mayor que $1 + \varepsilon$ o menor que $1 - \varepsilon$, para todo n suficientemente grande.

10. Obtener del criterio del problema 6 el criterio de la raíz n-ésima.

11. Establecer el siguiente criterio de comparación: Si la serie Σb_ν de términos positivos converge y si

$$\left|\frac{a_{n+1}}{a_n}\right| < \frac{b_{n+1}}{b_n}$$

a partir de cierto término, la serie Σa_ν es absolutamente convergente. Si Σb_ν diverge y si

$$\left|\frac{a_{n+1}}{a_n}\right| > \frac{b_{n+1}}{b_n}$$

a partir de cierto término, la serie Σa_ν no es absolutamente convergente.

***12.** Por comparación con la serie $\sum_{\nu=1}^{\infty} \frac{1}{\nu^{\alpha}}$, establecer el *criterio de Raabe:* La serie $\Sigma |a_\nu|$ converge o diverge según que

$$n\left(\frac{|a_n|}{|a_{n+1}|} - 1\right)$$

sea mayor que $1 + \varepsilon$ o menor que $1 - \varepsilon$ para todo n suficientemente grande y para algún $\varepsilon > 0$ independiente de n.

13. Por comparación con $\Sigma \frac{1}{\nu(\log \nu)^{\alpha}}$, establecer el siguiente criterio: La serie $\Sigma |a_\nu|$ converge o diverge según que

$$n \log n \left(\frac{|a_n|}{|a_{n+1}|} - 1 - \frac{1}{n}\right)$$

sea mayor que $1 + \varepsilon$ o menor que $1 - \varepsilon$ para todo n suficientemente grande.

14. Probar el *criterio de Gauss:*

Si $$\frac{|a_n|}{|a_{n+1}|} = 1 + \frac{\mu}{n} + \frac{R_n}{n^{1+\varepsilon}},$$

donde $|R_n|$ es acotado y $\varepsilon > 0$ es independiente de n, la serie $\Sigma |a_\nu|$ converge si $\mu > 1$, y diverge si $\mu \leq 1$.

15. Investigar la convergencia o divergencia de las siguientes series "hipergeométricas":

a) $\frac{\alpha}{\beta} + \frac{\alpha(\alpha+1)}{\beta(\beta+1)} + \frac{\alpha(\alpha+1)(\alpha+2)}{\beta(\beta+1)(\beta+2)} + \cdots$.

b) $1 + \frac{\alpha \cdot \beta}{1 \cdot \gamma} + \frac{\alpha(\alpha+1) \cdot \beta(\beta+1)}{1 \cdot 2 \cdot \gamma(\gamma+1)} + \frac{\alpha(\alpha+1)(\alpha+2) \cdot \beta(\beta+1)(\beta+2)}{1 \cdot 2 \cdot 3 \cdot \gamma(\gamma+1)(\gamma+2)} + \cdots$.

SECCION 7.4, página 545

1. La sucesión $f_n(x)$, $n = 1, 2, \ldots$, está definida en el intervalo $0 \leq x \leq 1$ por las ecuaciones

$$f_0(x) \equiv 1, \qquad f_n(x) = \sqrt{x f_{n-1}(x)}.$$

a) Demostrar que en el intervalo $0 \leq x \leq 1$ la sucesión converge a un límite continuo.

**b)* Probar que la convergencia es uniforme.

***2.** Sea $f_0(x)$ una función continua en el intervalo $0 \leq x \leq a$, y sea la sucesión de funciones $f_n(x)$ definida por

$$f_n(x) = \int_0^x f_{n-1}(t)\, dt, \qquad n = 1, 2, \ldots .$$

Demostrar que en cualquier intervalo fijo $0 \leq x \leq a$ la sucesión converge uniformemente a 0.

***3.** Sea $f_n(x)$, $n = 1, 2, \ldots$, una sucesión de funciones con derivadas continuas en el intervalo $a \leq x \leq b$. Demostrar que si $f_n(x)$ converge en cada punto del intervalo y que si la desigualdad $|f_n'(x)| < M$ (donde M es una constante) se satisface para todos los valores de n y de x, entonces la convergencia es uniforme.

4. *a*) Demostrar que la serie $\sum_{\nu=1}^{\infty} \frac{1}{\nu^x}$ converge uniformemente para $x \geq 1 + \varepsilon$, siendo $\varepsilon > 0$ cualquier número fijo.

b) Demostrar que la serie derivada $-\sum \frac{\log \nu}{\nu^x}$ converge uniformemente para $x \geq 1 + \varepsilon$, donde ε es un número positivo fijo.

***5.** Demostrar que la serie $\sum \frac{\cos \nu x}{\nu^\alpha}$, $\alpha > 0$, converge uniformemente para $\varepsilon \leq x \leq 2\pi - \varepsilon$ donde ε es cualquier valor positivo suficientemente pequeño.

6. La serie

$$\frac{x-1}{x+1} + \frac{1}{3}\left(\frac{x-1}{x+1}\right)^3 + \frac{1}{5}\left(\frac{x-1}{x+1}\right)^5 + \cdots$$

converge uniformemente para $\varepsilon \leq x \leq N$, siendo ε y N números positivos fijos.

7. Encontrar en qué regiones son convergentes las siguientes series:

a) $\sum x^{\nu!}$.

b) $\sum \frac{(\nu!)^2 x^\nu}{(2\nu)!}$.

c) $\sum \frac{a^\nu}{\nu^x}$, $a < 1$.

d) $\sum \frac{a^\nu}{\nu^x}$, $a > 1$.

e) $\sum \frac{\log \nu}{\nu^x}$.

f) $\sum \frac{x^\nu}{1 - x^\nu}$.

***8.** Demostrar que si la serie de Dirichlet, $\sum \frac{a_\nu}{\nu^x}$, converge para $x = x_0$, converge también para cualquier $x > x_0$; y si diverge para $x = x_0$, diverge también para cualquier $x < x_0$. Existe así una "abscisa de convergencia" tal que para cualquier valor de x mayor la serie converge y para cualquier valor de x menor la serie diverge.

9. Si $\sum \frac{a_\nu}{\nu^x}$ converge para $x = x_0$, la serie derivada, $-\sum \frac{a_\nu \log \nu}{\nu^x}$, converge para cualquier $x > x_0$.

SECCION 7.5, página 556

1. Si el intervalo de convergencia de la serie de potencias $\Sigma a_n x^n$ es $|x| < \rho$, y el de $\Sigma b_n x^n$ es $|x| < \rho'$, siendo $\rho < \rho'$, ¿cuál es el intervalo de convergencia de $\Sigma(a_n + b_n)x^n$?

2. Si $a_\nu > 0$ y Σa_ν converge, demostrar que entonces

$$\lim_{x \to 1-0} \sum a_\nu x^\nu = \sum a_\nu.$$

3. Si $a_\nu > 0$ y Σa_ν diverge, demostrar que entonces

$$\lim_{x \to 1-0} \sum a_\nu x^\nu = \infty.$$

***4.** Demostrar el teorema de Abel:
Si $\Sigma a_\nu X^\nu$ converge, entonces $\Sigma a_\nu x^\nu$ converge uniformemente para $0 \leq x \leq X$.

***5.** Si $\Sigma a_\nu X^\nu$ converge, entonces $\lim_{x \to X-0} \Sigma a_\nu x^\nu = \Sigma a_\nu X^\nu$. Demostrarlo.

***6.** Por multiplicación de series de potencias demostrar que
a) $e^x e^y = e^{x+y}$. *b*) $\operatorname{sen} 2x = 2 \operatorname{sen} x \cos x$.

7. Usando la serie binomial calcular $\sqrt{2}$ con cuatro cifras decimales.

8. Sea a_n una sucesión de números reales y S el conjunto de todos los puntos límite de a_n. Denótese el extremo superior p de S por $p = \overline{\lim}\, a_n$. Mostrar que la serie de potencias $\sum_{n=0}^{\infty} c_n x^n$ converge para $|x| < p$ y diverge para $|x| > p$, donde

$$\rho = \frac{1}{\lim \sqrt[n]{|c_n|}}.$$

APENDICE, página 570

1. Demostrar que la serie de potencias para $\sqrt{(1-x)}$ sigue siendo convergente para $x = 1$.

2. Demostrar que para todo número positivo ε existe un polinomio en x que representa a $\sqrt{(1-x)}$ en el intervalo $0 \leq x \leq 1$ con un error menor que ε.

3. Haciendo $x = 1 - t^2$ en el problema 2, demostrar que para todo número positivo ε existe un polinomio en t que representa a $|t|$ en el intervalo $-1 \leq t \leq 1$ con un error menor que ε.

4. *a*) Demostrar que si $f(x)$ es continua para $a \leqq x \leqq b$, entonces para todo $\varepsilon > 0$ existe una función poligonal $\varphi(x)$ (esto es, una función continua cuya gráfica consiste en un número finito de segmentos rectilíneos unidos entre sí formando esquinas) tal que $|f(x) - \varphi(x)| < \varepsilon$ para todo valor de x en el intervalo.

b) Probar que toda función poligonal $\varphi(x)$ puede ser representada mediante una suma $\varphi(x) = a + bx + \Sigma c_i |x - x_i|$, donde las x_i son las abscisas de las esquinas.

5. Teorema de aproximación de Weierstrass. Demostrar, sobre la base del aserto anterior, que si $f(x)$ es continua en $a \leq x \leq b$, entonces para todo número positivo ε existe un polinomio $P(x)$ tal que $|f(x) - P(x)| < \varepsilon$ para todos los valores de x en el intervalo $a \leq x \leq b$.

Sugerencia: Aproxímese $f(x)$ por una combinación lineal de la forma $(x - x_r) + |x - x_r|$.

6. Demostrar que los siguientes productos infinitos convergen:

a) $\prod_{n=1}^{\infty} (1 + (\tfrac{1}{2})^{2n})$;

b) $\prod_{n=2}^{\infty} \frac{n^3 - 1}{n^3 + 1}$;

c) $\prod_{n=1}^{\infty} \left(1 - \frac{z^2}{n!}\right)$,

si $|z| < 1$.

7. Probar, usando los métodos del texto, que $\prod_{n=1}^{\infty} \left(1 + \frac{1}{n}\right)$ diverge.

8. Demostrar la identidad

$$\prod_{\nu=1}^{\infty} (1 + x^{2\nu}) = \frac{1}{1 - x}$$

para $|x| < 1$.

***9.** Considérense todos los números naturales cuya representación en el sistema decimal no contenga el número 9 entre sus dígitos. Demostrar que la suma de los inversos de esos números converge.

10. *a*) Demostrar que para $s > 1$ es

$$1 - \frac{1}{2^s} + \frac{1}{3^s} - \frac{1}{4^s} + \cdots = (1 - 2^{1-s})\zeta(s),$$

donde $\zeta(s)$ es la función Zeta definida en la p. 575.

b) Usese esta identidad para probar que $\lim_{s \to 1+} (s - 1)\zeta(s) = 1$.

11. *Criterio integral de convergencia*

a) Sea $f(x)$ una función positiva y decreciente para $x \geqslant 1$. Demostrar que la integral impropia $\int_1^{\infty} f(x)\,dx$ y la serie infinita $\sum_{k=1}^{\infty} f(k)$ convergen o divergen simultáneamente.

b) Demostrar que en cualquiera de esos casos el límite

$$\lim_{n \to \infty} \left(\int^n f(x)\,dx - \sum_{k=1}^{n} f(k) \right)$$

existe.

c) Aplíquese este criterio para probar que la serie

$$\sum_{n=2}^{\infty} \frac{1}{n \log^{\alpha} n}$$

converge cuando $\alpha > 1$ y diverge cuando $\alpha \leq 1$.

8

Series trigonométricas

Las funciones representadas por series de potencias o, como Lagrange las llamó, las *"funciones analíticas"*, desempeñan sin duda un papel muy importante en el Análisis. No obstante, la clase de funciones analíticas es demasiado restringida en muchos casos. Fue por ello de suma importancia para la Matemática, y para una gran variedad de aplicaciones, el que Fourier, en su "Théorie analytique de la chaleur",[1] observara e ilustrara mediante gran número de ejemplos el hecho de que *series trigonométricas* convergentes de la forma

$$(1) \qquad f(x) = \frac{a_0}{2} + \sum_{\nu=1}^{\infty} (a_\nu \cos \nu x + b_\nu \operatorname{sen} \nu x),$$

con coeficientes constantes a_ν, b_ν, fuesen capaces de representar una amplia clase de funciones $f(x)$ "arbitrarias", clase que incluye esencialmente todas las funciones de interés específico, ya sea que estén definidas geométricamente por medios mecánicos o de alguna otra manera. Más aún, es posible expresar en esa forma funciones que tienen discontinuidades de salto o que obedecen leyes de formación diferentes en diferentes intervalos.

Poco después del dramático descubrimiento de Fourier, las "series de Fourier" fueron reconocidas, no sólo como una herramienta de gran utilidad en la Física y en la Mecánica, sino también como una de las fuentes más fructíferas de gran número de bellos resultados puramente matemáticos. Cauchy y, muy especialmente, Dirichlet, entre los años 1820 y 1830 pusieron sobre bases sólidas el razonamiento algo heurístico e incompleto de Fourier, haciendo el tema tan accesible como su importancia lo requería.

A pesar de lo "arbitrario" de las funciones expresables mediante series trigonométricas, aquéllas están esencialmente sujetas a satisfacer la condición de periodicidad con período 2π, puesto que cada término de la serie tiene

[1] Puede verse también, por ejemplo, *The Analytical Theory of Heat*, de Joseph Fourier, versión en inglés, Dover Publications, 1955.

este período. Sin embargo, como veremos después, esta restricción no es esencial, siempre que se considere una función restringida a un intervalo finito a partir del cual pueda hacerse fácilmente la extensión como una función periódica.

Este capítulo constituye una introducción elemental a la teoría de las series de Fourier, dejándose a un lado refinamientos más avanzados.

Después de una discusión preliminar de las funciones periódicas, demostraremos el teorema principal, en el cual se establece la validez del desarrollo trigonométrico para una amplia clase de funciones.

En las secciones subsiguientes discutiremos tópicos algo más avanzados, tales como la convergencia uniforme y absoluta de las series de Fourier y la aproximación mediante polinomios de funciones continuas arbitrarias. En el Apéndice discutiremos la teoría de los polinomios de Bernoulli y sus aplicaciones.

8.1 Funciones periódicas

a. Comentarios generales. Extensión periódica de una función

Las funciones sen nx y cos nx son funciones periódicas de x con período común 2π. Así también, cualquier suma finita, o infinita pero convergente, del tipo (1) es también una función periódica con período 2π. Haremos ahora algunas observaciones generales relativas a las funciones periódicas, ampliando aquéllas que se hicieron en el capítulo 4, p. 355.

La periodicidad de una función $f(x)$ con período T se expresa mediante la ecuación

$$(2a) \qquad f(x + T) = f(x),$$

válida para todos los valores de x.[1] El tener período T implica que $f(x)$ tiene también los períodos $\pm T, \pm 2T, \ldots, \pm mT, \ldots$, y

$$(2b) \qquad f(x \pm mT) = f(x)$$

para todos los enteros m.

En ciertos casos especiales puede ocurrir también que $f(x)$ tenga un período más corto. Por ejemplo, la función sen $(4\pi x/T)$ tiene el período T así como también el período más pequeño $T/2$.

[1] Al representar funciones periódicas, es a menudo conveniente considerar la variable independiente x como un punto sobre una circunferencia, en vez de hacerlo sobre una recta. Para una función $f(x)$ con período 2π, consideramos x como el ángulo en el centro de un círculo de radio unitario, medido entre un radio inicial arbitrario y el radio correspondiente a un punto variable sobre la circunferencia. En esa forma la periodicidad de $f(x)$ significa que a cada punto de la circunferencia le corresponde uno y sólo un valor de la función, aunque el ángulo x mismo está determinado solamente a menos de múltiplos de 2π.

Como vimos en el capítulo 4, p. 355, una función $f(x)$ definida en un intervalo cerrado $a \leq x \leq b$ puede ser desarrollada como una función periódica con período $T = b - a$ para todos los valores de x, definiendo la función en intervalos adyacentes sucesivos de longitud T fuera del intervalo original $a \leq x \leq b$ por la relación de periodicidad

$$f(x + nT) = f(x), \qquad n = \pm 1, \pm 2, \ldots . \tag{2c}$$

La función extendida no está definida en forma única ni es necesariamente continua en los puntos extremos $x = a + nT = b + (n - 1)T$ de los intervalos de longitud T. Podemos admitir funciones $f(x)$ con discontinuidades de salto en los puntos $x = \xi$, que sean continuas a uno y otro lado de ξ pero no necesariamente definidas o continuas en el punto ξ mismo.

Establecido lo anterior, damos las siguientes notaciones y la definición de $f(\xi)$, que serán de utilidad en todo este capítulo: denotamos los límites derecho e izquierdo de $f(x)$ en $x = \xi$, respectivamente, por

$$f(\xi + 0) = \lim_{\varepsilon \to 0} f(\xi + \varepsilon^2), \tag{3a}$$

$$f(\xi - 0) = \lim_{\varepsilon \to 0} f(\xi - \varepsilon^2); \tag{3b}$$

Es conveniente asignar a $f(\xi)$, el valor de f en el punto de discontinuidad, el *valor medio*

$$f(\xi) = \tfrac{1}{2}[f(\xi + 0) + f(\xi - 0)] \tag{4}$$

Con esta convención no quedan ya restricciones para extender periódicamente la función original, del intervalo cerrado $a \leq x \leq b$ a todos los valores de x, aun en los casos en que $f(a) \neq f(b)$. Sólo se necesita tener cuidado con los valores que $f(x)$ toma en los puntos de discontinuidad de salto, en particular, si los valores definidos originalmente para $f(a)$ y $f(b)$ no coinciden. Para definir la extensión periódica tiene que usarse el valor medio $\frac{1}{2}[f(a) + f(b)]$ en lugar de los valores $f(a)$ y $f(b)$.

b. Integrales calculadas sobre un período

La gráfica de una función periódica tiene claramente la misma forma en dos intervalos consecutivos cualesquiera correspondientes a un período. Esto tiene una consecuencia importante: para una función periódica $f(x)$ de período T y para una constante arbitraria a,

$$\int_{-a}^{T-a} f(x)\,dx = \int_0^T f(x)\,dx, \tag{5}$$

o, en palabras: la integral de una función periódica evaluada sobre un intervalo "período" de longitud T tiene siempre el mismo valor, independientemente de la localización del intervalo.

Para demostrar esto sólo se necesita observar que, en virtud de la ecuación $f(\xi - T) = f(\xi)$, la substitución $x = \xi - T$, para α y β cualesquiera, conduce a

$$\int_\alpha^\beta f(x)\,dx = \int_{\alpha+T}^{\beta+T} f(\xi)\,d\xi = \int_{\alpha+T}^{\beta+T} f(x)\,dx.$$

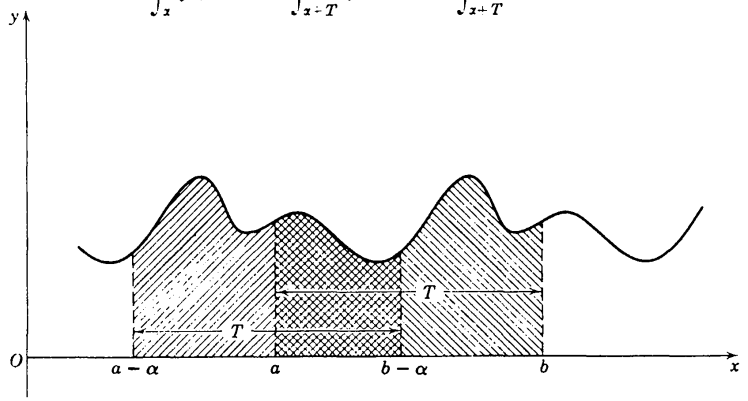


Figura 8.1 Se ilustra la integral calculada sobre un período completo.

En particular, para $\alpha = -a$ y $\beta = 0$ se tiene

$$\int_{-a}^{0} f(x)\,dx = \int_{T-a}^{T} f(x)\,dx$$

y, por tanto,

$$\begin{aligned}\int_{-a}^{T-a} f(x)\,dx &= \int_{-a}^{0} f(x)\,dx + \int_0^{T-a} f(x)\,dx \\ &= \int_{T-a}^{T} f(x)\,dx + \int_0^{T-a} f(x)\,dx \\ &= \int_0^T f(x)\,dx,\end{aligned}$$

como se afirmó. Este resultado se comprende fácilmente por medio de la figura 8.1. Basta recordar el significado geométrico de la integral.

c. Vibraciones armónicas

Las funciones periódicas más simples, a partir de las cuales construiremos las más generales, son las funciones $a \operatorname{sen} \omega x$ y $a \cos \omega x$, o bien, con

mayor generalidad, $a \operatorname{sen} \omega(x - \xi)$ y $a \cos \omega(x - \xi)$, en donde $a\ (\geq 0)$, $\omega\ (> 0)$ y ξ son constantes. Estas funciones representan *"vibraciones sinusoidales"* o *vibraciones (u oscilaciones) armónicas simples.*[1] El *período de vibración* es $T = 2\pi/\omega$. El número ω se llama la *frecuencia angular* o *circular* de las vibraciones.[2] Puesto que $1/T = \omega/2\pi$ es el número de vibraciones en la unidad de tiempo, esto es, la frecuencia, ω es el número de vibraciones en el tiempo 2π. El número a se llama la *amplitud* de la vibración; representa el máximo valor de la función $a \operatorname{sen} \omega(x - \xi)$ o bien de $a \cos \omega(x - \xi)$, pues tanto el seno como el coseno tienen valor máximo igual a 1. Los números $\omega(x - \xi)$ y $\omega\xi$ se llaman, respectivamente, la *fase* y el desplazamiento o *corrimiento de fase.*

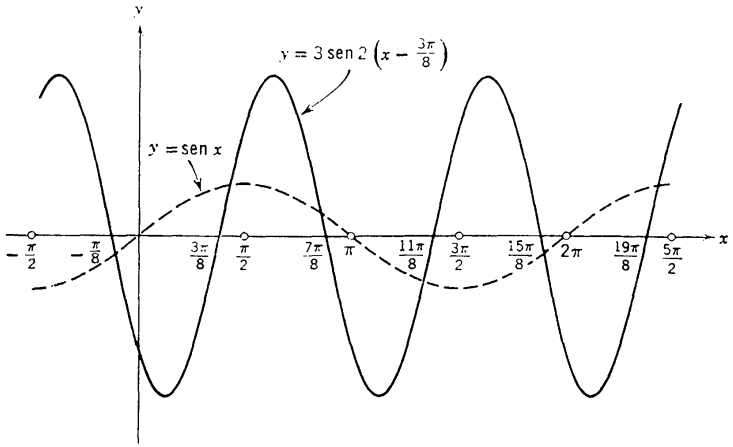


Figura 8.2 Vibraciones sinusoidales.

Las funciones $a \operatorname{sen} \omega(x - \xi)$ se obtienen gráficamente "estirando" la curva *seno* en las proporciones $1 : \omega$ a lo largo del eje x y $a : 1$ a lo largo del eje y, y trasladando después la curva una distancia ξ en el sentido positivo del eje x (ver Fig. 8.2).

Mediante las fórmulas de adición para funciones trigonométricas es también posible representar las vibraciones armónicas anteriores por $\alpha \cos \omega x + \beta \operatorname{sen} \omega x$ y $\beta \cos \omega x - \alpha \operatorname{sen} \omega x$, respectivamente, siendo $\alpha = -a \operatorname{sen} \omega\xi$ y $\beta = a \cos \omega\xi$. Recíprocamente, toda función de la forma $\alpha \cos \omega x + \beta \operatorname{sen} \omega x$ representa una vibración sinusoidal $a \operatorname{sen} \omega(x - \xi)$ de amplitud $a = \sqrt{\alpha^2 + \beta^2}$

[1] Cualquiera de estas dos fórmulas representa por sí misma (para todos los valores de a y ξ) el conjunto de todas las vibraciones sinusoidales. Las dos fórmulas son equivalentes, pues $a \operatorname{sen} \omega(x - \xi) = a \cos \omega \left[x - \left(\xi + \frac{\pi}{2\omega} \right) \right]$.

[2] Obsérvese que distinguimos entre *"frecuencia"* y *"frecuencia angular o circular"*.

y *corrimiento de fase* $\omega\xi$ dados por las ecuaciones $\alpha = -a \operatorname{sen} \omega\xi$, y $\beta = a \cos \omega\xi$. Usando la expresión $\alpha \cos \omega x + \beta \operatorname{sen} \omega x$ podemos escribir inmediatamente la suma de dos o más de tales funciones que tengan la misma frecuencia angular ω en la forma de otra vibración con frecuencia angular ω.

Como vimos anteriormente, las funciones periódicas aparecen cuando se desea representar en forma paramétrica una curva cerrada. Naturalmente, tales funciones pueden usarse para representar fenómenos inducidos por un movimiento circular, por ejemplo, un proceso que se repite periódicamente en sincronía con un volante. Además, están asociadas a todos los fenómenos de vibración.

8.2 Superposición de vibraciones armónicas

a. Armónicas. Polinomios trigonométricos

Aunque muchas vibraciones son puramente sinusoidales (ver p. 422), la mayoría de los movimientos periódicos tienen un carácter más complejo, siendo obtenidos por "superposición" de varias vibraciones sinusoidales. Matemáticamente, el movimiento de un punto sobre una recta, cuya coordenada x es una función del tiempo, puede representarse mediante una función que sea la suma de cierto número de funciones periódicas puras del tipo considerado antes. Las componentes armónicas de la función se superponen (esto es, sus ordenadas se suman). En esta superposición se considera que las frecuencias angulares (y, claro está, los períodos) de las vibraciones superpuestas son todas diferentes, pues la superposición de dos vibraciones sinusoidales de la misma frecuencia angular conduce a otra vibración sinusoidal de igual frecuencia, como se demostró con anterioridad.

Para la superposición de dos vibraciones sinusoidales con diferentes frecuencias angulares, ω_1 y ω_2, existen dos posibilidades fundamentalmente distintas, dependiendo de que ω_1/ω_2 sea o no racional, o, como también se dice, de que las frecuencias sean *conmensurables* o *inconmensurables.*

Como un ejemplo de la primera posibilidad consideremos el caso en que la segunda frecuencia angular es dos veces la primera: $\omega_2 = 2\omega_1$. El período de la segunda vibración es entonces la mitad del período de la primera, $2\pi/2\omega_1 = T_2 = T_1/2$, de manera que aquélla no sólo tiene el período T_2 sino que también tiene el período doble, T_1, pues la función se repite después de este doble período. La función formada por superposición deberá, de la misma manera, tener el período T_1. La segunda vibración, cuya frecuencia angular es doble y cuyo período es la mitad del primero, se llama una *primera armónica* de la primera vibración (la fundamental).

Asertos correspondientes son válidos si se introduce una tercera vibración de frecuencia angular $\omega_3 = 3\omega_1$. Aquí nuevamente la función $\operatorname{sen} 3\omega_1 x$ necesariamente se repite con período $2\pi/\omega_1 = T_1$. Esta vibración se llama

una *segunda armónica* de la vibración dada. Análogamente, podemos considerar las armónicas tercera, cuarta, ..., $(n-1)$-ésima, con frecuencias angulares $\omega_4 = 4\omega_1$, $\omega_5 = 5\omega_1$, ..., $\omega_n = n\omega_1$, y corrimientos de fase cualesquiera. Cualquiera de tales armónicas necesariamente se repite después del período $T_1 = 2\pi/\omega_1$ y, en consecuencia, toda función obtenida por superposición de cierto número de vibraciones, armónicas de una fundamental dada de frecuencia angular ω_1, es a su vez una función periódica con período $2\pi/\omega_1 = T_1$. Superponiendo vibraciones cuyas frecuencias angulares varíen desde la de la fundamental hasta la de la $(n-1)$-ésima armónica, se obtiene una función periódica en la forma de un *polinomio trigonométrico,*

$$S_n(x) = \frac{a_0}{2} + \sum_{\nu=1}^{n} (a_\nu \cos \nu\omega x + b_\nu \operatorname{sen} \nu\omega x). \tag{6}$$

(La constante $a_0/2$, que no afecta la periodicidad, se deja indeterminada por las razones de conveniencia que veremos después.) Como esta función contiene $2n + 1$ constantes arbitrarias, a_ν, b_ν, estamos en condiciones de generar curvas que en nada se parezcan a las sinusoides originales. Las figuras 8.3 a 8.5 constituyen ilustraciones gráficas.

El término "armónica" hace alusión a la Acústica,[1] en donde una vibración fundamental de frecuencia angular ω corresponde a un tono de cierta altura, y las armónicas primera, segunda, tercera, etc., corresponden a la sucesión de armónicas de la fundamental, es decir, a la primera octava, a la segunda octava, etc.

En general, cuando se superponen vibraciones cuyas frecuencias angulares guardan relaciones racionales entre sí, es posible representar dichas frecuencias como múltiplos enteros de una frecuencia fundamental común.

La superposición de dos vibraciones de frecuencias angulares ω_1 y ω_2 inconmensurables representa, sin embargo, un fenómeno diferente. En este caso la superposición de vibraciones sinusoidales ya no es periódica. Sin entrar en una discusión detallada, diremos que tales funciones tienen un carácter "aproximadamente periódico" o, como a veces se dice, son *cuasi periódicas.*

*b. Batidos

En un comentario final acerca de la superposición de vibraciones sinusoidales, nos referiremos al fenómeno de los llamados *batidos.* Si se superponen dos vibraciones de amplitud unitaria pero de frecuencias angulares ω_1 y ω_2, diferentes entre sí, y si, por simplicidad, se toma el mismo valor

[1] En Acústica se usa también el término *sobretono.*

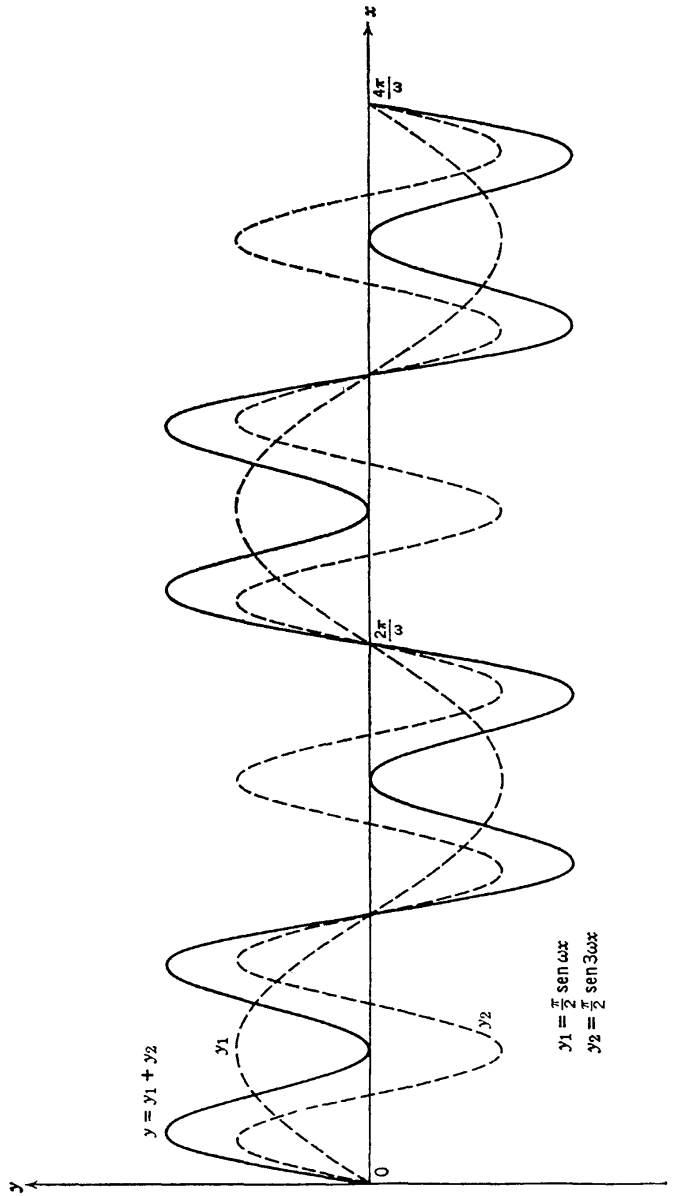


Figura 8.3 Combinación de vibraciones. (Las proporciones de la figura corresponden a la hipótesis $\omega = 1$).

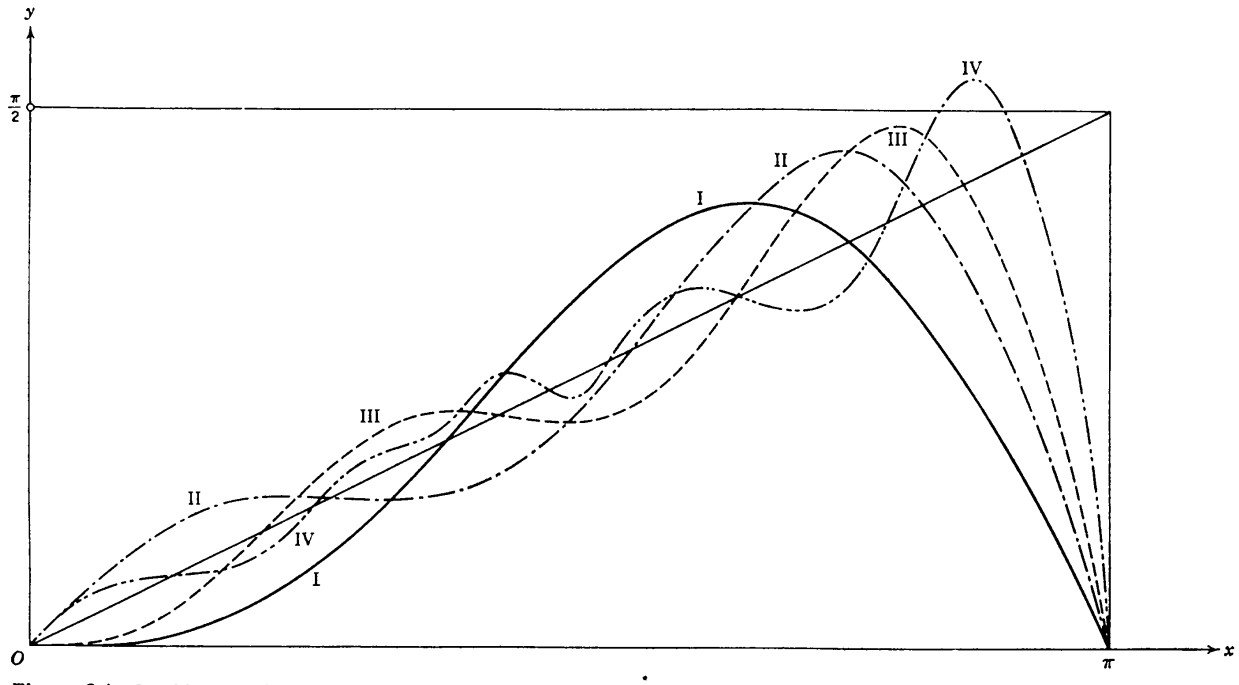


Figura 8.4 Combinación de vibraciones: I: $\operatorname{sen} x - (\operatorname{sen} 2x)/2$; II: $\operatorname{sen} x - (\operatorname{sen} 2x)/2 + (\operatorname{sen} 3x)/3$; III: $\operatorname{sen} x - (\operatorname{sen} 2x)/2 + (\operatorname{sen} 3x)/3 - (\operatorname{sen} 4x)/4$; IV: $\operatorname{sen} x - (\operatorname{sen} 2x)/2 + (\operatorname{sen} 3x)/3 - (\operatorname{sen} 4x)/4 + (\operatorname{sen} 5x)/5 - (\operatorname{sen} 6x)/6 + (\operatorname{sen} 7x)/7$.

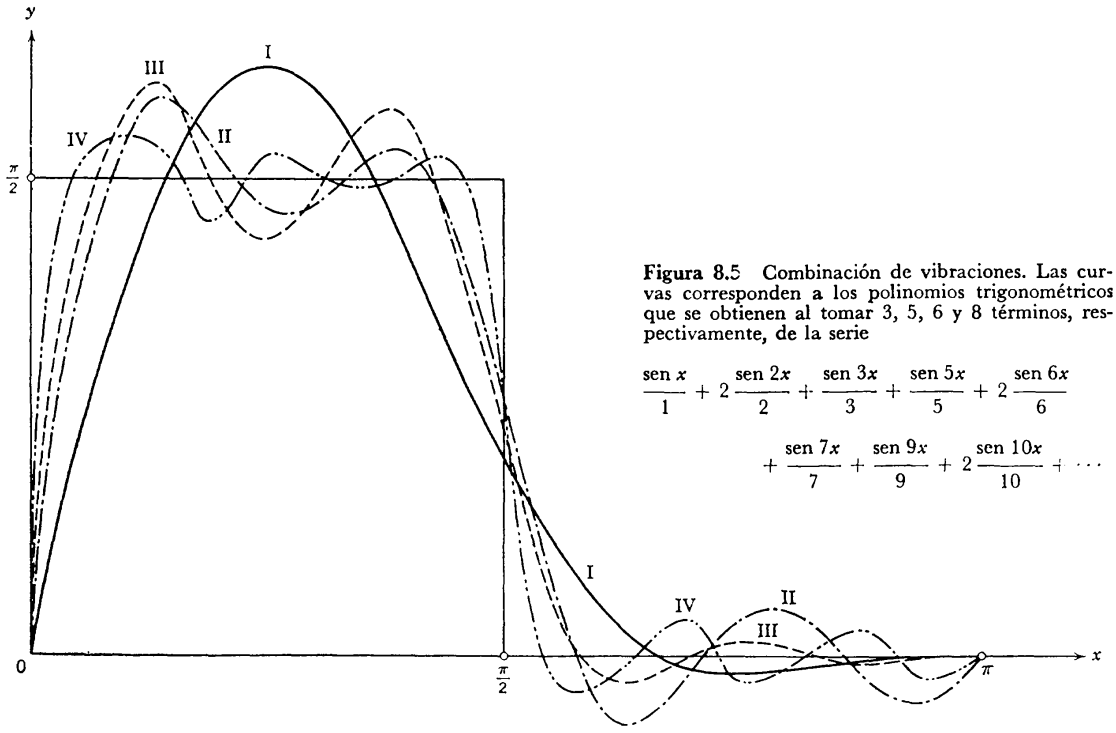


Figura 8.5 Combinación de vibraciones. Las curvas corresponden a los polinomios trigonométricos que se obtienen al tomar 3, 5, 6 y 8 términos, respectivamente, de la serie

$$\frac{\operatorname{sen} x}{1} + 2\frac{\operatorname{sen} 2x}{2} + \frac{\operatorname{sen} 3x}{3} + \frac{\operatorname{sen} 5x}{5} + 2\frac{\operatorname{sen} 6x}{6} + \frac{\operatorname{sen} 7x}{7} + \frac{\operatorname{sen} 9x}{9} + 2\frac{\operatorname{sen} 10x}{10} + \cdots$$

de ξ (ver p. 591) para ambas (la generalización al caso de fases arbitrarias se deja al lector), resulta entonces la función

$$y = \operatorname{sen} \omega_1 x + \operatorname{sen} \omega_2 x \qquad (\omega_1 > \omega_2 > 0).$$

Por una bien conocida fórmula trigonométrica se tiene

$$y = 2 \cos [\tfrac{1}{2}(\omega_1 - \omega_2)x] \operatorname{sen} [\tfrac{1}{2}(\omega_1 + \omega_2)x].$$

Esta ecuación representa un fenómeno que se describe como sigue: considérese una vibración de frecuencia angular $\frac{1}{2}(\omega_1 + \omega_2)$ y de período

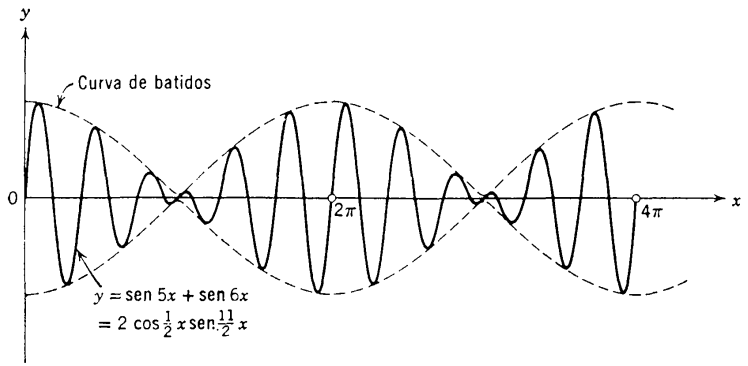


Figura 8.6 Batidos.

$4\pi/(\omega_1 + \omega_2)$. Tal vibración no tiene una amplitud constante sino una "amplitud" variable dada por la expresión

$$2 \cos [\tfrac{1}{2}(\omega_1 - \omega_2)x],$$

que varía con un período mayor: $4\pi/(\omega_1 - \omega_2)$. Esta descripción es particularmente útil cuando las dos frecuencias angulares, ω_1 y ω_2, son relativamente grandes mientras que su diferencia $(\omega_1 - \omega_2)$, es relativamente pequeña. En este caso la amplitud $2 \cos [\frac{1}{2}(\omega_1 - \omega_2)x]$ de la vibración, cuyo período es $4\pi/(\omega_1 + \omega_2)$, varía lentamente en comparación con el período de vibración, y este cambio de amplitud se repite en forma periódica con un período, $4\pi/(\omega_1 - \omega_2)$, grande. Tales cambios rítmicos de amplitud se denominan *batidos*. Este fenómeno se presenta con frecuencia en Acústica y en Electrónica. En radiotransmisión, las frecuencias ω_1 y ω_2 caen, por regla general, en un rango muy por encima del que puede detectar el oído humano, en tanto que la diferencia $\omega_1 - \omega_2$ cae en el rango de notas audibles. Los batidos producen entonces un tono audible aunque las vibraciones originales sean imperceptibles por el oído.

En la Fig. 8.6 se ilustra un caso en que se tienen batidos.

8.3 Notación compleja

a. Comentarios generales

Las operaciones con funciones trigonométricas se simplifican a menudo usando números complejos de acuerdo con la relación de Euler:

$$\cos \theta + i \operatorname{sen} \theta = e^{i\theta},$$

o sea,

$$(7a) \qquad \cos \theta = \tfrac{1}{2}(e^{i\theta} + e^{-i\theta}),$$

$$(7b) \qquad \operatorname{sen} \theta = \frac{1}{2i}(e^{i\theta} - e^{-i\theta}).$$

(Compárese con capítulo 7, p. 566.) Según esto, es posible representar las vibraciones sinusoidales en términos de las cantidades complejas $e^{i\omega x}$, $e^{-i\omega x}$, o $ae^{i\omega(x-\xi)}$, $ae^{-i\omega(x-\xi)}$, siendo a, ω y $\omega\xi$ la amplitud, la frecuencia angular y el corrimiento de fase, respectivamente. Como cabría esperar, las vibraciones reales se obtienen a partir de las expresiones complejas simplemente separando partes real e imaginaria.

Una de las ventajas de la notación compleja consiste en que las derivadas con respecto al tiempo se obtienen derivando la función exponencial compleja como si i fuese una constante real. La fórmula

$$\begin{aligned} \frac{d}{dx} a[\cos \omega(x - \xi) + i \operatorname{sen} \omega(x - \xi)] \\ = a\omega[-\operatorname{sen} \omega(x - \xi) + i \cos \omega(x - \xi)] \\ = ia\omega[\cos \omega(x - \xi) + i \operatorname{sen} \omega (x - \xi)], \end{aligned}$$

que se obtiene a partir de las fórmulas para las derivadas de las funciones seno y coseno, puede escribirse en la forma concisa

$$(8) \qquad \frac{d}{dx} ae^{i\omega(x-\xi)} = ia\omega e^{i\omega(x-\xi)}.$$

La integral de una función compleja $\gamma(x)$, digamos $\gamma(x) = p(x) + iq(x)$, está definida en forma natural por

$$\int \gamma(x)\, dx = \int p(x)\, dx + i \int q(x)\, dx.$$

Según esto, para $n \neq 0$ se tiene

$$\begin{aligned} \int e^{inx}\, dx &= \int \cos nx\, dx + i \int \operatorname{sen} nx\, dx \\ &= \frac{1}{n} \operatorname{sen} nx - \frac{i}{n} \cos nx = \frac{1}{in} e^{inx}. \end{aligned}$$

En particular, para cualquier entero n resulta

$$\int_{-\pi}^{\pi} e^{inx}\,dx = \begin{cases} 0 & \text{para } n \neq 0 \\ 2\pi & \text{para } n = 0. \end{cases}$$

Más generalmente, teniendo en cuenta que $e^{inx}e^{-imx} = e^{i(n-m)x}$, se obtiene para m y n enteros cualesquiera

$$\int_{-\pi}^{\pi} e^{inx}e^{-imx}\,dx = \begin{cases} 0, & \text{para } n \neq m \\ 2\pi & \text{para } n = m. \end{cases} \tag{9}$$

Estas relaciones son simplemente expresiones condensadas de las relaciones de ortogonalidad entre funciones trigonométricas (ver p. 295).

*b. Aplicación al caso de corrientes alternas

Ilustramos aquí las ideas anteriores mediante un ejemplo importante. Denotaremos la variable independiente, el tiempo, con t, en vez de hacerlo con x.

Consideremos un circuito eléctrico con resistencia R e inductancia L, al cual se aplica una "fuerza electromotriz" (voltaje) externa E. Si se trata de corriente "continua" o directa, el voltaje E es constante y la corriente I está dada por la ley de Ohm: $E = RI$. En el caso de corriente alterna, sin embargo, E (y, por lo tanto, I) es una función del tiempo t, y la *ley de Ohm* toma la *forma generalizada* (ver p. 653)

$$E - L\frac{dI}{dt} = RI. \tag{10}$$

Supongamos que la fuerza electromotriz externa, E, es sinusoidal, con frecuencia angular ω, y está dada por $\varepsilon\cos\omega t$, o bien por $\varepsilon\,\text{sen}\,\omega t$, y combinemos formalmente ambas posibilidades en la forma compleja

$$E = \varepsilon e^{i\omega t} = \varepsilon\cos\omega t + i\varepsilon\,\text{sen}\,\omega t,$$

donde ε representa la amplitud. Es a menudo útil admitir valores *complejos* también para la *amplitud:*

$$\varepsilon = |\varepsilon|\,e^{-i\eta}.$$

Entonces,

$$E = |\varepsilon|\,e^{i(\omega t-\eta)} = |\varepsilon|\,\{\cos(\omega t - \eta) + i\,\text{sen}\,(\omega t - \eta)\}.$$

Podemos operar con este *"voltaje complejo"* E y la correspondiente *"corriente compleja"* I como si i fuese un parámetro real. El significado de la relación compleja entre las cantidades complejas E e I es entonces que la corriente que corresponde a una fuerza electromotriz $\varepsilon\cos\omega t$ es la parte real de I, en tanto que la correspondiente a una fuerza electromotriz $\varepsilon\,\text{sen}\,\omega t$ es la parte imaginaria de I. La corriente compleja está dada por una expresión de la forma

$$I = \alpha e^{i\omega t} = \alpha(\cos\omega t + i\,\text{sen}\,\omega t),$$

que también es sinusoidal con frecuencia angular ω. La derivada de I está dada entonces formalmente por

$$\frac{dI}{dt} = i\,\alpha\omega e^{i\omega t}$$

$$= \alpha\omega(-\,\text{sen}\,\omega t + i\cos\omega t) = i\omega I.$$

Substituyendo estas cantidades en la forma *generalizada* de la *ley de Ohm* (Ec. 10), y dividiendo por $e^{i\omega t}$, se obtiene la ecuación

$$\varepsilon - \alpha Li\omega = R\alpha,$$

o sea,

$$\alpha = \frac{\varepsilon}{R + i\omega L},$$

así como también

$$E = (R + i\omega L)I = WI.$$

Podemos considerar esta última ecuación como la *ley de Ohm* en forma compleja para corrientes alternas, llamando a la cantidad

$$W = R + i\omega L$$

resistencia compleja del circuito. La ley de Ohm tiene entonces la misma forma para corrientes "continuas": la corriente es igual al voltaje dividido por la resistencia.

Escribiendo la resistencia compleja W en la forma

$$W = we^{i\delta} = w\cos\delta + iw\,\text{sen}\,\delta,$$

donde

$$|W| = w = \sqrt{(R^2 + L^2\omega^2)}, \qquad \tan\delta = \frac{\omega L}{R},$$

obtenemos

$$I = \frac{\varepsilon}{w}e^{i(\omega t-\delta)} = \frac{E}{W}.$$

De acuerdo con esta fórmula, la corriente tiene el mismo período (y la misma frecuencia angular) que el voltaje. La amplitud, α, de la corriente está relacionada con la amplitud, ε, de la fuerza electromotriz, para ε real, mediante la expresión

$$\alpha = \frac{\varepsilon}{w}$$

Además, existe una diferencia de fase entre la corriente y el voltaje. La corriente no alcanza su máximo al mismo tiempo que el voltaje sino un tiempo δ/ω más tarde y, por supuesto, lo mismo es cierto para el mínimo. En ingeniería eléctrica la cantidad $w = \sqrt{R^2 + L^2\omega^2}$ se llama a menudo la *impedancia* o *resistencia en corriente alterna* del circuito, para la frecuencia angular ω. El corrimiento de fase, dado por lo general en grados, se llama algunas veces *defasaje.*

Si la amplitud ε es compleja y tiene la forma

$$\varepsilon = |\varepsilon|\,e^{-i\eta},$$

nada esencial cambia en la forma de la ley de Ohm, excepto que η es un corrimiento de fase adicional y se tiene entonces

$$E = |\varepsilon|\,e^{i(\omega t-\eta)},$$

$$I = \frac{E}{W} = \frac{|\varepsilon|}{|W|}e^{i\omega t}e^{-i(\delta+\eta)}.$$

c. Notación compleja para polinomios trigonométricos

Una vibración compuesta del tipo

$$S_n(x) = \tfrac{1}{2}a_0 + \sum_{\nu=1}^{n}(a_\nu\cos\nu x + b_\nu\,\text{sen}\,\nu x) \tag{11}$$

(donde, por brevedad, se ha tomado $\omega = 1$) puede reducirse a la forma compleja mediante las substituciones

$$\cos \nu x = \tfrac{1}{2}(e^{i\nu x} + e^{-i\nu x}) \qquad \text{y} \qquad \operatorname{sen} \nu x = -\tfrac{1}{2}i(e^{i\nu x} - e^{-i\nu x}).$$

La primera expresión toma entonces la forma más simple

$$S_n(x) = \sum_{\nu=-n}^{n} \alpha_\nu e^{i\nu x}, \tag{12}$$

en la que los números complejos α_ν están relacionados con los números reales a_0, a_ν y b_ν por medio de las ecuaciones

$$\begin{cases} \alpha_\nu = \tfrac{1}{2}(a_\nu - ib_\nu), \\ \alpha_{-\nu} = \tfrac{1}{2}(a_\nu + ib_\nu), \qquad \text{para} \quad \nu = 1, 2, \ldots, n, \\ \alpha_0 = \tfrac{1}{2}a_0. \end{cases} \tag{13a}$$

Al resolver estas ecuaciones para los a_ν y b_ν, se encuentra que

$$\begin{aligned} a_\nu &= \alpha_\nu + \alpha_{-\nu}, \\ b_\nu &= i(\alpha_\nu - \alpha_{-\nu}). \end{aligned} \tag{13b}$$

(Queda incluido el caso $\nu = 0$.)

Recíprocamente, cualquier expresión de la forma

$$\sum_{\nu=-n}^{n} \alpha_\nu e^{i\nu x}$$

puede considerarse como una función que representa la superposición de vibraciones escritas en forma compleja. El resultado de esta superposición es real si, y sólo si, $\alpha_\nu + \alpha_{-\nu}$ es real y $\alpha_\nu - \alpha_{-\nu}$ es imaginario puro; o sea, si α_ν y $\alpha_{-\nu}$ son números complejos conjugados.

d. Una fórmula trigonométrica

Como aplicación de la notación compleja, demostraremos la identidad siguiente:

$$\begin{aligned} \sigma_n(\alpha) &= \tfrac{1}{2} + \cos\alpha + \cos 2\alpha + \cdots + \cos n\alpha \\ &= \frac{\operatorname{sen}(n + \tfrac{1}{2})\alpha}{2 \operatorname{sen} \tfrac{1}{2}\alpha}, \end{aligned} \tag{14}$$

que necesitaremos más adelante en este mismo capítulo. La fórmula tiene sentido solamente cuando $\operatorname{sen} \tfrac{1}{2}\alpha \neq 0$, es decir, cuando α no tiene alguno de los valores $0, \pm 2\pi, \pm 4\pi, \ldots$. Sin embargo, una vez establecida la fórmula para $\operatorname{sen} \tfrac{1}{2}\alpha \neq 0$, es posible concluir que la expresión $[\operatorname{sen}(n + \tfrac{1}{2})\alpha]/(2 \operatorname{sen} \tfrac{1}{2}\alpha)$ es una función continua de α, para todo α, si se define su valor en los puntos excepcionales como el valor de $\sigma_n(\alpha)$, es decir, $n + \tfrac{1}{2}$.

Para la demostración, reemplazamos la función coseno por su expresión exponencial [ver fórmula (13*a*) con $a_\nu = 1$, $b_\nu = 0$]:

$$\sigma_n(\alpha) = \tfrac{1}{2} \sum_{\nu=-n}^{n} e^{i\nu\alpha}.$$

En el lado derecho tenemos una progresión geométrica con el cociente común $q = e^{i\alpha} = \cos\alpha + i \operatorname{sen}\alpha$. Por tanto, q puede tomar el valor 1 sólo si $\cos\alpha = 1$ y $\operatorname{sen}\alpha = 0$, o sea, si α tiene uno de los valores excepcionales $0, \pm 2\pi, \pm 4\pi, \ldots$. Cuando α tiene cualquier otro valor la fórmula ordinaria para la suma conduce a

$$\begin{aligned} \sigma_n(\alpha) &= \tfrac{1}{2}e^{-in\alpha}\frac{1 - q^{2n+1}}{1 - q} \\ &= \tfrac{1}{2} \cdot \frac{e^{-in\alpha} - e^{(n+1)i\alpha}}{1 - e^{i\alpha}}. \end{aligned}$$

Multiplicando el numerador y el denominador por $e^{-i\alpha/2}$ se obtiene

$$\sigma_n(\alpha) = \frac{\operatorname{sen}(n + \tfrac{1}{2})\alpha}{2 \operatorname{sen} \tfrac{1}{2}\alpha},$$

como se afirmó antes.

Al integrar $\sigma_n(t)$ sobre $0 \leq t \leq \pi$, se encuentra, independientemente de n, el útil resultado siguiente:

$$\begin{aligned} \int_0^\pi \frac{\operatorname{sen}(n + \tfrac{1}{2})t}{2 \operatorname{sen} \tfrac{1}{2}t}\, dt &= \int_0^\pi \left(\tfrac{1}{2} + \sum_{\nu=1}^{n} \cos \nu t \right) dt, \\ &= \tfrac{1}{2}\pi \end{aligned} \tag{15}$$

pues la integral de cada término de la serie se anula.

8.4 Series de Fourier

a. Coeficientes de Fourier

Los *polinomios trigonométricos* de orden n,

$$f(x) = S_n(x) = \tfrac{1}{2}a_0 + \sum_{\nu=1}^{n} (a_\nu \cos \nu x + b_\nu \operatorname{sen} \nu x) \tag{16}$$

dependen de los $2n + 1$ coeficientes a_ν y b_ν. Es notable que estos *"coeficientes de Fourier"* puedan ser expresados en forma sencilla en términos de los valores $f(x)$ de la suma, mediante las fórmulas

$$a_\mu = \frac{1}{\pi}\int_{-\pi}^{\pi} f(x) \cos \mu x\, dx, \; b_\mu = \frac{1}{\pi}\int_{-\pi}^{\pi} f(x) \operatorname{sen} \mu x\, dx. \tag{17}$$

La demostración se realiza multiplicando (16) por $\cos \mu x$ o por sen μx, e integrando después. Las relaciones de ortogonalidad (ver p. 295) conducen de inmediato a estas fórmulas, puesto que sólo los términos con $\nu = \mu$ dan una contribución no nula.

En la terminología de los números complejos,

$$(16a) \qquad f(x) = S_n(x) = \sum_{\nu=-1}^{n} \alpha_\nu e^{i\nu x},$$

con $$a_\nu = \alpha_\nu + \alpha_{-\nu}; \qquad b_\nu = i(\alpha_\nu - \alpha_{-\nu}).$$

Las expresiones correspondientes para los coeficientes complejos de Fourier son

$$(17a) \qquad \alpha_\nu = \frac{1}{2\pi}\int_{-\pi}^{\pi} f(x) e^{-i\nu x}\, dx,$$

como se deduce también de las relaciones complejas de ortogonalidad [(9), p. [599].

De paso se nota que el factor $\frac{1}{2}$ en la notación para el término constante $\frac{1}{2}a_0$ de (16) sirve sólo para hacer que la fórmula (17) sea válida también para $\nu = 0$.

El teorema principal relativo a las series de Fourier surge ahora de una pregunta natural: al hacer que el grado n del polinomio de Fourier (16) tienda a infinito ¿se tiene la posibilidad de representar funciones $f(x)$ periódicas con período 2π pero, por lo demás, esencialmente arbitrarias?

El resultado de los siguientes párrafos será sin duda que: *Cualquier función periódica $f(x)$, continua a trozos y con derivadas de primero y segundo orden continuas a trozos, puede representarse mediante una "serie de Fourier" infinita,*

$$f(x) = \frac{a_0}{2} + \sum_{\nu=1}^{\infty} (a_\nu \cos \nu x + b_\nu \operatorname{sen} \nu x),$$

o bien, en notación compleja,

$$f(x) = \sum_{\nu=-\infty}^{\infty} \alpha_\nu e^{i\nu x};$$

con coeficientes dados por (17) y (17a).

b. Lema básico

En primer lugar, recordemos que, por definición, una función *continua a trozos* en un intervalo es una función continua excepto en un número finito de discontinuidades por salto dentro del intervalo.

Recordemos además que el valor de una función periódica $f(x)$ en un punto de discontinuidad se ha definido [Ec. (4), p. 589] como el promedio de los valores límite derecho e izquierdo.

Una función $f(x)$ es *continua a trozos* y tiene primera y segunda derivadas continuas a trozos si es posible dividir el intervalo total en un número finito de subintervalos, de manera tal que f, f' y f'' sean continuas en cada subintervalo abierto y se aproximen a límites definidos en los puntos extremos.

La clave para la demostración del teorema principal es un hecho muy sencillo, establecido por el siguiente Lema.

LEMA. *Si una función, $k(x)$, y su primera derivada, $k'(x)$, son continuas a trozos en el intervalo $a \leq x \leq b$, entonces la integral*

$$K_\lambda = \int_a^b k(x) \operatorname{sen} \lambda x\, dx$$

tiende a cero cuando $\lambda \to \infty$.

DEMOSTRACIÓN. Para demostrar este lema hacemos uso de la integración por partes. Supóngase que k y k' son continuas en $a \leq x \leq b$. Se tiene entonces

$$(18) \qquad K_\lambda = \int_a^b k(x) \operatorname{sen} \lambda x\, dx$$
$$= \frac{1}{\lambda}\left[k(a) \cos \lambda a - k(b) \cos \lambda b + \int_a^b k'(x) \cos \lambda x\, dx \right]$$

Es claro que cuando λ crece el lado derecho tiende a cero. Si $k(x)$ o $k'(x)$ tuviesen discontinuidades de salto (ξ) en el intervalo, dividiríamos éste en partes por medio de ξ, aplicaríamos nuestro argumento a cada parte y luego sumaríamos los resultados.

Sin demostrarlo, diremos que el lema sigue siendo válido, en realidad, sin necesidad de la hipótesis que se refiere a la existencia de la derivada $k'(x)$. La continuidad a trozos de k es suficiente. La demostración con estas condiciones más débiles se basa en el hecho de que para $\lambda \neq 0$ la función sen λx es alternadamente positiva y negativa en intervalos consecutivos de longitud π/λ. Para valores grandes de λ las contribuciones a la integral provenientes de intervalos contiguos casi se cancelan entre sí como consecuencia de la continuidad de $k(x)$.

c. Demostración de que $\displaystyle\int_0^\infty \frac{\operatorname{sen} z}{z}\, dz = \frac{\pi}{2}$

Como una aplicación del Lema, evaluaremos la integral

$$(19) \qquad I = \int_0^\infty \frac{\operatorname{sen} z}{z}\, dz.$$

Esta integral impropia está definida por la relación

$$I = \lim_{M \to \infty} I_M$$

en la que

$$I_M = \int_0^M \frac{\operatorname{sen} z}{z}\, dz.$$

La convergencia de la integral impropia I, es decir, la existencia del límite de I_M para $M \to \infty$, ha sido demostrada ya en la p. 329. Dicha demostración se basó en la integración por partes y puede reproducirse aquí. Si, por ejemplo, $0 < M < N$, se tiene

$$(20) \qquad \begin{aligned} |I_N - I_M| &= \left| \int_M^N \frac{\operatorname{sen} z}{z}\, dz \right| \\ &= \left| -\frac{\cos z}{z} \Big|_M^N + \int_M^N \frac{\cos z}{z^2}\, dz \right| \\ &\leq \frac{1}{M} + \frac{1}{N} + \int_M^N \frac{dz}{z^2} = \frac{2}{M}. \end{aligned}$$

Puesto que I_N e I_M difieren en una cantidad arbitrariamente pequeña cuando M y N son suficientemente grandes, la existencia de $I = \lim_{M\to\infty} I_M$ queda garantizada por el criterio de convergencia de Cauchy. Además, haciendo que N tienda a infinito en (20) se encuentra una estimación para la "rapidez" con que I_M se acerca a su límite I:

$$(20a) \qquad |I - I_M| \leq \frac{2}{M}.$$

La expresión para I puede reescribirse de manera tal que I aparezca como un límite de integrales sobre un intervalo finito fijo. Sea p un número positivo arbitrario. Para $M = \lambda p$, con la substitución $z = \lambda x$, $dz = \lambda\, dx$ se obtiene

$$I_{\lambda p} = \int_0^{\lambda p} \frac{\operatorname{sen} z}{z}\, dz = \int_0^p \frac{\operatorname{sen} \lambda x}{x}\, dx.$$

Ahora bien, puesto que $\lambda p \to \infty$ para $\lambda \to \infty$ y un número positivo fijo p, resulta claro que

$$I = \lim_{\lambda\to\infty} \int_0^p \frac{\operatorname{sen} \lambda x}{x}\, dx.$$

O, en forma más precisa, de (20*a*) se concluye que

$$\left| I - \int_0^p \frac{\operatorname{sen} \lambda x}{x}\, dx \right| < \frac{2}{\lambda p}.$$

Así, para cualquier número positivo p las expresiones

$$\int_0^p \frac{\operatorname{sen} \lambda x}{x}\, dx$$

tienden a uno y el mismo límite, I, cuando $\lambda \to \infty$. Además, la convergencia es uniforme en p para valores de p mayores que cierto número positivo fijo, P. Sin duda, la diferencia entre la integral y el límite I es entonces menor que ε para $\lambda > 2/P\varepsilon$.

Apliquemos ahora el Lema de la p 604 a la función

$$k(x) = \frac{1}{x} - \frac{1}{2 \operatorname{sen} (x/2)}.$$

Definiendo $k(0) = 0$, la función $k(x)$ es continua y tiene una primera derivada continua para $0 \leq x < 2\pi$ (ver p. 483). Por tanto, el Lema muestra que

$$\int_0^p \operatorname{sen} \lambda x \left(\frac{1}{x} - \frac{1}{2 \operatorname{sen} (x/2)} \right) dx$$

tiende a cero para $\lambda \to \infty$, siempre que $0 \leq p < 2\pi$. Además, por (18), la convergencia es uniforme si $0 \leq p < \pi$, puesto que $|k(x)|$ y $|k'(x)|$ son acotadas en el intervalo $0 \leq x \leq \pi$. Del resultado anterior se deduce que para cualquier p en el intervalo $0 < p < 2\pi$ se tiene

$$\lim_{\lambda\to\infty} \int_0^p \frac{\operatorname{sen} \lambda x}{2 \operatorname{sen} (x/2)}\, dx = I,$$

y también que la convergencia es uniforme en p sin $P \leq p \leq \pi$, siendo P un número positivo fijo.

Ahora bien, esta integral ya ha sido evaluada [ver fórmula (15), p. 602] en el caso en que $p = \pi$ y $\lambda = n + \frac{1}{2}$; y el resultado es el valor $\pi/2$, independientemente de n. Haciendo que λ tienda a infinito a través de valores de la forma $\lambda = n + \frac{1}{2}$, se encuentra entonces para I el valor $\pi/2$:

$$(21) \qquad \int_0^\infty \frac{\operatorname{sen} z}{z}\, dz = \frac{\pi}{2}.$$

Se ha demostrado, además, que

$$(21a) \qquad \lim_{\lambda\to\infty} \int_0^p \frac{\operatorname{sen} \lambda x}{x}\, dx = \frac{\pi}{2},$$

cuya convergencia es uniforme para $P \leq p$, donde P es un número positivo fijo; y también que

$$(21b) \qquad \lim_{\lambda\to\infty} \int_0^p \frac{\operatorname{sen} \lambda x}{2 \operatorname{sen} (x/2)}\, dx = \frac{\pi}{2},$$

cuya convergencia es uniforme para $P \leq p \leq \pi$.

d. Desarrollo de Fourier para la función $\phi(x) = x$

El resultado anterior conduce directamente al desarrollo de Fourier de dos funciones periódicas lineales a trozos, $\phi(x)$ y $\chi(x)$, relacionadas entre sí, que se definen en el intervalo $-\pi < x < \pi$ por medio de las ecuaciones

$$\phi(x) = x$$

y

$$\chi(x) = \begin{cases} \pi - x & \text{para } x > 0 \\ 0 & \text{para } x = 0. \\ -\pi - x & \text{para } x < 0. \end{cases} \tag{22}$$

(Ver Figs. 8.7 y 8.8.)

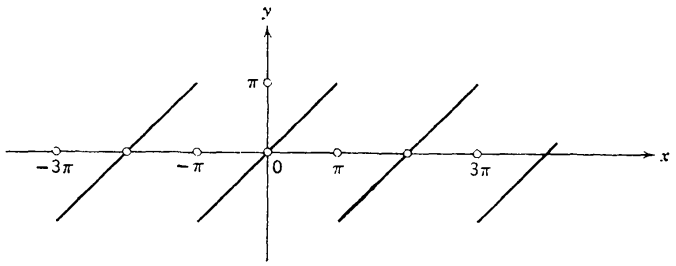


Figura 8.7 La función $\phi(x)$.

La primera función, ϕ, extendida periódicamente fuera del intervalo $-\pi < x < +\pi$, presenta discontinuidades de salto en los puntos extremos, mientras que $\chi(x)$ presenta un salto de 2π en $x = 0$. Es claro que las dos funciones extendidas periódicamente están relacionadas entre sí mediante la ecuación

$$\chi(x) = \phi(\pi - x).$$

El desarrollo de Fourier para $\chi(x)$, cuando $0 < x \leq \pi$, se deduce inmediatamente de las fórmulas (14) de p. 601 y (21*b*) de p. 606, con $\lambda = n + \frac{1}{2}$ y $p = x$, bastando para ello pasar al límite con $n \to \infty$. Resulta así la serie de Fourier

$$\chi(x) = 2\left(\operatorname{sen} x + \frac{\operatorname{sen} 2x}{2} + \frac{\operatorname{sen} 3x}{3} + \cdots\right). \tag{23a}$$

Lo mismo es válido también para $-\pi \leq x < 0$, pues ambos lados son funciones impares de x. La serie es uniformemente convergente para $\varepsilon < |x| \leq \pi$, donde ε es un número positivo arbitrariamente pequeño. Para $x = 0$ todos

los términos de la serie son nulos y, por consiguiente, también lo es la suma, cosa que concuerda con la definición de $\chi(0)$. La identidad (23*a*) es válida para todo x por tener ambos lados el período 2π.

Puede confirmarse fácilmente que los coeficientes del desarrollo son efectivamente los coeficientes de Fourier definidos por la fórmula (17), p. 602.

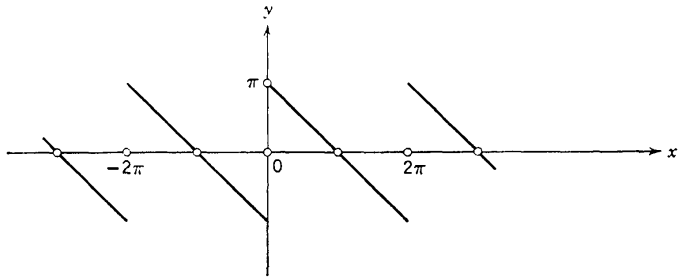


Figura 8.8 La función $\chi(x)$.

El desarrollo de Fourier para $\phi(x)$ puede obtenerse ahora directamente de la relación $\phi(x) = \chi(\pi - x)$:

$$\phi(x) = 2\sum_{\nu=1}^{\infty} (-1)^{\nu+1} \frac{\operatorname{sen} \nu x}{\nu} \tag{23b}$$

$$= 2(\operatorname{sen} x - \tfrac{1}{2} \operatorname{sen} 2x + \tfrac{1}{3} \operatorname{sen} 3x - + \cdots).$$

En este caso la convergencia es uniforme siempre que el punto x se mantenga separado de los puntos de discontinuidad $x = \pm\pi$, mediante la condición $|x| < \pi - \varepsilon$.

Para $x = \pi/2$ se obtiene nuevamente la serie de Leibnitz:

$$\frac{\pi}{2} = 2(1 - \tfrac{1}{3} + \tfrac{1}{5} - \cdots + -).$$

Debe mencionarse que las series para χ y ϕ no convergen absolutamente. En efecto, los valores absolutos para $x = \pi/2$ forman la serie

$$2\sum_{1}^{\infty} \frac{1}{2\nu - 1},$$

que es divergente.

La fórmula (23*b*) es notable como el ejemplo de una serie infinita de funciones continuas que converge para todo x pero que tiene por suma una función discontinua, a saber, la función lineal a trozos $\phi(x)$. Cualquier suma parcial de la serie es continua, puesto que la suma de cualquier número finito de funciones continuas es a su vez continua. Debido a que una serie

infinita uniformemente convergente de funciones continuas tiene una suma continua, la serie de Fourier no puede ser uniformemente convergente en la vecindad de un punto x en el cual ϕ es discontinua, es decir, para $x = \pm\pi, \pm 3\pi, \ldots$. La Fig. 8.4, p. 595, muestra cómo las sumas parciales sucesivas, que son polinomios trigonométricos (funciones continuas), aproximan uniformemente, en un intervalo de continuidad, a la función lineal por trozos $\frac{1}{2}\phi(x)$, pero que cerca del punto extremo las funciones cambian con rapidez cada vez mayor.

e. El teorema principal sobre desarrollos de Fourier

Los coeficientes de Fourier. Después de los preparativos anteriores, se puede fácilmente asegurar la posibilidad de desarrollar en serie de Fourier una amplia clase de funciones. La forma de tal desarrollo para una función $f(x)$ con período 2π es

$$f(x) = \tfrac{1}{2}a_0 + \sum_{\nu=1}^{\infty} (a_\nu \cos \nu x + b_\nu \operatorname{sen} \nu x), \tag{24a}$$

o bien, en notación compleja,

$$f(x) = \sum_{\nu=-\infty}^{\infty} \alpha_\nu e^{i\nu x}. \tag{24b}$$

Supondremos, para comenzar, que se dispone de desarrollos uniformemente convergentes de las formas (24a) o (24b) para la función $f(x)$. Es posible entonces determinar los coeficientes a_ν, b_ν y α_ν de esos desarrollos multiplicando el primero por $\cos \mu x$, o bien $\operatorname{sen} \mu x$ y el segundo por $e^{-i\mu x}$ e integrando desde $-\pi$ hasta π, haciendo uso de las relaciones de ortogonalidad (ver pp. 295 y 599):

$$\int_{-\pi}^{\pi} \operatorname{sen} \nu x \operatorname{sen} \mu x \, dx = \int_{-\pi}^{\pi} \cos \nu x \cos \mu x \, dx = \begin{cases} 0 & \text{si } \mu \neq \nu \\ \pi & \text{si } \mu = \nu \neq 0, \end{cases}$$

$$\int_{-\pi}^{\pi} \operatorname{sen} \nu x \cos \mu x \, dx = 0, \tag{25}$$

$$\int_{-\pi}^{\pi} e^{i\nu x} e^{-i\mu x} \, dx = \begin{cases} 0 & \text{si } \mu \neq \nu \\ 2\pi & \text{si } \mu = \nu. \end{cases}$$

Representando por t la variante de integración, se obtienen de inmediato las fórmulas

$$\int a_\mu = \frac{1}{\pi}\int_{-\pi}^{\pi} f(t) \cos \mu t \, dt, \qquad b_\mu = \frac{1}{\pi}\int_{-\pi}^{\pi} f(t) \operatorname{sen} \mu t \, dt \tag{26a}$$

para $\mu = 0, 1, 2, \ldots$, y

$$\alpha_\mu = \frac{1}{2\pi}\int_{-\pi}^{\pi} f(t) e^{-i\mu x} \, dt \tag{26b}$$

para $\mu = 0, \pm 1, \pm 2, \ldots$.

Así, si $f(x)$ puede desarrollarse en una serie, ya sea de la forma (24a) o (24b), uniformemente convergente, entonces los coeficientes sólo pueden tener los valores determinados por las fórmulas (26a) y (26b). Pero, aún sin una justificación rigurosa de este procedimiento, es claro que las fórmulos (26a) o (26b) definen una sucesión de números, a_ν, b_ν y α_ν, llamados los *coeficientes de Fourier,* para toda función $f(x)$ que sea continua o continua por trozos en el intervalo $-\pi \leq x \leq \pi$.

Dada una función $f(x)$, formamos con los coeficientes definidos por (26a, b) las *sumas parciales de Fourier*

$$S_n(x) = \tfrac{1}{2}a_0 + \sum_{\nu=1}^{n} (a_\nu \cos \nu x + b_\nu \operatorname{sen} \nu x),$$

o, lo que es lo mismo,

$$S_n(x) = \sum_{\nu=-n}^{\nu=n} \alpha_\nu e^{i\nu x}.$$

Nuestra tarea consiste en demostrar que estas sumas de Fourier son realmente convergentes para $n \to \infty$ y que el límite es la función $f(x)$.

Enunciamos ahora el

TEOREMA PRINCIPAL. *La serie de Fourier*

$$\tfrac{1}{2}a_0 + \sum_{\nu=1}^{\infty} (a_\nu \cos \nu x + b_\nu \operatorname{sen} \nu x), \tag{27a}$$

o sea,

$$\sum_{\nu=-\infty}^{\infty} \alpha_\nu e^{i\nu x}, \tag{27b}$$

formada con los coeficientes de Fourier (26a) *o* (26b), *converge al valor* $f(x)$ *para cualquier función continua por trozos,* $f(x)$, *de período* 2π *y con derivadas de primero y segundo órdenes continuas por trozos.*[1] *Aquí el valor de* $f(x)$ *en un punto de discontinuidad debe definirse como*

$$f(x) = \tfrac{1}{2}[f(x + 0) + f(x - 0)]. \tag{27c}$$

DEMOSTRACIÓN.[2] Para la demostración, introducimos en el n-ésimo "polinomio de Fourier",

$$S_n(x) = \tfrac{1}{2}a_0 + \sum_{\nu=1}^{\infty} (a_\nu \cos \nu x + b_\nu \operatorname{sen} \nu x),$$

[1] Nuevamente indicamos que este teorema puede demostrarse para clases de funciones mucho más generales (ver, por ejemplo, la sección 8.6). Sin embargo, el resultado que aquí se formula es suficiente para la mayoría de las aplicaciones.

[2] Damos aquí únicamente la demostración para el desarrollo de f en la serie (27a). La serie (27b) se deduce entonces por la substitución dada en la Ec. (13b), página 601.

las expresiones integrales (26*a*) de los coeficientes e intercambiamos la integración y la suma; y así resulta

$$S_n(x) = \frac{1}{\pi}\int_{-\pi}^{\pi} f(t)\left[\tfrac{1}{2} + \sum_{\nu=1}^{\infty}(\cos \nu t \cos \nu x + \operatorname{sen} \nu t \operatorname{sen} \nu x)\right] dt,$$

o, usando el teorema de adición para el coseno,

$$S_n(x) = \frac{1}{\pi}\int_{-\pi}^{\pi} f(t)\left[\tfrac{1}{2} + \sum_{\nu=1}^{n}\cos \nu(t-x)\right] dt.$$

Por la fórmula de suma (14) de la p. 603, se llega a

$$S_n(x) = \frac{1}{2\pi}\int_{-\pi}^{\pi} f(t)\,\frac{\operatorname{sen}[(n+\tfrac{1}{2})(t-x)]}{\operatorname{sen}\tfrac{1}{2}(t-x)}\,dt. \tag{28}$$

Finalmente, haciendo $\tau = t - x$ y recordando que la periodicidad permite desplazar el intervalo de integración en la cantidad x (ver p. 590), resulta

$$S_n(x) = \frac{1}{2\pi}\int_{-\pi}^{\pi} f(x+\tau)\,\frac{\operatorname{sen}(n+\tfrac{1}{2})\tau}{\operatorname{sen}\tfrac{1}{2}\tau}\,d\tau, \tag{28a}$$

donde, por supuesto, x tiene un valor fijo.

Ahora demostraremos que $S_n(x)$ tiende a $f(x)$ cuando $n \to \infty$. En efecto,

$$\lim_{n\to\infty} S_n(x) = \lim_{n\to\infty}\frac{1}{2\pi}\int_{-\pi}^{\pi} f(x+t)\,\frac{\operatorname{sen}(n+\tfrac{1}{2})t}{\operatorname{sen}\tfrac{1}{2}t}\,dt = f(x). \tag{29}$$

Como $f(x) = \tfrac{1}{2}[f(x+0) + f(x-0)]$ para todo x, se tiene [ver fórmula (15), p. 602]

$$S_n(x) - f(x) = \frac{1}{\pi}\int_0^{\pi}\frac{[f(x+t) - f(x+0)]}{2\operatorname{sen}\tfrac{1}{2}t}\operatorname{sen}(n+\tfrac{1}{2})t\,dt + \frac{1}{\pi}\int_{-\pi}^{0}\frac{[f(x+t) - f(x-0)]}{2\operatorname{sen}\tfrac{1}{2}t}\operatorname{sen}(n+\tfrac{1}{2})t\,dt.$$

Si podemos demostrar que las funciones $[f(x+t) - f(x+0)]/(2\operatorname{sen}\tfrac{1}{2}t)$ y $[f(x+t) - f(x-0)]/(2\operatorname{sen}\tfrac{1}{2}t)$ de la variable t, junto con sus primeras derivadas, son continuas a trozos en los intervalos $0 \le t \le \pi$ y $-\pi \le t \le 0$ respectivamente, entonces, por el Lema básico (p. 603), las dos integrales que aparecen en el lado derecho de la ecuación anterior tienden a cero cuando $n \to \infty$, quedando así establecida la validez de la fórmula (29).

Por consiguiente, el teorema principal quedará demostrado si podemos demostrar que para un valor fijo de x la función de t definida por

$$\phi(t) = \frac{f(x+t) - f(x+0)}{2\operatorname{sen}\tfrac{1}{2}t} \quad \text{para } 0 < t < \pi$$

$$\phi(t) = \frac{f(x+t) - f(x-0)}{2\operatorname{sen}\tfrac{1}{2}t} \quad \text{para } -\pi < t < 0$$

es continua a trozos y tiene primera y segunda derivadas que también lo son, siempre que f, f' y f'' sean continuas a trozos.

Para indagar si el cociente $\phi(t)$ satisface estas condiciones, observamos primero que el denominador se anula solamente para $t = 0$ y que, por consiguiente, ϕ y su primera derivada son continuas a trozos excepto, posiblemente, cerca de $t = 0$. Sólo en el punto singular $t = 0$ podría no tenerse derivabilidad. Todo lo que tenemos que hacer, por consiguiente, es demostrar que $\phi(t)$ y su derivada $\phi'(t)$ se aproximan a ciertos límites si t tiende a cero, respectivamente, desde valores positivos y negativos. Demostraremos, en efecto, que estos límites existen y que tienen los valores

$$\phi(+0) = f'(x+0), \quad \psi(-0) = f'(x-0)$$

y

$$\phi'(+0) = \tfrac{1}{2}f''(x+0), \quad \phi'(-0) = \tfrac{1}{2}f''(x-0),$$

respectivamente.

Para la demostración, introduzcamos la función $g(t)$ por medio de la ecuación $\phi(t) = g(t)h(t)$, donde el factor $h(t)$ está definido por

$$h(t) = \frac{t}{2\operatorname{sen}(t/2)} \quad \text{para } t \neq 0, \quad h(0) = 1.$$

Tenemos (ver capítulo 5, p. 482) en $h(t)$ una función continua con una derivada continua en todo el intervalo $-\pi \le t \le \pi$, siendo $h(0) = 1$, y $h'(0) = 0$. Por consiguiente, al tomar límites haciendo $t \to 0$, los valores de $g(t)$ y $\phi(t)$ coinciden entre sí, y lo mismo ocurre con $g'(t)$ y $\phi'(t) = gh' + g'h$.

Ahora bien, en el intervalo $0 < t < \pi$ (ver comentarios generales acerca de expresiones indeterminadas en el capítulo 5, p. 481) el teorema del valor medio del Cálculo asegura que

$$g(t) = \frac{f(x+t) - f(x+0)}{t} = f'(x+\xi),$$

siendo ξ un valor entre 0 y t. Entonces, cuando t (y por tanto, también cuando ξ) tiende a cero, resulta

$$g(+0) = f'(x+0).$$

Y para la derivada,

$$g'(t) = \frac{tf'(x+t) + f(x+0) - f(x+t)}{t^2},$$

expresión en la cual también el numerador y el denominador son funciones que tienden a cero para $t \to 0$, y cuyas derivadas son, respectivamente, $tf''(x+t)$ y $2t$. Para determinar el límite cuando $t \to 0$ hacemos uso del teorema generalizado del valor medio (ver p. 243); y así encontramos

$$g'(t) = \frac{\eta f''(x+\eta)}{2\eta} = \tfrac{1}{2}f''(x+\eta),$$

donde η es un valor entre 0 y t. Si $t \to 0$, entonces $\eta \to 0$, y, por tanto, como se dijo antes, $g'(+0) = \phi'(+0) = \frac{1}{2}f''(x+0)$.

El mismo razonamiento es aplicable a valores negativos de t. En consecuencia, está justificada la aplicación del Lema, y el teorema principal queda así demostrado.

Puede decirse que el resultado obtenido es lo suficientemente amplio como para cubrir todas las necesidades que surgen en el Cálculo y sus aplicaciones. Sin embargo, el interés teórico de los matemáticos, comenzando con el trabajo original de Dirichlet, se orientó frecuentemente hacia una mayor generalidad, es decir, el desarrollo de funciones pertenecientes a una clase más amplia.[1] Estos esfuerzos han estimulado un análisis más refinado de los conceptos de función e integral y conducido al desarrollo del análisis avanzado de Fourier, campo atractivo y especializado que está fuera del objeto de este libro.

8.5 Ejemplos de series de Fourier

a. Comentarios preliminares

En lo que sigue supondremos que el período de las funciones $f(x)$ es 2π.

Si $f(x)$ es una función par (ver p. 53), es claro entonces que $f(x)$ sen νx es impar y $f(x)\cos \nu x$ es par, de manera que

$$b_\nu = \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)\,\operatorname{sen}\,\nu x\,dx = 0,$$

y se obtiene una "serie de cosenos". Si, por otra parte, la función $f(x)$ es impar, entonces

$$a_\nu = \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)\cos \nu x\,dx = 0,$$

y lo que resulta es una "serie de senos".[2]

[1] Debe hacerse notar que existen ejemplos de funciones continuas que no son desarrollables en serie de Fourier. Además, existen también ejemplos de funciones $f(x)$ representadas por series trigonométricas convergentes que, no obstante, no son series de Fourier con coeficientes dados por las expresiones (26). Tales ejemplos muestran que debe establecerse una distinción entre las series trigonométricas, en general, y las series de Fourier, en particular, sobre todo cuando se trata de investigaciones refinadas. Además, tales ejemplos ilustran el hecho de que, sin duda, condiciones aún más restrictivas que la continuidad son convenientes; no obstante que las restricciones supuestas en nuestro teorema general y en una extensión que se da en la sección 8.6 son bastante más fuertes que las realmente necesarias. (Para la teoría general véase: *Trigonometrical Series,* por A. Zygmund, Chelsea Publishing Co., 1952.)

[2] En consecuencia, si la función $f(x)$ está dada inicialmente sólo en el intervalo $0 < x < \pi$, podemos extenderla al intervalo $-\pi < x < 0$, ya sea como una función impar o como una función par. Así, para el intervalo $0 < x < \pi$, más pequeño, resulta una serie de senos o una serie de cosenos.

b. Desarrollo de la función $\phi(x) = x^2$

Para la función par x^2 se obtiene, después de una doble integración por partes,

$$a_\nu = \frac{2}{\pi}\int_0^{\pi} x^2\cos \nu x\,dx = (-1)^\nu \frac{4}{\nu^2}, \qquad (\nu > 0),$$

$$a_0 = \frac{2\pi^2}{3},$$

de manera que resulta el desarrollo

$$x^2 = \frac{\pi^2}{3} - 4\left(\frac{\cos x}{1^2} - \frac{\cos 2x}{2^2} + \frac{\cos 3x}{3^2} + \cdots\right). \tag{30}$$

Derivando término a término esta serie y dividiendo por 2 el resultado, se obtiene formalmente la serie (23*b*) de p. 608, que antes obtuvimos para $\phi(x) = x$.

c. Desarrollo de $x\cos x$

(Véase Fig. 8.9.) Para esta función impar tenemos que

$$a_\nu = 0, \qquad b_\nu = \frac{2}{\pi}\int_0^{\pi} x\cos x\,\operatorname{sen}\,\nu x\,dx.$$

Usando la fórmula

$$\int_0^{\pi} x\,\operatorname{sen}\,\mu x\,dx = (-1)^{\mu+1}\frac{\pi}{\mu}, \qquad (\mu = 1, 2, \ldots,),$$

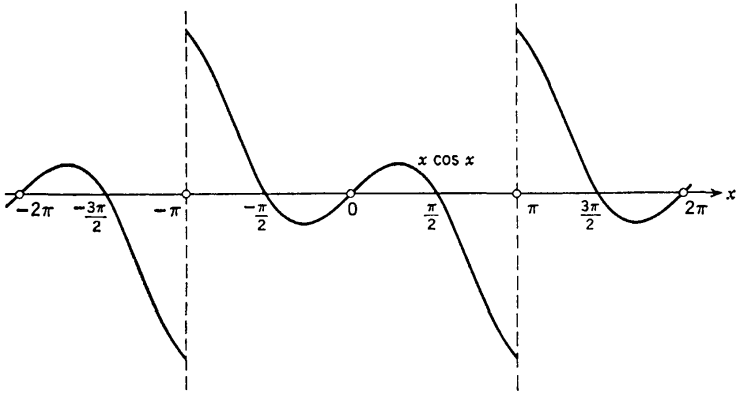


Figura 8.9

encontramos que

$$b_\nu = \frac{2}{\pi}\int_0^\pi x\cos x \operatorname{sen} \nu x\, dx$$
$$= \frac{1}{\pi}\int_0^\pi x[\operatorname{sen}(\nu+1)x + \operatorname{sen}(\nu-1)x]\, dx$$
$$= (-1)^\nu\left(\frac{2\nu}{\nu^2-1}\right) \qquad (\nu = 2, 3, \ldots)$$
$$b_1 = -\tfrac{1}{2}.$$

Por consiguiente, la serie buscada es

$$x\cos x = -\tfrac{1}{2}\operatorname{sen} x + 2\sum_{\nu=2}^{\infty}\frac{(-1)^\nu \nu}{\nu^2-1}\operatorname{sen}\nu x. \tag{31}$$

Si a esta serie le sumamos la (23*b*) de p. 608, que obtuvimos para $\phi(x) = x$, obtenemos el resultado

(31*a*)
$$x(1+\cos x) = \tfrac{3}{2}\operatorname{sen} x + 2\left(\frac{\operatorname{sen} 2x}{1\cdot 2\cdot 3} - \frac{\operatorname{sen} 3x}{2\cdot 3\cdot 4} + \frac{\operatorname{sen} 4x}{3\cdot 4\cdot 5} - + \cdots\right).$$

Cuando la función que coincide con $x\cos x$ en el intervalo $-\pi < x < \pi$ se extiende periódicamente más allá de este intervalo, aparecen las mismas discontinuidades (ver Fig. 8.7) que las que exhibe la función $\phi(x)$ considerada antes en la sección 8.4*d*. Por otra parte, la función $x(1+\cos x)$, extendida periódicamente, sigue siendo continua en los puntos extremos del intervalo y, de hecho, su derivada también sigue siendo continua, pues las discontinuidades son eliminadas por el factor $1+\cos x$, el cual, lo mismo que su derivada, se anula en los puntos extremos. Esto equivale al hecho de que la serie (31) converge uniformemente para todo x, lo cual es evidente si se compara con la serie de términos constantes $\frac{1}{1^3} + \frac{1}{2^3} + \frac{1}{3^3} + \cdots$.

d. La función $f(x) = |x|$

Para esta función par se tiene $b_\nu = 0$ y $a_\nu = \frac{2}{\pi}\int_0^\pi x\cos \nu x\, dx$. Integrando por partes se encuentra de inmediato que

$$\int_0^\pi x\cos \nu x\, dx = \frac{1}{\nu}x\operatorname{sen}\nu x\Big|_0^\pi - \frac{1}{\nu}\int_0^\pi \operatorname{sen}\nu x\, dx$$
$$= \begin{cases} 0, & \text{si } \nu \text{ es par y} \neq 0, \\ -\dfrac{2}{\nu^2}, & \text{si } \nu \text{ es impar.} \end{cases}$$

En consecuencia,

$$|x| = \tfrac{1}{2}\pi - \frac{4}{\pi}\left(\cos x + \frac{\cos 3x}{3^2} + \frac{\cos 5x}{5^2} + \cdots\right). \tag{32}$$

Haciendo $x = 0$ resulta la notable fórmula

$$\frac{\pi^2}{8} = 1 + \frac{1}{3^2} + \frac{1}{5^2} + \cdots. \tag{32a}$$

e. Una función constante por trozos

La función definida por las ecuaciones

$$f(x) = \operatorname{sgn} x = \begin{cases} -1, & \text{para } -\pi < x < 0, \\ 0, & \text{para } x = 0, \\ +1, & \text{para } 0 < x < \pi, \end{cases}$$

como se indica en la Fig. 1.22, p. 56, es impar. Por consiguiente $a_\nu = 0$ y

$$b_\nu = \frac{2}{\pi}\int_0^\pi \operatorname{sen}\nu x\, dx = \begin{cases} 0 & \text{si } \nu \text{ es par,} \\ \dfrac{4}{\pi\nu}, & \text{si } \nu \text{ es impar,} \end{cases}$$

de manera que la serie de Fourier para esta función es

$$f(x) = \frac{4}{\pi}\left(\frac{\operatorname{sen} x}{1} + \frac{\operatorname{sen} 3x}{3} + \cdots\right). \tag{33}$$

En particular, para $x = \frac{1}{2}\pi$ se obtiene nuevamente la serie de Leibnitz.

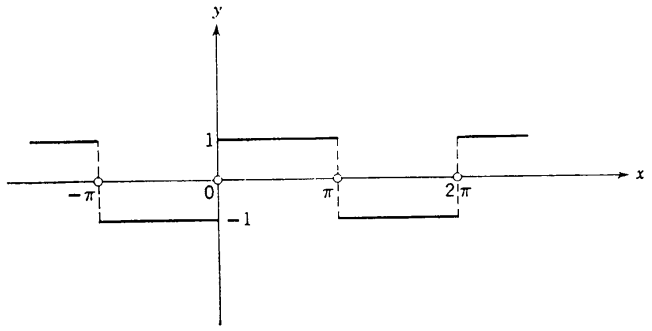


Figura 8.10

La serie (33) puede derivarse formalmente de la serie para $|x|$ dada en (32), por medio de la derivación término a término.

f. La función $|\operatorname{sen} x|$

La función par $f(x) = \operatorname{sen} x$ puede desarrollarse en una "serie de cosenos", estando dados los coeficientes a_ν por los siguientes cálculos:

$$\begin{aligned}\tfrac{1}{2}\pi a_\nu &= \int_0^\pi \operatorname{sen} x \cos \nu x \, dx \\ &= \tfrac{1}{2}\int_0^\pi [\operatorname{sen}(\nu+1)x - \operatorname{sen}(\nu-1)x]\, dx \\ &= \begin{cases} 0 & \text{si } \nu \text{ es impar,} \\ \dfrac{-2}{\nu^2-1} & \text{si } \nu \text{ es par.}\end{cases}\end{aligned}$$

En esta forma resulta, escribiendo 2ν en vez de ν,

$$|\operatorname{sen} x| = \frac{2}{\pi} - \frac{4}{\pi}\sum_{\nu=1}^{\infty} \frac{\cos 2\nu x}{4\nu^2 - 1}. \tag{34}$$

g. Desarrollo de $\cos \mu x$. Descomposición de la cotangente en fracciones parciales. El producto infinito para el seno

La función $f(x) = \cos \mu x$ para $-\pi < x < \pi$, en la cual μ no es un entero, es una función par; en consecuencia, $b_\nu = 0$ y

$$\begin{aligned}\tfrac{1}{2}\pi a_\nu &= \int_0^\pi \cos \mu x \cos \nu x \, dx \\ &= \tfrac{1}{2}\int_0^\pi [\cos(\mu+\nu)x + \cos(\mu-\nu)x]\, dx \\ &= \tfrac{1}{2}\left[\frac{\operatorname{sen}(\mu+\nu)\pi}{\mu+\nu} + \frac{\operatorname{sen}(\mu-\nu)\pi}{\mu-\nu}\right] \\ &= \frac{\mu(-1)^\nu}{\mu^2-\nu^2}\operatorname{sen}\mu\pi.\end{aligned}$$

En esta forma tenemos que

$$\cos \mu x = \frac{2\mu \operatorname{sen}\mu\pi}{\pi}\left(\frac{1}{2\mu^2} - \frac{\cos x}{\mu^2-1^2} + \frac{\cos 2x}{\mu^2-2^2} + - \cdots\right). \tag{35}$$

Esta función con período 2π, extendida periódicamente fuera del intervalo $-\pi < x < \pi$, sigue siendo continua en los puntos $x = \pm\pi$. Haciendo $x = \pi$, dividiendo ambos lados de la ecuación por $\operatorname{sen}\mu\pi$ y escribiendo x en vez de μ, obtenemos la ecuación

$$\cot \pi x = \frac{2x}{\pi}\left(\frac{1}{2x^2} + \frac{1}{x^2-1^2} + \frac{1}{x^2-2^2} + \cdots\right). \tag{36}$$

Ésta es la descomposición de la cotangente en fracciones parciales (por analogía con la descomposición finita de funciones racionales en fracciones parciales, discutida en el capítulo 3, p. 306). Tal descomposición constituye un resultado muy importante en el Análisis.

Podemos escribir la serie anterior en la forma

$$\cot \pi x - \frac{1}{\pi x} = -\frac{2x}{\pi}\left[\frac{1}{1^2-x^2} + \frac{1}{2^2-x^2} + \cdots\right].$$

Si x está en el intervalo $0 \le x \le q < 1$, el valor absoluto del n-ésimo término de la derecha es menor que $2/[\pi(n^2-q^2)]$. Por tanto, la serie converge uniformemente en este intervalo y puede ser integrada término a término. Multiplicando ambos lados por π e integrando resulta en el lado izquierdo

$$\pi\int_0^x \left(\cot \pi t - \frac{1}{\pi t}\right) dt = \log\frac{\operatorname{sen}\pi x}{\pi x} - \lim_{a\to 0}\log\frac{\operatorname{sen}\pi a}{\pi a} = \log\frac{\operatorname{sen}\pi x}{\pi x}$$

y en el derecho

$$\log\left(1 - \frac{x^2}{1^2}\right) + \log\left(1 - \frac{x^2}{2^2}\right) + \cdots = \lim_{n\to\infty}\sum_{\nu=1}^{n}\log\left(1 - \frac{x^2}{\nu^2}\right).$$

Así,

$$\begin{aligned}\log\frac{\operatorname{sen}\pi x}{\pi x} &= \lim_{n\to\infty}\sum_{\nu=1}^{n}\log\left(1 - \frac{x^2}{\nu^2}\right) \\ &= \lim_{n\to\infty}\log\prod_{\nu=1}^{n}\left(1 - \frac{x^2}{\nu^2}\right) = \log\lim_{\nu=1}^{n}\prod\left(1 - \frac{x^2}{\nu^2}\right).\end{aligned}$$

Al pasar del logaritmo a la función exponencial se obtiene

$$\operatorname{sen}\pi x = \pi x\left(1 - \frac{x^2}{1^2}\right)\left(1 - \frac{x^2}{2^2}\right)\left(1 - \frac{x^2}{3^2}\right)\cdots. \tag{36a}$$

Hemos así obtenido la famosa expresión para el seno, consistente en un producto infinito.[1]

De este resultado haciendo $x = \frac{1}{2}$, se obtiene el producto de Wallis

$$\tfrac{1}{2}\pi = \prod_{\nu=1}^{\infty}\frac{2\nu}{2\nu-1}\cdot\frac{2\nu}{2\nu+1} = \frac{2}{1}\cdot\frac{2}{3}\cdot\frac{4}{3}\cdot\frac{4}{5}\cdots$$

tal como se obtuvo en la p. 301.

[1] Esta fórmula es particularmente interesante porque exhibe en forma directa la anulación de la función $\operatorname{sen}\pi x$ en los puntos $x = 0, \pm 1, \pm 2, \ldots$. En este aspecto, se asemeja a la factorización de un polinomio cuando se conocen sus raíces.

h. Ejemplos adicionales

Mediante breves cálculos análogos a los anteriores es posible obtener otros ejemplos de desarrollos.

La función $f(x)$ definida por la ecuación $f(x) = \operatorname{sen} \mu x$ para $-\pi < x < \pi$ puede desarrollarse según la serie

$$(37)\quad \operatorname{sen} \mu x = \frac{2 \operatorname{sen} \mu\pi}{\pi}\left(\frac{\operatorname{sen} x}{\mu^2 - 1^2} - \frac{2 \operatorname{sen} 2x}{\mu^2 - 2^2} + \frac{3 \operatorname{sen} 3x}{\mu^2 - 3^2} - + \cdots\right).$$

Haciendo $x = \frac{1}{2}\pi$ y usando la relación $\operatorname{sen} \mu\pi = 2 \operatorname{sen} \frac{1}{2}\mu\pi \cos \frac{1}{2}\mu\pi$ obtenemos la descomposición de la secante, es decir, de la función $1/\cos \frac{1}{2}\mu\pi$, en fracciones parciales. Tal desarrollo es

$$\pi \sec \pi x = \frac{\pi}{\cos \pi x} = 4 \sum_{\nu=1}^{\infty} \frac{(-1)^{\nu}(2\nu - 1)}{4x^2 - (2\nu - 1)^2},$$

donde en lugar de $\frac{1}{2}\mu$ hemos escrito x.

Las series análogas a la (35) y a la (36) para las funciones hiperbólicas $\cosh \mu x$ y $\operatorname{senh} \mu x (-\pi < x < \pi)$ son

$$\cosh \mu x = \frac{2\mu}{\pi} \operatorname{senh} \mu\pi \left(\frac{1}{2\mu^2} - \frac{\cos x}{\mu^2 + 1^2} + \frac{\cos 2x}{\mu^2 + 2^2} - \frac{\cos 3x}{\mu^2 + 3^2} + \cdots\right),$$

$$\operatorname{senh} \mu x = \frac{2}{\pi} \operatorname{senh} \mu\pi \left(\frac{\operatorname{sen} x}{\mu^2 + 1^2} - \frac{2 \operatorname{sen} 2x}{\mu^2 + 2^2} + \frac{3 \operatorname{sen} 3x}{\mu^2 + 3^2} - + \cdots\right).$$

8.6 Discusión adicional sobre la convergencia

a. Resultados

Un examen más minucioso de los coeficientes de Fourier, a_ν y b_ν, conduce fácilmente a los siguientes corolarios del teorema principal de la sección 8.4*e*, p. 609.

(*a*) La serie de Fourier (27) de p 610 converge a $f(x)$ si $f(x)$ es una función periódica que satisface las condiciones menos estrictas de ser ella misma y su *primera* derivada, $f'(x)$, continuas a trozos o, como también se dice, si $f(x)$ es una función *"lisa" por trozos.*

(*b*) Si la función periódica lisa por trozos, $f(x)$, es continua, la convergencia es absoluta y uniforme.

(*c*) Si la función lisa por trozos, $f(x)$, presenta discontinuidades de salto, la convergencia es uniforme en cualquier intervalo cerrado que no contenga un punto de discontinuidad.

La demostración de (*b*) se basa en una sencilla desigualdad debida a Bessel; en tanto que para demostrar (*a*) y (*c*) se usarán los resultados de la sección 8.4*d*, p. 608.

b. Desigualdad de Bessel

Esta desigualdad proporciona cotas para los coeficientes de Fourier de cualquier función continua a trozos pero no necesariamente derivable. Establece que

$$(38)\quad \tfrac{1}{2}a_0^2 + \sum_{\nu=1}^{n} (a_\nu^2 + b_\nu^2) \leq M^2;$$

donde la cota $M^2 = \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)^2\, dx$ es un número fijado por la función $f(x)$, que no depende de los coeficientes de Fourier individuales, a_ν, b_ν, ni del número n. Con los coeficientes complejos de Fourier, α_ν [ver (13*a*), p. 601], la desigualdad de Bessel puede escribirse de inmediato en la forma

$$(38a)\quad \sum_{\nu=-n}^{n} |\alpha_\nu|^2 \leq \frac{1}{2\pi}\int_{-\pi}^{\pi} f(x)^2\, dx = \tfrac{1}{2}M^2.$$

La desigualdad es una consecuencia directa de otra cuya validez es casi evidente:

$$\frac{1}{\pi}\int_{-\pi}^{\pi} \left[f(x) - \tfrac{1}{2}a_0 - \sum_{\nu=1}^{n} (a_\nu \cos \nu x + b_\nu \operatorname{sen} \nu x) \right]^2 dx \geq 0.$$

La integral aquí se evalúa desarrollando el cuadrado que aparece como integrando y teniendo en cuenta las relaciones de ortogonalidad (25) (p. 609), así como también las definiciones (17) (p. 602) de los coeficientes de Fourier. Al integrar los términos individualmente se llega de inmediato a la desigualdad de Bessel en la forma (38) dada.

Puesto que el lado izquierdo de la desigualdad de Bessel crece monótonamente con n en tanto que la cota superior M^2 está fija, podemos pasar al límite con $n \to \infty$ e inferir la validez de la desigualdad

$$(39)\quad 2\sum_{\nu=-\infty}^{\infty} |\alpha_\nu|^2 = \tfrac{1}{2}a_0^2 + \sum_{\nu=1}^{\infty} (a_\nu^2 + b_\nu^2) \leq M^2$$

Esta última se cumple para los coeficientes de Fourier de una función continua a trozos $f(x)$ aun cuando f pueda no ser representable por las series (27*a*) o (27*b*).

En la sección 8.7*d* demostraremos que la desigualdad de essel (39) sigue siendo válida si se reemplaza el signo de desigualdad por el de igualdad.

*c. Demostración de los corolarios (a), (b) y (c)

Suponiendo que $f(x)$ es continua, aplicamos la desigualdad de Bessel a su derivada, $g(x) = f'(x)$, que es continua a trozos y cuyos coeficientes de Fourier son $c_\nu = +\nu b_\nu$, $d_\nu = -\nu a_\nu$, como puede verificarse fácilmente in-

tegrando por partes (ya que los términos integrados se anulan mutuamente):

$$c_\nu = \frac{1}{\pi}\int_{-\pi}^{\pi} f'(x)\cos \nu x\, dx = +\int_{-\pi}^{\pi} \nu f(x)\,\text{sen}\,\nu x\, dx = +\nu b_\nu,$$

y análogamente para d_ν. [Aquí se ha usado la propiedad de que $f(x)$ es continua y periódica.] Se tiene, en consecuencia,

$$\sum_{\nu=1}^{n} \nu^2(a_\nu^2 + b_\nu^2) = \sum_{\nu=1}^{n} (c_\nu^2 + d_\nu^2)$$

$$\leq \frac{1}{\pi}\int_{-\pi}^{\pi} g(x)^2\, dx = \frac{1}{\pi}\int_{-\pi}^{\pi} f'(x)^2\, dx = M^2.$$

Este resultado permite construir para la serie de Fourier de $f(x)$ una serie mayorante con términos positivos constantes, la cual, de acuerdo con la p. 551, asegura la convergencia absoluta y uniforme que se menciona en (b). En efecto, para la ν-ésima oscilación armónica la desigualdad de Cauchy-Schwarz (ver p. 39) da

$$|a_\nu \cos \nu x + b_\nu \,\text{sen}\, \nu x|^2 \leq (a_\nu^2 + b_\nu^2)(\cos^2 \nu x + \text{sen}^2\, \nu x) = a_\nu^2 + b_\nu^2;$$

y usando entonces la desigualdad

$$pq \leq \tfrac{1}{2}(p^2 + q^2)$$

con $p = 1/\nu$, $q = \nu\sqrt{a_\nu^2 + b_\nu^2}$, se obtiene que para todo ν es

$$|a_\nu \cos \nu x + b_\nu \,\text{sen}\, \nu x| \leq \frac{1}{\nu}\,\nu\sqrt{a_\nu^2 + b_\nu^2}$$

$$\leq \frac{1}{2}\left[\frac{1}{\nu^2} + \nu^2(a_\nu^2 + b_\nu^2)\right].$$

La suma sobre ν del lado izquierdo de esta última desigualdad converge y, por consiguiente, constituye una mayorante. Por tanto, la serie de Fourier

$$\tfrac{1}{2}a_0 + \sum_{\nu=1}^{\infty} (a_\nu \cos \nu x + b_\nu \,\text{sen}\, \nu x)$$

converge uniformemente. Su suma es entonces $s(x)$, que es una función continua de x. Para demostrar que realmente $s(x) = f(x)$ usamos el artificio consistente en considerar la función integrada

$$F(x) = \int_{-\pi}^{x} (f(t) - \tfrac{1}{2}a_0)\, dt.$$

Es claro que $F(x)$ es continua para $-\pi \leq x \leq \pi$; además, F tiene en $x = -\pi$ el mismo valor que en $x = \pi$, pues

$$F(\pi) = \int_{-\pi}^{+\pi} f(t)\, dt - \pi a_0 = 0 = F(-\pi).$$

Por tanto, la extensión periódica de F es continua. Puesto que las dos primeras derivadas de F son continuas a trozos, la función F está representada por su serie de Fourier. Por el mismo argumento que se usó antes, basado en la integración por partes, los coeficientes de Fourier de F son $-(1/\nu)b_\nu$ y $(1/\nu)a_\nu$ para $\nu \neq 0$, de manera que

$$F(x) = \tfrac{1}{2}A_0 + \sum_{\nu=1}^{\infty} \frac{1}{\nu}(-b_\nu \cos \nu x + a_\nu \,\text{sen}\, \nu x),$$

siendo A_0 algún coeficiente constante. Ahora sabemos que la serie obtenida formalmente por derivación término a término es uniformemente convergente. En consecuencia, es válido derivar término a término formalmente (ver p. 555), resultando así la relación deseada:

$$F'(x) = f(x) - \tfrac{1}{2}a_0 = \sum_{\nu=1}^{\infty} (a_\nu \cos \nu x + b_\nu \,\text{sen}\, \nu x).$$

Con el fin de demostrar los asertos restantes en el caso de una función periódica continua a trozos, f, cuya derivada, f', es también continua a trozos, recordemos que, por los resultados anteriores, dichos asertos son válidos para la función periódica $\chi(x)$ de la sección 8.4*d* y, en consecuencia, también para la función $\chi(x - \xi)$, que presenta un salto de 2π en el punto ξ. Si ahora la función $f(x)$ presenta saltos de $\beta_1, \beta_2, \ldots, \beta_m$ en los puntos $\xi_1, \xi_2, \ldots, \xi_m$, entonces $f^*(x) = f(x) - \dfrac{1}{2\pi}\sum_{i=1}^{m} \beta_i \chi(x - \xi_i)$ satisface las condiciones de (b) y, por tanto, tiene una serie de Fourier uniformemente convergente; con lo cual se demuestra la validez de (a) y (c) para $f(x)$.

d. Orden de magnitud de los coeficientes de Fourier. Derivación de series de Fourier

Las discusiones anteriores relativas a la convergencia ilustran una propiedad general: Los coeficientes de Fourier a_ν, b_ν convergen más rápidamente a cero, conforme $n \to \infty$, cuando $f(x)$ es más lisa, es decir, cuando es mayor el número de derivadas de la función periódica $f(x)$ que son continuas. En forma equivalente, la convergencia de la serie de Fourier es mejor a medida que la función es más lisa. Enunciado en forma precisa: Si la función periódica $f(x)$ tiene derivadas continuas hasta el orden k y una derivada continua a trozos de orden $k + 1$, existe entonces una cota B, que sólo depende de $f(x)$ y de k, tal que

$$|a_\nu|, |b_\nu| < \frac{B}{\nu^{k+1}}. \tag{40}$$

La demostración es casi inmediata (véase lo anterior) si se aplica la integración por partes. Por brevedad escribimos en notación compleja

$$a_\nu - ib_\nu = 2\alpha_\nu$$

e integramos por partes sucesivamente hasta que aparezca como integrando el factor $f^{(k+1)}(x)$. Debido a la periodicidad y continuidad de $f(x)$, $f'(x)$, etc., los términos extremos se anulan mutuamente y resulta

$$2\pi\alpha_\nu = \int_{-\pi}^{\pi} f(x)e^{-i\nu x}\,dx = -\frac{i}{\nu}\int_{-\pi}^{\pi} f'(x)e^{-i\nu x}\,dx$$
$$= \cdots = \left(\frac{-i}{\nu}\right)^{k+1}\int_{-\pi}^{\pi} f^{(k+1)}(x)e^{-i\nu x}\,dx.$$

Por tanto, si $\frac{1}{2}B$ es una cota superior para $|f^{(k+1)}(x)|$, entonces $|\alpha_\nu| \leq \frac{1}{2}B/\nu^{k+1}$, lo cual implica las desigualdades (40).

Otro resultado notable es que para $k > 2$ la serie de Fourier puede derivarse término a término $k-1$ veces, y en esa forma resulta la serie de Fourier para la función derivada. Esto se demuestra observando que todas estas series derivadas tienen como mayorante la serie convergente $B\sum_{\nu=1}^{\infty}\frac{1}{\nu^2}$, por lo que ellas mismas son absoluta y uniformemente convergentes (ver los criterios del capítulo 7, p. 559).

*8.7 Aproximación mediante polinomios trigonométricos y racionales

a. Comentario general sobre las representaciones de funciones

Desde los primeros años del Cálculo se presentó el difícil problema de establecer la forma en que debería modificarse el concepto de función para dar lugar a la posibilidad de "expresiones explícitas". A menudo las funciones no se dan en forma analítica, sino más bien por construcción geométrica o mecánica, o por la descripción *geométrica* de sus gráficas, las que podrían ser de naturaleza diferente en distintos intervalos.

El descubrimiento de las series de Fourier a principios del siglo XIX constituyó un paso de suma trascendencia en la búsqueda de una respuesta para la vieja pregunta. Revelaba que, en efecto, muchas funciones "arbitrarias", sin duda mucho menos restringidas que las "analíticas", podían expresarse como series de Fourier convergentes. Pero las series de Fourier no cubren todas las funciones continuas: como se dijo antes, sin que se demostrara, es posible definir funciones continuas cuyas series de Fourier, formadas con los coeficientes de Fourier, no convergen.

Es de lo más notable el que, abandonando el principio de las series infinitas, según el cual la aproximación se realiza por la sola adición de términos de orden superior, sea posible construir para cualquier función continua, $f(x)$, polinomios de aproximación trigonométricos o racionales, $P_n(x)$, de grado n, tales que cuando $n \to \infty$ convergen uniformemente a la función dada $f(x)$ en un intervalo cerrado.

b. Teorema de aproximación de Weierstrass

Demostraremos los siguientes teoremas, que están estrechamente relacionados entre sí.

a) Si $f(x)$ es una función continua en un intervalo cerrado, I, contenido dentro del intervalo mayor $-\pi < x < \pi$, entonces f puede aproximarse uniformemente en I por medio de un polinomio trigonométrico de período 2π y de orden n suficientemente grande.

b) Cualquier función $f(x)$ continua en un intervalo cerrado, I, puede aproximarse uniformemente en I por medio de un polinomio en x, $P(x)$. Este enunciado, debido a Weierstrass, puede complementarse (ver p. 555) con el corolario:

**c*) Si $f(x)$ posee una derivada continua en I, entonces los polinomios de aproximación pueden escogerse de manera tal que los polinomios derivados, $P_n'(x)$, aproximen uniformemente a la derivada $f'(x)$.

La demostración de *a*) es bastante sencilla. Primero aproximamos $f(x)$ mediante una función lineal por trozos cuya gráfica es un polígono, $L_n(x)$, inscripto en la gráfica de $f(x)$ (ver Fig. 8.11). Es claro que el valor absoluto de la diferencia entre $L_n(x)$ y $f(x)$ es menor que cualquier margen dado arbitrariamente pequeño, $\varepsilon/2$, siempre que los vértices del polígono estén sobre los puntos equidistantes $x_1, x_2, \ldots, x_n$ y que la constante $h = x_{\nu+1} - x_\nu$ sea suficientemente pequeña. Esto es una consecuencia de la continuidad uniforme de la función continua f en el intervalo I (ver p. 123).

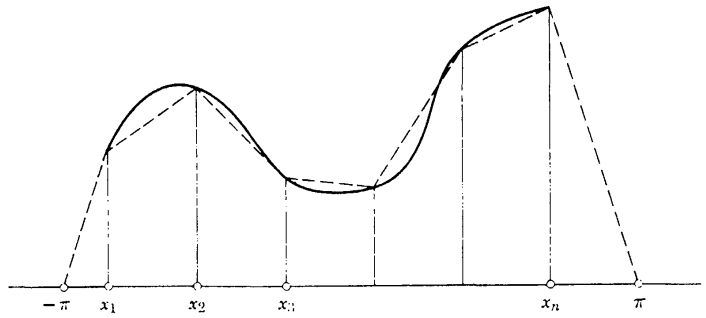


Figura 8.11 Aproximación uniforme de una función continua mediante un polígono.

El siguiente paso es unir, como se indica en la figura, los puntos extremos $-\pi$ y π del intervalo mayor por medio de líneas rectas, extendiendo así $L_n(x)$ a una función lineal por trozos, que nuevamente llamaremos $L_n(x)$, definida en el intervalo cerrado $-\pi \leq x \leq \pi$. Esta función, que se anula

en ambos puntos extremos, puede ahora extenderse periódicamente y, de acuerdo con la sección 8.6*a*, desarrollarse en una serie de Fourier uniformemente convergente, cuya sección polinomial, $S_m(x)$, difiera de $L_n(x)$, en valor absoluto, en una cantidad menor que $\varepsilon/2$ siempre que m sea suficientemente grande. En estas condiciones es $|S_m - f| \leq |S_m - L_n| + |L_n - f| < \varepsilon$, y así *a*) queda demostrado.[1]

Para demostrar (*b*) reemplacemos en cada término de la suma finita $S_m(x)$, según la sección 5.5*b*, p. 471, las funciones trigonométricas $\cos \nu x$ y $\operatorname{sen} \nu x$ por polinomios de Taylor cuyo residuo sea uniformemente pequeño. A continuación, combinando estas últimas aproximaciones construimos un polinomio, $P_N(x)$, para el cual $|P_N(x) - S_m(x)| < \varepsilon/2$, donde N debe escogerse suficientemente grande para asegurar la aproximación dada por $\varepsilon/2$. Combinando aproximaciones, se tiene que en el intervalo más pequeño es $|P_N(x) - f(x)| < \varepsilon$ siempre que m se escoja de manera que

$$|S_m(x) - f(x)| < \varepsilon/2.$$

*c. Aproximación trigonométrica de Fejer para los polinomios de Fourier, usando medias aritméticas

El teorema *a*) de la sección 8.7*b* puede demostrarse de manera muy simple por medio de una construcción directa y más bien explícita del polinomio de aproximación, proporcionada por el notable teorema, debido a L. Fejer, que enunciamos a continuación.

TEOREMA. *Si $S_n(x)$ es el n-ésimo polinomio de Fourier de una función periódica continua, $f(x)$, entonces la media aritmética*

$$F_n(x) = \frac{S_0(x) + \cdots + S_n(x)}{n+1}$$

converge uniformemente a $f(x)$ cuando $n \to \infty$.

El teorema garantiza la convergencia promediando cualquier oscilación perturbadora que pudiera estar presente en la aproximación de Fourier ordinaria.

DEMOSTRACIÓN. La demostración es similar a la del teorema principal sobre desarrollos de Fourier, pero es más simple debido a que el "núcleo" oscilante $\dfrac{\operatorname{sen}(n+\frac{1}{2})x}{2\operatorname{sen}\frac{1}{2}x}$ que allí aparece se reemplaza en este caso por el "núcleo de Fejer" positivo $S_n(t) = \left(\dfrac{\operatorname{sen}\frac{1}{2}(n+1)t}{2\operatorname{sen}\frac{1}{2}t}\right)^2 \cdot \dfrac{2}{n+1}$. Observemos primero que la función $\sigma_n(\alpha) = \frac{1}{2} + \cos\alpha + \cdots + \cos n\alpha$ de la p. 601 puede escribirse en la forma

$$\sigma_n(\alpha) = \frac{\operatorname{sen}(n+\frac{1}{2})\alpha}{2\operatorname{sen}\frac{1}{2}\alpha} = \frac{\operatorname{sen}\frac{1}{2}\alpha \operatorname{sen}(n+\frac{1}{2})\alpha}{2\operatorname{sen}^2\frac{1}{2}\alpha}$$
$$= \frac{1}{2}\frac{\cos n\alpha - \cos(n+1)\alpha}{1-\cos\alpha},$$

haciendo uso de las fórmulas de adición para el coseno. Así resulta la fórmula

$$\frac{\sigma_0(\alpha) + \sigma_1(\alpha) + \cdots + \sigma_n(\alpha)}{n+1} = \frac{1}{2(n+1)}\frac{1-\cos(n+1)\alpha}{1-\cos\alpha}$$
$$= \frac{1}{2(n+1)}\left(\frac{\operatorname{sen}[(n+1)\alpha/2]}{\operatorname{sen}(\alpha/2)}\right)^2$$
$$= s_n(\alpha).$$

Puesto que, por la definición de $\sigma_n(\alpha)$ [ver (14), p. 601], se tiene

$$\frac{1}{\pi}\int_{-\pi}^{\pi} \sigma_k(\alpha)\, d\alpha = 1,$$

debe tenerse también que

$$\frac{1}{\pi}\int_{-\pi}^{\pi} s_n(\alpha)\, d\alpha = 1.$$

Por otra parte [ver (28*a*), p. 611],

$$S_n(x) = \frac{1}{\pi}\int_{-\pi}^{\pi} f(x+t)\sigma_n(t)\, dt$$

y, en consecuencia,

$$F_n(x) = \frac{1}{\pi(n+1)}\int_{-\pi}^{\pi} f(x+t)[\sigma_0(t) + \cdots + \sigma_n(t)]\, dt$$
$$= \frac{1}{\pi}\int_{-\pi}^{\pi} f(x+t)s_n(t)\, dt.$$

Si δ es cualquier número positivo, entonces

$$f(x) - F_n(x) = \frac{1}{\pi}\int_{-\pi}^{\pi} [f(x) - f(x+t)]s_n(t)\, dt$$
$$= \frac{1}{\pi}\int_{-\delta}^{\delta} [f(x) - f(x+t)]s_n(t)\, dt$$
$$+ \frac{1}{\pi}\int_{\delta}^{\pi} [f(x) - f(x+t)]s_n(t)\, dt$$
$$+ \frac{1}{\pi}\int_{-\pi}^{-\delta} [f(x) - f(x+t)]s_n(t)\, dt.$$

[1] El mismo resultado se obtiene cuando I es el intervalo total $-\pi \leq x \leq \pi$, si se supone que $f(\pi) = f(-\pi)$. En este caso se escoge, como antes, un polígono de aproximación, $L_n(x)$, bastando tomar $L_n(-\pi) = L_n(\pi) = f(-\pi) = f(\pi)$.

Para $f(x)$ continua, la continuidad es uniforme y es posible encontrar un δ tal que $|f(x) - f(x+t)| < \frac{1}{2}\varepsilon$ para todo x en $[-\pi, \pi]$ y para $|t| < \delta$. Además, f es acotada: $|f| < M$. Puesto que, por definición,

$$|s_n(t)| \leq \frac{1}{2(n+1)\,\text{sen}^2\,(\delta/2)}, \quad \text{para} \quad \delta \leq |t| \leq \pi,$$

usando $s_n \geq 0$ se encuentra que

$$\begin{aligned}|f(x) - F_n(x)| &\leq \frac{\varepsilon}{3\pi}\int_{-\delta}^{\delta} |s_n(t)|\,dt + \frac{2\pi}{\pi}\int_{\delta}^{\pi} |s_n(t)|\,dt + \frac{2\pi}{\pi}\int_{-\pi}^{-\delta} |s_n(t)|\,dt \\ &\leq \frac{\varepsilon}{3\pi}\int_{-\pi}^{\pi} s_n(t)\,dt + \frac{2\pi}{\pi}\cdot\frac{2\pi}{2(n+1)\,\text{sen}^2\,(\delta/2)} \\ &= \frac{\varepsilon}{3} + \frac{2\pi}{(n+1)\,\text{sen}^2\,(\delta/2)};\end{aligned}$$

y entonces es claro que

$$|f(x) - F_n(x)| \leq \varepsilon$$

para n suficientemente grande. El teorema queda así demostrado.

*d. Aproximación en media y relación de Parseval

La proximidad de dos funciones, $g(x)$ y $h(x)$, en su intervalo cerrado, I, en el cual son continuas, puede medirse, con miras a la convergencia uniforme, por medio del valor máximo de $|g(x) - h(x)|$. Llamando *norma máxima* al máximo valor absoluto de una función continua, $\phi(x)$ en I, podemos expresar la convergencia uniforme de una sucesión de funciones f_n a una función f cuando $n \to \infty$ diciendo que la norma máxima de la diferencia $f - f_n$, o $f_n - f$, tiende a cero: $||f - f_n|| \to 0$.

Para aproximaciones de Fourier (así como para otras importantes teorías matemáticas que rebasan los objetivos de este volumen) es natural considerar otra medida o "norma" para la desviación entre dos funciones o, lo que es suficiente, para la "distancia" de la función idénticamente cero a una función $\phi(x)$. Es ésta la "media cuadrática" o *"norma media cuadrática"*, $\mu = ||\phi||$, definida por el valor promedio

$$\mu^2 = \frac{1}{l}\int_I \phi(x)^2\,dx = ||\phi||^2$$

donde l es la longitud del intervalo I. Esta medida es menos precisa que la de la amplitud máxima, debido a que su pequeñez no significa necesariamente que la función es pequeña en todas partes.

Como un ejemplo, la amplitud de x^n en el intervalo I: $0 \leq x \leq 1$ tiene el valor $(2n+1)^{-\frac{1}{2}}$, el cual puede hacerse arbitrariamente pequeño con

sólo escoger n suficientemente grande, en tanto que la función x^n es igual a 1 para $x = 1$.

Si la norma cuadrática $||f_n - f||$ tiende a cero cuando $n \to \infty$, decimos entonces que f_n tiende a f *"en media cuadrática"*.

Puede ser de utilidad designar como "distancia" a la norma cuadrática, debido a la llamada desigualdad del triángulo, que corresponde a la que es válida para números (ver p. 38). Esta desigualdad, $||f + g|| \leq ||f|| + ||g||$, para las dos funciones f y g, es fácil de deducir: aplicando la desigualdad $pq \leq \frac{1}{2}(p^2 + q^2)$ con $p = \dfrac{f(x)}{||f||}$, $q = \dfrac{g(x)}{||g||}$, e integrando sobre I se obtiene

$$\frac{1}{l}\int_I f(x)g(x)\,dx \leq ||f|| \cdot ||g||.$$

Por tanto,

$$\begin{aligned}||f + g||^2 &= \frac{1}{l}\int_I [f(x) + g(x)]^2\,dx \\ &= ||f||^2 + ||g||^2 + \frac{1}{l}\int_I 2f(x)g(x)\,dx \\ &\leq (||f|| + ||g||)^2,\end{aligned}$$

o sea,

$$||f + g|| \leq ||f|| + ||g||.$$

Con estos conceptos podemos hacer más clara la desigualdad de Bessel de la sección 8.6*b*. Primero demostraremos que: *La aproximación en media cuadrática más precisa para una función continua a trozos, $f(x)$, mediante un polinomio trigonométrico de orden n,*

$$\begin{aligned}T_n &= \frac{c_0}{2} + \sum_{\nu=1}^{n} c_\nu \cos \nu x + d_\nu \,\text{sen}\, \nu x \\ &= \sum_{\nu=-n}^{n} \beta_\nu e^{i\nu x},\end{aligned}$$

[con $\beta_0 = c_0/2$, $\beta_\nu + \beta_{-\nu} = c_\nu$, $i(\beta_\nu - \beta_{-\nu}) = d_\nu$], *en la cual la elección de los coeficientes c_ν, d_ν es arbitraria, está dada por el polinomio de Fourier*

$$\begin{aligned}S_n &= \frac{a_0}{2} + \sum_{\nu=1}^{n} a_\nu \cos \nu x + b_\nu \,\text{sen}\, \nu x \\ &= \sum_{\nu=-n}^{n} \alpha_\nu e^{i\nu x},\end{aligned}$$

donde a_ν, b_ν y α_ν son, respectivamente, precisamente los coeficientes reales y complejos de Fourier determinados a partir de f por medio de las fórmulas (26) *de p.* 609.

La demostración, escrita por brevedad en notación compleja, se obtiene fácilmente usando las relaciones de ortogonalidad (25) en el intervalo $I = [-\pi, \pi]$ para las funciones e^{ivx}:

$$\begin{aligned}
&\frac{1}{l}\int_I \left(f(x) - \sum_{v=-n}^{n} \beta_v e^{ivx}\right)^2 dx \\
&\quad = \frac{1}{l}\int_I \left[f(x)^2 - 2\sum_{v=-n}^{n} \beta_v f(x) e^{ivx} + \left(\sum_{v=-n}^{n} \beta_v e^{ivx}\right)^2\right] dx \\
&\quad = \|f\|^2 - 2\sum_{v=-n}^{n} \beta_v \alpha_{-v} + \sum_{v=-n}^{n} \beta_v \beta_{-v} \\
&\quad = \|f\|^2 - \sum_{v=-n}^{n} \alpha_v \alpha_{-v} + \sum_{v=-n}^{n} (\alpha_v - \beta_v)(\alpha_{-v} - \beta_{-v}) \\
&\quad = \|f\|^2 - \sum_{v=-n}^{n} \alpha_v \bar{\alpha}_v + \sum_{v=-n}^{n} (\alpha_v - \beta_v)(\bar{\alpha}_v - \bar{\beta}_v) \\
&\quad = \|f\|^2 - \sum_{v=-n}^{n} |\alpha_v|^2 + \sum_{v=-n}^{n} |\alpha_v - \beta_v|^2;
\end{aligned}$$

Es claro que esta última expresión alcanza su mínimo cuando los β_v se toman como los coeficientes de Fourier α_v, es decir, $\alpha_v = \beta_v$ o, lo que es lo mismo, $c_v = a_v$, $d_v = b_v$.

Ahora podemos demostrar el *teorema de Parseval* usando los resultados de aproximación obtenidos antes.

La *desigualdad de Bessel,*

$$\tfrac{1}{2}a_0^2 + \sum_{v=1}^{n} (a_v^2 + b_v^2) \leq \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)^2\, dx$$

cuando $n \to \infty$ *se convierte en la igualdad de Parseval,*

$$\tfrac{1}{2}a_0^2 + \sum_{v=1}^{\infty} (a_v^2 + b_v^2) = \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)^2\, dx,$$

válida para cualquier función $f(x)$ *de período* 2π *que sea continua para todo* x.

DEMOSTRACIÓN. Por el teorema de aproximación de Weierstrass para polinomios trigonométricos, podemos escoger una sucesión de polinomios, T_n tal que $[f(x) - T_n(x)] \to 0$ uniformemente en x. Por tanto, también

$$\frac{1}{2\pi}\int_{-\pi}^{\pi} [f(x) - T_n(x)]^2\, dx \to 0 \quad \text{cuando } n \to \infty.$$

Sin embargo, de acuerdo con nuestro último resultado, el polinomio de Fourier

$$S_n(x) = \frac{a_0}{2} + \sum_{v=1}^{n} (a_v \cos vx + b_v \operatorname{sen} vx)$$

constituye, entre todos los polinomios de Fourier de orden n, la mejor aproximación en media para $f(x)$, de manera que

$$\frac{1}{2\pi}\int_{-\pi}^{\pi} [f(x) - S_n(x)]^2\, dx = \frac{1}{2\pi}\int_{-\pi}^{\pi} [f(x) - T_n(x)]^2\, dx.$$

Por tanto,

$$\lim_{n\to\infty} \frac{1}{2\pi}\int_{-\pi}^{\pi} [f(x) - S_n(x)]^2\, dx = 0.$$

Al cuadrar el integrando como se hizo en la p. 628, se obtiene la relación de Parseval.

Finalmente, hacemos la observación de que la relación de Parseval sigue siendo válida cuando $f(x)$ tiene un cierto número de discontinuidades de salto. La demostración, aunque sencilla, no la daremos aquí.

Apéndice I

*A.I.1 Alargamiento del intervalo del período. Teorema de la integral de Fourier

El intervalo base, $-\pi \leq x \leq \pi$, de las funciones periódicas podría reemplazarse por cualquier intervalo $-B \leq x \leq B$. Mediante la transformación $y = \pi x/B$ este intervalo de longitud $2B$ se transforma en el intervalo $-\pi \leq y \leq \pi$, y una función $f(x)$ con período $2B$ se transforma en una función $g(y) = f(By/\pi) = f(x)$ con período 2π. El teorema principal, escrito en forma compleja [ver fórmula (27*a*), p. 610], implica que

$$g(y) = \sum_{v=-\infty}^{\infty} \frac{1}{2\pi}\int_{-\pi}^{\pi} g(t)\, e^{-iv(t-y)}\, dt;$$

y por la transformación se tiene entonces que

$$f(x) = \frac{1}{2\pi}\sum_{v=-\infty}^{\infty} \frac{\pi}{B}\int_{-B}^{B} f(s)\, e^{-iv\pi(s-x)/B}\, ds, \tag{41}$$

donde la variable de integración se ha reemplazado por $s = Bt/\pi$.

La relación (41) es válida para toda función lisa a trozos en $-B \leq x \leq B$.

Hagamos $\pi/B = h$, $v\pi/B = vh = u_v$ y escribamos (41) en la forma

$$\begin{aligned}
f(x) &= \frac{1}{2\pi}\sum_{v=-\infty}^{\infty} h \int_{-B}^{B} f(s)\, e^{-iu_v(s-x)}\, ds \\
&= \frac{1}{2\pi}\sum_{v=-\infty}^{\infty} h e^{iu_v x} H_v,
\end{aligned}$$

donde $H_\nu = \int_{-B}^{B} e^{-iu_\nu s} f(s)\, ds$. El paso formal al límite para $B \to \infty$, o $\Delta u = h \to 0$, es obvio y conduce a

$$f(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{iux}\, du \int_{-\infty}^{-\infty} e^{-ius} f(s)\, ds. \tag{42}$$

Esta es la fórmula de la *integral de Fourier*, que será demostrada rigurosamente en el Volumen II para una amplia clase de funciones f. Tal fórmula puede escribirse en una forma simétrica más clara, en términos de un par de relaciones integrales recíprocas entre una función $f(x)$ y su "transformada de Fourier", $F(u)$:

$$F(u) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} f(s) e^{-ius}\, ds \tag{43}$$

$$f(x) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} F(u) e^{iux}\, dx. \tag{43a}$$

La fórmula de la integral de Fourier, (42), puede escribirse en una forma que no requiere el uso de exponentes imaginarios. Para ello sólo se necesita hacer uso de las expresiones

$$e^{iux}e^{-ius} = e^{iu(x-s)} = \cos u(s-x) - i \operatorname{sen} u(s-x).$$

Puesto que $\operatorname{sen} u(s-x)$ es una función impar de u y $\cos u(s-x)$ una función par, la integración con respecto a u, desde $-\infty$ hasta $+\infty$, del término con seno da una contribución nula, en tanto que al integrar el término con coseno se obtiene el doble del valor obtenido al integrar desde 0 hasta ∞. Por tanto,

$$f(x) = \frac{1}{\pi}\int_0^{\infty} du \int_{-\infty}^{+\infty} f(s) \cos u(s-x)\, ds. \tag{43b}$$

*A.I.2 Fenómeno de Gibbs en puntos de discontinuidad

La convergencia de la serie de Fourier en la vecindad de una discontinuidad de salto exhibe una característica notable que Gibbs descubrió examinando las gráficas de los polinomios de Fourier

$$S_n(x) = \frac{a_0}{2} + \sum_{\nu=1}^{n} (a_\nu \cos \nu x + b_\nu \operatorname{sen} \nu x).$$

Como se hizo resaltar en el capítulo 7, p. 547, la convergencia no uniforme de una sucesión convergente en la vecindad de una discontinuidad de la función límite puede ser visualizada por la forma en que las gráficas continuas de las funciones de aproximación dejan de ser buenas aproximaciones para la gráfica discontinua de la función límite.

En un desarrollo de Fourier estas gráficas no aproximan en forma simple a la gráfica de $f(x)$ aumentada por los segmentos verticales $x = \xi$ que unen los dos puntos extremos en el sitio, ξ, del salto. En vez de ello, las gráficas de S_n presentan ondas que en la proximidad de ξ exceden a las ordenadas $f(\xi + 0)$ y $f(\xi - 0)$, a uno y otro lado, en casi un 9% de la altura total del salto. Así, las gráficas de aproximación se aproximan a la gráfica de $f(x)$ aumentada por un segmento de recta vertical en $x = \xi$, no sólo uniendo los dos puntos extremos de la gráfica de $f(x)$ sino rebasando el segmento de unión en ambos extremos (ver Figs. 8.4 y 8.5, pp. 595 y 596).

El análisis matemático de esta situación es sencillo y bastará discutirlo solamente para la discontinuidad de la función $\chi(x)$ de la sección 8.6*d*, caso al que se redujeron (p. 624) todas las discontinuidades de salto.

La función $\frac{1}{2}\chi(x)$ para x positivo [ver fórmula (23*a*), p. 607] está dada por

$$\tfrac{1}{2}\chi(x) = \tfrac{1}{2}(\pi - x) = \sum_{\nu=1}^{\infty} \frac{\operatorname{sen} \nu x}{\nu}, \quad 0 < x < \pi.$$

Integrando la fórmula (14) de p. 601, se encuentra que

$$S_n(x) = \sum_{\nu=1}^{n} \frac{\operatorname{sen} \nu x}{\nu} = -\tfrac{1}{2}x + \int_0^x \frac{\operatorname{sen} (n + \frac{1}{2})t}{2 \operatorname{sen} \frac{1}{2}t}\, dt$$

Por tanto, el residuo $r_n(x) = \frac{1}{2}\chi(x) - S_n(x)$ toma la forma

$$r_n(x) = \tfrac{1}{2}\pi - \int_0^x \frac{\operatorname{sen} (n + \frac{1}{2})t}{t}\, dt + \rho_n(x),$$

donde

$$\rho_n(x) = \int_0^x \frac{2 \operatorname{sen} \frac{1}{2}t - t}{2t \operatorname{sen} \frac{1}{2}t} \operatorname{sen} (n + \tfrac{1}{2})t\, dt.$$

Puesto que la expresión $(2 \operatorname{sen} \frac{1}{2}t - t)/2t \operatorname{sen} \frac{1}{2}t$ es continua a trozos y tiene una primera derivada que también lo es, del lema de la p. 604 inferimos que $\rho_n(x)$ tiende a cero uniformemente cuando $n \to \infty$ para $0 < x < \pi$. Además,

$$\sigma_n(x) = \tfrac{1}{2}\pi - \int_0^x \frac{\operatorname{sen} (n + \frac{1}{2})t}{t}\, dt = \tfrac{1}{2}\pi - \int_0^{(n+\frac{1}{2})x} \frac{\operatorname{sen} t}{t}\, dt$$

tiende a cero cuando $n \to \infty$ para cualquier valor de x positivo individualmente considerado (ver p. 605). La convergencia, sin embargo, no es uniforme. Es claro que la derivada de $\sigma_n(x)$ se anula en los puntos $x_k = 2k\pi/(2n+1)$, para $k = 1, 2, 3, \ldots$. En forma más precisa, es fácil ver que $\sigma_n(x)$ tiene mínimos en los puntos $x_1, x_3, x_5, \ldots$ y máximos en $x_2, x_4, \ldots$. Más aún, los valores de σ_n en los puntos de mínimos forman una sucesión creciente. Así, $\sigma_n(x)$ tiene como mínimo "absoluto", para x positivos, el valor

$$\sigma_n(x_1) = \tfrac{1}{2}\pi - \int_0^{\pi} \frac{\operatorname{sen} t}{t}\, dt$$

$$= \tfrac{1}{2}\pi - \int_0^{\pi} \left(1 - \frac{1}{3!}t^2 + \frac{1}{5!}t^4 - \cdots\right) dt$$

$$= \pi\left(\tfrac{1}{2} - 1 + \frac{\pi^2}{2\cdot 3\cdot 3} - \frac{\pi^4}{2\cdot 3\cdot 4\cdot 5\cdot 5} + \frac{\pi^6}{2\cdot 3\cdot 4\cdot 5\cdot 6\cdot 7\cdot 7} - \cdots\right)$$

$$\approx -0.090\cdots\pi.$$

Cuando n es grande el residuo r_n es aproximadamente igual a σ_n. En consecuencia, para n grande el polinomio de aproximación S_n excede a la función χ aproximadamente en $(9/100)\pi$, esto es, en cerca de un 9% de la diferencia entre los valores límite derecho e izquierdo de la función en el origen. Así, la rama oscilante de la gráfica de $S_n(x)$ sobrepasa efectivamente la altura de la gráfica de $\chi(x)$ y exhibe el fenómeno de límite descripto anteriormente.

Es fácil ver que los valores medios de Fejer para la suma $S_n(x)$ están libres del fenómeno de Gibbs.

*A.I.3 Integración de series de Fourier

En general, como hemos visto (p. 552), una serie infinita puede integrarse término a término si es uniformemente convergente. Pero en el caso de las series de Fourier tenemos el notable resultado de que la integración por términos siempre es posible. Afirmamos entonces lo siguiente: *Si $f(x)$ es una función continua a trozos en $-\pi \le x \le \pi$, cuyo desarrollo formal de Fourier es*

$$\tfrac{1}{2}a_0 + \sum_{\nu=1}^{\infty} (a_\nu \cos \nu x + b_\nu \operatorname{sen} \nu x),$$

entonces, para dos puntos x_1 y x_2 cualesquiera,

$$\int_{x_1}^{x_2} f(x)\, dx = \int_{x_1}^{x_2} \tfrac{1}{2}a_0\, dx + \sum_{\nu=1}^{\infty} \int_{x_1}^{x_2} (a_\nu \cos \nu x + b_\nu \operatorname{sen} \nu x)\, dx,$$

es decir, la serie de Fourier puede integrarse término a término. Además, la serie del lado derecho converge uniformemente con respecto a x_2 para x_1 fijo.

La parte notable de este teorema es que no sólo no se necesita garantizar la convergencia uniforme de la serie sino que tampoco se necesita hacer uso de su convergencia.

Para demostrar el teorema defínase, como en la p. 623,

$$F(x) = \int_{-\pi}^{x} [f(t) - \tfrac{1}{2}a_0]\, dt.$$

La función $F(x)$ es continua y tiene derivada continua a trozos; además, satisface la condición $F(\pi) = F(-\pi) = 0$, de manera que sigue siendo continua cuando se extiende periódicamente. Así, la serie de Fourier para $F(x)$,

$$\tfrac{1}{2}A_0 + \sum_{\nu=1}^{\infty} (A_\nu \cos \nu x + B_\nu \operatorname{sen} \nu x),$$

converge uniformemente a $F(x)$. Integrando por partes se obtienen, para $\nu \neq 0$, los valores

$$A_\nu = \frac{1}{\pi}\int_{-\pi}^{\pi} F(t) \cos \nu t\, dt = -\frac{1}{\pi}\int_{-\pi}^{\pi} f(t)\, \frac{\operatorname{sen} \nu t}{\nu}\, dt = -\frac{b_\nu}{\nu};$$

$$B_\nu = \frac{1}{\pi}\int_{-\pi}^{\pi} F(t) \operatorname{sen} \nu t\, dt = \frac{1}{\pi}\int_{-\pi}^{\pi} f(t)\, \frac{\cos \nu t}{\nu}\, dt = \frac{a_\nu}{\nu},$$

para los coeficientes de Fourier. Por tanto, la serie

$$F(x_2) - F(x_1) = \sum_{\nu=1}^{\infty} [A_\nu(\cos \nu x_2 - \cos \nu x_1) + B_\nu(\operatorname{sen} \nu x_2 - \operatorname{sen} \nu x_1)]$$

$$= \sum_{\nu=1}^{\infty} \left[-\frac{b_\nu}{\nu}(\cos \nu x_2 - \cos \nu x_1) + \frac{a_\nu}{\nu}(\operatorname{sen} x\nu_2 - \operatorname{sen} \nu x_1)\right]$$

converge uniformemente en ν. Reemplazando $F(x)$ por $\int_{\pi}^{x} [f(x) - \tfrac{1}{2}a_0]\, dx$, obtenemos la relación

$$\int_{x_1}^{x_2} [f(x) - \tfrac{1}{2}a_0]\, dx = \sum_{\nu=1}^{\infty} \int_{x_1}^{x_2} (a_\nu \cos \nu x + b_\nu \operatorname{sen} \nu x)\, dx,$$

como se afirmó.

Apéndice II

*A.II.1 Polinomios de Bernoulli y sus aplicaciones

a. Definición y desarrollo de Fourier

En la obtención de la serie de Taylor (p. 467) desempeñaron un papel importante los polinomios en x de grado $n \ge 1$: $P_n(x) = (x - \xi)^n/n!$, con parámetro ξ. La sucesión de estos polinomios está caracterizada por las condiciones de que todo polinomio P_{n+1} es una función primitiva de P_n, es decir, $P'_{n+1}(x) = P_n(x)$, y además que $P_n(\xi) = 0$ y $P_0(x) = 1$.

Construiremos ahora otra notable sucesión de polinomios mediante integraciones sucesivas: los *polinomios de Bernoulli,* que luego extenderemos como funciones periódicas y desarrollaremos en serie de Fourier.

Los polinomios de Bernoulli, $\phi_n(x)$, para $0 \leq x \leq 1$, están definidos por las siguientes relaciones de recurrencia:

$$\phi_n'(x) = \phi_{n-1}(x), \qquad \phi_0(x) = 1 \tag{44a}$$

$$\int_0^1 \phi_n(x)\, dx = 0, \qquad \text{para } n > 0. \tag{44b}$$

Cuando $\phi_0, \phi_1, \ldots, \phi_n$ son conocidas, la condición (44*a*) determina ϕ_n a menos de una constante de integración arbitraria. Esta constante se calcula después usando la condición (44*b*). Por inducción vemos de inmediato que ϕ_n es un polinomio de grado n cuyos coeficientes son números racionales. Los primeros polinomios de Bernoulli son fáciles de calcular:

$$\phi_0(x) = 1,$$

$$\phi_1(x) = x - \frac{1}{2},$$

$$\phi_2(x) = \frac{1}{2}x^2 - \frac{1}{2}x + \frac{1}{12},$$

$$\phi_3(x) = \frac{1}{6}x^3 - \frac{1}{4}x^2 + \frac{1}{12}x,$$

$$\phi_4(x) = \frac{1}{24}x^4 - \frac{1}{12}x^3 + \frac{1}{24}x^2 - \frac{1}{720}.$$

Para $n > 1$ las relaciones (44*a, b*) dan

$$\phi_n(1) - \phi_n(0) = \int_0^1 \phi_n'(t)\, dt = 0.$$

Por consiguiente, los polinomios ϕ_n pueden ser extendidos, partiendo del intervalo básico $0 \leq x \leq 1$, a todos los valores de x, originándose así funciones *periódicas* $\psi_n(x)$ con período igual a 1. Éstas son las llamadas *funciones de Bernoulli.* La función $\psi_1(x)$ coincide con la función discontinua $\frac{1}{2\pi}\phi(2\pi x - \pi)$ y puede representarse [ver fórmula (23*b*), p. 608] como una serie de Fourier:

$$\psi_1(x) = \frac{1}{\pi}\left(\frac{\operatorname{sen} 2\pi x}{1} + \frac{\operatorname{sen} 4\pi x}{2} + \frac{\operatorname{sen} 6\pi x}{3} + \cdots\right). \tag{45a}$$

Mediante integraciones sucesivas se obtiene entonces

$$\psi_n(t) = (-1)^{(n/2)+1} \cdot \frac{2}{(2\pi)^n} \sum_{k=1}^{\infty} \frac{\cos 2\pi kt}{k^n}, \qquad \text{para } n \text{ par} \tag{45b}$$

$$\psi_n(t) = (-1)^{(n+1)/2} \cdot \frac{2}{(2\pi)^n} \sum_{k=1}^{\infty} \frac{\operatorname{sen} 2\pi kt}{k^n}, \qquad \text{para } n \text{ impar.} \tag{45c}$$

En el intervalo original, $0 \leq x \leq 1$, las funciones periódicas $\psi_n(t)$ son idénticas a los polinomios de Bernoulli $\phi_n(t)$.

Para n, par, ψ_n es una función par; para n impar, ψ_n es impar. En forma equivalente

$$\psi_n(-x) = (-1)^n \psi_n(x). \tag{45d}$$

Los términos constantes en los polinomios de Bernoulli sucesivos forman una notable sucesión de números racionales

$$b_n = \phi_n(0). \tag{46a}$$

$$= \begin{cases} \psi_n(0) \text{ para } n \neq 1, \\ -\frac{1}{2} \text{ para } n = 1. \end{cases}$$

Del desarrollo de Fourier obtenemos inmediatamente

$$b_n = 0 \qquad \text{para } n = 3, 5, \ldots, \tag{46b}$$

$$b_n = (-1)^{(n/2)+1} \cdot \frac{2}{(2\pi)^n} \sum_{k=1}^{\infty} \frac{1}{k^n}, \qquad \text{para } n = 2, 4, \ldots. \tag{46c}$$

Además, es evidente que para n par ($n = 2m$) los signos de b_{2m} se alternan.

En lugar de los números b_n, que decrecen rápidamente al crecer n, Jacobo Bernoulli introdujo los números siguientes, que resultan más adecuados:

$$B_m = (-1)^{m-1}(2m)!\, b_{2m}. \tag{47}$$

Éstos son los llamados *números de Bernoulli.* (La igualdad entre los números $B_{2m}^* = (-1)^{m-1}B_m$ y los números de Bernoulli introducidos en la p. 577 se hará evidente después.) En particular,

$$B_1 = \frac{1}{6}, \qquad B_2 = \frac{1}{30}, \qquad B_3 = \frac{1}{42}, \qquad B_4 = \frac{1}{30},$$

$$B_5 = \frac{5}{66}, \qquad B_6 = \frac{691}{2730}, \qquad B_7 = \frac{7}{6}, \ldots.$$

Como una consecuencia de la fórmula (46*c*) se tiene

$$\sum_{k=1}^{\infty} \frac{1}{k^{2n}} = (-1)^{n-1}(2\pi)^{2n}\tfrac{1}{2}b_{2n} = \frac{(2\pi)^{2n}}{2(2n)!}B_n, \tag{48}$$

la cual es una representación explícita de la función ζ de Riemann, $\zeta(s)$, para enteros $s = 2n$ (ver p. 575), en términos de números conocidos. Con ella es posible obtener, por ejemplo, fórmulas tan notables como

$$1 + \frac{1}{2^2} + \frac{1}{3^2} + \frac{1}{4^2} + \cdots = \frac{\pi^2}{6} = \zeta(2)$$

y

$$1 + \frac{1}{2^4} + \frac{1}{3^4} + \frac{1}{4^4} + \cdots = \frac{\pi^4}{90} = \zeta(4).$$

Cuando $n \to \infty$, los números b_n y B_n tienden, respectivamente, a cero e infinito. En efecto, vemos en seguida que

$$1 < \sum_{k=1}^{\infty} \frac{1}{k^{2n}} \leq \sum_{k=1}^{\infty} \frac{1}{k^2} = \frac{\pi^2}{6} < 2.$$

Y, por consiguiente,

$$2(2\pi)^{-2n} < |b_{2n}| < 4(2\pi)^{-2n}.$$

Puesto que $2\pi > 1$ y $(2\pi)^{-2n} \to 0$ cuando $n \to \infty$, resulta que $b_{2n} \to 0$, en tanto que $b_{2n+1} = 0$. Además

$$B_n = (2\pi)!\,|b_{2n}| > 2(2\pi)!\,(2\pi)^{-2n};$$

y, como es fácil ver, el lado derecho tiende a infinito al crecer n.

*b. La función generadora. La serie de Taylor para la cotangente trigonométrica e hiperbólica

Los números y los polinomios de Bernoulli conducen de manera elegante al desarrollo de Taylor para la cotangente y funciones relacionadas. Estos desarrollos se deducen de la manera más fácil por medio de la llamada *función generadora* de las funciones de Bernoulli, a saber, la función

$$F(t, z) = \sum_{n=0}^{\infty} \psi_n(t) z^n. \tag{49}$$

Es ésta una serie de potencias en z cuyos coeficientes son las funciones de Bernoulli que dependen del parámetro t. Sobre la base del desarrollo de Fourier, de las Ecs. (45) se llega a la estimación

$$|\psi_n(t)| \leq \left[\frac{2}{(2\pi)^n}\right] \sum_{k=1}^{\infty} \frac{1}{k^n} \leq \left[\frac{2}{(2\pi)^n}\right] \sum_{k=1}^{\infty} \frac{1}{k^2}$$
$$= \frac{\pi^2}{3(2\pi)^n} < \frac{4}{(2\pi)^n}$$

para todo t y todo $n \geq 2$. Por tanto, el valor absoluto del n-ésimo término de la serie para $F(t, z)$ es menor que $4(|z|/2\pi)^n$. Así, para todo t el radio de convergencia de la serie de potencias en z es por lo menos igual a 2π, como se ve comparando con la serie

$$4 \sum_{m=1}^{\infty} \left(\frac{|z|}{2\pi}\right)^n.$$

Puesto que para un valor fijo de z tal que $|z| < 2\pi$ la serie para $F(z, t)$ tiene una serie mayorante convergente, independiente de t, de la teoría general (ver p. 551) se deduce que la serie converge uniformemente para todo t. Puede entonces ser integrada término a término en este dominio, y su derivación término a término será posible siempre que la serie resultante sea también uniformemente convergente. Usaremos esta propiedad para encontrar una forma explícita que dé $F(t, z)$ (ver p. 555). Derivando término a término con respecto a t obtenemos formalmente para $0 < t < 1$ (si $t = 0$ o $t = 1$, $\psi_1(t)$ no tiene derivada):

$$\frac{d}{dt} F(t, z) = \sum_{n=1}^{\infty} \psi_{n-1}(t) z^n$$
$$= z \sum_{n=1}^{\infty} \psi_{n-1}(t) z^{n-1}$$
$$= z \sum_{n=0}^{\infty} \psi_n(t) z^n$$
$$= zF(t, z).$$

Esta serie tiene la misma forma que la original y es, en efecto, uniformemente convergente, de manera que la derivación término a término se justifica. En consecuencia, para todo valor fijo de z tal que $|z| < 2\pi$ y para $0 < t < 1$ la función generadora $F(t, z)$ satisface la ecuación diferencial $dF/dt = zF(t, z)$. La solución general de esta ecuación es $F = ce^{zt}$, donde c es un factor cuyo valor depende del parámetro z (ver p. 244). Para determinar c integramos la serie de $F(t, z)$ con respecto a t entre 0 y 1:

$$\int_0^1 F(t, z)\, dt = c \int_0^1 e^{zt}\, dt$$
$$= c\, \frac{e^z - 1}{z}$$
$$= \int_0^1 \sum_{n=0}^{\infty} z^n \psi_n(t)\, dt$$
$$= 1 + \sum_{n=1}^{\infty} z^n \int_0^1 \psi_n(t)\, dt = 1.$$

Por consiguiente, $c = z/(e^z - 1)$, y así llegamos al resultado final:

$$F(t, z) = \frac{ze^{zt}}{e^z - 1}. \tag{50}$$

Haciendo $t \to 0$ en esta expresión se obtiene la serie de Taylor para la función $z/(e^z - 1)$:

$$\lim_{t \to 0} F(t, z) = \frac{z}{e^z - 1} = 1 + \sum_{n=1}^{\infty} b_n z^n.$$

Puesto que $b_1 = -\frac{1}{2}$, al sumar $\frac{1}{2}z$ a ambos lados de esa ecuación resulta

$$\frac{z}{e^z - 1} + \frac{z}{2} = 1 + \sum_{n=2}^{\infty} b_n z^n. \tag{51}$$

De paso se nota que esta fórmula prueba que los números $B_n^* = n!b_n$ son los números de Bernoulli introducidos en la p. 577. Como $b_0 = 1$ y $b_n = 0$ para n impar, se tiene que

$$\begin{aligned} \frac{e^z + 1}{e^z - 1} \cdot \frac{z}{2} &= \frac{z}{2} \cdot \frac{e^{z/2} + e^{-z/2}}{e^{z/2} - e^{-z/2}} \\ &= \frac{z}{2} \cdot \frac{2\cosh \frac{1}{2}z}{2 \operatorname{senh} \frac{1}{2}z} \qquad (52) \\ &= \tfrac{1}{2}z \coth \tfrac{1}{2}z = \sum_{n=0}^{\infty} b_{2n} z^{2n} = \sum \frac{B_{2n}^*}{(2n)!} z^{2n}. \end{aligned}$$

En esta forma se obtiene la serie de Taylor para la cotangente hiperbólica, que se dio en la p. 578. Los coeficientes de Taylor están relacionados de manera muy sencilla con los números de Bernoulli. Hemos, pues, demostrado que el desarrollo es válido para todo $|z| < 2\pi$.

De manera semejante se obtiene la serie de Taylor para la cotangente ordinaria (trigonométrica). Para $|z| < 2\pi$ y $0 < t < 1$, la función generadora

$$G(t, z) = \sum_{n=0}^{\infty} (-1)^n \psi_{2n}(t) z^{2n} \tag{53}$$

se deriva dos veces y así se encuentra que G satisface la ecuación diferencial $d^2G/dt^2 + z^2G = 0$, cuya solución general es $G = a\cos(zt) + b\,\mathrm{sen}\,(zt)$, donde a y b no dependen de t pero sí posiblemente del parámetro z. Para determinar a y b se usan dos condiciones. La primera, $\int_0^1 G(t, z)\,dt = 1$, se encuentra por integración término a término. La segunda,

$$\lim_{t \to 0} \frac{dG(t, z)}{dt} = \tfrac{1}{2}z^2$$

para todo z, se encuentra por derivación término a término, para lo cual se usa el hecho de que si $n > 1$ es

$$\psi_{2n}'(0) = \psi_{2n-1}(0) = b_{2n-1} = 0.$$

Estas condiciones implican que

$$a = \frac{z}{2}\cot\frac{z}{2}, \qquad b = \frac{z}{2},$$

de manera que para $|z| < 2\pi$ y $0 < t < 1$ se tiene

$$G(t, z) = \frac{z}{2} \frac{\cos(zt - z/2)}{\mathrm{sen}\,(z/2)}.$$

Los detalles se dejan al lector como ejercicio.

Haciendo $t \to 0$ en esta fórmula, se obtiene el desarrollo de Taylor para la cotangente (ver p. 578) con $|z| < 2\pi$:

$$G(0, z) = \sum_{n=0}^{\infty} (-1)^n b_{2n} z^{2n} = \tfrac{1}{2}z \cot \tfrac{1}{2}z. \tag{54}$$

c. La fórmula de la suma de Euler-Maclaurin

En la sección 5.4*b* obtuvimos la fórmula de Taylor por medio de sucesivas integraciones por partes. A continuación presentamos una derivación análoga para una fórmula famosa debida a Euler, en la que los polinomios de Bernoulli, o, mejor dicho, sus extensiones periódicas ,toman el lugar que antes correspondió a los polinomios $(t - b)^n/n!$. (Reemplazamos entonces a y b en la p. 467 por 0 y 1; esto siempre puede hacerse transformando la variable t en la variable $s = (t - a)/(b - a)$, por lo que el cambio no es esencial.)

En vez de comenzar con la relación

$$f(1) - f(0) = \int_0^1 f'(t)\,dt, \tag{55}$$

lo que correspondería a la deducción anterior de la fórmula de Taylor, empezamos ahora con la relación

$$\int_0^1 f(t)\,dt = \int_0^1 f(t)\psi_0(t)\,dt, \tag{56}$$

que conducirá a una mayor simetría. Por ser

$$\psi_0(t) = \psi_1'(t), \qquad \psi_1(+0) = -\tfrac{1}{2},$$

y $\psi_1(1 - 0) = \frac{1}{2}$, la fórmula de integración por partes,

$$\int_0^1 u\,dv = uv\,\Big|_0^1 - \int_x^1 v\,du,$$

con $u = f(t)$, $v = \psi_1(t)$, $f(0) = f_0$, $f(1) = f_1$ conduce a (ver también capítulo 3, p. 298):

$$\int_0^1 f(t)\,dt = \tfrac{1}{2}(f_0 + f_1) - \int_0^1 f'(t)\psi_1(t)\,dt,$$

o sea,

$$(57) \qquad \tfrac{1}{2}(f_0 + f_1) = \int_0^1 f(t)\,dt + \int_0^1 f'(t)\psi_1(t)\,dt,$$

que es una expresión explícita para la desviación de la suma del lado izquierdo respecto de la integral $\int_0^1 f(t)\,dt$.

Puesto que una fórmula correspondiente es válida para cualquier intervalo entre dos enteros sucesivos, debido a la periodicidad de $\psi_1(t)$ se obtiene de inmediato la ecuación

$$(58) \qquad \tfrac{1}{2}f_0 + f_1 + f_2 + \cdots f_{n-1} + \tfrac{1}{2}f_n = \int_0^n f(x)\,dx + \int_0^n f'(x)\psi_1(x)\,dx,$$

o bien, para cualquier intervalo $a \leq x \leq b$ con a y b enteros,

$$(58a) \qquad f_a + f_{a+1} + \cdots + f_{b-1} = \int_a^b f(x)\,dx + \int_a^b f'(x)\psi_1(x)\,dx, -\tfrac{1}{2}(f_b - f_a).$$

En esta forma se obtiene una expresión exacta para la diferencia entre la suma del lado izquierdo (el área de los rectángulos inscritos, en el caso de una función creciente) y el primer término del lado derecho (el área bajo la curva). La ecuación (58a) constituye la expresión más sencilla de la *fórmula de la suma de Euler-Maclaurin.*

Es natural tratar de mejorar este resultado repitiendo la integración por partes. Integrando la expresión $\int_a^b f'(x)\psi_1(x)\,dx$ y haciendo $u = f'(x)$, $dv = \psi_1(x)\,dx$ se obtiene

$$\int_a^b f'(x)\psi_1(x)\,dx = f'(x)\psi_2(x)\Big|_a^b - \int_a^b f''(x)\psi_2(x)\,dx.$$

Ahora bien, puesto que

$$\psi_2(b) = \psi_2(a) = \psi_2(0) = b_2,$$

el primer término toma la forma

$$b_2[f'(b) - f'(a)];$$

y el segundo término puede integrarse por partes nuevamente, dando

$$-b_3[f''(b) - f''(a)] + \int_a^b f'''(x)\psi_3(x)\,dx.$$

Como $b_3 = 0$, el primer término de esta última expresión se anula, y si se repite la integración por partes resulta

$$b_4[f'''(b) - f'''(a)] - \int_a^b f''''(x)\psi_4(x)\,dx.$$

Prosiguiendo de esta manera hasta llegar a ψ_{2k} se obtiene la *forma general* de la *fórmula de la suma de Euler,*

$$(59) \qquad f_a + f_{a+1} + \cdots + f_{b-1} = \int_a^b f(x)\,dx - \tfrac{1}{2}[f(b) - f(a)] + \sum_{n=1}^{k} b_{2n}[f^{(2n-1)}(b) - f^{(2n-1)}(a)] + R_k,$$

en la que el residuo R_k puede escribirse en una de las formas siguientes:

$$(60) \qquad R_k = -\int_a^b f^{(2k)}(x)\psi_{2k}(x)\,dx$$

o bien

$$(60a) \qquad R_k = \int_a^b f^{(2k+1)}(x)\psi_{2k+1}(x)\,dx.$$

d. Aplicaciones. Desarrollos asintóticos

Desarrollos convergentes. La fórmula de la suma de Euler puede aplicarse bajo condiciones diferentes. En primer lugar, si $R_k \to 0$ cuando $k \to \infty$, la serie infinita

$$\sum_{n=1}^{\infty} b_{2n}[f^{(2n-1)}(b) - f^{(2n-1)}(a)]$$

converge, y así la fórmula permite expresar en forma exacta la suma de la serie correspondiente. También proporciona desarrollos en series para determinadas funciones.

Desarrollos no convergentes. Como segunda posibilidad, y acaso de mayor importancia, considérese el caso en que el residuo R_k puede no tender a cero cuando $k \to \infty$. Entonces la serie anterior no necesariamente converge. Sin embargo, puede ocurrir que al principio los valores absolutos $|R_k|$ decrezcan al crecer k y que $|R_k|$ sea muy pequeño para valores de k escogidos adecuadamente, pero que posteriormente (para valores de k mayores) $|R_k|$ empiece a crecer marcadamente. En este caso la fórmula de suma puede ser una herramienta importante para fines de cálculo numérico. Aunque no es posible obtener una precisión arbitrariamente grande, como con las series convergentes, se puede calcular el valor del lado izquierdo con un error que en el peor de los casos es igual al mínimo valor de $|R_k|$, lo que a menudo constituye una aproximación excelente. Examinaremos a continuación ejemplos de ambos fenómenos.

Ejemplo. Funciones exponenciales. Consideremos primero la función $f(x) = e^{zx}$ para algún z fijo. Con $a = 0$ y $b = 1$ obtenemos, para cualquier número k, la relación

$$f_0 = \int_0^1 f(x)\,dx - \tfrac{1}{2}[f(1) - f(0)] + \sum_{n=1}^{k} b_{2n}[f^{(2n-1)}(b) - f^{(2n-1)}(a)] + R_k.$$

En consecuencia,

$$1 = \frac{e^z - 1}{z} - \tfrac{1}{2}(e^z - 1) + \sum_{n=1}^{k} b_{2n} z^{2n-1}(e^z - 1) + R_k$$

$$= \frac{e^z - 1}{z}\cdot\left[1 - \frac{z}{2} + \sum_{n=1}^{k} b_{2n} z^{2n}\right] + R_k,$$

donde

$$R_k = -\int_0^1 z^{2k} e^{zx} \psi_{2k}(x)\,dx.$$

Puesto que $|\psi_{2k}(x)| \leqslant 4/(2\pi)^{2k}$ (p. 639), se encuentra que

$$|R_k| \leq |z|^{2k}\cdot e^{|z|}\cdot\frac{4}{(2\pi)^{2k}} = 4e^{|z|}\left(\frac{|z|}{2\pi}\right)^{2k},$$

o sea, $R_k \to 0$, al menos para $|z| < 2\pi$. Por tanto, para estos valores de z puede hacerse que k crezca más allá de toda cota en la fórmula de la suma obteniéndose

$$\frac{z}{e^z - 1} = 1 - \tfrac{1}{2}z + \sum_{n=1}^{\infty} b_{2n} z^{2n} \tag{61}$$

para la función $z/(e^z - 1)$. Esta fórmula ya ha sido obtenida por otros métodos (p. 638). Obsérvese que el intervalo de convergencia es nuevamente $|z| < 2\pi$.

e. Sumas de potencias. Fórmula de recurrencia para los números de Bernoulli

Un ejemplo aún más sencillo de una fórmula de suma de Euler convergente se presenta cuando la serie del lado derecho contiene sólo un número finito de términos, especialmente si $f(x)$ es un polinomio de grado $r \geq 1$, de manera que $f^{(r+1)}(x)$ se anule idénticamente. Escogemos $f(x) = x^r$, $a = 0$, $b = n$ y $k > \frac{1}{2}r$. Por simplicidad, introducimos de nuevo la sucesión B_n^* de números de Bernoulli, definida antes (p. 638), en la forma $B_n^* = n!\, b_n$ para todo n. Observando que

$$B_0^* = 1,\ B_1^* = -\tfrac{1}{2},\ B_3^* = B_5^* = B_7^* = \cdots = B_{2n+1}^* = 0,$$

vemos que la fórmula de Euler (59) se convierte en

$$\begin{aligned} 1 + 2^r + 3^r + \cdots + (n-1)^r &= \int_0^n x^r\,dx + \sum_{\nu=1}^{r} \frac{B_\nu^*}{\nu!}\cdot(f^{(\nu-1)}(n) - f^{(\nu-1)}(0)) \\ &= \frac{n^{r+1}}{r+1} + \sum_{\nu=1}^{r} \frac{B_\nu^*}{\nu!}\, r(r-1)\cdots(r-\nu+2)n^{r-\nu+1} \\ &= \frac{1}{r+1}\left\{n^{r+1} + \sum_{\nu=1}^{r}\binom{r+1}{\nu} B_\nu^* n^{(r+1)-\nu}\right\} \\ &= \frac{1}{r+1}\left\{\sum_{\nu=0}^{r+1}\binom{r+1}{\nu} n^{(r+1)-\nu} B_\nu^* - B_{r+1}^*\right\}. \end{aligned}$$

Esta fórmula puede escribirse simbólicamente en la forma

$$1 + 2^r + 3^r + \cdots + (n-1)^r = \frac{1}{r+1}\{(n + B^*)^{r+1} - B^{*r+1}\}, \tag{62}$$

en la que el término que aparece dentro del paréntesis deberá desarrollarse formalmente haciendo uso del teorema del binomio, y cada una de las "potencias" B^{*k} se reemplazará por el correspondiente número de Bernoulli, B_k^*. Por ejemplo,

$$\begin{aligned} 1 + 2^2 + 3^2 + \cdots + (n-1)^2 &= \tfrac{1}{3}(n^3 + 3n^2 B_1 + 3nB_2) \\ &= \tfrac{1}{6}(2n^3 - 3n^2 + n) \\ 1 + 2^3 + 3^3 + \cdots + (n-1)^3 &= \tfrac{1}{4}n^2(n-1)^2 \end{aligned}$$

(ver p. 81).

Haciendo $n = 1$ la fórmula (62) se reduce a

$$\frac{1}{r+1}\{(1 + B^*)^{r+1} - B^{*r+1}\} = 0,$$

o sea,

$$(1 + B^*)^{r+1} = B^{*r+1} \quad \text{para todo } r = 1; \tag{62a}$$

que es precisamente la fórmula de recurrencia para B_k^* dada en la p. 579.

f. Constante de Euler y serie de Stirling

Un ejemplo de la aplicación de la fórmula de Euler-Maclaurin en el segundo caso, o sea, cuando se tiene divergencia, está dado por la función $f(x) = 1/x$ con $a = 1$ y $b = n$. Por (58a) se ve que

$$\begin{aligned} 1 + \tfrac{1}{2} + \tfrac{1}{3} + \cdots + \frac{1}{n-1} &= \int_1^n \frac{dx}{x} + \frac{1}{2}\left(1 - \frac{1}{n}\right) - \int_1^n \frac{\psi_1(x)}{x^2}\,dx \\ &= \log n + \tfrac{1}{2} - \frac{1}{2n} - \int_1^n \frac{\psi_1(x)}{x^2}\,dx, \end{aligned} \tag{63}$$

o sea,

$$1 + \tfrac{1}{2} + \tfrac{1}{3} + \tfrac{1}{4} + \cdots + \frac{1}{n} - \log n = \tfrac{1}{2} + \frac{1}{2n} - \int_1^n \frac{\psi_1(x)}{x^2}\,dx.$$

Cuando $n \to \infty$ la integral de la derecha converge, pues $|\psi_1(x)| \leqslant \frac{1}{2}$ para todo x. Así, el valor absoluto del integrando es siempre menor que el de la integral convergente $\int_1^\infty dx/x^2$. Resulta entonces la relación

$$\lim_{n\to\infty}\left[\sum_{k=1}^{n}\frac{1}{k} - \log n\right] = \tfrac{1}{2} - \int_1^\infty \frac{\psi_1(x)}{x^2}\,dx = C \tag{64}$$

donde C es una constante definida, la constante de Euler, que se introdujo ya en la p. 542.

Tenemos entonces dos resultados: 1) La serie armónica crece con la misma rapidez que la función logaritmo (ambas divergen a infinito); y 2) existe una expresión explícita para la diferencia entre las dos,

$$\sum_{k=1}^{n}\frac{1}{k} - \log n - C = R_n = \frac{1}{2n} + \int_n^\infty \frac{\psi_1(x)}{x^2}\,dx.$$

Obsérvese que R_n se anula cuando $n \to \infty$ por lo menos en primer orden.

Haciendo $f(x) = \log x$, $a = 1$, $b = n$ en la fórmula (59) de p. 642, se obtiene una aplicación más importante. En este caso,

$$\log 1 + \log 2 + \cdots + \log(n-1) = n\log n - n + 1 - \tfrac{1}{2}\log n$$
$$- \sum_{m=1}^{k} b_{2m}(2m-2)!\left(1 - \frac{1}{n^{2m-1}}\right) + \int_1^n \frac{(2k)!}{x^{2k+1}}\psi_{2k+1}(x)\,dx.$$

Sumando $\log n$ a ambos lados de esta ecuación, resulta

$$\log n! = (n + \tfrac{1}{2})\log n - n + c_k + \sum_{m=1}^{k}\frac{(2m-2)!}{n^{2m-1}}b_{2m} - r_k(n), \tag{65}$$

donde

$$c_k = 1 - \sum_{m=1}^{k} b_{2m}(2m-2)! + \int_1^\infty \frac{(2k)!}{x^{2k+1}}\psi_{2k+1}(x)\,dx$$

$$r_k(n) = \int_n^\infty \frac{(2k)!}{x^{2k+1}}\psi_{2k+1}(x)\,dx.$$

Las integrales impropias convergen para $k > 0$, puesto que las funciones $\psi_{2k+1}(x)$ son periódicas y, por consiguiente, acotadas para todo x (ver p. 326). El valor de la constante c_k puede encontrarse observando que cuando $n \to \infty$ es

$$c_k = \lim_{n\to\infty}\log\left(\frac{n!\,e^n}{n^{n+1/2}}\right)$$

De la fórmula de Stirling (14) de p. 521 (o directamente del producto de Wallis para π dado en la p. 300) se concluye que $c_k = \log\sqrt{2\pi}$. Si los números de Bernoulli b_{2m} se expresan en la forma b_{2m} como $(-1)^{m-1}B_m/(2m)!$ (ver fórmula (47), p. 636), se obtiene la llamada *serie de Stirling:*

$$\log\left(\frac{n!}{\sqrt{2\pi}\,n^{n+1/2}e^{-n}}\right) = \sum_{m=1}^{k}\frac{(-1)^{m-1}B_m}{2m(2m-1)n^{2m-1}} - r_k(n).$$

Esta fórmula es un refinamiento de la "fórmula de Stirling". Para cualquier entero positivo fijo, k, y n suficientemente grande, los términos de la suma se aproximan a cero según los órdenes de aproximación $1/n$, $1/n^3$, $1/n^5$, ..., $1/n^{2k-1}$, respectivamente. El término residual, $r_k(n)$, se aproxima a cero con la misma rapidez que $1/n^{2k}$, pues $\psi_{2k+1}(x)$ es una función acotada. Así, para un valor de k fijo y un valor de n muy grande, cada término de la suma resulta ser muy grande comparado con los términos siguientes, y el residuo será más pequeño que todos los términos de la suma. Se obtiene entonces una fórmula de aproximación de la forma

$$\log\left(\frac{n!}{\sqrt{2\pi}\,n^{n+1/2}e^{-n}}\right) = \frac{B_1}{1\cdot 2}\frac{1}{n} - \frac{B_2}{3\cdot 4}\frac{1}{n^3} + \frac{B_3}{5\cdot 6}\frac{1}{n^5} - + \cdots \tag{66}$$
$$= \frac{1}{12}\frac{1}{n} - \frac{1}{360}\frac{1}{n^3} + \frac{1}{1260}\frac{1}{n^5} - \frac{1}{1680}\frac{1}{n^7} + \frac{1}{1188}\frac{1}{n^9} - + \cdots$$

Este desarrollo no debe ser considerado, sin embargo, como una serie infinita convergente. Es sólo *asintóticamente* correcto, en el sentido de que si cortamos la serie después de un número dado de términos, digamos, k términos, el error r_k es pequeño comparado con todos los términos que se conservan, siempre que n sea suficientemente grande. En ningún caso es posible lograr que el error sea arbitrariamente pequeño tomando más y más términos, una vez que se ha fijado n. De hecho, la serie infinita (66) diverge, como puede deducirse fácilmente de la estimación dada en la p. 637 para los números de Bernoulli. Dado un valor de n grande, existe un número de términos de la serie que podrían usarse. Así, para un valor de n moderadamente grande se tiene la aproximación

$$n! \approx \sqrt{2\pi}n^{n+1/2}e^{n+1/12n},$$

y cuando n es muy grande, la fórmula

$$n! \approx \sqrt{2\pi}n^{n+1/2}e^{n+1/12n-1/360n^3}$$

proporciona una aproximación más precisa; etc.

PROBLEMAS

SECCION 8.1, página 588

1. El *período fundamental*, T, de una función periódica, f, se define como el extremo inferior ("inf") de los períodos positivos de f. Demostrar que:

a) Si $T \neq 0$, entonces T es un período.

b) Si $T \neq 0$, entonces cualquier otro período es un múltiplo entero de T.

c) Si $T = 0$ y f es continua en cualquier punto, entonces f es una función constante.

2. Demostrar que si f tiene períodos inconmensurables, T_1 y T_2, entonces el período fundamental, T, es cero. Dar un ejemplo de una función no constante cuyos períodos sean inconmensurables.

3. Supóngase que los períodos fundamentales de f y g son, respectivamente, a y b. Si a y b son conmensurables, por ejemplo $a/b = q/p$, donde p y q son enteros primos entre sí, demostrar entonces que, por ejemplo, $f + g$ puede tener un período fundamental de cualquier valor m/n, siendo $m = aq = bp$ y n cualquier número natural.

SECCION 8.5, página 613

1. Obtener la serie de Fourier para la función $f(x) = \pi x$ en el intervalo $0 \leq x \leq 1$, en forma de una serie de senos pura y de una serie de cosenos pura.

2. Mostrar cómo puede representarse una función definida en un intervalo acotado arbitrario por medio de una serie de Fourier.

3. Obtener el producto infinito para el coseno a partir de la relación

$$\cos \pi x = \frac{\operatorname{sen} 2\pi x}{2 \operatorname{sen} \pi x}.$$

4. Usando los productos infinitos para el seno y para el coseno, evaluar

a) $\dfrac{2}{1} \cdot \dfrac{2}{3} \cdot \dfrac{6}{5} \cdot \dfrac{6}{7} \cdot \dfrac{10}{9} \cdot \dfrac{10}{11} \cdot \dfrac{14}{13} \cdots$

b) $2 \cdot \dfrac{2}{3} \cdot \dfrac{4}{3} \cdot \dfrac{8}{9} \cdot \dfrac{10}{9} \cdot \dfrac{14}{15} \cdot \dfrac{16}{15}$.

5. Expresar en fracciones parciales la cotangente hiperbólica.

6. Determínense las propiedades especiales de los coeficientes de Fourier de los desarrollos para funciones pares e impares tales que $f(x) = f(\pi - x)$.

SECCION 8.6, página 619

1. Investigar la convergencia del desarrollo de Fourier

$$\cos x + \frac{\cos 2x}{2} + \frac{\cos 3x}{3} + \cdots$$

de la función $-\log 2 \left| \operatorname{sen} \dfrac{x}{2} \right|$.

SECCION 8.7, página 623

1. Demostrar la relación de Parseval para una función lisa por trozos, f, que puede tener cierto número de discontinuidades.

APENDICE II.1, página 634

1. Demostrar que

$$\phi_n(t) = \frac{1}{n!} \sum_{k=0}^{n} \binom{n}{k} B_k^* t^{n-k}.$$

2. Demostrar que para $n \geq 1$ vale

$$\phi_n(t) = (-1)^n \phi_n(1 - t).$$

3. Usando la expresión en fracciones parciales para la cotangente, desarrollar en una serie de potencias en x la función $\pi x \cot \pi x$. Comparando ésta con la serie dada en las pp. 640-641, demostrar que

$$\sum_{\nu=1}^{\infty} \frac{1}{\nu^{2m}} = (-1)^{m-1} \frac{(2\pi)^{2m}}{2 \cdot (2m)!} B_{2m}^*.$$

4. Demostrar que

$$\sum_{\nu=1}^{\infty} \frac{1}{(2\nu - 1)^{2m}} = \frac{(-1)^{m-1}(2^{2m} - 1)\pi^{2m}}{2(2m)!} B_{2m}^*.$$

5. Demostrar que

$$\sum_{\nu=1}^{\infty} \frac{(-1)^{\nu}}{\nu^{2m}} = \frac{(-1)^m (2^{2m} - 2)\pi^{2m}}{2 \cdot (2m)!} B_{2m}^*.$$

6. Usando los productos infinitos para el seno y para el coseno, demostrar que

a) $\log \left(\dfrac{\operatorname{sen} x}{x} \right) = -\sum_{\nu=1}^{\infty} \dfrac{(-1)^{\nu-1} 2^{2\nu-1} B_{\nu}^*}{(2\nu)! \, \nu} x^{2\nu}$;

b) $\log \cos x = -\sum_{\nu=1}^{\infty} \dfrac{(-1)^{\nu-1} 2^{2\nu-1} (2^{2\nu} - 1) B_{\nu}^*}{(2\nu)! \, \nu} x^{2\nu}$.

7. Verificar que

a) $\displaystyle\int_0^1 \frac{\log x}{1 - x} dx = -\frac{\pi^2}{6}$;

b) $\displaystyle\int_0^1 \frac{\log x}{1 + x} dx = -\frac{\pi^2}{12}$.

9

Ecuaciones diferenciales para los tipos más simples de vibraciones

En varias ocasiones anteriores hemos encontrado ecuaciones diferenciales, esto es, ecuaciones a partir de las cuales es posible determinar una función desconocida, y tales que involucran no sólo la función misma sino también sus derivadas.

El problema más simple de este tipo es el de encontrar la integral indefinida de una función dada $f(x)$, o sea, encontrar una función $y = F(x)$ que satisfaga la ecuación diferencial $y' - f(x) = 0$. En el capítulo 3, p. 244, se demostró que una ecuación de la forma $y' = \alpha y$ es satisfecha por una función exponencial $y = ce^{\alpha x}$, y las funciones trigonométricas se caracterizaron por ecuaciones diferenciales (p. 331). Como se vio en el capítulo 4 (pp. 423-424), las ecuaciones diferenciales surgen en relación con los problemas de mecánica, y, en realidad, muchas ramas de las Matemáticas puras y la mayor parte de las Matemáticas aplicadas dependen de las ecuaciones diferenciales. En este capítulo, sin profundizar en la teoría general, consideraremos las ecuaciones diferenciales de los tipos más simples de vibraciones. Éstas no son solamente de interés teórico sino que también son sumamente importantes en las Matemáticas aplicadas.

Será conveniente recordar las siguientes ideas y definiciones generales. Por una *solución* de una ecuación diferencial entendemos una función que, cuando se substituye en la ecuación diferencial, satisface ésta "idénticamente"; esto es, "para todos los valores considerados de la variable independiente". A menudo se usa el término *"integral"* en vez del término *"solución"*. En primer lugar, debido a que el problema es más o menos una generalización del problema ordinario de integrar; y en segundo lugar, debido a que frecuentemente ocurre que la solución se encuentra en realidad por integración.

9.1 Problemas de vibración en Mecánica y en Física

a. Las vibraciones mecánicas más simples

En el capítulo 4 (p. 423) ya se han considera los tipos más simples de vibraciones mecánicas. En ese capítulo se consideró una partícula de masa m, que está en libertad de moverse sobre el eje x y que tiende a regresar siempre a su posición inicial, $x = 0$, bajo la acción de una fuerza restauradora. La magnitud de esta fuerza la consideramos proporcional al desplazamiento x y, en efecto, la hicimos igual a $-kx$, donde k es una constante positiva y el signo negativo expresa el hecho de que la fuerza siempre está dirigida hacia el origen. En esta ocasión supondremos que también está presente una fuerza de fricción y que tal fuerza es proporcional a la velocidad de la partícula y opuesta a ella. Dicha fuerza está entonces dada por una expresión de la forma $-r\dot{x}$, donde r es una constante de fricción positiva. Finalmente, supondremos también que la partícula está sujeta a la acción de una fuerza $f(t)$, que es una función del tiempo t. Entonces, por la ley fundamental de Newton, el producto de la masa m y la aceleración $\ddot{x}$ debe ser igual a la fuerza total, es decir, la fuerza elástica más la fuerza de fricción más la fuerza externa. Esto se expresa mediante la ecuación

$$m\ddot{x} + r\dot{x} + kx = f(t). \tag{1}$$

Esta ecuación gobierna el movimiento de la partícula. Si recordamos los ejemplos anteriores de ecuaciones diferenciales, tales como el problema de integración para $\dot{x} = dx/dt = f(t)$, cuya solución es $x = \int f(t)\,dt + c$, o bien la solución de la ecuación diferencial particular $m\ddot{x} + kx = 0$ de la p. 424, observaremos que estos problemas tienen un número infinito de soluciones diferentes. En el presente caso también encontraremos que existe un número infinito de soluciones, lo que se expresa de la siguiente manera: Es posible encontrar una *solución general* o *integral completa* $x(t)$ de la ecuación diferencial, que dependa no sólo de la variable independiente t sino también de dos parámetros arbitrarios, c_1 y c_2, llamados *constantes de integración*. Asignando valores especiales a estas constantes se obtiene una solución particular, y *toda* solución puede encontrarse dando valores especiales a dichas constantes.

Lo anterior es perfectamente comprensible (ver también p. 422). No puede esperarse que la ecuación diferencial por sí sola determine por completo el movimiento. Por el contrario, es razonable pedir que para un instante dado, digamos, en el tiempo $t = 0$, sea posible escoger la posición inicial, $x(0) = x_0$, y la velocidad inicial, $\dot{x}(0) = \dot{x}_0$ (brevemente, el *"estado" inicial*) de manera arbitraria. En otras palabras, en el tiempo $t = 0$ la partícula debe poder iniciar su movimiento con una posición y una velocidad completamente arbitrarias. Una vez que esto se logra, cabe esperar que el movimiento subsiguiente esté completamente determinado.

Las dos constantes arbitrarias c_1 y c_2 en la solución general permiten seleccionar la solución particular que satisfaga las condiciones iniciales. En la siguiente sección veremos que esto sólo puede realizarse de una manera.

Si no hay fuerzas externas presentes, es decir, si $f(t) = 0$, el movimiento se llama *movimiento libre.* La ecuación diferencial se dice entonces que es *homogénea.* Si $f(t)$ es diferente de 0 para todos los valores de t, decimos que el movimiento es *forzado* y que la ecuación diferencial es *no homogénea.* El término $f(t)$ en ocasiones se llama también *término de perturbación.*

b. Oscilaciones eléctricas

Un sistema mecánico del tipo sencillo que se ha descrito puede realizarse físicamente sólo de manera aproximada. Un ejemplo lo tenemos en el péndulo, siempre que sus oscilaciones sean pequeñas. Las oscilaciones de una aguja magnética, las del núcleo de un teléfono o del diafragma de un micrófono, y otras vibraciones mecánicas, pueden representarse con un buen grado de aproximación por sistemas como el descripto. Sin embargo, existe otro tipo de fenómenos que corresponden de manera más precisa a la ecuación diferencial (1). Este es el caso, por ejemplo, de un circuito eléctrico oscilatorio.

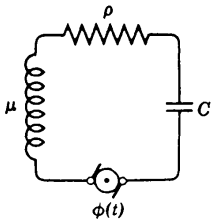


Figura 9.1 Circuito eléctrico oscilatorio.

Consideremos el circuito dibujado en la Fig. 9.1, con inductancia μ, resistencia ρ y capacitancia $C = 1/\kappa$. Supongamos que ese circuito está accionado por una fuerza electromotriz externa, $\phi(t)$, que es una función conocida del tiempo t, tal como el voltaje proporcionado por una dínamo o el voltaje producido por ondas eléctricas. Para describir el proceso que tiene lugar en el circuito denotemos el voltaje en el condensador por E y su carga por Q. Estas cantidades están entonces relacionadas por la ecuación $CE = E/\kappa = Q$. La corriente I, que, como el voltaje E, es una función del tiempo, se define como la rapidez de variación de la carga, por unidad de tiempo; es decir, como la rapidez con la cual disminuye la carga en el condensador: $I = -\dot{Q} = -dQ/dt = -\dot{E}/\kappa$. La ley de Ohm establece que el producto de la corriente por la resistencia es igual a la fuerza electromotriz (voltaje), o sea, es igual al voltaje del condensador, E, menos la fuerza contraelectromotriz debida a la autoinducción, más la fuerza electromotriz externa $\phi(t)$. En esta forma llegamos a la ecuación $I\rho = E - \mu I + \phi(t)$, o $-(\rho/\kappa)\dot{E} = E + (\mu/\kappa)\ddot{E} + \phi(t)$, es decir, $\mu\ddot{E} + \rho\dot{E} + \kappa E = -\kappa\phi(t)$, que el voltaje en el circuito satisface. Vemos, por consiguiente, que resulta una ecuación diferencial precisamente del tipo (1). En vez de la masa, tenemos la inductancia; en vez de la fuerza de fricción, la resistencia; y en vez de la constante elástica, la inversa de la capacidad; en tanto que la fuerza electromotriz externa (excepto por un factor constante) corresponde a la fuerza externa. Si la fuerza electromotriz es 0, la ecuación diferencial es homogénea.

Multiplicando ambos lados de la ecuación diferencial por $-1/\kappa$ y derivando con respecto al tiempo se obtiene para la corriente I la ecuación equivalente:

$$\mu\ddot{I} + \rho\dot{I} + \kappa I = \phi(t),$$

que difiere de la ecuación para el voltaje solamente por los términos del lado derecho, y que cuando se trata de oscilaciones libres ($\phi = 0$) tiene una forma idéntica.

9.2 Solución de la ecuación homogénea. Oscilaciones libres

a. La solución formal

Es fácil obtener una solución de la ecuación homogénea (1), $m\ddot{x} + r\dot{x} + kx = 0$, en la forma de una función exponencial, determinando una constante de manera tal que la expresión $e^{\lambda t} = x$ sea una solución, Si esta solución de prueba y sus derivadas $\dot{x} = \lambda e^{\lambda t}$, $\ddot{x} = \lambda^2 e^{\lambda t}$ se substituyen en la ecuación diferencial y se cancela luego el factor común $e^{\lambda t}$, se obtiene la ecuación cuadrática:

$$m\lambda^2 + r\lambda + k = 0 \tag{2}$$

para λ. Las raíces de esta ecuación son

$$\lambda_1 = -\frac{r}{2m} + \frac{1}{2m}\sqrt{r^2 - 4mk}, \qquad \lambda_2 = -\frac{r}{2m} - \frac{1}{2m}\sqrt{r^2 - 4mk}.$$

Cada una de las expresiones $x = e^{\lambda_1 t}$ y $x = e^{\lambda_2 t}$ es, al menos formalmente, una solución particular de la ecuación diferencial, como puede verse realizando las operaciones en sentido inverso. Pueden presentarse tres casos diferentes:

1. $r^2 - 4mk > 0$. Las dos raíces λ_1 y λ_2 son en este caso reales, negativas y diferentes, y la ecuación diferencial tendrá dos soluciones:

$$u_1 = e^{\lambda_1 t} \quad \text{y} \quad u_2 = e^{\lambda_2 t}.$$

Con ayuda de estas dos soluciones es posible construir de inmediato una solución en la que aparezcan dos constantes arbitrarias. En efecto, después de derivar vemos que

$$x = c_1 u_1 + c_2 u_2 \tag{3}$$

es también una solución de la ecuación diferencial. En la sección 9.3 demostraremos que esta expresión es, de hecho, la solución más general de la

ecuación; es decir, a partir de ella podemos obtener *cualquier* solución de la ecuación substituyendo para ello c_1 y c_2 por valores numéricos adecuados.

2. $r^2 - 4mk = 0$. La ecuación cuadrática tiene una raíz doble. Así se tiene inicialmente, excepto por un factor constante, sólo una solución: $x = w_1 = e^{-rt/2m}$. Pero en este caso es fácil verificar que la función

$$x = w_2 = te^{-rt/2m}$$

es también una solución de la ecuación diferencial.[1] En efecto, se encuentra que

$$\dot{x} = \left(1 - \frac{r}{2m}t\right)e^{-rt/2m}, \qquad \ddot{x} = \left(\frac{r^2}{4m^2}t - \frac{r}{m}\right)e^{-rt/2m},$$

y por substitución vemos que la ecuación diferencial

$$m\ddot{x} + r\dot{x} + \frac{r^2}{4m}x = m\ddot{x} + r\ddot{x} + kx = 0$$

se satisface. Por tanto, la expresión

$$x = c_1e^{-rt/2m} + c_2te^{-rt/2m} \tag{4}$$

da también una solución de la ecuación diferencial, con dos constantes de integración arbitrarias c_1 y c_2.

3. $r^2 - 4mk < 0$. Haciendo $r^2 - 4mk = -4m^2\nu^2$ obtenemos dos soluciones complejas de la ecuación diferencial, dadas por las expresiones $x = u_1 = e^{-rt/2m+i\nu t}$ y $x = u_2 = e^{-rt/2m-i\nu t}$. La fórmula de Euler

$$e^{\pm i\nu t} = \cos \nu t \pm i \operatorname{sen} \nu t$$

nos da, por una parte, las siguientes expresiones para las partes real e imaginaria de la solución compleja u_1:

$$v_1 = e^{-rt/2m}\cos \nu t, \qquad v_2 = e^{-rt/2m}\operatorname{sen} \nu t;$$

y, por otra parte, la representación

$$v_1 = \frac{u_1 + u_2}{2}, \qquad v_2 = \frac{u_1 - u_2}{2i}.$$

Por la segunda forma de la representación, vemos que v_1 y v_2 son soluciones (reales) de la ecuación diferencial. Verificar esto directamente por derivación y substitución es un ejercicio muy sencillo.

A partir de las dos soluciones particulares podemos construir una solución general,

$$x = c_1v_1 + c_2v_2 = (c_1\cos \nu t + c_2 \operatorname{sen} \nu t)e^{-rt/2m}, \tag{5}$$

[1] Obtenemos esta solución, en forma natural, por medio del siguiente proceso de límite: si $\lambda_1 \neq \lambda_2$, entonces la expresión $(e^{\lambda_1 t} - e^{\lambda_2 t})/(\lambda_1 - \lambda_2)$ también representa una solución. Haciendo ahora λ_1 tender a λ_2 y escribiendo λ en vez de λ_1, λ_2, nuestra expresión se convierte en $d(e^{\lambda t})/d\lambda = te^{\lambda t}$.

con dos constantes arbitrarias, c_1 y c_2. Lo anterior puede también escribirse en la forma

$$x = ae^{-rt/2m}\cos \nu(t - \delta), \tag{6}$$

donde se ha escrito $c_1 = a\cos \nu\delta$, $c_2 = a \operatorname{sen} \nu\delta$, y donde a, δ son dos nuevas constantes.

Como se recordará, esta solución ya ha sido obtenida antes para el caso especial $r = 0$ (sección 5.4).

b. Interpretación de la solución

En los casos $r > 2\sqrt{mk}$ y $r = 2\sqrt{mk}$ la solución está dada por la curva exponencial o por la gráfica de la función $te^{-rt/2m}$, que para valores grandes de t se asemeja a la curva exponencial, y también por la superposición de dichas curvas. En estos casos el proceso es aperiódico, es decir, cuando el tiempo aumenta la "distancia" x se aproxima asintóticamente a cero sin haber oscilaciones en la proximidad del valor $x = 0$. El movimiento no es, por tanto, oscilatorio. El efecto de la fricción o *amortiguamiento* es tan grande que impide que la fuerza elástica produzca movimientos oscilatorios.

Algo completamente diferente ocurre para $r < \sqrt{2mk}$, en cuyo caso el amortiguamiento es tan pequeño que aparecen raíces complejas, λ_1, λ_2. La expresión $x = a\cos \nu(t - \delta)e^{-rt/2m}$ da entonces *oscilaciones armónicas amortiguadas*. Son estas oscilaciones que siguen la ley sinusoidal y que tienen una frecuencia angular $\nu = \sqrt{k/m - r^2/4m^2}$, pero cuya amplitud, en lugar de ser constante, está dada por la expresión $ae^{-rt/2m}$. Esto es, la amplitud disminuye exponencialmente; y mientras mayor sea la expresión $r/2m$, más rápido será el decrecimiento. En la literatura de la Física este factor de amortiguamiento se llama a menudo la *constante de atenuación* de la oscilación amortiguada. El término indica que el logaritmo de la amplitud decrece con la razón de decrecimiento $r/2m$. En la Fig. 9.2 se ilustra una oscilación amortiguada de esta clase. Como hicimos antes, llamaremos a la cantidad $T = 2\pi/\nu$ *período* de la oscilación y a la cantidad $\nu\delta$ el *corrimiento de fase*. Para el caso especial $r = 0$ se obtienen oscilaciones armónicas simples con frecuencia $\nu_0 = \sqrt{k/m}$, la llamada *frecuencia natural* del sistema oscilatorio no amortiguado.

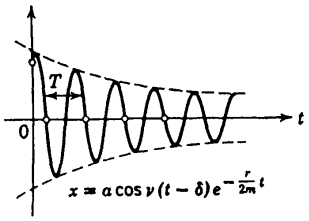


Figura 9.2 Oscilación armónica amortiguada.

c. Satisfacción de condiciones iniciales dadas. Unicidad de la solución

Queda todavía por demostrar que la solución en la que aparecen las dos constantes c_1 y c_2 puede hacerse corresponder a cualquier estado inicial prefijado, y también que tal solución representa todas las soluciones posibles de la ecuación. Supongamos que se ha encontrado una solución que en el tiempo $t = 0$ satisface las condiciones iniciales $x(0) = x_0$ $\dot{x}(0) = \dot{x}_0$, donde los números x_0 y $\dot{x}_0$ pueden tener cualquier valor. Entonces en el caso 1 de la sección 9.2*a* (p. 652) debemos hacer

$$c_1 + c_2 = x_0,$$
$$c_1\lambda_1 + c_2\lambda_2 = \dot{x}_0.$$

Tenemos así dos ecuaciones lineales para las constantes c_1 y c_2, y dichas ecuaciones admiten como soluciones únicas los valores

$$c_1 = \frac{\dot{x}_0 - \lambda_2 x_0}{\lambda_1 - \lambda_2}, \qquad c_2 = \frac{\dot{x}_0 - \lambda_1 x_0}{\lambda_2 - \lambda_1}.$$

En el caso 2 el mismo procedimiento da las dos ecuaciones lineales

$$c_1 = x_0,$$
$$\lambda c_1 + c_2 = \dot{x}_0 \left(\lambda = -\frac{r}{2m}\right),$$

a partir de las cuales es posible obtener nuevamente c_1 y c_2. Por último, en el caso 3 las ecuaciones que determinan las constantes toman la forma

$$a \cos \nu\delta = x_0,$$
$$a\left(\nu \operatorname{sen} \nu\delta - \frac{r}{2m}\cos \nu\delta\right) = \dot{x}_0,$$

cuyas soluciones son:

$$\delta = \frac{1}{\nu}\operatorname{arc\,cos}\frac{x_0}{a}, \qquad a = \frac{1}{\nu}\sqrt{\left[\nu^2 x_0^2 + \left(\dot{x}_0 + \frac{r}{2m}x_0\right)^2\right]}.$$

En esa forma hemos demostrado que las soluciones generales pueden satisfacer condiciones iniciales arbitrarias. Falta aún demostrar que no existe otra solución. Para ello basta probar que dado cierto estado inicial no existen dos soluciones diferentes.

Supongamos que existen dos soluciones, $u(t)$ y $v(t)$, tales que $u(0) = x_0$, $u(0) = \dot{x}_0$ y $v(0) = x_0$, $\dot{v}(0) = \dot{x}_0$. Entonces la diferencia es también una solución, por lo que se tendrá $w = u - v$. Esta solución corresponde, por consiguiente, a un estado inicial de reposo, es decir, a un estado tal que en

el tiempo $t = 0$ la partícula se encuentra en su posición de reposo y tiene velocidad cero. Pero en este caso la partícula nunca se pondría en movimiento. Para demostrarlo multipliquemos ambos lados de la ecuación diferencial $m\ddot{w} + r\dot{w} + kw = 0$ por $2\dot{w}$ y recordemos que $2\dot{w}\ddot{w} = (d/dt)\dot{w}^2$ y $2w\dot{w} = (d/dt)w^2$. Obtenemos así

$$\frac{d}{dt}(m\dot{w}^2) + \frac{d}{dt}(kw^2) + 2r\dot{w}^2 = 0.$$

Integrando entre los límites $t = 0$ y $t = \tau$ y usando la condición inicial $w(0) = 0$, $\dot{w}(0) = 0$, se encuentra que

$$m\dot{w}^2(\tau) + kw^2(\tau) + 2r\int_0^\tau \left(\frac{dw}{dt}\right)^2 dt = 0.$$

Esta ecuación, sin embargo, encierra una contradicción cuando la función w es diferente de 0 para cualquier instante $\tau > 0$, pues el lado izquierdo es entonces positivo (ya que m, k y r son positivos) en tanto que el lado derecho es cero. Por tanto, $w = u - v$ es siempre igual a 0; lo que prueba que la solución es única.

9.3 La ecuación no homogénea. Oscilaciones forzadas

a. Comentarios generales. Superposición

Antes de proceder a la solución del problema en el caso de una fuerza externa $f(t)$, es decir, a la solución de la ecuación no homogénea, hacemos el siguiente comentario.

Si w y v son dos soluciones de la ecuación no homogénea, la diferencia $u = w - v$ satisface la ecuación homogénea. Esto puede verse de inmediato por substitución. Recíprocamente, si u es una solución de la ecuación homogénea y v lo es de la ecuación no homogénea, entonces $w = u + v$ es también una solución de esta última ecuación. Por tanto, a partir de una solución [1] de la ecuación no homogénea obtenemos *todas* las soluciones de dicha ecuación sumando la integral completa de la ecuación homogénea. Sólo necesitamos, por consiguiente, encontrar *una* solución de la ecuación no homogénea. Físicamente esto significa que si tenemos una oscilación forzada debido a una fuerza externa, al superponerle una oscilación libre arbitraria, representada por una solución de la ecuación homogénea, obtenemos un fenómeno que satisface la misma ecuación no homogénea que la oscilación forzada original. Si se trata de una fuerza de fricción, el movimiento libre en el caso oscilatorio debe tender a desaparecer a medida que el tiempo transcurre, debido al factor de amortiguamiento $e^{-rt/2m}$. Por

[1] Llamada a menudo *integral o solución particular*.

consiguiente, cuando se tiene una vibración forzada y existe una fuerza de fricción, es indiferente cuál sea la vibración libre superpuesta: el movimiento tenderá siempre hacia el mismo estado final al transcurrir el tiempo.

En segundo lugar, observamos que el efecto de una fuerza $f(t)$ puede partirse de la misma manera que la fuerza misma. Con esto queremos decir lo siguiente: si $f_1(t)$, $f_2(t)$ y $f(t)$ son tres funciones tales que

$$f_1(t) + f_2(t) = f(t),$$

y si $x_1 = x_1(t)$ es una solución de la ecuación diferencial $m\ddot{x} + r\dot{x} + kx = f_1(t)$, y $x_2 = x_2(t)$ es una solución de $m\ddot{x} + r\dot{x} + kx = f_2(t)$, entonces $x(t) = x_1(t) + x_2(t)$ es una solución de la ecuación

$$m\ddot{x} + r\dot{x} + kx = f(t). \tag{7}$$

Un resultado semejante se tiene, por supuesto, cuando $f(t)$ está compuesta por cualquier número de términos. Este resultado simple pero importante se llama *principio de superposición*. Su demostración es casi inmediata a partir de la ecuación misma. Subdividiendo la función $f(t)$ en dos o más términos es posible, de la misma manera, dividir la ecuación diferencial en varias ecuaciones, cuyo manejo, en determinadas circunstancias, puede resultar más sencillo.

El caso más importante es el de una fuerza externa, $f(t)$, periódica. Tal fuerza externa puede descomponerse en componentes periódicas puras, mediante un desarrollo en serie de Fourier, y en esa forma [1] puede aproximarse con la precisión deseada por medio de la suma de un número finito de tales funciones. Es suficiente, por tanto, encontrar la solución de la ecuación diferencial bajo la hipótesis de que su lado derecho tiene la forma

$$a \cos \omega t \quad \text{o,} \quad b \operatorname{sen} \omega t,$$

donde a, b y ω son constantes arbitrarias.

En vez de trabajar con estas funciones trigonométricas, podemos obtener la solución de manera más sencilla y elegante si usamos la notación compleja. Haciendo $f(t) = ce^{i\omega t}$, el principio de superposición muestra que sólo necesitamos considerar la ecuación diferencial

$$m\ddot{x} + r\dot{x} + kx = ce^{i\omega t}, \tag{8}$$

en la que c representa una constante arbitraria real o compleja. Tal ecuación diferencial representa en realidad dos ecuaciones diferenciales reales. Pues si partimos el lado derecho en dos términos, tomando, por ejemplo, $c = 1$, y escribimos $e^{i\omega t} = \cos \omega t + i \operatorname{sen} \omega t$, entonces x_1 y x_2, las soluciones de las dos ecuaciones diferenciales *reales* $m\ddot{x} + r\dot{x} + kx = \cos \omega t$ y $m\ddot{x} + r\dot{x} + kx = \operatorname{sen} \omega t$, se combinan para dar la solución $x = x_1 + x_2$ de la ecuación diferencial *compleja*. Recíprocamente, resolviendo primero la ecuación diferencial en forma compleja, la parte real de su solución nos dará la función x_1, y la parte imaginaria la función x_2.

[1] Suponiendo que sea continua y lisa por trozos (p. 621), el caso más importante en Física.

b. Solución de la ecuación no homogénea

Resolveremos la ecuación (8) mediante un artificio que surge en forma natural de la intuición. Supongamos que c es real y (para determinado instante) que $r \neq 0$. Adivinamos entonces que existirá un movimiento cuyo ritmo es el mismo que el de la fuerza periódica externa, y de acuerdo con esta hipótesis intentamos encontrar una solución de la ecuación diferencial, en la forma

$$x = \sigma e^{i\omega t}, \tag{9}$$

en la que sólo falta determinar el factor σ, que es independiente del tiempo. Substituyendo esta expresión y sus derivadas, $\dot{x} = i\omega\sigma e^{i\omega t}$, $\ddot{x} = -\omega^2\sigma e^{i\omega t}$, en la ecuación diferencial y eliminando el factor común $e^{i\omega t}$ obtenemos la ecuación

$$-m\omega^2\sigma + ir\omega\sigma + k\sigma = c,$$

o sea,

$$\sigma = \frac{c}{-m\omega^2 + ir\omega + k}. \tag{10}$$

Recíprocamente, observamos que para este valor de σ la expresión es realmente una solución de la ecuación diferencial. Sin embargo, para expresar con claridad lo que este resultado significa es necesario realizar algunas transformaciones.

Comenzaremos escribiendo el factor complejo σ en la forma

$$\sigma = c\frac{k - m\omega^2 - ir\omega}{(k - m\omega^2)^2 + r^2\omega^2} = c\alpha e^{-i\omega\delta}, \tag{11}$$

donde el "factor de distorsión" positivo α y el "corrimiento de fase" $\omega\delta$ son expresados en términos de las cantidades dadas m, r, k mediante las ecuaciones

$$\alpha^2 = \frac{1}{(k - m\omega^2)^2 + r^2\omega^2}, \qquad \operatorname{sen} \omega\delta = r\omega\alpha, \qquad \cos \omega\delta = (k - m\omega^2)\alpha.$$

Con esta notación nuestra solución toma la forma

$$x = c\alpha e^{i\omega(t-\delta)},$$

cuyo significado es el siguiente: a la fuerza $c \cos \omega t$ le corresponde el "efecto" $c\alpha \cos \omega(t - \delta)$, y a la fuerza $c \operatorname{sen} \omega t$ le corresponde el efecto $c\alpha \operatorname{sen} \omega(t - \delta)$.

Vemos por lo anterior que el efecto es una función del mismo tipo que la fuerza, esto es, una *oscilación* no amortiguada (sostenida). Esta oscilación difiere de la oscilación que representa a la fuerza en que su amplitud ha aumentado en la proporción $\alpha : 1$ y la fase ha variado según el ángulo $\omega\delta$. Por supuesto, el mismo resultado puede obtenerse fácilmente sin hacer uso de la notación compleja, aunque a costa de cálculos algo más complicados.

De acuerdo con el comentario que se hizo al principio de esta sección, una vez encontrada esta solución tenemos completamente resuelto el problema; pues superponiendo cualquier oscilación libre podemos obtener la oscilación forzada más general.

Reuniendo los resultados anteriores podemos establecer lo siguiente:

La integral completa[1] *de la ecuación diferencial*

$$m\ddot{x} + r\dot{x} + kx = ce^{i\omega t}$$

(en la que $x \neq 0$*) es* $x = c\alpha e^{i\omega(t-\delta)} + u$*, donde* u *es la integral completa de la ecuación homogénea* $m\ddot{x} + r\dot{x} + kx = 0$, y las cantidades α y δ están definidas por las ecuaciones

$$\alpha^2 = \frac{1}{(k - m\omega^2)^2 + r^2\omega^2}; \qquad \operatorname{sen}\omega\delta = r\omega\alpha, \qquad \cos\omega\delta = (k - m\omega^2)\alpha.$$

Las constantes de esta solución general permiten hacer que la solución satisfaga un estado inicial arbitrario, es decir, dados cualesquiera valores de x_0 y $\dot{x}_0$, las constantes pueden escogerse de manera tal que $x(0) = x_0$ y $\dot{x}(0) = \dot{x}_0$.

c. La curva de resonancia

Para formarnos una idea de la solución que hemos obtenido y de lo que significa en las aplicaciones, estudiaremos el "factor de distorsión", α, como función de la "frecuencia de excitación", ω, es decir, la función

$$\phi(\omega) = \frac{1}{\sqrt{(k - m\omega^2)^2 + r^2\omega^2}}. \tag{13}$$

Tal investigación en detalle es motivada por el hecho de que, dadas las constantes k, m y r o, como a veces se dice, dado un "sistema oscilatorio", es posible considerar éste como sujeto a fuerzas de excitación periódicas cuyas frecuencias angulares son muy diferentes entre sí; siendo importante, por tanto, considerar la solución de la ecuación diferencial para este amplio rango de fuerzas de excitación. Con el fin de describir en forma conveniente la función, introducimos la cantidad $\omega_0 = \sqrt{k/m}$. Este número, ω_0, es la

[1] O *general*.

frecuencia angular que tendrían las oscilaciones libres del sistema si la fricción fuese nula; o, más brevemente, la *frecuencia (angular) natural del sistema no amortiguado* (ver p. 655). En realidad, la frecuencia del sistema *libre,* en presencia de la función r, no es igual a ω_0 sino que tiene el valor

$$\nu = \sqrt{\frac{k}{m} - \frac{r^2}{4m^2}},$$

donde suponemos que $4km - r^2 > 0$. (Si éste no es el caso, el sistema libre no tiene frecuencia: es aperiódico.)

La función $\phi(\omega)$ tiende asintóticamente al valor cero a medida que la frecuencia de excitación tiende a infinito y, de hecho, tal función se anula con orden $1/\omega^2$. Además, $\phi(0) = 1/k$; en otras palabras, una fuerza excitadora de frecuencia cero y magnitud igual a 1, o sea, una fuerza constante de magnitud uno, da origen a un desplazamiento $1/k$ del sistema oscilatorio. En la región de valores positivos de ω la derivada $\phi'(\omega)$ no puede anularse excepto allí donde se anula la derivada de la expresión $(k - m\omega^2)^2 + r^2\omega^2$, esto es, para un valor $\omega = \omega_1 > 0$ que satisfaga la ecuación

$$-4m\omega(k - m\omega^2) + 2r^2\omega = 0.$$

Es claro que para que tal valor exista deberá tenerse que $2km - r^2 > 0$; y en este caso

$$\omega_1 = \sqrt{\frac{k}{m} - \frac{r^2}{2m^2}} = \sqrt{\omega_0^2 - \frac{r^2}{2m^2}}.$$

Puesto que la función $\phi(\omega)$ es positiva en todas partes, crece monótonamente para valores pequeños de ω y se anula en el infinito, el valor ω_1 debe dar un máximo. Llamamos a esta frecuencia ω_1 la "frecuencia de resonancia" del sistema.

Substituyendo la expresión para ω_1 encontramos que el valor del máximo es

$$\phi(\omega_1) = \frac{1}{r\sqrt{(k/m - r^2/4m^2)}}.$$

Cuando $r \to 0$ dicho valor máximo crece más allá de toda cota. Para $r = 0$, es decir, para un sistema oscilatorio no amortiguado, la función $\phi(\omega)$ tiene una discontinuidad infinita en el valor $\omega = \omega_1$. Es éste un caso límite, que consideraremos más adelante en forma especial.

La gráfica de la función $\phi(\omega)$ se llama *curva de resonancia* del sistema. Para $\omega = \omega_1$ (y, en consecuencia, para valores pequeños de r en la vecindad de la frecuencia "natural") la distorsión de amplitud, $\alpha = \phi(\omega)$, es particularmente grande. Este resultado es la expresión matemática del "fenómeno de resonancia", el cual, para valores fijos de m y k, es más evidente a medida que r decrece.

En la Fig. 9.3 se muestra una familia de curvas de resonancias correspondientes, todas ellas, a los valores $m = 1$ y $k = 1$ (y, por tanto, a $\omega_0 = 1$), pero con valores $D = \frac{1}{2}r$ diferentes. Observamos que para valores pequeños de D aparece una resonancia muy marcada cerca de $\omega = 1$; y en el caso límite, $D = 0$, $\phi(\omega)$ tendría para $\omega = 1$ una discontinuidad infinita en vez de un máximo. Al crecer D, el máximo se mueve hacia la izquierda, y para el valor $D = 1/\sqrt{2}$ tenemos $\omega_1 = 0$. En este último caso el punto en el que la tangente es horizontal se ha movido hacia el origen y el máximo ha desaparecido. Si $D > 1/\sqrt{2}$ el cero de $\phi'(\omega)$ no existe, y la curva de resonancia deja de tener máximo y la resonancia desaparece.

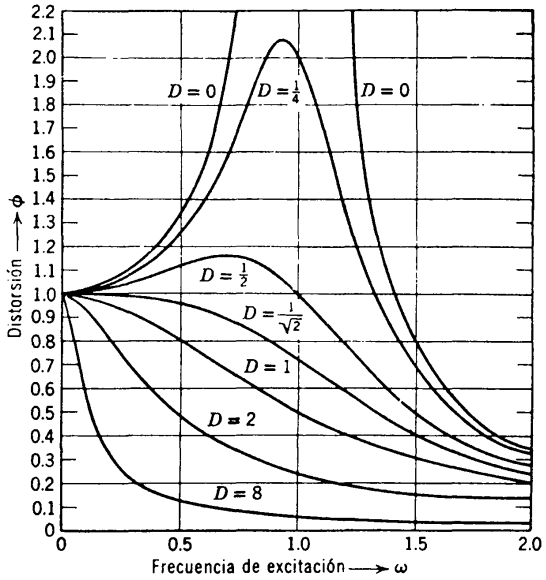


Figura 9.3 Curvas de resonancia.

En general, el fenómeno de resonancia desaparece tan pronto como se cumple la condición

$$2km - r^2 \leq 0.$$

En el caso de la igualdad, la curva de resonancia alcanza su máxima altura, $\phi(0) = 1/k$ para $\omega_1 = 0$; y su tangente es horizontal allí; y la curva, después de un camino inicialmente casi horizontal, decrece hacia el valor cero.

d. Una discusión adicional sobre las oscilaciones

La discusión anterior no es suficiente. Para entender en realidad el fenómeno del movimiento forzado es necesario hacer hincapié en lo siguiente. La integral particular $c\alpha e^{i\omega(t-\delta)}$ debe considerarse como un *estado límite* hacia el cual tiende cada vez más, a medida que el tiempo transcurre, la integral completa

$$x(t) = c\alpha e^{i\omega(t-\delta)} + c_1u_1 + c_2u_2.$$

Esto es así porque la oscilación libre $c_1u_1 + c_2u_2$, superpuesta a la integral particular, se desvanece al transcurrir el tiempo. Este desvanecimiento será lento si r es pequeño y será rápido si r es grande.

Supongamos, por ejemplo, que al iniciarse el movimiento, esto es, en el tiempo $t = 0$, el sistema está en reposo, de manera que $x(0) = 0$ y $\dot{x}(0) = 0$. Con estos datos podemos determinar las constantes c_1 y c_2, las cuales, como puede verse de inmediato, *no* se anulan simultáneamente. Aunque la frecuencia de excitación es aproximada o exactamente igual a ω_1, de manera que existe resonancia, la relativamente gran amplitud $\alpha = \phi(\omega_1)$ no aparece al principio. Por el contrario, quedará encubierta por la función $c_1u_1 + c_2u_2$ y aparecerá por primera vez cuando aquélla se haya desvanecido; es decir, aparece más lentamente cuando r es menor.

Para el sistema no amortiguado, esto es, para $r = 0$, la solución "falla" si la frecuencia de excitación es igual a la frecuencia natural $\omega_0 = \sqrt{k/m}$, pues entonces $\phi(\omega_0)$ es infinita. Por consiguiente, no es posible en este caso obtener una solución de la ecuación $m\ddot{x} + kx = e^{i\omega t}$ en la forma $\sigma e^{i\omega t}$. Es posible, sin embargo, obtener inmediatamente una solución particular de la forma $x = \sigma te^{i\omega t}$. Substituyendo esta expresión en la ecuación diferencial y recordando que

$$\dot{x} = \sigma e^{i\omega t}(1 + i\omega t), \qquad \ddot{x} = \sigma e^{i\omega t}(2i\omega - t\omega^2),$$

se tiene

$$\sigma(2im\omega - m\omega^2 t + kt) = 1,$$

y, puesto que $m\omega^2 = k$,

$$\sigma = \frac{1}{2im\omega}$$

Así, *cuando existe resonancia en un sistema no amortiguado la solución es de la forma*

$$x = \frac{t}{2im\omega}e^{i\omega t} = \frac{t}{2i\sqrt{km}}e^{i\omega t}.$$

Usando la notación real, cuando $f(t) = \cos\omega t$ se tiene que

$$x = \frac{1}{2}\frac{t}{\sqrt{km}}\operatorname{sen}\omega t$$

y cuando $f(t) = \operatorname{sen} \omega t$ el resultado es

$$x = -\frac{1}{2}\frac{t}{\sqrt{km}}\cos \omega t.$$

Hemos, pues, encontrado una función que podríamos llamar "oscilación", pero cuya amplitud crece linealmente con el tiempo. La oscilación libre superpuesta no se desvanece, pues no es amortiguada; sin embargo, conserva su amplitud original y llega a no ser importante en comparación con la amplitud creciente de la oscilación forzada especial. El hecho de que en este caso la solución oscile hacia arriba y hacia abajo entre cotas positivas y negativas que crecen continuamente con el tiempo representa el significado real de la discontinuidad infinita que presenta la función de resonancia para un sistema no amortiguado.

e. Comentarios sobre la construcción de instrumentos de registro

En una gran variedad de aplicaciones en la Física y en la Ingeniería resulta de suma importancia la discusión precedente. En el caso de muchos instrumentos, tales como galvanómetros, sismógrafos, circuitos osciladores de radiorreceptores y en diafragmas de micrófonos, el problema consiste en registrar un desplazamiento x producido por una fuerza externa periódica. En tales casos la cantidad x satisface nuestra ecuación diferencial, por lo menos en primera aproximación.

Si T es el período de oscilación de la fuerza externa, es posible desarrollar ésta en una serie de Fourier de la forma

$$f(t) = \sum_{l=-\infty}^{\infty} \gamma_l e^{il(2\pi/T)t},$$

o, mejor aún, considerarla como representada con suficiente aproximación por la suma trigonométrica $\sum_{l=-N}^{N} \gamma_l e^{il(2\pi/T)t}$ compuesta solamente por un número finito de términos. Por el principio de superposición (p. 657), la solución $x(t)$ de la ecuación diferencial estará representada (no considerando la oscilación libre superpuesta) por una serie infinita [1] de la forma

$$x(t) = \sum_{l=-\infty}^{\infty} \sigma_l e^{il(2\pi/T)t},$$

o aproximadamente por una expresión finita de la forma

$$x(t) = \sum_{l=-N}^{N} \sigma_l e^{il(2\pi/T)t}.$$

En vista de nuestros resultados anteriores,

$$\sigma_l = \gamma_l \alpha_l e^{-i\delta_l(2\pi l/T)}$$

y

$$\alpha_l^2 = \frac{1}{\left(k - ml^2\dfrac{4\pi^2}{T^2}\right)^2 + r^2 l^2 \dfrac{4\pi^2}{T^2}}, \quad \tan\frac{2\pi l}{T}\delta_l = \frac{2\pi l r}{T\left(k - m\dfrac{4\pi^2 l^2}{T^2}\right)}$$

[1] Las cuestiones de convergencia no se discutirán aquí.

Podemos entonces describir la acción de cualquier fuerza externa periódica de la siguiente manera: si descomponemos la fuerza de excitación en componentes periódicas puras, los términos individuales de la serie de Fourier, entonces cada componente experimenta una distorsión de amplitud y un corrimiento de fase propios, y los efectos separados pueden luego superponerse por adición. Si sólo nos interesa la distorsión de amplitud (en las aplicaciones el corrimiento de fase es sólo de importancia secundaria [2] y además puede tratarse de la misma manera que la distorsión de amplitud), un estudio de las curvas de resonancia nos da plena información acerca de la forma en que los movimientos del aparato de registro "siguen" a la fuerza de excitación externa. Para valores muy grandes de l, o de $\omega[=(2\pi/T)l]$, el efecto de la frecuencia de excitación sobre el desplazamiento x difícilmente será perceptible. Por otra parte, *todas* las frecuencias de excitación en la vecindad de ω_1, la frecuencia (angular) de resonancia, afectarán marcadamente a la cantidad x.

En la construcción de aparatos de medición y registro de cantidades físicas, las constantes m, r y k están a nuestra disposición, por lo menos dentro de límites amplios. Estos valores podrían escogerse de manera que la forma de la curva de resonancia se adapte lo mejor posible a las necesidades especiales de las mediciones que interesan. Intervienen aquí dos consideraciones de importancia. En primer lugar, es deseable que el aparato tenga la máxima sensibilidad posible; esto es, que para todas las frecuencias ω en cuestión el valor de α sea tan grande como las condiciones lo permitan. Para valores pequeños de ω, como ya hemos visto, α es aproximadamente proporcional a $1/k$, de manera que el número $1/k$ es una medida de la sensibilidad del instrumento para frecuencias de excitación pequeñas. La sensibilidad puede, por consiguiente, mejorarse aumentando $1/k$, o sea, haciendo más débil la fuerza restauradora.

El otro punto importante se refiere a la necesidad de una *ausencia de distorsión relativa*, o "indiferencia" (con respecto a las frecuencias). Supongamos que la representación $f(t) = \sum_{l=-N}^{N} \gamma_l e^{il(2\pi/T)t}$ es la adecuada para la fuerza de excitación. Decimos entonces que el aparato registra la fuerza de excitación con ausencia de distorsión relativa o con "indiferencia" si para todas las frecuencias angulares $\omega \leq N(2\pi/T)$ el factor de distorsión tiene aproximadamente el mismo valor. Esta condición es indispensable cuando se desea obtener conclusiones acerca del proceso de excitación directamente a partir del comportamiento del aparato, por ejemplo, si se desean reproducir sonidos musicales tanto graves con o agudos por medio de un radio o una grabadora de cinta, y con una relación de intensidades aproximadamente correcta. Una reproducción "sin distorsión" nunca puede lograrse en forma total, pues la curva de resonancia no tiene porción horizontal alguna. Puede, sin embargo, intentarse escoger las constantes m, k y r del aparato de manera tal que no exista una marcada resonancia, y también para que la curva tenga una tangente horizontal en su comienzo, con lo cual $\varphi(\omega) = \alpha$ será aproximadamente constante para valores pequeños de ω. Como ya sabemos, esto puede lograrse haciendo

$$2km - r^2 = 0.$$

Dadas las constantes m y k, es posible reunir los requisitos mencionados ajustando adecuadamente el valor de la fricción, r, lo cual se logra, en el caso de un circuito eléctrico, por ejemplo, insertando una resistencia de valor conveniente. La curva de resonancia nos muestra que entonces el sistema no amortiguado prácticamente no introduce distorsión para frecuencias que van desde 0 hasta la frecuencia natural, ω_0; y que por encima de esta frecuencia el amortiguamiento es considerable. Obtenemos, por consiguiente, una ausencia de distorsión relativa en un intervalo dado de frecuencias escogiendo, primero, m tan pequeña y k tan grande que la frecuencia natural ω_0 del sistema no amortiguado sea mayor que cualquiera de las frecuencias de excitación consideradas, y luego escogiendo el factor de amortiguamiento, r, de acuerdo con la ecuación $2km - r^2 = 0$.

[2] Ya que, por ejemplo, es imperceptible por el oído humano.

Lista de fechas biográficas

Abel, Niels Henrik (1802-1829)
Archimedes (287?-212 a.C.)
Barrow, Isaac (1630-1677)
Bernoulli, Jakob (1654-1705)
Bernoulli, Johannes (1667-1748)
Bessel, Friedrich Wilhelm (1784-1846)
Bolzano, Bernhard (1781-1848)
Brahe, Tycho (1546-1601)
Briggs, Henry (1556?-1630)
Cantor, Georg (1845-1918)
Cauchy, Augustin (1789-1857)
Coulomb, Charles Augustin de (1736-1806)
Darboux, Gaston (1842-1917)
Dedekind, Richard (1831-1916)
De Moivre, Abraham (1667-1754)
Descartes, (Cartesius) René (1596-1650)
Dirichlet, Gustav Lejeune (1805-1859)
Einstein, Albert (1879-1955)
Euclid (alrededor de 300 a.C.)
Euler, Leonhard (1707-1783)
Fejer, Lipot (1880-1959)
Fermat, Pierre de (1601-1665)
Fourier, Joseph (1768-1830)
Fresnel, Augustin (1788-1827)
Gauss, Carl Friedrich (1777-1855)
Gibbs, Josiah Willard (1839-1903)
Gregory, James (1638-1675)
Guldin, Paul (1577-1643)
Hermite, Charles (1822-1901)

Hölder, Otto (1860-1937)
Huygens, Christian (1629-1695)
Jensen, J. L. W. V. (1859-1925)
Kepler, Johannes (1571-1630)
Lagrange, Joseph Louis (1736-1813)
Lambert, Johann Heinrich (1728-1777)
Landau, Edmund (1877-1938)
Leibnitz, Gottfried Wilhelm von (1646-1716)
L'Hôpital, Guillaume, François Antoine de (1661-1704)
Lipschitz, Rudolf Otto (1832-1903)
Lorentz, Hendrik Antoon (1853-1928)
Maclaurin, Colin (1698-1746)
Michelson, Albert Abraham (1852-1931)
Morley, Edmund Williams (1838-1923)
Napier, John (1550-1617)
Newton, Isaac (1642-1727)
Ohm, Georg Simon (1787-1854)
Parseval, Marc Anton (A. ?-1836)
Ptolemy (Claudius Ptolemaeus) (Siglo II a.C.)
Raabe, Joseph Ludwig (1801-1859)
Riemann, Bernhard (1826-1866)
Rolle, Michael (1652-1719)
Schwarz, Hermann Amandus (1843-1921)
Seidel, Philipp Ludwig von (1821-1896)
Simpson, Thomas (1710-1761)
Stirling, James (1692-1770)
Taylor, Brook (1685-1731)
Vega, George (1754-1802)
Wallis, John (1616-1703)
Weierstrass, Karl (1815-1897)

Indice

A

Abel, criterio de, 534
  teorema de, 586
Abierto, intervalo, 28
Absoluto, valor, 27, 128
Aceleración, 191, 413
  componente normal de, 413
  componente tangencial de, 413
Acotada, 27
Acotada, sucesión, 94
Acotado, 27
Acotado, conjunto de números reales, 120
Adición, teorema de (de la trigonometría), 333
Afín, función o "mapeo", 44
Algebraica, ecuación, 127
  función, 72
Algebraico, número, 103
Alternas, corrientes, 599
Amplitud, de oscilación, 428, 598-599
  de una partición, 214
Analítica, función, 561
Angulo, entre dos direcciones, 359
  de dirección, 182
  de inclinación, 359
  polar, 402
Antecedente, 42
Aproximación, mediante polinomios trigonométricos y racionales, 623
  de Weierstrass, 584, 624
  en media cuadrática, 628
  sucesiva, 513
  trigonométrica de Fejer, 625
Arc cot x, 236
Arco, simple dirigido, 352, 356
  orientado, 352
  sentido del, 352
  simple, 352
Arc sen x, 235
  valor principal del, 233
Arc tan x, 236, 458, 462, 568
Area, 125, 413
  Limitada por curvas cerradas, 446
Area, en coordenadas polares, 388
  bajo una curva, 143
  dentro de curvas cerradas, 382
  orientada, 383, 385
Aritmética, media, 163, 164
  media, 40, 132, 161, 211
  media geométrica, 136
Aritmético, promedio, 163
Armónica, media, 131
  oscilación, 654
  serie, 644
  simple, 423, 593
Armónica, 593
Asociativa, ley, 25
Astroide, 446, 452,
Atenuación, constante de, 654
Atracción gravitacional, 430, 439
Autoinducción, 249
Axioma, 110
  de continuidad, 33

B

Batidos, 593
Bessel, desigualdad de, 620, 629
Binaria, representación, 35
Bionomiales, coeficientes, 82, 133
  coeficientes, generales, 474
  serie, 473, 486, 653
  teorema, 82

C

Cadena, regla de la cadena para la derivación, 239
Caída de los cuerpos, 184, 191

Caída libre, 419
Capacidad, 441
Cardioide, 452
Casti dodos, números de una sucesión infinita, 113
Catenaria, 395
Catenoide, 395
Cauchy, criterio de convergencia de, 97, 120, 528
  criterio de Cauchy para la convergencia uniforme, 550
  desigualdad de Cauchy, 39, 131
  fórmula de Cauchy para el resto en la fórmula de Taylor, 466, 469
Centro, de gravedad, 164
  de masa, 390
Cerradas, curvas, 358
Cíclica, permutación, 358
Cicloide, 429, 453
  "común", 347, 362, 365, 394
  evoluta de, 445
Cinética, energía, 393, 437
Círculo de convergencia, 569
Circunferencia,
  de curvatura, 378, 477, 478, 493
  osculadora, 378, 477
  representación racional, 312
Coeficientes de Fourier, 602
  indeterminados, 562
Coeficientes, orden de magnitud, 622
Compacto, 119
Complejo, conjugado, 128
  número, 126, 127
Componente, 398
Composición, función, 76
Compuesta, función, 239
Cóncavo, 257, 375
Condensador, 441
Confocal, 453
Conjugado, número complejo, 128
Conjunto, 25
  de números naturales, 25, 26
  de números racionales, 26, 28, 32, 129
  de números reales, 31, 36
  numerables, 121
Conjunto, numerable, 121
Conmensurable, 29
Conmutativa ley, 26
Contacto, 377
  de curvas, 475
  de enésimo orden, 475
  de orden infinito, 479
Continua, seccionalmente o por trozos, 603
  uniformemente, 123
Continuidad, 47, 55, 123
  axioma de, 33
  definición de, 57
  de funciones compuestas, 78
  de funciones derivables, 199
  de Hölder, 140
  de Lipschitz, 67
  módulo de, 65, 199
  principio, 118
  uniforme, 64
Continuo, de números, 25
  espacio-tiempo, 382
Continuo de números reales su plenitud, 118
Convergencia, criterio de Cauchy para la, 98, 528
  círculo de, 569
  criterio integral para, 585
  de integrales impropias, 224
  de series, 98
  uniforme, 545, 548, 550, 551
Convergente absolutamente, 573
  condicionalmente, 573
Convexa, 257
  función, 375
Corriente, 249, 599
Cota, inferior máxima, 120
  integral, 61
  superior mínima, 120-121
Cotangente, resolución de fracciones parciales, 617
  desarrollo de Taylor para la, 640
Coulomb, ley de, 439
Criterio de comparación, 537
Crítico, punto, 262, 263, 365
Cuadrática, aproximación en media, 627, 628
  función, 72
Cuadratura, 500
Cúbica, parábola, 52
Curva, área bajo, 144
  Centro de masa de, 392
  cerrada, 352, 354, 358
  Cerrada simple, 358, 360
  contacto de, 475
  distancia a lo largo de, 370
  en coordenadas polares, 345
  esquinas de, 365

lados positivos y negativo de una, 360, 364
longitud de, 366
movimiento restringido a una, 417
normal de una, 363, 442
orientada, 358, 359
osculante, 378
paralela, 356
pendiente de una, 180, 182
polar, 355
representación paramétrica, 343
secante a una, 178
simple, 352
tangente a una, 178
Curvatura, 372, 413
centro de, 376, 442, 447
circunferencia de, 376, 477, 478, 493
invariancia de, 380
radio de, 376, 441
signo de, 373
Cúspide, 189, 362
de astroide, 446
de evoluta, 441

D

Darboux, integral de, 220
Decimal, fracción periódica, 91
representación decimal de números reales, 33
Dedekind, corte de, 113
Definida, expresión cuadrática, 303
De Moivre, teorema de, 128, 567
Densidad, 29
de números racionales, 116
Dependiente, variable, 42
Derivabilidad de funciones, 187, 280
Derivable, función, 182, 201
Derivación, 178
regla de la cadena, 240
reglas para, 223
Derivada, 180
de un producto, 224
de una suma, 224
de un cociente, 225
hacia adelante, 189
hacia atrás, 189
superior, 191, 225
Descartes, hoja de, 453
Desigualdad, de Bessel, 620, 629
de Cauchy, 39, 131
de Schwartz, 218
del triángulo, 627
Desigualdades, 36
de Cauchy-Schwartz, 39, 218
representación geométrica, 54
triángulo, 38, 627
Desplazamiento, 380
Diferencial, 200, 224
de cos x, 331
de función exponencial, 244
de sen x, 331
ecuación, 649
homogénea, 651
integral de, 649
Diferencias, 489, 490
cociente de, 180, 211
Dígito, 33
Dirección, 401
ángulo de, 183, 358, 400
Directores, cosenos, 400, 412
Dirichlet, integral de, 328, 573
serie de, 583
Dirigidas, líneas rectas, 358
Discontinuidad, infinita, 59
de salto, 55, 59
Discriminante, 303
Distancia entre dos funciones, 627
Distritibutiva, ley, 26
Divergente, sucesión, 94
serie, 98
Dominio de una función, 42, 45

E

Ecuación, algebraica, 127
Ejes de coordenadas, cambio de, 378
Eléctricas, oscilaciones, 651
Electricidad, cantidad de, 441
Eléctrico, circuito, 249, 599
Electromotriz, fuerza, 249, 599
Eliminable, singularidad, 59, 63, 470
Elipse, 396
área encerrada, por 387
evoluta de, 446
longitud de, 454
representación paramétrica racional, 346
Elíptica, función, 319
integral, 318, 340, 396, 428, 454, 565
Encaje de intervalos, 31, 32
Energía, conservación de, 423, 437, 438

cinética, 393, 438
potencial, 439
Entorno (vecindad), 36
Envolvente, 442
Enfriamiento, ley de, 247
Epicicloide, 348
por redondeo, 504
por truncamiento, 504
Escalar, 397
Escape, velocidad de, 435, 440
Esquinas, 365
Estacionario, punto, 261, 365
Estereográfica, proyección, 45
Euler, constante de, 543, 644
fórmula de, 567
Evoluta, 377, 441, 445
cúspide de, 442
de cicloide, 445
de elipse, 446
Exponencial, función, 74, 173, 174, 238, 270-271, 472
ecuación diferencial de la, 244
orden de magnitud de la, 270
serie de potencias para, 562
Extensión, 48
Extrapolación, 493
Extremo, 259
inferior *(infimum)*, 121
puntos relativos, 259, 261
superior *(supremum)*, 121

F

Factorial, 79, 328
Fase, 593
desplazamiento o corrimiento de, 591, 598, 654
Fejer, aproximación trigonométrica de, 625
núcleo de, 627
Fermat, principio de mínimo tiempo de, 267
Fijo, punto, 517
Fourier, coeficiente de, 602, 610, 620
fórmula integral de, 632
integración de, 634
serie de, 589, 603
transformada de, 631
Fraccionaria, parte, 356
Fracción, decimal, 33
parcial, 306
Frecuencia, circular, 591, 598
de oscilación, 423
natural, 654
resonancia, 660
Fresnel, integral, de, 329
Fuerza, 416, 417
elástica, 422
resultante, 415
Función, 41
acotada, 124
afín, 44
algebraica, 72
analítica, 561, 569
composición, 76, 239
compuesta, 239
cóncava, 357
continua, 123, 124, 187, 199
continuación periódica, de, 357
continua en el sentido de Lipschitz, 67
continua en el sentido de Lispchitz, monótona, 53
convexa, 375
cuadrática, 72
derivabilidad de, 182, 187, 199, 201, 280
derivada de una, 240
derivadas superiores de una, 241
de una variable compleja, 569
dominio de una, 42, 45
elemental, 109, 282
elíptica, 319
en sobre, 43
explícita, 282
exponencial, 74, 173, 238, 470, 562
gamma, 327
gráfica de una, 42, 43
hiperbólica, 250, 381, 568
"Hölder-continua", 68
identidad, 78
impar, 53
integrable, 150
inversa, 69, 77
límite de, 105
lineal, 72
monótonamente creciente, decreciente, 199
multivalente (multiforme), 49
par, 53
periódica, 355, 588

peso, 164
primitiva, 208, 210
promedio de una, 161
racional, 71
rango de, 42, 43, 45
trigonométrica, 73, 186, 295, 318, 567
univalente (unívoca, uniforme), 49
"uno a uno", 45
vectorial, 411
zeta, 577
Fundamental, teorema, del cálculo, 206, 208, 209

G

Gamma, función, 327
Gauss, criterio de, 583
Geométrica, media, 40, 137
serie, 90, 91
Gibbs, fenómeno de, 631
Gráfica de una función, 42, 43
Gravedad, 416
centro de, 164
Gravitacional, aceleración, 430
constante, 430
Guldin, regla de, 391

H

Hipérbola, área encerrada, por, 390
rectangular, 51, 252
representación racional, 293, 312
Hiperbólica, cotangente, serie de Taylor para, 638
expresiones exponenciales para, 668
función, 250, 381
inversa, 253
teorema de adición para, 252
Hipocicloide, 348, 453
Hölder, condición de, 67
continuidad de, 67, 140
exponente de, 67
Homogénea, ecuación, 652

I

Identidad, transformación, función o "mapeo", 78
Imagen, 43
Impar, función, 53
Impedancia, 600
Impropia, integral, 572
Inclinación, ángulo, de, 359
Inconmensurable, 29
frecuencias, 592
Indefinida, integral, 206, 209, 210
Independiente, variable, 42
Indeterminadas, expresiones, 483
Indeterminadas, formas, derivadas de, 483
Indice, 448, 451
Inductancia, 599
Inducción, 80
matemática, 80
Inercia, momento de, 392
Infimo extremo inferior, 120
Infinita, sucesión, 79
serie, 98
Inflexión, punto de, 258, 375, 477
tangente de, 258
Inicial, condición, 332, 417, 655
estado, 650
Instantánea dirección de movimiento, 412
Integrable, función, 150
Integración, por partes, 295
constantes de, 652
de funciones racionales, 302
Integral(es), 144
aditividad, 158
cotas, 161
cálculo de, 502
criterio de convergencia para integrales impropias, 324
criterio, para convergencia de series, 585
de Darboux, 220
de Dirichlet, 330, 573
de ecuaciones diferenciales, 649, 650
definición analítica, 145
definidas, 166
de Fresnel, 329
de línea, 385
de Riemann, 150, 220
elementales, 363
elípticas, 318, 340, 453, 565
impropias, 320, 330, 572
indefinida, 165, 206, 210, 211
notación, de Leibnitz, para, 147

representaciones, 451
signo de, 147
Integrando, 148
seccionalmente continuo, 321
Intermedio, valor, propiedad del, 132
teorema del, 68, 124
teorema del, cálculo diferencial, 194, 212, 243
Interpolación, 487
error de, 491
lineal, 203
polinomio de, 489
Intervalos, 28, 46-47
abierto, 28
acotado, 47
cerrado, 28
encaje de, 31
Invariancia, 378
Inversa, 69, 77
derivada de, 228
Involuta, 443, 445, 446
Irracional, número, 30, 114, 129
Isócranas, 429
Iteración, método de, 516

J

Jensen, desigualdad de, 336

K

Kepler, tercera ley de, 433

L

Lagrange, forma de, para el resto en la fórmula de Taylor, 466, 469
fórmula de interpolación, de, 493, 494
Leibnitz, criterio de convergencia de, 531
notación de, para la integral, 147
regla de, 225, 334
Leibnitz-Gregory, serie de, 462, 608
Lemniscata, 125
área en, 390, 396
L'Hospital, regla de, 483
Límite, definición de, 93
de una función, 105
de una sucesión, 84, 93, 115
lateral derecho, 589
lateral izquierdo, 589
operaciones de, 95
punto, 118
Línea integral de, 385
Lineal, función, 72
inerpolación, 203
Lipschitz, condición de, 67
continuidad de, 67
Logaritmo, 74, 207, 272
cálculo de, 512
de base cualquiera, 175
desarrollo de, 459
función, 167
natural, 167
orden de magnitud de, 270
teorema de adición, 169
Longitud, 412
como un parámetro, 370
de curva en coordenadas polares, 369
de elipse, 454
definición alternativa, 368
invariancia de, 368
Lorentz, transformación de, 379

M

Maclaurin, teorema de, 470
Magnitud, orden de, 269
"Mapeo" (función), 42
afín, 44
en sobre, 43
identidad, 78
perspectivo, 45
"uno a uno", 45, 53, 78, 79
Máximo, 259, 478
absoluto, 260, 261
estricto, 260, 265
existencia de, 125
norma, 627
relativo, 259
valor, 262
Mayorante, 538, 551
Media aritmética, 40, 161
altura media, 390
aritmética, geométrica, 136
armónica, 131, 132
geométrica, 40, 131
velocidad media, 183, 212
Medio (*ver* valor medio),

Mínimo, 124, 259, 478
absoluto, 261, 262
relativo, 259
valor, 261
Módulo, 128
de continuidad, 64, 199
Momento, 390
de inercia, 392
Monótona, función, 53, 198
sucesión, 97, 119
Movimiento, circular, 432
de cuerpos en caída, 415
ecuación de, 416
forzado, 651
ley de Newton del, 414
oscilatorio, 427
restringido a una curva vinculado, 418
sobre una curva dada, 423
uniforme, 184
Multiplicación, ley de, 175

N

Natural, frecuencia, 654
logaritmo, 167
números naturales, 25, 46
Newton, 141, 180, 183, 414
fórmula de interpolación de, 488, 490
ley de la gravitación de, 430
ley de movimiento de, 414, 418
método de, 512, 519
Normal, positiva, 364
a una curva, 363
Norma, máxima, 627
media cuadrática, 627
Núcleo de Féjer, 625
Nula, sucesión, 113
Numerable, conjuhto, 121
Número(s), algebraico(s), 102
complejo, 126, 127
complejo conjugado, 127
continuo de, 25, 31
eje de, 27
irracional, 30, 114, 129
natural, 25
racional, 26, 129
real, 31, 114

O

Odógrafa, 455
Ohm, ley de, 249, 600, 651
Operaciones, racionales, 26
8 con límites, 95, 112
Orden, 115
de puntos, 358, 359
Orden de magnitud, 269
de anulación de una función, 273
de una función, 273
notación "O" y "o", 274
Orientación, 357
en sentido contrario a las manecillas del reloj, 361
Ortogonales, direcciones, 407
Oscilación, 591
amplitud de, 423, 429
armónica amortiguada, 654
eléctrica, 651
forzada, 657
frecuencia de, 423
libre, 652
mecánica, 650
período de, 654
Osculante circunferencia, 477
Osculantes, parábolas, 476, 493
O y o, notaciones, 274

P

Parábola, 52, 72
cúbica, 53
de Neil, 190
osculante, 476, 493
Paraleta, curva, 455
Paralelo, desplazamiento, 379, 397
Parámetro, cambio de, 345
tiempo como, 347
Parcial, fracción, 306
suma, 98
Par, función, 53
Parseval, ecuación de, 647
teorema de, 630
Partición, "amplitud" de, 214
Pendiente de curva, 180
Péndulo, cicloidal, 429, 445
oscilación del, 428
período de oscilación del, 566
simple, 428

Periódicas, fracciones decimales, 91
funciones, 588
Periódico, movimiento, 433
Período, 355, 588
de función periódica, 355, 588, 630, 647
de movimiento, 433
de oscilación, 654
integral sobre un, 589
intervalo de, 630
Perspectiva(o), función, transformación, o "mapeo", 45
Peso, factores de, 164
de un cuerpo, 416
función de, 164
π, 103
producto de Wallis para, 300
Polar, ángulo, 401
Polares, coordenadas, 126
área en, 388
longitud de curva en, 369
Polinomios, 71
de interpolación, 487
trigonométricos, 595
Postulados, 110
Potencial energía, 439-440
Potencias, series de, 458, 467, 556, 569
intervalo de convergencia, 557
para función exponencial, 562
Potencias, sumas de, 643
con exponentes arbitrarios, 174
Presión atmosférica, 247
Primitiva, función, 208, 211
Primos, números,
serie de inversos de, 576
Producto, infinito, 574
simbólico, 76, 239, 403
Promedio aritmético, 163
altura promedio, 390
de una función, 161
pesado, 164
razón incremental promedio,
velocidad promedio, 183, 212
Promedio pesado, 164
Propiedades, globales, 366
locales, 366
Proyección, estereográfica, 45, 311
Punto, estacionario, 261, 365
de inflexión, 375
racional, 28

R

Racionales, funciones, 71
numerables, 121
números, 26, 129
operaciones, 26
puntos, 28
Radianes, medida en, 74
Radio de convergencia, de una serie de potencias, 569
Radio de curvatura (*véase* curvatura),
Raíces de la unidad, 128, 129
Rango de una función, 42, 43, 45
Rapidez, 371, 412, 413
Razón incremental, promedio, 182
instantánea, 182, 183
Razón y raíz, criterios de, 538
Reales, números, 31, 114
no numerables, 121
plenitud de los, 118
representación binaria, 35
representación decimal, 33
Rectangular, hipérbola, 51, 252
Rectificabilidad, 367, 453
Redondeo, error por, 504
Referencia, sistema de, 378, 382
Reflexión, ley de, 266
Refracción, ley de, 266
Regla de posición falsa, 515
Relatividad, teoría especial de, 381
Resistencia, 559
Resonancia, curva de, 660
frecuencia de, 660
Resorte, 440
Restricción, 48
Riemann, integral de, 150, 220
función zeta de, 636
suma de, 150, 320
Rolle, teorema general de, 487
teorema de, 196
Rotación de ejes, 410
sentido de, 359, 360

S

Salto, discontinuidad de, 55, 59
Schwarz, desigualdad de, 218
Secante de curva, 178
Seccionalmente continua, 573
suave, 619

Sentido de rotación, 359
positivo, 352
sobre un arco, 352
Serie(s), absolutamente convergente(s), 528, 533, 535
armónica, 530
binomial, 69, 486, 563
comparación de, 436
condicionalmente convergente, 528, 534, 535
convergente, 98, 528
de Dirichlet, 583
de Fourier, 587
de números primos recíprocos, 576
de potencias, 556
diferenciación de, 554
divergente, 74
geométrica, 90, 91
hipergeométrica, 582
infinita, 98, 472
integración de, 552
mayorantes, 538
operaciones con, 436
reordenamiento de términos, 535
suma de infinitas, 527
trigonométricas, 588
Series, convergencia uniforme de, 548, 550, 551
Sgn x, 55, 59
Simbólico, producto, 76, 239, 402
Simple, péndulo, 418, 419, 428
Simpson, regla de, 503, 505
Seno, producto infinito para, 617
serie de potencias para, 472
Singularidad eliminable, 59, 63
Snell, ley de refracción de, 269
Stirling, fórmula de, 521, 646
serie de, 646
Subsucesión, 119
Substitución, regla de, para integrales, 286, 288
Sucesión, acotada, 94
convergente, 93
de intervalos racionales encajados, 113
divergente, 94
infinita, 79
límite de, 84, 116
monótona, 97, 119
nula, 113
Sucesiva, aproximación, 512
Suma, por parte, 533
símbolo de, 98
Superficie de revolución, 391
Superposición de vibraciones, 592
principio de, 657
Supremo extremo superior, 121

T

Tangente, 178, 362, 477
a una curva, 178, 183
cosenos directores de, 363, 364, 372
dirección de, 412
ecuación de, 362
fórmula, 501, 504
positiva, 364
Taylor, fórmula de, 463, 465, 466
polinomio de, 464, 476
serie de, 468, 561
teorema de, 468
Topología, 360
Trabajo, 435, 437
diagrama de, 436
Tractriz, 454
Transformación, afín, 44
de Lorentz, 381
identidad, 78
Trapezoide, fórmula, 501
regla del, 505
Transcendentes, números, 102
Traslación, 379, 398
Triángulo, desigualdad del, 38
Trigonométrica, función, 73, 237, 318
serie, 587
derivación, de, 227
expresiones exponenciales para, 567
funciones trigonométricas, ecuación diferencial de, 331
inversas, 232
relaciones de ortogonalidad para, 295
Trigonométrico, polinomio, 593
Trocoide, 351, 452, 454
Truncamiento, error por, 504

U

Unidad, raíces de la, 128, 129
Uniforme, continuidad, 65, 123
convergencia, 545, 548

V

Valor medio (intermedio), 163
teorema del, del cálculo diferencial, 194, 212, 243
teorema del, del cálculo integral, 163
teorema del, generalizado del cálculo integral, 164
vecindad (en torno), 36
Vectores, 398
ángulo entre, 404
construcción paralela para suma de, 403
coordenadas de, 408
definición de, 398
de posición, 399
derivada de, 411
integral de, 411
longitud de, 399
opuestos, 399, 401
producto escalar de, 405
producto exterior de, 405
resultante de, 402
Vectores, suma de, 402
unitarios, 408
Velocidad, 412
componentes de, 379
cuerpos en caída libre, 175
promedio, 183
Vibración, 650
amplitud de, 591
armónica, 591
elástica, 422
eléctrica, 651
mecánica, 650
período de, 591
sinusoidal, 591
superposición de, 592
Voltaje, 599

W

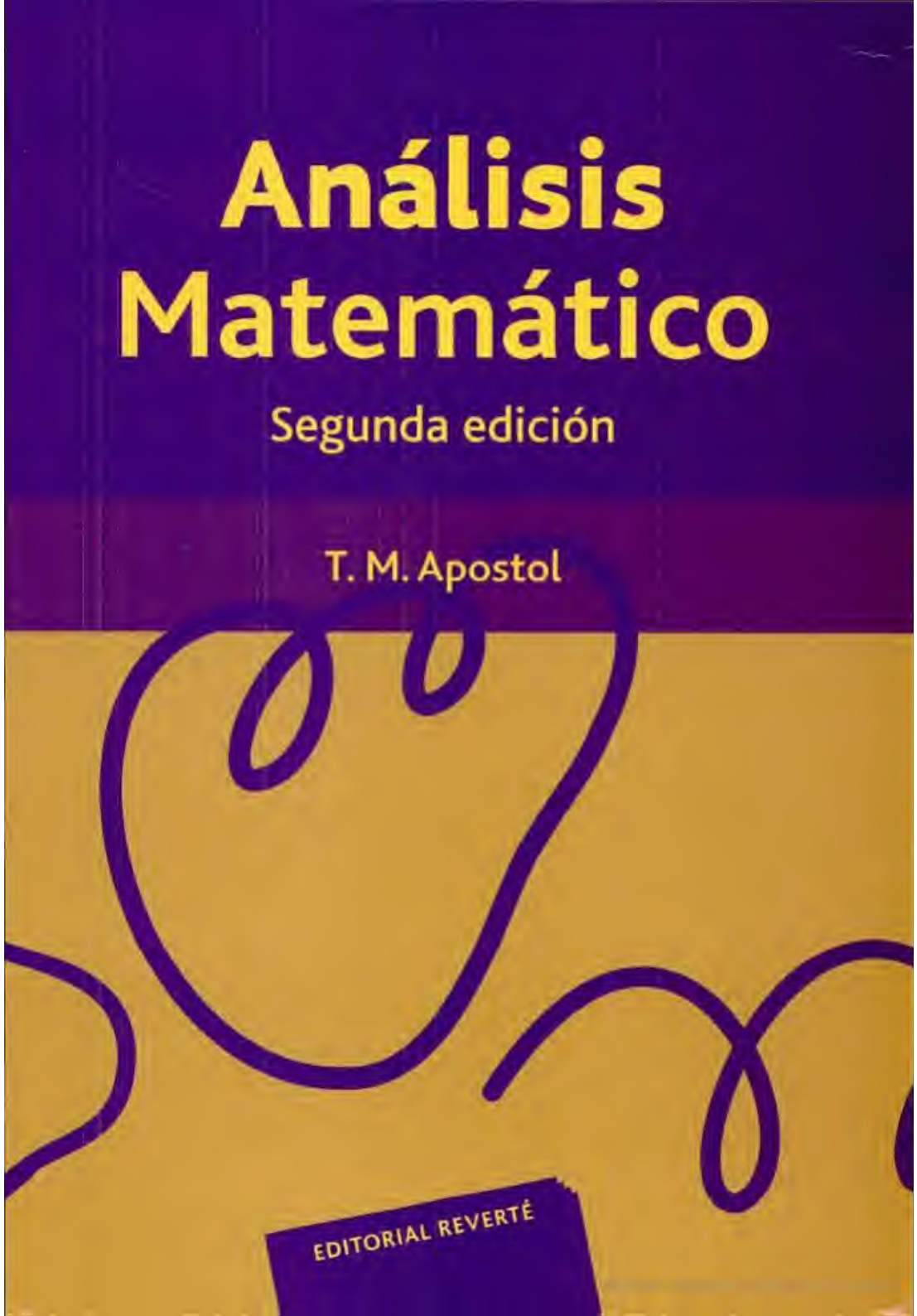
Wallis, fórmula de, 300, 302
Weierstrass, 584, 624
principio de, 118, 119

Z

Zeta, función zeta de Riemann, 575, 585, 637
como producto infinito, 575

—ooo—

La edición, composición, diseño e impresión de esta obra fueron realizados
bajo la supervisión de GRUPO NORIEGA EDITORES
Balderas 95, Col. Centro. México, D.F. C.P. 06040
Tel.: (5)521 48 49 y (5)512 30 09 Fax: (5)512 29 03 y (5)510 94 15
Impreso en Programas Educativos, S.A. de C.V. • 117754 000 06 99 517
Certificado por ISO-9002:1994/NMX-CC-004:1995 con el No. de registro RSC-048
e-mail: limusa@noriega.com.mx • www.noriega.com.mx



Título de la obra original:
Mathematical Analysis

Edición original en lengua inglesa publicada por:
Addison-Wesley Publishing Company, Inc. Reading, Massachusetts, U.S.A.

Copyright © Addison-Wesley Publishing Company, Inc

Versión española por:
Dr. José Pla Carrera
Doctor en Matemáticas
Profesor de la Facultad de Matemáticas en la Universidad de Barcelona

Revisada por:
Dr. Enrique Linés Escardó
Catedrático de la Facultad de Ciencias de la Universidad de Madrid

Propiedad de:
EDITORIAL REVERTÉ, S. A.
Loreto, 13-15, Local B
08029 Barcelona
Tel: (34) 93 419 33 36
Fax: (34) 93 419 51 89
e-mail: reverte@reverte.com
www.reverte.com

REIMPRESIÓN: ABRIL DE 2006

Reservados todos los derechos. La reproducción total o parcial de esta obra, por cualquier medio o procedimiento, comprendidos la reprografía y el tratamiento informático, queda rigurosamente prohibida, salvo excepción prevista en la ley. Asimismo queda prohibida la distribución de ejemplares mediante alquiler o préstamo públicos, la comunicación pública y la transformación de cualquier parte de esta publicación (incluido el diseño de la cubierta) sin la previa autorización de los titulares de la propiedad intelectual y de la Editorial. La infracción de los derechos mencionados puede ser constitutiva de delito contra la propiedad intelectual (Art. 270 y siguientes del Código Penal). El Centro Español de Derechos Reprográficos (CEDRO) vela por el respeto a los citados derechos.

Edición en español

© Editorial Reverté, S. A., 1976

ISBN: 84-291-5004-8
Depósito legal: B-17304-2006

Impreso en España - Printed in Spain

Impresión: Domingraf Impressors

Material protegido por derechos de autor

Indice analítico

Capítulo 1 El sistema de los números reales y el de los complejos

1.1 Introducción 1
1.2 Los axiomas de cuerpo 2
1.3 Los axiomas de orden 2
1.4 Representación geométrica de los números reales 4
1.5 Intervalos 4
1.6 Los enteros 5
1.7 Teorema de descomposición única para enteros 6
1.8 Los números racionales 8
1.9 Los números irracionales 8
1.10 Cotas superiores; elemento máximo, cota superior mínima (supremo) 10
1.11 El axioma de completitud 11
1.12 Algunas propiedades del supremo 12
1.13 Propiedades de los enteros deducidas del axioma de completitud 13
1.14 La propiedad arquimediana del sistema de los números reales 13
1.15 Los números racionales con representación decimal finita 13
1.16 Aproximaciones decimales finitas de los números reales 14
1.17 Representaciones decimales infinitas de los números reales 15
1.18 Valor absoluto y desigualdad triangular 16
1.19 La desigualdad de Cauchy-Schwarz 17
1.20 Más y menos infinito y la extensión $\mathbf{R}^*$ del sistema de los números reales 18
1.21 Los números complejos 19
1.22 Representación geométrica de los números complejos 21
1.23 La unidad imaginaria 22
1.24 Valor absoluto de un número complejo 22
1.25 Imposibilidad de ordenar los números complejos 23
1.26 Exponenciales complejas 24
1.27 Otras propiedades de las exponenciales complejas 25
1.28 El argumento de un número complejo 25
1.29 Potencias enteras y raíces de números complejos 26
1.30 Los logaritmos complejos 28
1.31 Potencias complejas 29
1.32 Senos y cosenos complejos 30
1.33 Infinito y el plano complejo ampliado $\mathbf{C}^*$ 30
Ejercicios 31

Capítulo 2 Algunas nociones básicas de la teoría de conjuntos 39
2.1 Introducción 39
2.2 Notaciones 39
2.3 Pares ordenados 40
2.4 Producto cartesiano de dos conjuntos 41
2.5 Relaciones y funciones 41
2.6 Más terminología referente a funciones 42
2.7 Funciones uno a uno e inversas 43
2.8 Funciones compuestas 45
2.9 Sucesiones 45
2.10 Conjuntos coordinables (equipotentes) 46
2.11 Conjuntos finitos e infinitos 46
2.12 Conjuntos numerables y no numerables 47
2.13 El conjunto de los números reales no es numerable 48
2.14 Algebra de conjuntos 49
2.15 Colecciones numerables de conjuntos numerables 51
Ejercicios 52

Capítulo 3 Elementos de topología en conjuntos de puntos 57
3.1 Introducción 57
3.2 El espacio euclídeo $\mathbf{R}^n$ 57
3.3 Bolas abiertas y conjuntos abiertos de $\mathbf{R}^n$ 60
3.4 La estructura de los conjuntos abiertos de $\mathbf{R}^1$ 61
3.5 Conjuntos cerrados 63
3.6 Puntos adherentes. Puntos de acumulación 63
3.7 Conjuntos cerrados y puntos adherentes 65
3.8 Teorema de Bolzano-Weierstrass 66
3.9 Teorema de encaje de Cantor 68
3.10 Teorema del recubrimiento de Lindelöf 68
3.11 Teorema del recubrimiento de Heine-Borel 70
3.12 Compacidad en $\mathbf{R}^n$ 71
3.13 Espacios métricos 73
3.14 Topología en espacios métricos 74
3.15 Subconjuntos compactos de un espacio métrico 77
3.16 Frontera de un conjunto 78
Ejercicios 78

Capítulo 4 Límites y continuidad 85
4.1 Introducción 85
4.2 Sucesiones convergentes en un espacio métrico 86
4.3 Sucesiones de Cauchy 88
4.4 Espacios métricos completos 90
4.5 Límite de una función 90
4.6 Límites de funciones con valores complejos 92
4.7 Límites de funciones con valores vectoriales 93
4.8 Funciones continuas 95

Material protegido por derechos de autor
Material protegido por derechos de autor

4.9 La continuidad de las funciones compuestas 96
4.10 Funciones complejas y funciones vectoriales continuas 97
4.11 Ejemplos de funciones continuas 97
4.12 Continuidad y antiimágenes de conjuntos abiertos y cerrados 98
4.13 Funciones continuas sobre conjuntos compactos 100
4.14 Aplicaciones topológicas (homeomorfismos) 102
4.15 Teorema de Bolzano 102
4.16 Conexión 104
4.17 Componentes de un espacio métrico 106
4.18 Conexión por arcos 107
4.19 Continuidad uniforme 109
4.20 Continuidad uniforme y conjuntos compactos 110
4.21 Teorema del punto fijo para contracciones 111
4.22 Discontinuidades de las funciones reales 113
4.23 Funciones monótonas 115
Ejercicios 116

Capítulo 5 Derivadas 125
5.1 Introducción 125
5.2 Definición de derivada 125
5.3 Derivadas y continuidad 126
5.4 Álgebra de derivadas 127
5.5 La regla de la cadena 128
5.6 Derivadas laterales y derivadas infinitas 129
5.7 Funciones con derivada no nula 130
5.8 Derivadas cero y extremos locales 131
5.9 Teorema de Rolle 132
5.10 Teorema del valor medio para derivadas 132
5.11 Teorema del valor intermedio para las derivadas 134
5.12 Fórmula de Taylor con resto 136
5.13 Derivadas de funciones vectoriales 137
5.14 Derivadas parciales 138
5.15 Diferenciación de funciones de una variable compleja 140
5.16 Ecuaciones de Cauchy-Riemann 142
Ejercicios 146

Capítulo 6 Funciones de variación acotada y curvas rectificables 153
6.1 Introducción 153
6.2 Propiedades de las funciones monótonas 153
6.3 Funciones de variación acotada 154
6.4 Variación total 156
6.5 Propiedad aditiva de la variación total 157
6.6 La variación total $[a, x]$, como función de x 158
6.7 Funciones de variación acotada expresadas como diferencia de dos funciones crecientes 159
6.8 Funciones continuas de variación acotada 159
6.9 Curvas y caminos 161

6.10 Caminos rectificables y longitud de un arco 161
6.11 Propiedades de aditividad y de continuidad de la longitud de arco 163
6.12 Caminos equivalentes. Cambios de parámetros 164
Ejercicios 165

Capítulo 7 La integral de Riemann-Stieltjes 169
7.1 Introducción 169
7.2 Notación 170
7.3 La definición de la integral de Riemann-Stieltjes 171
7.4 Propiedades lineales 171
7.5 Integración por partes 174
7.6 Cambio de variable en una integral de Riemann-Stieltjes 175
7.7 Reducción de una integral de Riemann 176
7.8 Funciones escalonadas como integradores 177
7.9 Reducción de una integral de Riemann-Stieltjes a una suma finita 179
7.10 Fórmula de sumación de Euler 181
7.11 Integradores monótonos crecientes. Integrales superior e inferior 181
7.12 Propiedades aditiva y lineal de las integrales superior e inferior 185
7.13 Condición de Riemann 186
7.14 Teoremas de comparación 187
7.15 Integradores de variación acotada 189
7.16 Condiciones suficientes para la existencia de las integrales de Riemann-Stieltjes 193
7.17 Condiciones necesarias para la existencia de las integrales de Riemann-Stieltjes 194
7.18 Teoremas del valor medio para las integrales de Riemann-Stieltjes 195
7.19 La integral como función del intervalo 196
7.20 El segundo teorema fundamental del Cálculo integral 197
7.21 Cambio de variable en una integral de Riemann 199
7.22 Segundo teorema del valor medio para integrales de Riemann 200
7.23 Integrales de Riemann-Stieltjes dependientes de un parámetro 201
7.24 Derivación bajo el signo integral 203
7.25 Intercambio en el orden de integración 203
7.26 Criterio de Lebesgue para la existencia de las integrales de Riemann 205
7.27 Integrales complejas de Riemann-Stieltjes 211
Ejercicios 212

Capítulo 8 Series infinitas y productos infinitos 223
8.1 Introducción 223
8.2 Sucesiones convergentes y divergentes de números complejos 223
8.3 Límite superior y límite inferior de una sucesión real 224
8.4 Sucesiones monótonas de números reales 225
8.5 Series infinitas 226
8.6 Introducción y supresión de paréntesis 227
8.7 Series alternadas 229
8.8 Convergencia absoluta y condicional 230

Material protegido por derechos de autor
Material protegido por derechos de autor

8.9 Parte real y parte imaginaria de una serie compleja 231
8.10 Criterios de convergencia para las series de términos positivos 231
8.11 La serie geométrica 232
8.12 El criterio de la integral 232
8.13 Las notaciones O grande y o pequeña 234
8.14 El criterio del cociente y el criterio de la raíz 235
8.15 Criterios de Dirichlet y de Abel 236
8.16 Sumas parciales de la serie geométrica Σz^n sobre el círculo unidad $|z|=1$ 237
8.17 Reordenación de series 238
8.18 Teorema de Riemann para series condicionalmente convergentes 240
8.19 Series parciales 241
8.20 Sucesiones dobles 243
8.21 Series dobles 244
8.22 Teorema de reordenación para series dobles 245
8.23 Una condición suficiente para la igualdad de series reiteradas 247
8.24 Multiplicación de series 248
8.25 Sumabilidad de Césaro 250
8.26 Productos infinitos 252
8.27 Producto de Euler para la función zeta de Riemann 255
Ejercicios 256

Capítulo 9 Sucesiones de funciones 265
9.1 Convergencia puntual de sucesiones de funciones 265
9.2 Ejemplos de sucesiones de funciones reales 266
9.3 Definición de convergencia uniforme 268
9.4 Convergencia uniforme y continuidad 269
9.5 La condición de Cauchy para la convergencia uniforme 270
9.6 Convergencia uniforme de series infinitas de funciones 271
9.7 Una curva que llena todo el espacio 272
9.8 Convergencia uniforme e integración de Riemann-Stieltjes 274
9.9 Sucesiones convergentes con convergencia no uniforme que pueden ser integradas término a término 275
9.10 Convergencia uniforme y diferenciación 278
9.11 Condiciones suficientes para la convergencia uniforme de series 280
9.12 Convergencia uniforme y sucesiones dobles 281
9.13 Convergencia en media 282
9.14 Serie de potencias 284
9.15 Multiplicación de series de potencias 289
9.16 El teorema de sustitución 290
9.17 Recíproca de una serie de potencias 291
9.18 Series reales de potencias 292
9.19 Serie de Taylor generada por una función 293
9.20 Teorema de Bernstein 294
9.21 La serie binómica 297
9.22 Teorema del límite de Abel 298
9.23 Teorema de Tauber 300
Ejercicios 301

Capítulo 10 La integral de Lebesgue 307
10.1 Introducción 307
10.2 Integral de una función escalonada 308
10.3 Sucesiones monótonas de funciones escalonadas 309
10.4 Funciones superiores y sus integrales 312
10.5 Las funciones integrales de Riemann como ejemplo de las funciones superiores 316
10.6 La clase de las funciones integrables de Lebesgue en un intervalo general 318
10.7 Propiedades básicas de la integral de Lebesgue 319
10.8 Integración de Lebesgue y conjuntos de medida cero 323
10.9 Teoremas de convergencia monótona de Levi 323
10.10 Teorema de convergencia dominada de Lebesgue 330
10.11 Aplicaciones del teorema de convergencia dominada de Lebesgue 333
10.12 Integrales de Lebesgue sobre intervalos no acotados como límite de integrales sobre intervalos acotados 335
10.13 Integrales de Riemann impropias 337
10.14 Funciones medibles 340
10.15 Continuidad de funciones definidas por medio de integrales de Lebesgue 342
10.16 Diferenciación bajo signo integral 345
10.17 Intercambio en el orden de integración 349
10.18 Conjuntos medibles de la recta real 352
10.19 La integral de Lebesgue en subconjuntos arbitrarios de **R** 355
10.20 Integrales de Lebesgue de funciones complejas 356
10.21 Productos interiores y normas 357
10.22 El conjunto $L^2(I)$ de las funciones de cuadrado integrable 358
10.23 El conjunto $L^2(I)$ como espacio semimétrico 360
10.24 Un teorema de convergencia para series de funciones de $L^2(I)$ 360
10.25 Teorema de Riesz-Fischer 362
Ejercicios 363

Capítulo 11 Series de Fourier e integrales de Fourier 373
11.1 Introducción 373
11.2 Sistemas ortogonales de funciones 373
11.3 El teorema de óptima aproximación 374
11.4 Serie de Fourier de una función relativa a un sistema ortonormal 376
11.5 Propiedades de los coeficientes de Fourier 377
11.6 Teorema de Riesz-Fischer 378
11.7 Los problemas de convergencia y representación para series trigonométricas 380
11.8 Lema de Riemann-Lebesgue 381
11.9 Integrales de Dirichlet 383
11.10 Una representación integral para las sumas parciales de una serie de Fourier 386
11.11 Teorema de localización de Riemann 387

Material protegido por derechos de autor
Material protegido por derechos de autor

16.13 Desarrollo en serie de potencias de las funciones analíticas 546
16.14 Desigualdades de Cauchy. Teorema de Liouville 548
16.15 Separación de los ceros de una función analítica 549
16.16 El teorema de identidad para funciones analíticas 551
16.17 Módulos máximo y mínimo de una función analítica 551
16.18 El teorema de la aplicación abierta 553
16.19 Desarrollos de Laurent para funciones analíticas en un anillo 554
16.20 Singularidades aisladas 557
16.21 Residuo de una función en un punto singular aislado 559
16.22 Teorema de Cauchy del residuo 560
16.23 Números de ceros y de polos en una región 561
16.24 Cálculo de integrales reales por medio de residuos 562
16.25 Cálculo de la suma de Gauss por el método de los residuos 565
16.26 Aplicación del teorema del residuo a la fórmula de inversión para transformadas de Laplace 570
16.27 Aplicaciones conformes 572
Ejercicios 575
Índice de símbolos especiales 585
Índice alfabético 589

Análisis matemático

Material protegido por derechos de autor

Material protegido por derechos de autor

CAPÍTULO 1

El sistema de los números reales y el de los complejos

1.1 INTRODUCCIÓN

El Análisis matemático estudia conceptos relacionados de alguna manera con los números reales; por ello empezaremos nuestro estudio del Análisis con una discusión del sistema de los números reales.

Existen diversos métodos para introducir los números reales. Uno de ellos parte de los enteros positivos 1, 2, 3, ..., que considera conceptos no definidos, utilizándolos para construir un sistema más amplio, los *números racionales* positivos (cocientes de enteros positivos), los negativos y el cero. Los números racionales son utilizados, a su vez, para construir los *números irracionales*, números reales como $\sqrt{2}$ y π, que no son racionales. El sistema de los números reales lo constituye la reunión de los números racionales e irracionales.

A pesar de que estas cuestiones constituyen una parte importante de los fundamentos de la Matemática, no las describiremos aquí con detalle. Es un hecho que, en la mayor parte del Análisis, nos interesarán solamente las *propiedades* de los números reales antes que los métodos utilizados para construirlos. Por lo tanto, consideraremos los números reales mismos como objetos no definidos, sometidos a ciertos axiomas de los que extraeremos ulteriores propiedades. Dado que el lector está, probablemente, familiarizado con la mayoría de las propiedades de los números reales que consideraremos en las páginas que siguen, la exposición será más bien breve. Su propósito es examinar las características más importantes y persuadir al lector de que, de ser necesario, todas las propiedades se podrían deducir a partir de los axiomas. Tratamientos más detallados podrán hallarse en las referencias del final de este capítulo.

Por conveniencia usaremos la notación y la terminología de la teoría de conjuntos elemental. Supongamos que S designa un conjunto (una colección de objetos). La notación $x \in S$ significa que x está en el conjunto S, escribiendo $x \notin S$ para indicar que x no está en S.

Un conjunto S es un *subconjunto* de T si cada elemento de S está también en T. Lo indicaremos escribiendo $S \subseteq T$. Un conjunto es *no vacío* si contiene, por lo menos, un elemento.

Material protegido por derechos de autor

Suponemos que existe un conjunto no vacío **R** de elementos, llamados números reales, que satisfacen los diez axiomas enumerados a continuación. Los axiomas se clasifican de manera natural en tres grupos a los que nos referiremos como *axiomas de cuerpo, axiomas de orden* y *axioma de completitud* (llamado también *axioma del supremo* o *axioma de continuidad*).

1.2 LOS AXIOMAS DE CUERPO

Junto con el conjunto **R** de los números reales admitimos la existencia de dos operaciones, llamadas *suma* y *multiplicación*, tales que, para cada par de números reales x e y, la *suma* $x + y$ y el *producto* xy son números reales determinados unívocamente por x e y, satisfaciendo los siguientes axiomas. (En los axiomas que a continuación se exponen, x, y, z representan números reales arbitrarios en tanto no se precise lo contrario.)

Axioma 1. $x + y = y + x,\ xy = yx$ *(leyes conmutativas).*

Axioma 2. $x + (y + z) = (x + y) + z,\ x(yz) = (xy)z$ *(leyes asociativas).*

Axioma 3. $x(y + z) = xy + xz$ *(ley distributiva).*

Axioma 4. *Dados dos números reales cualesquiera x e y, existe un número real z tal que $x + z = y$. Dicho número z se designará por $y - x$; el número $x - x$ se designará por 0. (Se puede demostrar que 0 es independiente de x.) Escribiremos $-x$ en vez de $0 - x$ y al número $-x$ lo llamaremos opuesto de x.*

Axioma 5. *Existe, por lo menos, un número real $x \neq 0$. Si x e y son dos números reales con $x \neq 0$, entonces existe un número z tal que $xz = y$. Dicho número z se desginará por y/x; el número x/x se designará por 1 y puede demostrarse que es independiente de x. Escribiremos x^{-1} en vez de $1/x$ si $x \neq 0$ y a x^{-1} lo llamaremos recíproco o inverso de x.*

De estos axiomas pueden deducirse todas las leyes usuales de la Aritmética; por ejemplo, $-(-x) = x,\ (x^{-1})^{-1} = x,\ -(x - y) = y - x,\ x - y = x + (-y)$, etc. (Para un desarrollo más detallado, ver Referencia 1.1.)

1.3 LOS AXIOMAS DE ORDEN

Suponemos también la existencia de una relación $<$ que establece una ordenación entre los números reales y que satisface los axiomas siguientes:

Material protegido por derechos de autor

Axioma 6. *Se verifica una y sólo una de las relaciones* $x = y$, $x < y$, $x > y$.

NOTA. $x > y$ significa lo mismo que $y < x$.

Axioma 7. *Si* $x < y$, *entonces, para cada* z, *es* $x + z < y + z$.

Axioma 8. *Si* $x > 0$ *e* $y > 0$, *entonces* $xy > 0$.

Axioma 9. *Si* $x > y$ e $y > z$, *entonces* $x > z$.

NOTA. Un número real x se llama *positivo* si $x > 0$ y *negativo* si $x < 0$. Designaremos por $\mathbf{R}^+$ el conjunto de todos los números reales positivos y por $\mathbf{R}^-$ el conjunto de todos los números reales negativos.

De estos axiomas pueden deducirse las reglas usuales que rigen las operaciones con desigualdades. Por ejemplo, si tenemos que $x < y$, entonces $xz < yz$ si z es positivo, mientras que $xz > yz$ si z es negativo. Además, si $x > y$ y $z > w$ con y y w positivos, entonces $xz > yw$. (Para una discusión más detallada de estas reglas ver Referencia 1.1.)

NOTA. El simbolismo $x \leq y$ se utiliza para abreviar la afirmación:

$$\text{"}x < y \quad \text{o} \quad x = y\text{."}$$

Resulta, pues, que $2 \leq 3$ ya que $2 < 3$; y $2 \leq 2$ ya que $2 = 2$. El símbolo $\geq$ se utiliza de forma análoga. Un número real x se llama *no negativo* si $x \geq 0$. Un par simultáneo de desigualdades tales como $x < y$, $y < z$ se abrevia por medio de la expresión $x < y < z$.

El teorema que sigue, que no es más que una consecuencia inmediata de los axiomas precedentes, se utiliza a menudo en las demostraciones del Análisis.

Teorema 1.1. *Sean* a *y* b *números reales tales que*

$$a \leq b + \varepsilon \text{ para cada } \varepsilon > 0. \tag{1}$$

Entonces $a \leq b$.

Demostración. Si $b < a$, entonces la desigualdad (1) no se satisface para $\varepsilon = (a - b)/2$ puesto que

$$b + \varepsilon = b + \frac{a - b}{2} = \frac{a + b}{2} < \frac{a + a}{2} = a.$$

Material protegido por derechos de autor

Por lo tanto, por el axioma 6, resulta que $a \leq b$.

El axioma 10, axioma de completitud, será enunciado en la sección 1.11.

1.4. REPRESENTACIÓN GEOMÉTRICA DE LOS NÚMEROS REALES

Los números reales son, a menudo, representados geométricamente como puntos de una recta (denominada *recta real* o *eje real*). Se elige un punto para que represente el 0 y otro a la derecha del 0 para que represente el 1, como muestra la Fig. 1.1. Esta elección determina la escala. Con un conjunto apropiado de axiomas para la Geometría euclídea a cada punto de la recta real corresponde un número real y uno sólo y, recíprocamente, cada número real está representado por un punto de la recta real y uno solo. Es usual referirse al *punto* x en vez de referirse al punto correspondiente al número real x.

Figura 1.1

La relación de orden admite una interpretación geométrica simple. Si $x < y$, el punto x está a la izquierda del punto y, como muestra la figura 1.1. Los números positivos están a la derecha del 0 y los números negativos están a la izquierda del 0. Si $a < b$, un punto x satisface las desigualdades $a < x < b$ si, y sólo si, x está *entre* a y b.

1.5 INTERVALOS

El conjunto de todos los puntos comprendidos entre a y b se denomina *intervalo*. A menudo es importante distinguir entre los intervalos que incluyen sus extremos y los intervalos que no los incluyen.

NOTACIÓN. La notación $\{x : x \text{ verifica } P\}$ designa el conjunto de todos los números reales x tales que satisfacen la propiedad P.

Definición 1.2. *Supongamos* $a < b$. *El intervalo abierto* (a, b) *se define por*

$$(a, b) = \{x : a < x < b\}.$$

El intervalo cerrado $[a, b]$ *es el conjunto* $\{x : a \leq x \leq b\}$. *Los intervalos semiabiertos* $(a, b]$ *y* $[a, b)$ *se definen análogamente utilizando, respectivamente, las desigualdades* $a < x \leq b$ *y* $a \leq x < b$. *Los intervalos infinitos se definen como sigue:*

$$(a, +\infty) = \{x : x > a\}, \qquad [a, +\infty) = \{x : x \geq a\},$$

Material protegido por derechos de autor

1.7 TEOREMA DE DESCOMPOSICIÓN ÚNICA PARA ENTEROS

Si n y d son enteros y si $n = cd$ para algún entero c, diremos que d es un *divisor* de n, o que n es un *múltiplo* de d, y escribiremos $d|n$ (se lee: d divide a n). Un entero n es *primo* si $n > 1$ y si los únicos divisores positivos de n son 1 y n. Si $n > 1$ y n no es primo, entonces n es *compuesto*. El entero 1 no es ni primo ni compuesto.

Esta sección expone algunos resultados elementales acerca de la descomposición de enteros, culminando con el *teorema de descomposición única*, llamado también *el teorema fundamental de la Aritmética*.

El teorema fundamental establece que (1) cada entero $n > 1$ puede ser representado como producto de factores primos y que (2) esta descomposición es única, salvo en el orden de los factores. Es fácil probar la parte (1).

Teorema 1.5. *Cada entero $n > 1$ es primo o producto de primos.*

Demostración. Utilizaremos la inducción sobre n. El teorema se verifica trivialmente para $n = 2$. Supongamos que es cierto para cada entero k con $1 < k < n$. Si n no es primo, admite un divisor d con $1 < d < n$. Por lo tanto, $n = cd$, con $1 < c < n$. Puesto que tanto c como d son $< n$, cada uno es primo o es producto de primos; luego n es un producto de primos.

Antes de probar la parte (2), la unicidad de la descomposición, introduciremos otros conceptos.

Si $d|a$ y $d|b$, diremos que d es un *divisor común* de a y b. El teorema que sigue demuestra que cada par de enteros a y b posee un divisor común que es combinación lineal de a y de b.

Teorema 1.6. *Cada par de enteros a y b admite un divisor común d de la forma*

$$d = ax + by$$

donde x e y son enteros. Además, cada divisor común de a y b divide a d.

Demostración. Supongamos primeramente que $a \geq 0$ y $b \geq 0$ y procedamos por inducción sobre $n = a + b$. Si $n = 0$, entonces $a = b = 0$ y podemos tomar $d = 0$ con $x = y = 0$. Supongamos entonces que el teorema ha sido probado para 0, 1, 2, ..., $n - 1$. Por simetría podemos suponer $a \geq b$. Si $b = 0$, entonces $d = a$, $x = 1$, $y = 0$. Si $b \geq 1$ podemos aplicar la hipótesis de inducción a $a - b$ y a b, ya que su suma es $a = n - b \leq n - 1$. Por lo tanto existe un divisor común d de $a - b$ y b de la forma $d = (a - b)x + by$. Este entero d

Material protegido por derechos de autor

índices de las q_s si es necesario, se puede suponer $p_1 \geq q_1$. Por lo tanto, $p_1 = q_1$ ya que tanto p_1 como q_1 son primos. En (2) simplificamos p_1 en ambos miembros y obtenemos

$$\frac{n}{p_1} = p_2 \cdots p_s = q_2 \cdots q_t.$$

Como n es compuesto, $1 < n/p_1 < n$; luego por la hipótesis de inducción las dos descomposiciones de n/p_1 son idénticas, si se prescinde del orden de los factores. Por lo tanto, lo mismo es cierto para (2) y la demostración está terminada.

1.8 LOS NÚMEROS RACIONALES

Los cocientes de enteros a/b (donde $b \neq 0$) se llamarán *números racionales*. Por ejemplo, 1/2, $-7/5$, y 6 son números racionales. El conjunto de los números racionales, que designaremos por **Q**, contiene a **Z** como subconjunto. Observe el lector que todos los axiomas de cuerpo y todos los axiomas de orden se verifican en **Q**.

Suponemos que el lector está familiarizado con ciertas propiedades elementales de los números racionales. Por ejemplo, si a y b son racionales, su media $(a + b)/2$ también lo es y está comprendida entre a y b. Así pues, entre dos números racionales hay una infinidad de números racionales, lo cual implica que, dado un número racional cualquiera, no sea posible hablar del número racional «inmediato superior».

1.9 LOS NÚMEROS IRRACIONALES

Los números reales que no son racionales se denominan *irracionales*. Por ejemplo, los números $\sqrt{2}$, e, π y e^r son irracionales.

En general no es fácil probar que un cierto número particular es irracional. No existe ninguna demostración simple de la irracionalidad de e^r, por ejemplo. Sin embargo, la irracionalidad de números tales como $\sqrt{2}$, $\sqrt{3}$ no es excesivamente difícil de establecer y, de hecho, probaremos fácilmente el siguiente:

Teorema 1.10. *Si n es un entero positivo que no sea un cuadrado perfecto, entonces $\sqrt{n}$ es irracional.*

Demostración. Suponemos en primer lugar que n no admite ningún divisor > 1 que sea cuadrado perfecto. Si admitimos que $\sqrt{n}$ es racional, llegamos a contradicción. Supongamos que $\sqrt{n} = a/b$, donde a y b son enteros sin divisores comunes. Entonces $nb^2 = a^2$ y, dado que el primer miembro de esta

Material protegido por derechos de autor

igualdad es un múltiplo de n, también lo será a^2. Sin embargo, si a^2 es múltiplo de n, a deberá serlo ya que n no admite divisores > 1 que sean cuadrados perfectos. (Esto se ve fácilmente examinando la descomposición de a en factores primos.) Todo ello significa que $a = cn$, donde c es un entero. Entonces la ecuación $nb^2 = a^2$ se transforma en $nb^2 = c^2n^2$, o $b^2 = nc^2$. El mismo argumento prueba que b debe ser asimismo múltiplo de n. Entonces a y b serían ambos múltiplos de n, lo cual contradice el hecho de que a y b carecen de divisores comunes. Esto finaliza la demostración en el caso de que n no admita un divisor > 1 que sea cuadrado perfecto.

Si n admite un factor que sea cuadrado perfecto, podremos escribir $n = m^2k$, donde $k > 1$ y k no admite divisores > 1 que sean cuadrados perfectos. Por lo tanto $\sqrt{n} = m\sqrt{k}$; y si $\sqrt{n}$ fuese racional, el número $\sqrt{k}$ sería también racional, contradiciendo lo que acabamos de demostrar.

Un tipo distinto de argumentación es preciso para probar que el número e es irracional. (Suponemos cierta familiaridad con la exponencial e^x del Cálculo elemental y su representación como serie infinita.)

Teorema 1.11. *Si* $e^x = 1 + x + x^2/2! + x^3/3! + \ldots + x^n/n! + \ldots$, *entonces el número* e *es irracional.*

Demostración. Probaremos que e^{-1} es irracional. La serie e^{-1} es una serie alternada con términos que decrecen constantemente en valor absoluto. En tales series el error cometido al cortar la serie por el n-ésimo término tiene el signo algebraico del primer término que se desprecia y, en valor absoluto, es menor que el del primer término que se desprecia. Por lo tanto, si $s_n = \sum_{k=0}^{n} (-1)^k/k!$, tenemos la desigualdad

$$0 < e^{-1} - s_{2k-1} < \frac{1}{(2k)!},$$

de la que se obtiene

$$0 < (2k-1)!\,(e^{-1} - s_{2k-1}) < \frac{1}{2k} \leq \frac{1}{2}, \tag{3}$$

para todo entero $k \geq 1$. Ahora bien $(2k-1)!s_{2k-1}$ es siempre un entero. Si e^{-1} fuese racional, entonces podríamos elegir k suficientemente grande para que $(2k-1)!e^{-1}$ fuese también un entero. A causa de (3) la diferencia entre ambos enteros debería ser un número comprendido entre 0 y $\frac{1}{2}$, lo cual es imposible. Luego e^{-1} no es racional y, por tanto, e tampoco lo es.

NOTA. Para una demostración de la irracionalidad de π, ver Ejercicio 7.33.

Los antiguos griegos sabían de la existencia de los números irracionales allá por el año 500 a.C. Sin embargo, una teoría satisfactoria de tales números

Material protegido por derechos de autor

no sería desarrollada hasta finales del siglo diecinueve en que tres teorías distintas son introducidas al mismo tiempo por Cantor, Dedekind y Weierstrass. En la Referencia 1.6 puede hallarse información acerca de las teorías de Dedekind y Cantor y sus equivalencias.

1.10 COTAS SUPERIORES; ELEMENTO MÁXIMO, COTA SUPERIOR MÍNIMA (SUPREMO)

Los números irracionales aparecen en Álgebra cuando se pretenden resolver ciertas ecuaciones cuadráticas. Por ejemplo, se desea un número real x tal que $x^2 = 2$. De los nueve axiomas enumerados anteriormente no puede deducirse si en $\mathbf{R}$ existe o no un número x, puesto que $\mathbf{Q}$ satisface también estos nueve axiomas y hemos probado que no existe ningún número racional cuyo cuadrado sea 2. El axioma de completitud nos permitirá introducir los números irracionales en el sistema de los números reales y proporcionar al sistema de los números reales una propiedad de continuidad que es fundamental en muchos de los teoremas de Análisis.

Antes de describir el axioma de completitud, es conveniente introducir una terminología y una notación adicionales.

Definición 1.12. *Sea* S *un conjunto de números reales. Si existe un número real* b *tal que* $x \leq b$ *para todo* x *de* S, *diremos que* b *es una cota superior de* S *y que* S *está acotado superiormente por* b.

Decimos *una* cota superior ya que cada número mayor que b también es una cota superior. Si una cota superior b es, además, un elemento de S, b se denomina *último elemento* o *elemento máximo* de S. A lo sumo habrá uno de tales b. Si existe tal número b, escribiremos

$$b = \text{máx } S.$$

Un conjunto carente de cotas superiores se denomina *no acotado superiormente.*

Las definiciones de los términos *cota inferior, acotado inferiormente, primer elemento* (o *elemento mínimo*) pueden formularse análogamente. Si S tiene un elemento mínimo, designaremos a dicho mínimo por mín S.

Ejemplos.

1. El conjunto $\mathbf{R}^+ = (0, +\infty)$ es un conjunto no acotado superiormente. No posee ni cotas superiores ni elemento máximo. Está acotado inferiormente por 0, pero no posee elemento mínimo.
2. El intervalo cerrado $S = [0, 1]$ está acotado superiormente por 1 e inferiormente por 0. De hecho, máx $S = 1$ y mín $S = 0$.

Material protegido por derechos de autor

3. El intervalo semiabierto $S = [0, 1)$ está acotado superiormente por 1, pero carece de elemento máximo. Su elemento mínimo es 0.

Para conjuntos como los del ejemplo 3 que están acotados superiormente pero que carecen de elemento máximo, existe un concepto que sustituye al de elemento máximo. Se denomina *extremo superior* o *supremo* del conjunto y se define como sigue:

Definición 1.13. *Sea S un conjunto de números reales acotado superiormente. Un número real b se denomina extremo superior de S si verifica las dos propiedades siguientes:*

a) *b es una cota superior de S.*
b) *Ningún número menor que b es cota superior de S.*

Ejemplos. Si $S = [0, 1]$ el elemento máximo 1 es asimismo extremo superior de S. Si $S = [0, 1)$, el número 1 es extremo superior de S, aun cuando S carece de elemento máximo.

Es fácil probar que un conjunto no puede tener dos extremos superiores distintos. Por lo tanto, si existe extremo superior de S, existe sólo uno y puede hablarse *del* extremo superior.

Es corriente, en la práctica, referirse al extremo superior de un conjunto por medio del término más breve de *supremo*, abreviado *sup*. Adoptamos esta convención y escribimos

$$b = \sup S,$$

para indicar que b es el supremo de S. Si S tiene un elemento máximo, entonces máx $S = \sup S$.

El *extremo inferior* o *ínfimo* de S, designado por inf S, se define de forma análoga.

1.11 EL AXIOMA DE COMPLETITUD

Nuestro último axioma del sistema de los números reales involucra la noción de supremo.

Axioma 10. *Todo conjunto no vacío S de números reales que esté acotado superiormente admite un supremo; es decir, existe un número real b tal que* $b = \sup S$.

Como consecuencia de este axioma se obtiene que todo conjunto no vacío de números reales acotado inferiormente admite un ínfimo.

Material protegido por derechos de autor

Teorema 1.16 (Propiedad de la comparación). *Dados dos subconjuntos no vacíos S y T de* $\mathbf{R}$ *tales que* $s \leq t$ *para todo s de S y todo t de T, si T tiene supremo, entonces S tiene supremo, y*

$$\sup S \leq \sup T.$$

1.13 PROPIEDADES DE LOS ENTEROS DEDUCIDAS DEL AXIOMA DE COMPLETITUD

Teorema 1.17. *El conjunto* $\mathbf{Z}^+$ *de los enteros positivos* 1, 2, 3, ..., *no está acotado superiormente.*

Demostración. Si $\mathbf{Z}^+$ estuviese acotado superiormente, entonces $\mathbf{Z}^+$ admitiría un supremo, tal como $a = \sup \mathbf{Z}^+$. Por el teorema 1.14 tendríamos que $a - 1 < n$ para algún n de $\mathbf{Z}^+$. Por lo tanto $n + 1 > a$ para esta n. Esto contradice el hecho de ser $a = \sup \mathbf{Z}^+$ ya que $n + 1 \in \mathbf{Z}^+$.

Teorema 1.18. *Para cada número real x existe un entero positivo n tal que* $n > x$.

Demostración. Si no fuese así, existiría un x que sería una cota superior para $\mathbf{Z}^+$, en contradicción con el teorema 1.17.

1.14 LA PROPIEDAD ARQUIMEDIANA DEL SISTEMA DE LOS NÚMEROS REALES

El teorema que sigue enuncia la propiedad arquimediana del sistema de los números reales. Geométricamente dice que todo segmento lineal, por largo que sea, puede recubrirse por medio de un número finito de segmentos lineales de longitud positiva dada, por pequeña que sea.

Teorema 1.19. *Si* $x > 0$ *y si y es un número real arbitrario, existe un entero positivo n tal que* $nx > y$.

Demostración. Aplicar el teorema 1.18 sustituyendo x por y/x.

1.15 LOS NÚMEROS RACIONALES CON REPRESENTACIÓN DECIMAL FINITA

Un número real de la forma

$$r = a_0 + \frac{a_1}{10} + \frac{a_2}{10^2} + \cdots + \frac{a_n}{10^n},$$

Material protegido por derechos de autor

En otras palabras, a_1 es el mayor entero que satisface las desigualdades

$$a_0 + \frac{a_1}{10} \leq x < a_0 + \frac{a_1 + 1}{10}.$$

En general, habiendo elegido $a_1, \ldots, a_{n-1}$ con $0 \leq a_i \leq 9$, sea a_n el mayor entero que satisfaga las desigualdades

$$a_0 + \frac{a_1}{10} + \cdots + \frac{a_n}{10^n} \leq x < a_0 + \frac{a_1}{10} + \cdots + \frac{a_n + 1}{10^n}. \tag{4}$$

Entonces $0 \leq a_n \leq 9$ y tendremos

$$r_n \leq x < r_n + \frac{1}{10^n},$$

donde $r_n = a_0 . a_1 a_2 \ldots a_n$. Esto completa la demostración. Es fácil verificar que x es, de hecho, el supremo del conjunto de los números racionales $r_1, r_2, \ldots$

1.17 REPRESENTACIONES DECIMALES INFINITAS DE LOS NÚMEROS REALES

Los enteros $a_0, a_1, a_2, \ldots$, obtenidos en la demostración del teorema 1.20 pueden utilizarse para definir una representación decimal infinita de x. Escribiremos

$$x = a_0 . a_1 a_2 \cdots$$

para indicar que a_n es el mayor entero que satisface (4). Por ejemplo, si $x = \frac{1}{8}$, obtendremos $a_0 = 0$, $a_1 = 1$, $a_2 = 2$, $a_3 = 5$ y $a_n = 0$ para todo $n \geq 4$. Por lo tanto, podemos escribir

$$\tfrac{1}{8} = 0{,}125000 \cdots$$

Si intercambiamos los signos de desigualdad $\leq$ y $<$ en (4), obtenemos una definición ligeramente diferente de representación decimal. Los decimales finitos r_n satisfacen $r_n < x \leq r_n + 10^{-n}$, sin embargo los dígitos $a_0, a_1, a_2, \ldots$, necesarios no son los mismos que en (4). Por ejemplo, si aplicamos esta segunda definición a $x = \frac{1}{8}$, obtenemos la representación decimal infinita

$$\tfrac{1}{8} = 0{,}124999 \cdots$$

El que un número real admita dos representaciones decimales distintas es un simple ejemplo del hecho de que dos conjuntos diferentes de números reales pueden tener el mismo supremo.

Material protegido por derechos de autor

1.18 VALOR ABSOLUTO Y DESIGUALDAD TRIANGULAR

En Análisis son bastante frecuentes los cálculos con desigualdades. Son de particular importancia las que se relacionan con la noción de *valor absoluto*. Si x es un número real, el valor absoluto de x, desginado por $|x|$, se define como sigue:

$$|x| = \begin{cases} x, & \text{si } x \geq 0, \\ -x, & \text{si } x \leq 0. \end{cases}$$

Una desigualdad importante concerniente a los valores absolutos viene dada por el siguiente:

Teorema 1.21. *Si $a \geq 0$, entonces tenemos la desigualdad $|x| \leq a$ si, y sólo si, $-a \leq x \leq a$.*

Demostración. De la definición de $|x|$ se obtiene la desigualdad $-|x| \leq x \leq |x|$, ya que $x = |x|$ o $x = -|x|$. Si suponemos que $|x| \leq a$, podemos escribir $-a \leq -|x| \leq x \leq |x| \leq a$ y la mitad del teorema queda demostrada. Recíprocamente, si suponemos $-a \leq x \leq a$, entonces, si $x \geq 0$, tenemos que $|x| = x \leq a$, mientras que si $x < 0$, tenemos que $|x| = -x \leq a$. En ambos casos obtenemos que $|x| \leq a$ y el teorema queda demostrado.

Podemos utilizar este teorema para demostrar la *desigualdad triangular*.

Teorema 1.22. *Para números reales arbitrarios x e y se verifica*

$$|x + y| \leq |x| + |y| \quad \textit{(desigualdad triangular)}$$

Demostración. Tenemos que $-|x| \leq x \leq |x|$ y que $-|y| \leq y \leq |y|$. Sumando obtenemos $-(|x| + |y|) \leq x + y \leq |x| + |y|$ y, en virtud del teorema 1.21, concluimos que $|x + y| \leq |x| + |y|$. Esto demuestra el teorema.

A menudo se utilizan otras formas de la desgiualdad triangular. Por ejemplo, si en el teorema 1.22 hacemos $x = a - c$ e $y = c - b$, resulta

$$|a - b| \leq |a - c| + |c - b|.$$

Asimismo, del teorema 1.22, obtenemos $|x| \geq |x + y| - |y|$. Haciendo $x = a + b$, $y = -b$, resulta

$$|a + b| \geq |a| - |b|.$$

Material protegido por derechos de autor

Intercambiando a y b obtendremos, además, $|a+b| \geq |b| - |a| = -(|a| - |b|)$, y por lo tanto

$$|a+b| \geq ||a| - |b||.$$

Por inducción podemos probar asimismo las generalizaciones

$$|x_1 + x_2 + \cdots + x_n| \leq |x_1| + |x_2| + \cdots + |x_n|$$

y

$$|x_1 + x_2 + \cdots + x_n| \geq |x_1| - |x_2| - \cdots - |x_n|.$$

1.19 LA DESIGUALDAD DE CAUCHY-SCHWARZ

Vamos a deducir ahora otra desigualdad usada a menudo en Análisis.

Teorema 1.23 (Desigualdad de Cauchy-Schwarz). *Si $a_1, \ldots, a_n$ y $b_1, \ldots, b_n$ son números reales cualesquiera, se tiene*

$$\left(\sum_{k=1}^{n} a_k b_k\right)^2 \leq \left(\sum_{k=1}^{n} a_k^2\right)\left(\sum_{k=1}^{n} b_k^2\right).$$

Además, la igualdad se verifica si, y sólo si, existe un número real x tal que $a_k x + b_k = 0$ para cada $k = 1, 2, \ldots, n$.

Demostración. Una suma de cuadrados no puede ser nunca negativa. Por lo tanto tenemos

$$\sum_{k=1}^{n} (a_k x + b_k)^2 \geq 0$$

para todo número real x, y es igualdad si, y sólo si, cada término es cero. Esta desigualdad puede escribirse en la forma

$$Ax^2 + 2Bx + C \geq 0,$$

donde

$$A = \sum_{k=1}^{n} a_k^2, \qquad B = \sum_{k=1}^{n} a_k b_k, \qquad C = \sum_{k=1}^{n} b_k^2.$$

Si $A > 0$, hacemos $x = -B/A$ a fin de obtener $B^2 - AC \leq 0$ que es la desigualdad deseada. Si $A = 0$, la demostración es trivial.

Material protegido por derechos de autor

NOTA. Utilizando notación vectorial, la desigualdad de Cauchy-Schwarz toma la forma

$$(\mathbf{a}\cdot\mathbf{b})^2 \leq \|\mathbf{a}\|^2 \|\mathbf{b}\|^2,$$

donde $\mathbf{a} = (a_1, \ldots, a_n)$, $\mathbf{b} = (b_1, \ldots, b_n)$ son dos vectores n-dimensionales,

$$\mathbf{a}\cdot\mathbf{b} = \sum_{k=1}^{n} a_k b_k,$$

es su producto escalar, y $\|\mathbf{a}\| = (\mathbf{a}\cdot\mathbf{a})^{1/2}$ es la longitud de $\mathbf{a}$.

1.20 MÁS Y MENOS INFINITO Y LA EXTENSIÓN $\mathbf{R}^*$ DEL SISTEMA DE LOS NÚMEROS REALES

En esta sección extenderemos el sistema de los números reales adjuntando dos «puntos ideales» designados por los símbolos $+\infty$ y $-\infty$ («más infinito» y «menos infinito»).

Definición 1.24. *Por sistema ampliado de los números reales, $\mathbf{R}^*$, entenderemos el conjunto de los números reales $\mathbf{R}$ junto con dos símbolos $+\infty$ y $-\infty$ que satisfagan las siguientes propiedades:*

a) *Si $x \in \mathbf{R}$, tenemos*

$$x + (+\infty) = +\infty, \qquad x + (-\infty) = -\infty,$$
$$x - (+\infty) = -\infty, \qquad x - (-\infty) = +\infty,$$
$$x/(+\infty) = x/(-\infty) = 0.$$

b) *Si $x > 0$, tenemos*

$$x(+\infty) = +\infty, \qquad x(-\infty) = -\infty.$$

c) *Si $x < 0$, tenemos*

$$x(+\infty) = -\infty, \qquad x(-\infty) = +\infty.$$

d)

$$(+\infty) + (+\infty) = (+\infty)(+\infty) = (-\infty)(-\infty) = +\infty,$$
$$(-\infty) + (-\infty) = (+\infty)(-\infty) = -\infty.$$

e) *Si $x \in \mathbf{R}$, entonces $-\infty < x < +\infty$.*

NOTACIÓN. Utilizaremos el símbolo $(-\infty, +\infty)$ para designar a $\mathbf{R}$ y $[-\infty, +\infty]$ para designar a $\mathbf{R}^*$. Los puntos de $\mathbf{R}$ se llaman «finitos» para distinguirlos de los puntos «infinitos» $+\infty$ y $-\infty$.

Material protegido por derechos de autor

Demostración. Solamente demostraremos la propiedad distributiva; las otras demostraciones son más simples. Si $x = (x_1, x_2)$, $y = (y_1, y_2)$ y $z = (z_1, z_2)$, entonces tenemos

$$\begin{aligned} x(y + z) &= (x_1, x_2)(y_1 + z_1, y_2 + z_2) \\ &= (x_1y_1 + x_1z_1 - x_2y_2 - x_2z_2, x_1y_2 + x_1z_2 + x_2y_1 + x_2z_1) \\ &= (x_1y_1 - x_2y_2, x_1y_2 + x_2y_1) + (x_1z_1 - x_2z_2, x_1z_2 + x_2z_1) \\ &= xy + xz. \end{aligned}$$

Teorema 1.28.

$$(x_1, x_2) + (0, 0) = (x_1, x_2), \qquad (x_1, x_2)(0, 0) = (0, 0),$$
$$(x_1, x_2)(1, 0) = (x_1, x_2), \qquad (x_1, x_2) + (-x_1, -x_2) = (0, 0).$$

Demostración. Las demostraciones son inmediatas a partir de las definiciones, lo mismo que en los teoremas 1.29, 1.30, 1.32 y 1.33.

Teorema 1.29. *Dados dos números complejos $x = (x_1, x_2)$ e $y = (y_1, y_2)$, existe un número complejo z tal que $x + z = y$. De hecho, $z = (y_1 - x_1, y_2 - x_2)$. Este z se designa por $y - x$. El número complejo $(-x_1, -x_2)$ se designa por $-x$.*

Teorema 1.30. *Para cualquier par de números complejos x e y, tenemos*

$$(-x)y = x(-y) = -(xy) = (-1, 0)(xy).$$

Definición 1.31. *Si $x = (x_1, x_2) \neq (0, 0)$ e y son números complejos, definimos $x^{-1} = [x_1/(x_1^2 + x_2^2), -x_2/(x_1^2 + x_2^2)]$, e $y/x = yx^{-1}$.*

Teorema 1.32. *Si x e y son números complejos con $x \neq (0, 0)$, existe un número complejo z tal que $xz = y$, a saber, $z = yx^{-1}$.*

Revisten especial interés las operaciones con números complejos cuya parte imaginaria es 0.

Teorema 1.33.

$$(x_1, 0) + (y_1, 0) = (x_1 + y_1, 0),$$
$$(x_1, 0)(y_1, 0) = (x_1y_1, 0),$$
$$(x_1, 0)/(y_1, 0) = (x_1/y_1, 0), \qquad \text{si } y_1 \neq 0.$$

NOTA. Es evidente, que en virtud del teorema 1.33, podemos realizar las operaciones aritméticas de los números complejos de parte imaginaria nula operan-

Material protegido por derechos de autor

dientemente, por Argand en 1806. Más tarde Gauss ideó la expresión un tanto desafortunada de «número complejo». Los números complejos admiten otras representaciones geométricas. En vez de utilizar puntos de un plano, se pueden utilizar puntos de otras superficies. Riemann encontró que la esfera es especialmente adecuada para este propósito. Se proyectan los puntos de la esfera desde el Polo Norte sobre el plano tangente a la esfera en el Polo Sur y entonces a cada punto del plano le corresponde un punto sobre la esfera. Con excepción del Polo Norte, a cada punto de la esfera le corresponde un punto sobre el plano y sólo uno. Esta correspondencia se denomina una *proyección estereográfica.* (Ver Fig. 1.3.)

1.23 LA UNIDAD IMAGINARIA

Conviene a veces considerar el número complejo (x_1, x_2) como un vector bidimensional de componentes x_1 y x_2. Sumar dos números complejos utilizando la definición 1.26 es lo mismo que sumar dos vectores componente a componente. El número complejo $1 = (1, 0)$ juega el mismo papel que el vector unitario de dirección horizontal. El análogo al vector unitario de dirección vertical vamos a introducirlo ahora.

Definición 1.34. *El número complejo $(0, 1)$ se representa por i y se llama unidad imaginaria.*

Teorema 1.35. *Cada número complejo $x = (x_1, x_2)$ puede representarse en la forma $x = x_1 + ix_2$.*

Demostración.

$$x_1 = (x_1, 0), \qquad ix_2 = (0, 1)(x_2, 0) = (0, x_2),$$
$$x_1 + ix_2 = (x_1, 0) + (0, x_2) = (x_1, x_2).$$

El próximo teorema expresa que el número complejo i proporciona una solución para la ecuación $x^2 = -1$.

Teorema 1.36. $i^2 = -1$.

Demostración.

$$i^2 = (0, 1)(0, 1) = (-1, 0) = -1.$$

1.24 VALOR ABSOLUTO DE UN NÚMERO COMPLEJO

Vamos a extender ahora el concepto de valor absoluto al sistema de los números complejos.

Material protegido por derechos de autor

Definición 1.37. *Si $x = (x_1, x_2)$, definimos el módulo, o valor absoluto, de x como el número real no negativo $|x|$ dado por*

$$|x| = \sqrt{x_1^2 + x_2^2}.$$

Teorema 1.38.

i) $|(0, 0)| = 0$, y $|x| > 0$ *si* $x \neq 0$.
ii) $|xy| = |x|\,|y|$.
iii) $|x/y| = |x|/|y|$, *si* $y \neq 0$.
iv) $|(x_1, 0)| = |x_1|$.

Demostración. Las afirmaciones (i) y (iv) son inmediatas. Para demostrar (ii), consideremos $x = x_1 + ix_2$, $y = y_1 + iy_2$, entonces $xy = x_1y_1 - x_2y_2 + i(x_1y_2 + x_2y_1)$. La afirmación (ii) se sigue de la relación

$$|xy|^2 = x_1^2y_1^2 + x_2^2y_2^2 + x_1^2y_2^2 + x_2^2y_1^2 = (x_1^2 + x_2^2)(y_1^2 + y_2^2) = |x|^2|y|^2.$$

La ecuación (iii) puede deducirse de (ii) escribiéndola en la forma $|x| = |y|\,|x/y|$.

Geométricamente, $|x|$ representa la longitud del segmento que une el origen con el punto x. En general, $|x - y|$ es la distancia entre los puntos x e y. Utilizando esta interpretación geométrica, el siguiente teorema establece que uno de los lados de un triángulo es menor que la suma de los otros dos lados.

Teorema 1.39. *Si x e y son números complejos, entonces*

$$|x + y| \leq |x| + |y| \quad \textit{(desigualdad triangular)}$$

La demostración se deja como ejercicio para el lector.

1.25 IMPOSIBILIDAD DE ORDENAR LOS NÚMEROS COMPLEJOS

Todavía no hemos definido ninguna relación de la forma $x < y$, si x e y son números complejos cualesquiera, ya que es imposible dar una definición de $<$ para los números complejos que satisfaga las propiedades dadas por los axiomas 6 al 8. Para justificarlo, supongamos que fuese posible definir una relación de orden $<$ que satisficiera los axiomas 6, 7 y 8. Entonces, como $i \neq 0$, se debiera tener $i > 0$ o $i < 0$, por el axioma 6. Supongamos que $i > 0$. Entonces tomando $x = y = i$ en el axioma 8, tendríamos $i^2 > 0$, o $-1 > 0$. Sumando 1 a ambos miembros (axioma 7), obtendríamos $0 > 1$. Por otro lado, aplicando el axioma 8 a $-1 > 0$, hallaríamos $1 > 0$. Tendríamos, pues, $0 > 1$ y también $1 > 0$, que, por el axioma 6, es imposible. Así pues, suponer que $i > 0$ lleva a contradicción. [¿Por qué la desigualdad $-1 > 0$ no era ya una

contradicción?] Un razonamiento análogo prueba que no es posible $i < 0$. Por lo tanto, los números complejos no pueden ser ordenados de tal suerte que se verifiquen los axiomas 6, 7 y 8.

1.26. EXPONENCIALES COMPLEJAS

La exponencial e^x (x real) ha sido mencionada anteriormente. Deseamos definir e^z para z complejo de tal suerte que las principales propiedades de la función exponencial real se conserven. Las citadas propiedades de e^x para x real son la ley de los exponentes, $e^{x_1}e^{x_2} = e^{x_1+x_2}$, y la ecuación $e^0 = 1$. Daremos una definición de e^z para z complejo que conserve estas propiedades y que se reduzca a la exponencial ordinaria cuando z sea real.

Si escribimos $z = x + iy$ (x, y reales), entonces para que se verifique la ley de los exponentes deberíamos tener $e^{x+iy} = e^xe^{iy}$. Queda entonces por definir lo que significa e^{iy}.

Definición 1.40. *Si $z = x + iy$, definimos $e^z = e^{x+iy}$ como el número complejo $e^z = e^x$ (cos $y + i$ sen y).*

Esta definición* coincide claramente con la función exponencial real cuando z es real (esto es, $y = 0$). Probaremos a continuación que la ley de los exponentes se cumple.

Teorema 1.41. *Si $z_1 = x_1 + iy_1$ y $z_2 = x_2 + iy_2$ son dos números complejos, entonces tenemos*

$$e^{z_1}e^{z_2} = e^{z_1+z_2}.$$

Demostración.

$$e^{z_1} = e^{x_1}(\cos y_1 + i \operatorname{sen} y_1), \qquad e^{z_2} = e^{x_2}(\cos y_2 + i \operatorname{sen} y_2),$$

$$e^{z_1}e^{z_2} = e^{x_1}e^{x_2}[\cos y_1 \cos y_2 - \operatorname{sen} y_1 \operatorname{sen} y_2 + i(\cos y_1 \operatorname{sen} y_2 + \operatorname{sen} y_1 \cos y_2)].$$

* Es posible dar muchos argumentos para motivar la ecuación $e^{iy} = \cos y + i$ sen y. Por ejemplo, escribamos $e^{iy} = f(y) + ig(y)$ e intentemos determinar las funciones de variable real f y g a fin de que las leyes usuales de las operaciones con exponenciales reales sean aplicables también a las exponenciales complejas. Diferenciando formalmente se obtiene $e^{iy} = g'(y) - if'(y)$, si suponemos que $(e^{iy})' = ie^{iy}$. Comparando estas dos expresiones para e^{iy}, vemos que f y g deben satisfacer las ecuaciones $f(y) = g'(y)$, $f'(y) = -g(y)$. La eliminación de g conduce a $f(y) = -f''(y)$. Como deseamos que $e^0 = 1$, debemos tener que $f(0) = 1$ y $f'(0) = 0$. Ello prueba que $f(y) = \cos y$ y $g(y) = -f'(y) = \operatorname{sen} y$. Por supuesto, este razonamiento *no prueba* nada, pero indica ostensiblemente que la definición $e^{iy} = \cos y + i$ sen y es razonable.

Material protegido por derechos de autor
Material protegido por derechos de autor

Ahora bien, $e^{x_1}e^{x_2} = e^{x_1+x_2}$, ya que x_1 y x_2 son ambos reales. Además,

$$\cos y_1 \cos y_2 - \operatorname{sen} y_1 \operatorname{sen} y_2 = \cos(y_1 + y_2)$$

y

$$\cos y_1 \operatorname{sen} y_2 + \operatorname{sen} y_1 \cos y_2 = \operatorname{sen}(y_1 + y_2),$$

y por lo tanto

$$e^{z_1}e^{z_2} = e^{x_1+x_2}[\cos(y_1 + y_2) + i \operatorname{sen}(y_1 + y_2)] = e^{z_1+z_2}.$$

1.27 OTRAS PROPIEDADES DE LAS EXPONENCIALES COMPLEJAS

En los teoremas siguientes z, z_1, z_2 designan números complejos.

Teorema 1.42. *e^z jamás es cero.*

Demostración. $e^z e^{-z} = e^0 = 1$. Por lo tanto, e^z no puede ser cero.

Teorema 1.43. *Si x es real, entonces $|e^{ix}| = 1$.*

Demostración. $|e^{ix}|^2 = \cos^2 x + \operatorname{sen}^2 x = 1$, y $|e^{ix}| > 0$.

Teorema 1.44. *$e^z = 1$ si, y sólo si, z es un múltiplo entero de $2\pi i$.*

Demostración. Si $z = 2\pi i n$, donde n es un entero, entonces

$$e^z = \cos(2\pi n) + i \operatorname{sen}(2\pi n) = 1.$$

Recíprocamente, supongamos que $e^z = 1$. Esto significa que $e^x \cos y = 1$ y $e^x \operatorname{sen} y = 0$. Como que $e^x \neq 0$, debe ser $\operatorname{sen} y = 0$, $y = k\pi$, donde k es un entero. Pero $\cos(k\pi) = (-1)^k$. Por lo tanto, $e^x = (-1)^k$ ya que $e^x \cos(k\pi) = 1$. Como $e^x > 0$, k debe ser par. Por lo tanto $e^x = 1$ y entonces $x = 0$. Esto prueba el teorema.

Teorema 1.45. *$e^{z_1} = e^{z_2}$ si, y sólo si, $z_1 - z_2 = 2\pi i n$ (donde n es un entero).*

Demostración. $e^{z_1} = e^{z_2}$ si, y sólo si, $e^{z_1-z_2} = 1$.

1.28 EL ARGUMENTO DE UN NÚMERO COMPLEJO

Si el punto $z = (x, y) = x + iy$ se representa en cordenadas polares r y θ, podemos escribir $x = r\cos\theta$ e $y = r \operatorname{sen}\theta$, es decir, $z = r\cos\theta + ir \operatorname{sen}\theta = re^{i\theta}$. Los dos números r y θ determinan a z de forma única. Recíprocamente, el nú-

$$z^0 = 1, \qquad z^{n+1} = z^n z, \qquad \text{si } n \geq 0,$$

$$z^{-n} = (z^{-1})^n, \qquad \text{si } z \neq 0 \text{ y } n > 0.$$

El teorema 1.50 establece que se verifican las reglas usuales de los exponentes. La demostración, que se puede hacer por inducción, se deja como ejercicio.

Teorema 1.50. *Dados dos enteros m y n, tenemos, para $z \neq 0$,*

$$z^n z^m = z^{n+m} \qquad \text{y} \qquad (z_1 z_2)^n = z_1^n z_2^n.$$

Teorema 1.51. *Si $z \neq 0$, y si n es un entero positivo, entonces existen exactamente n números complejos distintos $z_0, z_1, \ldots, z_{n-1}$ (llamados raíces n-ésimas de z), tales que*

$$z_k^n = z, \quad \text{para cada } k = 0, 1, 2, \ldots, n-1.$$

Además, estas raíces son dadas por las fórmulas

$$z_k = Re^{i\phi_k}, \qquad \text{donde} \qquad R = |z|^{1/n},$$

y

$$\phi_k = \frac{\arg(z)}{n} + \frac{2\pi k}{n} \qquad (k = 0, 1, 2, \ldots, n-1).$$

NOTA. Las n raíces n-ésimas de z están igualmente espaciadas sobre el círculo de radio $R = |z|^{1/n}$, con centro en el origen.

Demostración. Los n números complejos $Re^{i\phi_k}$, $0 \leq k \leq n-1$, son distintos y cada uno de ellos es una raíz n-ésima de z, ya que

$$(Re^{i\phi_k})^n = R^n e^{in\phi_k} = |z|e^{i[\arg(z)+2\pi k]} = z.$$

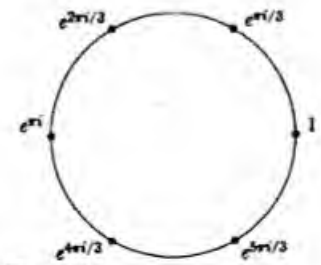


Figura 1.4

Material protegido por derechos de autor
Material protegido por derechos de autor

4. Si $x > 0$, Log $(x) = \log (x)$, ya que $|x| = x$ y arg $(x) = 0$.
5. Puesto que $|1 + i| = \sqrt{2}$ y arg $(1 + i) = \pi/4$, Log $(1 + i) = \log \sqrt{2} + i\pi/4$.

Teorema 1.54. *Si $z_1 z_2 \neq 0$, entonces*

$$\text{Log}\,(z_1 z_2) = \text{Log}\, z_1 + \text{Log}\, z_2 + 2\pi i n(z_1, z_2),$$

donde $n(z_1, z_2)$ es el entero definido en el teorema 1.48.

Demostración.

$$\begin{aligned}\text{Log}\,(z_1 z_2) &= \log |z_1 z_2| + i \arg (z_1 z_2) \\ &= \log |z_1| + \log |z_2| + i\,[\arg (z_1) + \arg (z_2) + 2\pi n(z_1, z_2)].\end{aligned}$$

1.31 POTENCIAS COMPLEJAS

Utilizando los logaritmos complejos, podemos dar ahora una definición de las potencias complejas de números complejos.

Definición 1.55. *Si $z \neq 0$ y si w es un número complejo cualquiera, definimos*

$$z^w = e^{w \text{Log}\, z}.$$

EJEMPLOS

1. $i^i = e^{i \text{Log}\, i} = e^{i((\pi/2))} = e^{-\pi/2}$.
2. $(-1)^i = e^{i \text{Log}\,(-1)} = e^{i(\pi)} = e^{-\pi}$.
3. Si n es un entero, entonces $z^{n+1} = e^{(n+1)\text{Log}\, z} = e^{n \text{Log}\, z} e^{\text{Log}\, z} = z^n z$, por lo que la definición 1.55 no se contradice con la definición 1.49.

Los dos teoremas siguientes nos suministran las reglas de cálculo con potencias complejas:

Teorema 1.56. *$z^{w_1} z^{w_2} = z^{w_1 + w_2}$, si $z \neq 0$.*

Demostración.

$$z^{w_1 + w_2} = e^{(w_1 + w_2)\text{Log}\, z} = e^{w_1 \text{Log}\, z} e^{w_2 \text{Log}\, z} = z^{w_1} z^{w_2}.$$

Teorema 1.57. *Si $z_1 z_2 \neq 0$, entonces*

$$(z_1 z_2)^w = z_1^w z_2^w e^{2\pi i w n(z_1, z_2)},$$

donde $n(z_1, z_2)$ es el entero definido en el teorema 1.48.

Demostración.

$$(z_1 z_2)^w = e^{w \text{Log}\,(z_1 z_2)} = e^{w[\text{Log}\, z_1 + \text{Log}\, z_2 + 2\pi i n(z_1, z_2)]}.$$

1.32 SENOS Y COSENOS COMPLEJOS

Definición 1.58. *Dado un número complejo z, definimos*

$$\cos z = \frac{e^{iz} + e^{-iz}}{2}, \qquad \text{sen}\, z = \frac{e^{iz} - e^{-iz}}{2i}.$$

NOTA. Cuando z es real, estas igualdades concuerdan con la definición 1.40.

Teorema 1.59. *Si $z = x + iy$, entonces tenemos*

$$\begin{aligned}\cos z &= \cos x \cosh y - i \,\text{sen}\, x \,\text{senh}\, y, \\ \text{sen}\, z &= \text{sen}\, x \cosh y + i \cos x \,\text{senh}\, y.\end{aligned}$$

Demostración.

$$\begin{aligned}2 \cos z &= e^{iz} + e^{-iz} \\ &= e^{-y}(\cos x + i \,\text{sen}\, x) + e^{y}(\cos x - i \,\text{sen}\, x) \\ &= \cos x(e^{y} + e^{-y}) - i \,\text{sen}\, x(e^{y} - e^{-y}) \\ &= 2 \cos x \cosh y - 2i \,\text{sen}\, x \,\text{senh}\, y.\end{aligned}$$

La demostración para sen z es análoga.

Más propiedades de los senos y cosenos se dan en los ejercicios.

1.33 INFINITO Y EL PLANO COMPLEJO AMPLIADO C*

A continuación extendemos el sistema de los números complejos adjuntando un punto ideal designado por el símbolo ∞.

Definición 1.60. *Por sistema de los números complejos ampliado* **C*** *entenderemos el plano complejo* **C** *junto con un símbolo ∞ que satisfaga las siguientes propiedades:*

a) *Si $z \in \mathbf{C}$, entonces se tiene $z + \infty = z - \infty = \infty$, $z/\infty = 0$.*
b) *Si $z \in \mathbf{C}$, pero $z \neq 0$, entonces $z(\infty) = \infty$ y $z/0 = \infty$.*
c) $\infty + \infty = (\infty)(\infty) = \infty$.

Material protegido por derechos de autor
Material protegido por derechos de autor

Definición 1.61. *Cada conjunto de* **C** *de la forma* $\{z : |z| > r \geq 0\}$ *se denomina entorno de* ∞*, o bola con centro en* ∞.

El lector puede preguntarse por qué a **R** le hemos adjuntado dos símbolos, $+\infty$ y $-\infty$, mientras que a **C** sólo le adjuntamos un símbolo, ∞. La respuesta radica en el hecho de que existe una relación de orden $<$ entre números reales, mientras que entre números complejos no sucede lo mismo. Para que ciertas propiedades de los números reales que involucran la relación $<$ se verifiquen sin excepción, es necesario disponer de dos símbolos, $+\infty$ y $-\infty$, tales como los definidos anteriormente. Ya hemos mencionado, por ejemplo, que cada conjunto no vacío tiene un sup en $\mathbf{R}^*$.

En **C** resulta más conveniente disponer de un solo punto ideal. A modo de ilustración, recordemos que la proyección estereográfica establece una correspondencia uno a uno entre los puntos del plano complejo y los puntos de la superficie de la esfera, distintos del Polo Norte. La aparente excepción del Polo Norte puede ser eliminada considerándolo la imagen geométrica del punto ideal ∞. Así conseguiremos una correspondencia uno a uno entre el plano complejo ampliado $\mathbf{C}^*$ y la superficie total de la esfera. Es evidente, desde un punto geométrico, que si el Polo Sur se coloca en el origen del plano complejo, el exterior de un «amplio» círculo en el plano se colocará, por proyección estereográfica, en un «pequeño» casquete esférico alrededor del Polo Norte. Ello ilustra con claridad por qué hemos definido un entorno de ∞ mediante una desigualdad de la forma $|z| > r$.

EJERCICIOS

Enteros

1.1 Demostrar que no existe un primo máximo. (Una demostración era conocida por Euclides.)

1.2 Si n es un entero positivo, probar la identidad algebraica

$$a^n - b^n = (a - b) \sum_{k=0}^{n-1} a^k b^{n-1-k}.$$

1.3 Si $2^n - 1$ es primo, probar que n es primo. Un primo de la forma $2^p - 1$, donde p es primo, se llama un *primo de Mersenne.*

1.4. Si $2^n + 1$ es primo, entonces n es una potencia de dos. Un primo de la forma $2^{2^m} + 1$ se llama un *primo de Fermat. Indicación:* Utilizar el ejercicio 1.2.

1.5 Los *números de Fibonacci* 1, 1, 2, 3, 5, 8, 13, ..., son definidos recursivamente por la fórmula $x_{n+1} = x_n + x_{n-1}$, con $x_1 = x_2 = 1$.
Probar que $(x_n, x_{n+1}) = 1$ y que $x_n = (a^n - b^n)/(a - b)$, donde a y b son las raíces de la ecuación $x^2 - x - 1 = 0$.

1.6 Probar que cada conjunto no vacío de números enteros positivos posee primer elemento. Este es el *principio de buena ordenación.*

Números racionaes e irracionales

1.7 Hallar el número racional cuya expresión decimal es 0,3344444...

1.8 Probar que la expresión decimal de x terminará en ceros (o en nueves) si, y sólo si, x es un número racional cuyo denominador es de la forma 2^n5^m, donde m y n son enteros no negativos.

1.9 Probar que $\sqrt{2} + \sqrt{3}$ es irracional.

1.10 Si a, b, c, d son racionales y si x es irracional, probar que $(ax + b)/(cx + d)$ es, en general, irracional. ¿Cuándo se dan las excepciones?

1.11 Dado un número real cualquiera $x > 0$, probar que hay un irracional entre 0 y x.

1.12 Si $a/b < c/d$ con $b > 0$, $d > 0$, probar que $(a + c)/(b + d)$ está entre a/b y c/d.

1.13 Sean a y b enteros positivos. Probar que $\sqrt{2}$ está siempre entre las dos fracciones a/b y $(a + 2b)/(a + b)$. ¿Qué fracción está más próxima a $\sqrt{2}$?

1.14 Probar que $\sqrt{n-1} + \sqrt{n+1}$ es irracional para todo entero $n \geq 1$.

1.15 Dado un número real x y un entero $N > 1$, probar que existen enteros h y k con $0 < k \leq N$ tales que $|kx - h| < 1/N$. *Indicación.* Considerar los $N + 1$ números $tx - [tx]$ para $t = 0, 1, 2, \ldots, N$ y probar que algún par difiere a lo más $1/N$.

1.16 Si x es irracional, probar que existe una infinidad de números racionales h/k con $k > 0$ tales que $|x - h/k| < 1/k^2$. *Indicación.* Suponer que sólo existe un número finito $h_1/k_1, \ldots, h_r/k_r$ y aplicar el ejercicio 1.15 para llegar a contradicción, con $N > 1/\delta$, donde δ es el menor de los números $|x - h_i/k_i|$.

1.17 Sea x un número racional positivo de la forma

$$x = \sum_{k=1}^{n} \frac{a_k}{k!},$$

donde cada a_k es un entero no negativo con $a_k \leq k - 1$ para $k \geq 2$ y $a_n > 0$. Sea $[x]$ el mayor entero contenido en x. Probar que $a_1 = [x]$, que $a_k = [k!\, x] - k\,[(k-1)!\, x]$ para $k = 2, \ldots, n$, y que n es el menor entero tal que $n!\, x$ es entero. Recíprocamente, probar que cada número racional positivo x puede ser expresado en esta forma de una manera y una sola.

Cotas superiores

1.18 Probar que el sup y el inf de un conjunto, si existen, son únicos.

1.19 Hallar el sup y el inf de cada uno de los siguientes conjuntos de números reales:

a) Todos los números de la forma $2^{-p} + 3^{-q} + 5^{-r}$, donde p, q y r toman todos los valores enteros positivos.

Material protegido por derechos de autor
Material protegido por derechos de autor

Números complejos

1.27 Expresar los siguientes números complejos en la forma $a + bi$.

a) $(1 + i)^3$ b) $(2 + 3i)/(3 - 4i)$,
c) $i^5 + i^{16}$, d) $\frac{1}{2}(1 + i)/(1 + i^{-8})$.

1.28 En cada caso, determinar todos los valores reales x e y que satisfacen la relación dada.

a) $x + iy = |x - iy|$, b) $x + iy = (x - iy)^2$, c) $\sum_{k=0}^{100} i^k = x + iy$.

1.29 Si $z = x + iy$, x e y reales, el complejo conjugado de z es el número complejo $\bar{z} = x - iy$. Probar que:

a) $\overline{z_1 + z_2} = \bar{z}_1 + \bar{z}_2$, b) $\overline{z_1 z_2} = \bar{z}_1 \bar{z}_2$, c) $z\bar{z} = |z|^2$,
d) $z + \bar{z}$ = al doble de la parte real de z.
e) $(z - \bar{z})/i$ = al doble de la parte imaginaria de z.

1.30 Describir geométricamente el conjunto de los números complejos z que satisfacen cada una de las condiciones siguientes:

a) $|z| = 1$, b) $|z| < 1$, c) $|z| \leq 1$,
d) $z + \bar{z} = 1$, e) $z - \bar{z} = i$, f) $\bar{z} + z = |z|^2$.

1.31 Dados tres números complejos z_1, z_2, z_3 tales que $|z_1| = |z_2| = |z_3| = 1$ y $z_1 + z_2 + z_3 = 0$, probar que estos tres números son los vértices de un triángulo equilátero inscrito en el círculo unidad y centrado en el origen.

1.32 Si a y b son números complejos, probar que:

a) $|a - b|^2 \leq (1 + |a|^2)(1 + |b|^2)$.

b) Si $a \neq 0$, entonces $|a + b| = |a| + |b|$ si, y sólo si, b/a es real y no negativo.

1.33 Si a y b son números complejos, probar que

$$|a - b| = |1 - \bar{a}b|$$

si, y sólo si, $|a| = 1$ o $|b| = 1$. ¿Para qué números a y b es válida la desigualdad $|a - b| < |1 - \bar{a}b|$?

1.34 Si a y c son números reales constantes, b es complejo, probar que la ecuación

$$az\bar{z} + b\bar{z} + \bar{b}z + c = 0 \qquad (a \neq 0,\ z = x + iy)$$

representa un círculo en el plano xy.

1.35 Recordemos la definición de la inversa de la tangente: dado un número real t, $\text{tg}^{-1}(t)$ es el único número real θ que satisface las dos condiciones siguientes:

$$-\frac{\pi}{2} < \theta < +\frac{\pi}{2}, \qquad \text{tg } \theta = t.$$

Material protegido por derechos de autor

iii) Si a es un entero, probar que la fórmula de (i) se verifica sin necesidad de imponer restricciones a θ. En este caso se conoce como el teorema de Moivre.

1.45 Utilizar el teorema de Moivre (ejercicio 1.44) para obtener las identidades trigonométricas

$$\text{sen } 3\theta = 3\cos^2\theta \text{ sen } \theta - \text{sen}^3\theta,$$
$$\cos 3\theta = \cos^3\theta - 3\cos\theta \text{ sen}^2\theta$$

válidas para todo θ real. ¿Son válidas si θ es complejo?

1.46 Definimos $\text{tg } z = (\text{sen } z)/(\cos z)$ y probar que, para $z = x + iy$, se tiene

$$\text{tg } z = \frac{\text{sen } 2x + i \text{ senh } 2y}{\cos 2x + \cosh 2y}$$

1.47 Sea w un número complejo dado. Si $w \neq \pm 1$, probar que existen dos valores de $z = x + iy$ que satisfacen las condiciones $\cos z = w$ y $-\pi < x \leq +\pi$. Hallar estos valores cuando $w = i$ y cuando $w = 2$.

1.48 Demostrar la identidad de Lagrange para números complejos:

$$\left|\sum_{k=1}^{n} a_k b_k\right|^2 = \sum_{k=1}^{n} |a_k|^2 \sum_{k=1}^{n} |b_k|^2 - \sum_{1 \leq k < j \leq n} |a_k b_j - a_j b_k|^2.$$

Utilizarla para deducir la desigualdad de Cauchy-Schwarz para números complejos.

1.49 a) Probar, utilizando la ecuación de la parte imaginaria de la fórmula de Moivre, que

$$\text{sen } n\theta = \text{sen}^n\theta \left\{ \binom{n}{1} \text{cotg}^{n-1}\theta - \binom{n}{3} \text{cotg}^{n-3}\theta + \binom{n}{5} \text{cotg}^{n-5}\theta - + \cdots \right\}.$$

b) Si $0 < \theta < \pi/2$, probar que

$$\text{sen } (2m + 1)\theta = \text{sen}^{2m+1}\theta P_m(\text{cotg}^2\theta)$$

donde P_m es el polinomio de grado m dado por

$$P_m(x) = \binom{2m+1}{1} x^m - \binom{2m+1}{3} x^{m-1} + \binom{2m+1}{5} x^{m-2} - + \cdots$$

Utilizar este resultado para demostrar que P_m tiene ceros en los m puntos distintos $x_k = \text{cotg}^2\{\pi k/(2m + 1)\}$ para $k = 1, 2, \ldots, m$.

c) Demostrar que la suma de los ceros de P_m viene dada por

$$\sum_{k=1}^{m} \text{cotg}^2 \frac{\pi k}{2m + 1} = \frac{m(2m - 1)}{3},$$

y que la suma de sus cuadrados viene dada por

$$\sum_{k=1}^{m} \text{cotg}^4 \frac{\pi k}{2m + 1} = \frac{m(2m - 1)(4m^2 + 10m - 9)}{45}.$$

Material protegido por derechos de autor

NOTA. Estas identidades pueden utilizarse para demostrar que $\sum_{n=1}^{\infty} n^{-2} = \pi^2/6$ y $\sum_{n=1}^{\infty} n^{-4} = \pi^4/90$. (Ver ejercicios 8.46 y 8.47.)

1.50 Probar que $z^n - 1 = \prod_{k=1}^{n} (z - e^{2\pi ik/n})$ para todo complejo z. Utilizar esto para deducir la fórmula

$$\prod_{k=1}^{n-1} \operatorname{sen} \frac{k\pi}{n} = \frac{n}{2^{n-1}} \quad \text{para } n \geq 2.$$

REFERENCIAS SUGERIDAS PARA POSTERIORES ESTUDIOS

1.1 Apostol, T. M., *Calculus*, Vol. 1, 2.ª ed. Ed. Reverté, S. A., Barcelona.

1.2 Birkhoff, G., y MacLane, S., *A Survey of Modern Algebra*, 3.ª ed. Macmillan, New York, 1965. (Hay traducción al castellano. Ed. Vicens Vives, Barcelona.)

1.3 Cohen, L., y Ehrlich, G., *The Structure of the Real-Number System*. Van Nostrand, Princeton, 1963.

1.4 Gleason, A., *Fundamentals of Abstract Analysis*. Addison-Wesley, Reading, 1966.

1.5 Hardy, G. H., *A Course of Pure Mathematics*, 10.ª ed. Cambridge University Press, 1952.

1.6 Hobson, E. W., *The Theory of Functions of a Real Variable and the Theory of Fourier's Series*, Vol. 1, 3.ª ed. Cambridge University Press, 1927.

1.7 Landau, E., *Foundations of Analysis*, 2.ª ed. Chelsea, New York, 1960.

1.8 Robinson, A., *Non-Standard Analysis*. North-Holland, Amsterdam, 1966.

1.9 Thurston, H. A., *The Number System*. Blackie, London, 1956.

1.10 Wilder, R. L., *Introduction to the Foundations of Mathematics*. 2.ª ed. Wiley, New York, 1965.

CAPÍTULO 2

Algunas nociones básicas de la teoría de conjuntos

2.1 INTRODUCCIÓN

Al estudiar las distintas ramas de la Matemática es útil manejar la notación y la terminología de la Teoría de conjuntos. Esta teoría, desarrollada por Boole y por Cantor a finales del siglo diecinueve, ha tenido una gran influencia en el desarrollo de las matemáticas del siglo veinte. Ha unificado muchas ideas, aparentemente desconexas, y ha ayudado a reducir muchos conceptos matemáticos a sus fundamentos lógicos de una manera elegante y metódica.

No daremos un desarrollo sistemático de la teoría de conjuntos; nos limitaremos a discutir algunos de sus conceptos básicos. El lector que desee explorar este terreno más ampliamente puede consultar las referencias del final de este capítulo.

Una colección de objetos, considerados como una sola entidad, se llamará *conjunto*. Los objetos de la colección se llamarán *elementos* o *miembros* del conjunto y diremos que *pertenecen al* conjunto o que *están* contenidos *en él*. El conjunto, a su vez se dice que, los *contiene* o *está compuesto* por sus elementos. Nuestro interés radica, principalmente, en los conjuntos de entes matemáticos; esto es, conjuntos de números, puntos, funciones, curvas, etc. Sin embargo, como la mayor parte de la teoría de conjuntos no depende de la naturaleza de los objetos individuales de la colección, supone una gran economía de imaginación estudiar conjuntos cuyos elementos puedan ser de cualquier tipo. Es a causa de esta cualidad de generalización por lo que la Teoría de conjuntos ha tenido un efecto tan grande en la mayor parte de los desarrollos matemáticos.

2.2 NOTACIONES

Los *conjuntos* los designaremos, usualmente, por medio de letras mayúsculas:

$$A, B, C, \ldots, X, Y, Z,$$

Material protegido por derechos de autor
Material protegido por derechos de autor

y los *elementos* por medio de letras minúsculas: $a, b, c, \ldots, x, y, z$. Se escribe $x \in S$ para indicar que «x es un elemento de S», o que «x pertenece a S». Si x no pertenece a S, se escribe $x \notin S$. A veces para designar un conjunto escribiremos sus elementos entre llaves; por ejemplo, el conjunto de los enteros pares positivos menores que 10 se expresa por medio de $\{2, 4, 6, 8\}$. Se escribe $S = \{x : x \text{ satisface a } P\}$ para indicar que S es la colección de los x para los cuales se verifica la propiedad P.

A partir de un conjunto dado es posible formar nuevos conjuntos, llamados *subconjuntos* del conjunto dado. Por ejemplo, el conjunto de todos los enteros positivos menores que 10 que son divisibles por 4, es decir, $\{4, 8\}$, es un subconjunto del conjunto de los enteros pares positivos menores que 10. En general, decimos que un conjunto A es subconjunto de B, y se escribe $A \subseteq B$, si todo elemento de A pertenece a B. La afirmación $A \subseteq B$ no elimina la posibilidad de que sea $B \subseteq A$. De hecho, son simultáneas $A \subseteq B$ y $B \subseteq A$ si, y sólo si, A y B tienen los mismos elementos. En este caso, decimos que A y B son iguales y escribimos $A = B$. Si A y B no son iguales, escribimos $A \neq B$. Si $A \subseteq B$, pero $A \neq B$, entonces se dice que A es un *subconjunto propio* de B.

Conviene considerar la posibilidad de un conjunto sin elementos; tal conjunto se llama *conjunto vacío* y se le considera, por convenio, subconjunto de todo conjunto. El lector puede hallar útil imaginar un conjunto como una caja que contiene ciertos objetos, sus elementos. El conjunto vacío es, entonces, una caja vacía. El conjunto vacío se designa por el símbolo $\emptyset$.

2.3 PARES ORDENADOS

Consideremos un conjunto de dos elementos a y b; es decir, el conjunto $\{a, b\}$. En virtud de nuestra definición de igualdad, este conjunto es igual al conjunto $\{b, a\}$, ya que no se halla involucrada la cuestión del orden. Sin embargo, es necesario considerar también conjuntos de dos elementos en los que el orden *sea* importante. Por ejemplo, en Geometría analítica plana, las coordenadas (x, y) de un punto representan un *par ordenado* de números. El *punto* $(3, 4)$ es distinto del punto $(4, 3)$, mientras que el *conjunto* $\{3, 4\}$ es el mismo que el conjunto $\{4, 3\}$. Cuando deseemos considerar un conjunto de dos elementos a y b, *ordenados*, escribiremos los elementos entre paréntesis: (a, b). Entonces a es el primer elemento y b el segundo. Es posible dar una definición de par ordenado de objetos (a, b) que involucre tan sólo el lenguaje de la teoría de conjuntos. Tal definición es la siguiente:

Definición 2.1.

$$(a, b) = \{\{a\}, \{a, b\}\}.$$

Esta definición establece que (a, b) es un conjunto que contiene dos elementos $\{a\}$ y $\{a, b\}$. Utilizando dicha definición se puede demostrar el siguiente teorema:

Teorema 2.2. *$(a, b) = (c, d)$ si, y sólo si, $a = c$ y $b = d$.*

Este teorema muestra que la definición 2.1 es una definición «razonable» de par ordenado, en el sentido de que el objeto a se distingue del objeto b. La demostración del teorema 2.2 es un ejercicio instructivo para el lector. (Ver ejercicio 2.1.)

2.4 PRODUCTO CARTESIANO DE DOS CONJUNTOS

Definición 2.3. *Dados dos conjuntos A y B, llamaremos producto cartesiano de A y B, y lo representaremos $A \times B$, al conjunto de todos los pares ordenados (a, b) tales que $a \in A$ y $b \in B$.*

Ejemplo. Si $\mathbf{R}$ representa el conjunto de todos los números reales, entonces a $\mathbf{R} \times \mathbf{R}$ le corresponde el conjunto de todos los números complejos.

2.5 RELACIONES Y FUNCIONES

Sean x e y números reales, de modo que el par ordenado (x, y) pueda ser interpretado como las coordenadas rectangulares de un punto del plano xy (o como un número complejo). Encontramos, con frecuencia, expresiones tales como

$$xy = 1, \qquad x^2 + y^2 = 1, \qquad x^2 + y^2 \leq 1, \qquad x < y. \tag{a}$$

Cada una de estas expresiones determina un cierto conjunto de pares ordenados (x, y) de números reales; es decir, el conjunto de todos los pares ordenados (x, y) para los que la expresión se satisface. Un tal conjunto de pares orde-

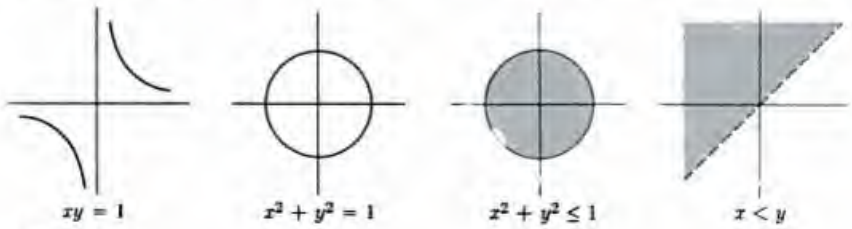


Figura 2.1

Material protegido por derechos de autor
Material protegido por derechos de autor

Si $\mathcal{D}(F)$ es un subconjunto de un producto cartesiano $A \times B$, entonces F es una *función de dos variables*. En este caso los valores de la función se designan por $F(a, b)$ en vez de $F((a, b))$. Una función de dos variables reales es aquella cuyo dominio es un subconjunto de $\mathbf{R} \times \mathbf{R}$.

Si S es un subconjunto de $\mathcal{D}(F)$, diremos que F está *definida en S*. En este caso, el conjunto de los $F(x)$ con $x \in S$ se denomina *imagen de S por F* y se designa por $F(S)$. Si T es un conjunto cualquiera que contenga a $F(S)$, entonces F se llama también *aplicación de S en T*. Esto se expresa, corrientemente, escribiendo

$$F: S \to T.$$

Si $F(S) = T$, se dice que la aplicación es *sobre T*. Una aplicación de S en sí mismo se denomina a veces *transformación*.

Consideremos, por ejemplo, la función de una variable compleja definida por la ecuación $F(z) = z^2$. Esta función aplica cada sector S de la forma $0 \leq \arg(z) \leq \alpha \leq \pi/2$ del plano complejo z sobre un sector $F(S)$ determinado por las desigualdades $0 \leq \arg[F(z)] \leq 2\alpha$. (Ver Fig. 2.2.)

Figura 2.2

Si dos funciones F y G satisfacen la relación de inclusión $G \subseteq F$, se dice que G es una *restricción* de F o que F es una *extensión* de G. En particular, si S es un subconjunto de $\mathcal{D}(F)$ y si G está definida por la ecuación

$$G(x) = F(x) \qquad \text{para todo } x \text{ de } S,$$

entonces se dice que G es la restricción de F a S. La función G consta de los pares de la forma $(x, F(x))$, con $x \in S$. Su dominio es S y su recorrido es $F(S)$.

2.7 FUNCIONES UNO A UNO E INVERSAS

Definición 2.7. *Sea F una función definida en S. Se dice que F es uno a uno en S si, y sólo si, para todo x e y de S,*

$$F(x) = F(y) \qquad \textit{implica } x = y.$$

Esto equivale a decir que una función que es uno a uno en S asigna valores distintos a elementos de S distintos. Estas funciones se llaman también *inyectivas*. Son importantes puesto que, como veremos en seguida, poseen *inversas*. Sin embargo, antes de establecer la definición de inversa de una función, conviene introducir una noción más general, que es la de *inversa* de una relación.

Definición 2.8. *Dada una relación S, la nueva relación* $\check{S}$ *definida por*

$$\check{S} = \{(a, b) : (b, a) \in S\}$$

se llama la inversa de S.

Así, un par ordenado (a, b) pertenece a $\check{S}$ si, y sólo si, el par con los elementos invertidos, (b, a), pertenece a S. Cuando S es una *relación plana*, esto significa, simplemente, que el grafo de $\check{S}$ es el simétrico del grafo de S con respecto a la recta $y = x$ como eje de simetría. En la relación definida por $x < y$, la relación inversa se define por $y < x$.

Definición 2. 9. *Supongamos que la relación F es una función. Consideremos la relación inversa* $\check{F}$*, que puede ser o no ser una función. Si* $\check{F}$ *es también una función, entonces* $\check{F}$ *se llama inversa de F y se designa por* F^{-1}.

La Figura 2.3(a) ilustra un ejemplo de una función F para la que $\check{F}$ no es función. En la Fig. 2.3(b) tanto F como su inversa son funciones.

El siguiente teorema nos dice que toda función que sea uno a uno en su dominio posee, siempre, una inversa.

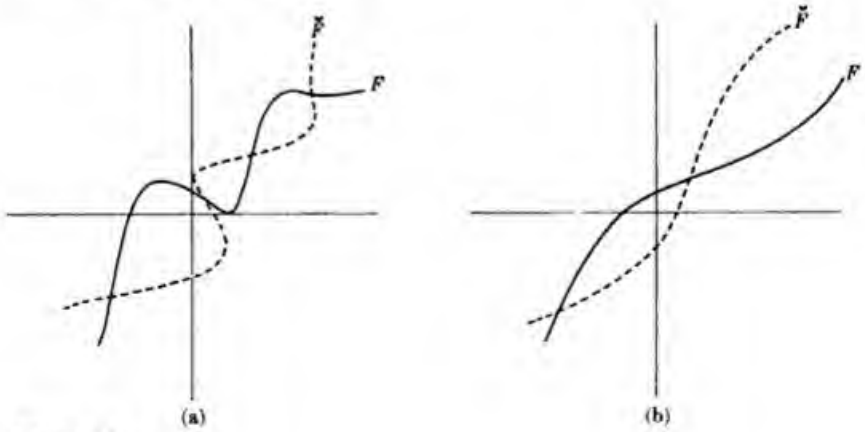


Figura 2.3

Material protegido por derechos de autor
Material protegido por derechos de autor

de F, esto es, el conjunto $\{F(1), F(2), F(3), \ldots\}$, *se designa también por* $\{F_1, F_2, F_3, \ldots\}$, *y el valor* F_n *se llama el término n-ésimo de la sucesión.*

Por motivos de brevedad, usaremos en ocasiones la notación $\{F_n\}$ para designar la sucesión infinita cuyo término n-ésimo es F_n.

Sea $s = \{s_n\}$ una sucesión infinita, y sea k una función cuyo dominio es el conjunto de los enteros positivos y cuyo recorrido es un subconjunto del conjunto de los enteros positivos. Supongamos que k «conserva el orden» o, con otras palabras, «es creciente», esto es, supongamos que

$$k(m) < k(n), \qquad \text{si } m < n.$$

La función compuesta $s \circ k$ está definida para todo entero $n \geq 1$, y para cada uno de tales n se tiene

$$(s \circ k)(n) = s_{k(n)}.$$

Una tal función compuesta se llama una *subsucesión* de s. De nuevo, por motivos de brevedad, utilizaremos a menudo, lo notación $\{s_{k(n)}\}$ o $\{s_{k_n}\}$ para designar la subsucesión de $\{s_n\}$ cuyo n-ésimo término es $s_{k(n)}$.

Ejemplo. Sea $s = \{1/n\}$ y sea k definida por $k(n) = 2^n$. Entonces $s \circ k = \{1/2^n\}$.

2.10 CONJUNTOS COORDINABLES (EQUIPOTENTES)

Definición 2.14. *Dos conjuntos A y B son coordinables, o equipotentes, y se escribe* $A \backsim B$ *si, y sólo si, existe una función uno a uno F cuyo dominio es el conjunto A y cuyo recorrido es el conjunto B.*

Se dice también que F establece una *correspondencia uno a uno* entre los conjuntos A y B. Es claro que cada conjunto A es coordinable consigo mismo (tomar como F la función «identidad» definida por $F(x) = x$ para todo x de A). Además, si $A \backsim B$ entonces $B \backsim A$, ya que si F es una función uno a uno que hace a A coordinable con B, entonces F^{-1} hará B coordinable con A. También, si $A \backsim B$ y si $B \backsim C$, entonces $A \backsim C$. (La demostración se deja al lector.)

2.11 CONJUNTOS FINITOS E INFINITOS

Se dice que un conjunto S es *finito* y que contiene n elementos si

$$S \sim \{1, 2, \ldots, n\}.$$

El entero n se llama *número cardinal* o simplemente *cardinal* de S. Es un ejercicio fácil demostrar que si $\{1, 2, \ldots, n\} \backsim \{1, 2, \ldots, m\}$ entonces $m = n$. Por

Material protegido por derechos de autor

lo tanto, el cardinal de un conjunto finito está bien definido. El conjunto vacío se considera también finito. Su cardinal se define por 0.

Los conjuntos que no son finitos se llaman *infinitos*. La diferencia principal entre ambos es que un conjunto infinito puede ser semejante a alguno de sus subconjuntos propios, mientras que un conjunto finito nunca podrá ser semejante a uno de sus subconjuntos propios. (Ver ejercicio 2.13.) Por ejemplo, el conjunto $\mathbf{Z}^+$ de todos los enteros positivos es semejante al subconjunto propio $\{2, 4, 8, 16, \ldots\}$ formado por las potencias de 2. La función uno a uno F que los hace semejantes está definida por $F(x) = 2^x$ para cada x de $\mathbf{Z}^+$.

2.12 CONJUNTOS NUMERABLES Y NO NUMERABLES

Un conjunto S se dice que es *infinito numerable* si es coordinable con el conjunto de todos los enteros positivos; esto es, si

$$S \sim \{1, 2, 3, \ldots\}.$$

En este caso existe una función f que establece una correspondencia uno a uno entre los enteros positivos y los elementos de S; por consiguiente, el conjunto S puede ser descrito como sigue:

$$S = \{f(1), f(2), f(3), \ldots\}.$$

A menudo se utilizan subíndices y $f(k)$ se designa por a_k (o por otra notación semejante) y se escribe, entonces, $S = \{a_1, a_2, a_3, \ldots\}$. Lo importante aquí es que la correspondencia nos permite utilizar los enteros positivos como «etiquetas» de los elementos de S. Un conjunto infinito numerable se dice que tiene cardinal $\aleph_0$ (léase: *álef subcero*).

Definición 2.15. *Un conjunto S es numerable si es o bien finito o bien infinito numerable. Un conjunto que no sea numerable se llama no numerable.*

Las palabras *numerable* y *no numerable* son sustituidas a veces por *contable* y *no contable*.

Teorema 2.16. *Todo subconjunto de un conjunto numerable es numerable.*

Demostración. Sea S un conjunto numerable dado y supongamos que $A \subseteq S$. Si A es finito, no hay nada que demostrar, por lo tanto podemos suponer que A es infinito (lo cual significa que S también lo es). Sea $s = \{s_n\}$ una sucesión infinita de términos todos distintos tal que

$$S = \{s_1, s_2, \ldots\}.$$

Material protegido por derechos de autor

Se define una función en el conjunto de los enteros positivos como sigue:

Sea $k(1)$ el menor entero positivo m tal que $s_m \in A$. Suponiendo que $k(1)$, $k(2)$, ..., $k(n-1)$ han sido definidas, sea $k(n)$ el menor entero positivo $m > k(n-1)$ tal que $s_m \in A$. Entonces k conserva el orden: $m > n$ implica $k(m) > k(n)$. Se forma entonces la función compuesta $s \circ k$. El dominio de $s \circ k$ es el conjunto de los enteros positivos y el recorrido de $s \circ k$ es A. Además, $s \circ k$ es uno a uno, ya que

$$s[k(n)] = s[k(m)],$$

implica

$$s_{k(n)} = s_{k(m)},$$

que significa $k(n) = k(m)$, y esto implica $n = m$. Esto prueba el teorema.

2.13 EL CONJUNTO DE LOS NÚMEROS REALES NO ES NUMERABLE

El siguiente teorema demuestra que existen conjuntos infinitos no numerables.

Teorema 2.17. *El conjunto de todos los números reales no es numerable.*

Demostración. Es suficiente demostrar que el conjunto de los x que satisfacen $0 < x < 1$ es no numerable. Si los números reales de este intervalo fuesen numerables, existiría una sucesión $s = \{s_n\}$ cuyos términos constituirían todo el intervalo. Probaremos que esto es imposible construyendo, dentro del intervalo, un número real que no sea término de esta sucesión. Una vez escritos los s_n como decimales infinitos:

$$s_n = 0.u_{n,1}u_{n,2}u_{n,3}\cdots,$$

donde cada $u_{n,i}$ es 0, 1, ..., o 9, consideramos el número real y cuya expresión decimal es

$$y = 0.v_1v_2v_3\cdots,$$

donde

$$v_n = \begin{cases} 1, & \text{si } u_{n,n} \neq 1, \\ 2, & \text{si } u_{n,n} = 1. \end{cases}$$

Entonces ningún término de la sucesión $\{s_n\}$ puede ser igual a y, ya que y difiere de s_1 en el primer decimal, de s_2 en el segundo decimal, ..., de s_n en el n-ésimo decimal. (Una situación como $s_n = 0{,}1999\ldots$ e $y = 0{,}2000\ldots$ no puede darse por la manera como han sido elegidas las v_n.) Como $0 < y < 1$, el teorema queda demostrado.

Material protegido por derechos de autor

Definición 2.21. *Si F es una colección arbitraria de conjuntos, la intersección de F se define como el conjunto cuyos elementos son los elementos que pertenecen a todos los conjuntos de F, y se designa por*

$$\bigcap_{A \in F} A.$$

La intersección de dos conjuntos A_1 y A_2 se desgina por $A_1 \cap A_2$ y consta de los elementos comunes a ambos conjuntos. Si A_1 y A_2 no tienen elementos comunes, entonces $A_1 \cap A_2$ es el conjunto vacío y A_1 y A_2 se llaman *disjuntos*. Si F es una colección finita (como más arriba), se escribe

$$\bigcap_{A \in F} A = \bigcap_{k=1}^{n} A_k = A_1 \cap A_2 \cap \cdots \cap A_n,$$

y si F es una colección numerable, se escribe

$$\bigcap_{A \in F} A = \bigcap_{k=1}^{\infty} A_k = A_1 \cap A_2 \cap \cdots$$

Si los conjuntos de la colección carecen de elementos comunes, su intersección es el conjunto vacío. Nuestras definiciones de reunión e intersección son, además, aplicables cuando F no es numerable. Por el modo como se han definido las reuniones y las intersecciones, las leyes conmutativas y asociativas se satisfacen automáticamente.

Definición 2.22. *El complemento de A relativamente a B, designado por $B - A$, se define como el conjunto*

$$B - A = \{x : x \in B, \text{ pero } x \notin A\}.$$

Nótese que $B - (B - A) = A$ siempre que $A \subseteq B$. Nótese también que $B - A = B$ si $B \cap A$ es vacío.

Las nociones de reunión, intersección, y complementario están ilustradas en la Fig. 2.4.

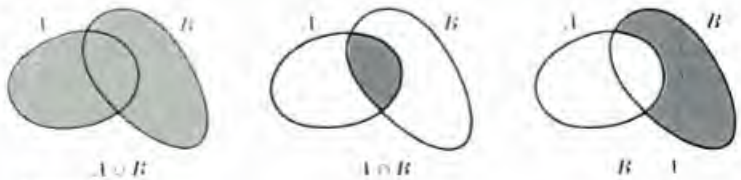


Figura 2.4

Material protegido por derechos de autor

Teorema 2.23. *Sea F una colección de conjuntos. Entonces para cada conjunto B, se tiene*

$$B - \bigcup_{A \in F} A = \bigcap_{A \in F} (B - A),$$

y

$$B - \bigcap_{A \in F} A = \bigcup_{A \in F} (B - A).$$

Demostración. Sea $S = \bigcup_{A \in F} A$, $T = \bigcap_{A \in F} (B - A)$. Si $x \in B - S$, entonces $x \in B$, pero $x \notin S$. Por lo tanto, no es cierto que x pertenezca a uno, por lo menos, de los A de F; por lo tanto x no pertenece a ninguno de los A de F. Luego, para cada A de F, $x \in B - A$. Pero esto implica que $x \in T$, luego $B - S \subseteq T$. Deshaciendo los pasos, se obtiene que $T \subseteq B - S$, y esto demuestra que $B - S = T$. Para demostrar la segunda afirmación, utilizar un argumento semejante.

2.15 COLECCIONES NUMERABLES DE CONJUNTOS NUMERABLES

Definición 2.24. *Si F es una colección de conjuntos tal que, cada dos conjuntos de F distintos, son disjuntos, se dice entonces que F es una colección de conjuntos disjuntos.*

Teorema 2.25. *Si F es una colección numerable de conjuntos disjuntos, tal como $F = \{A_1, A_2, \ldots\}$, en la que cada conjunto A_n es numerable, entonces la reunión $\bigcup_{k=1}^{\infty} A_k$ es también numerable.*

Demostración. Sea $A_n = \{a_{1,n}, a_{2,n}, a_{3,n} \ldots\}$, $n = 1, 2, \ldots$, y sea $S = \bigcup_{k=1}^{\infty} A_k$ Entonces todo elemento x de S está en uno de los conjuntos de F y, por lo tanto, $x = a_{m,n}$ para un cierto par de enteros (m, n). El par (m, n) está unívocamente determinado por x, ya que F es una colección de conjuntos disjuntos. Por lo tanto la función f definida por $f(x) = (m, n)$ si $x = a_{m,n}$, $x \in S$, tiene dominio S. El recorrido $f(S)$ es un subconjunto de $\mathbf{Z}^+ \times \mathbf{Z}^+$ (donde $\mathbf{Z}^+$ es el conjunto de los enteros positivos) y por lo tanto es numerable. Pero f es uno a uno y por consiguiente $S \backsim f(S)$, que equivale a decir que S es numerable.

Teorema 2.26. *Si $F = \{A_1, A_2, \ldots\}$ es una colección numerable de conjuntos, sea $G = \{B_1, B_2, \ldots\}$, donde $B_1 = A_1$ y, para $n > 1$,*

$$B_n = A_n - \bigcup_{k=1}^{n-1} A_k.$$

Material protegido por derechos de autor

Una relación que sea reflexiva, simétrica y transitiva se llama *relación de equivalencia.* Determinar cuál de estas propiedades satisface S, si S es el conjunto de todos los pares de números reales (x, y) tales que

a) $x \leq y$, b) $x < y$, c) $x < |y|$,
d) $x^2 + y^2 = 1$, e) $x^2 + y^2 < 0$, f) $x^2 + x = y^2 + y$.

2.3 Las siguientes funciones F y G están definidas para todo número real x por las ecuaciones dadas. En cada uno de los casos en que la función compuesta $G \circ F$ pueda definirse, dar el dominio de $G \circ F$ y una fórmula (o fórmulas) para $(G \circ F)(x)$.

a) $F(x) = 1 - x$, $G(x) = x^2 + 2x$.
b) $F(x) = x + 5$, $G(x) = |x|/x$, si $x \neq 0$, $G(0) = 1$.
c) $F(x) = \begin{cases} 2x, & \text{si } 0 \leq x \leq 1, \\ 1, & \text{en los casos restantes,} \end{cases}$ $G(x) = \begin{cases} x^2, & \text{si } 0 \leq x \leq 1, \\ 0, & \text{en los casos restantes.} \end{cases}$

Hallar $F(x)$ si $G(x)$ y $G[F(x)]$ vienen dados por:

d) $G(x) = x^3$, $G[F(x)] = x^3 - 3x^2 + 3x - 1$.
e) $G(x) = 3 + x + x^2$, $G[F(x)] = x^2 - 3x + 5$.

2.4 Dadas tres funciones F, G, H, ¿qué restricciones deben imponerse a sus dominios para que las cuatro funciones compuestas que siguen estén definidas?

$$G \circ F, \quad H \circ G, \quad H \circ (G \circ F), \quad (H \circ G) \circ F.$$

Suponiendo que sea posible definir $H \circ (G \circ F)$ y $(H \circ G) \circ F$, probar la ley asociativa

$$H \circ (G \circ F) = (H \circ G) \circ F.$$

2.5 Probar las siguientes identidades de la teoría de conjuntos para reuniones e intersecciones:

a) $A \cup (B \cup C) = (A \cup B) \cup C$, $A \cap (B \cap C) = (A \cap B) \cap C$.
b) $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$.
c) $(A \cup B) \cap (A \cup C) = A \cup (B \cap C)$.
d) $(A \cup B) \cap (B \cup C) \cap (C \cup A) = (A \cap B) \cup (A \cap C) \cup (B \cap C)$.
e) $A \cap (B - C) = (A \cap B) - (A \cap C)$.
f) $(A - C) \cap (B - C) = (A \cap B) - C$.
g) $(A - B) \cup B = A$ si, y sólo si, $B \subseteq A$.

2.6 Sea $f: S \to T$ una función. Si A y B son subconjuntos arbitrarios de S, probar que

$$f(A \cup B) = f(A) \cup f(B) \quad \text{y} \quad f(A \cap B) \subseteq f(A) \cap f(B).$$

Generalizar este resultado al caso de reuniones e intersecciones arbitrarias.

Material protegido por derechos de autor

2.18 Sea S la colección de todas las sucesiones cuyos términos sean los enteros 0 y 1. Probar que S es no numerable.

2.19 Probar que los siguientes conjuntos son numerables:

a) el conjunto de todos los círculos del plano complejo de radio racional y de centro de coordenadas racionales,

b) toda colección de intervalos disjuntos de longitud positiva.

2.20 Sea f una función a valores reales definida para todo x del intervalo $0 \leq x \leq 1$. Supongamos que existe un número positivo M que verifica la siguiente propiedad: para cada elección, con un número finito de puntos $x_1, x_2, \ldots, x_n$ del intervalo $0 \leq x \leq 1$, la suma

$$|f(x_1) + \cdots + f(x_n)| \leq M.$$

Sea S el conjunto de los x de $0 \leq x \leq 1$ para los que $f(x) \neq 0$. Probar que S es numerable.

2.21 Hallar la falacia de la siguiente «demostración» de que el conjunto de todos los intervalos de longitud positiva es numerable.

Sea $\{x_1, x_2, \ldots\}$ el conjunto numerable de todos los números racionales y sea I un intervalo de longitud positiva. Entonces I contiene una infinidad de puntos racionales x_n, pero de entre estos habrá uno que tendrá un *índice n mínimo*. Definimos una función F por medio de la ecuación $F(I) = n$, si x_n es el número racional de menor índice que pertenece al intervalo I. Esta función establece una correspondencia uno a uno entre el conjunto de todos los intervalos y un subconjunto de los enteros positivos. Por lo tanto el conjunto de todos los intervalos es numerable.

2.22 Sea S la colección de todos los subconjuntos de un conjunto dado T. Sea $f: S \to \mathbf{R}$ una función a valores reales definida en S. Se dice que la función f es *aditiva* si $f(A \cup B) = f(A) + f(B)$ siempre que A y B sean subconjuntos disjuntos de T. Si f es aditiva, probar que, para todo par de subconjuntos A y B, se tiene

$$f(A \cup B) = f(A) + f(B - A) \qquad \text{y} \qquad f(A \cup B) = f(A) + f(B) - f(A \cap B).$$

2.23 Aludimos al ejercicio 2.22. Suponemos que f es aditiva y suponemos además que las siguientes relaciones se verifican para dos subconjuntos particulares A y B de T:

$$f(A \cup B) = f(A') + f(B') - f(A')f(B')$$

$$f(A \cap B) = f(A)f(B), \qquad f(A) + f(B) \neq f(T),$$

donde $A' = T - A$, $B' = T - B$. Probar que estas relaciones determinan $f(T)$, y calcular el valor de $f(T)$.

REFERENCIAS SUGERIDAS PARA POSTERIORES ESTUDIOS

2.1 Boas, R. P., *A Primer of Real Funcions*. Carus Monograph No. 13. Wiley, New York, 1960.

2.2 Fraenkel, A., *Abstract Set Theory*, 3.ª ed. North-Holland, Amsterdam, 1965.
2.3 Gleason, A., *Fundamentals of Abstract Analysis*. Addison-Wesley, Reading, 1966.
2.4 Halmos, P. R., *Naive Set Theory*. Van Nostrand, New York, 1960. (Hay traducción francesa. Ed. Gauthier Villars. Hay traducción castellana).
2.5 Kamke, E., *Theory of Sets*. F. Bagemihl, translator. Dover, New York, 1950.
2.6 Kaplansky, I., *Set Theory and Metric Spaces*. Allyn and Bacon, Boston, 1972.
2.7 Rotman B., y Kneebone, G. T., *The Theory of Sets and Transfinite Numbers*. Elsevier, New York, 1968.

Material protegido por derechos de autor
Material protegido por derechos de autor

CAPÍTULO 3

Elementos de topología en conjuntos de puntos

3.1 INTRODUCCIÓN

La mayor parte del capítulo anterior trata de conjuntos «abstractos», esto es, conjuntos de objetos cualesquiera. En este capítulo consideraremos conjuntos de números reales, conjuntos de números complejos y, en general, conjuntos en espacios de más dimensiones.

En este estudio es conveniente y útil utilizar la terminología geométrica. Así, hablaremos de conjuntos de puntos de la recta real, conjuntos de puntos del plano, o conjuntos de puntos de espacios de mayor número de dimensiones. Más adelante estudiaremos funciones definidas en conjuntos de puntos, y es conveniente poseer un cierto conocimiento acerca de algunos tipos fundamentales de conjuntos de puntos, tales como conjuntos *abiertos*, conjuntos *cerrados* y conjuntos *compactos*, antes de abordar el estudio de las funciones. El estudio de estos conjuntos se llama *topología en conjuntos de puntos*.

3.2 EL ESPACIO EUCLÍDEO R^n

Un punto del espacio bidimensional es un par ordenado de números reales (x_1, x_2). Análogamente, un punto en un espacio tridimensional es una terna ordenada de números reales: (x_1, x_2, x_3). Es, pues, adecuado considerar una n-pla ordenada de números reales y referirnos a él como a un punto del espacio n-dimensional.

Definición 3.1. *Sea $n > 0$ un entero. Un conjunto ordenado de n números reales $(x_1, x_2, \ldots, x_n)$ se llama punto n dimensional o vector con n componentes. Los puntos o vectores se designarán por medio de una sola letra en negrita; por ejemplo,*

$$\mathbf{x} = (x_1, x_2, \ldots, x_n)$$

o

$$\mathbf{y} = (y_1, y_2, \ldots, y_n).$$

El número x_k se llama k-ésima coordenada del punto $\mathbf{x}$ o k-ésima componente del vector $\mathbf{x}$. El conjunto de todos los puntos n-dimensionales se llama espacio euclídeo n-dimensional o simplemente n-espacio, y se designa por $\mathbf{R}^n$.

Puede ocurrir que el lector se pregunte qué ventajas presenta trabajar en espacios de más de tres dimensiones. En realidad, el lenguaje de los n-espacios hace fácilmente comprensibles cuestiones más complicadas. El lector quizás esté lo suficientemente familiarizado con análisis vectorial de tres dimensiones, para percatarse de la ventaja que representa el poder escribir las ecuaciones de un movimiento que posee tres grados de libertad por medio de una sola ecuación vectorial en vez de tener que utilizar tres ecuaciones escalares. Existe una ventaja análoga cuando el sistema posee n grados de libertad.

Otra ventaja que se obtiene estudiando n-espacios para un n cualquiera es que, de una vez, se estudian todas las propiedades que son comunes a los 1-espacios, 2-espacios, 3-espacios, etc., esto es, propiedades independientes de la dimensión del espacio

Los espacios de más dimensiones se presentan como algo totalmente natural en campos tales como la Relatividad, y la Mecánica estadística y cuántica. Incluso espacios de infinitas dimensiones son corrientes en Mecánica cuántica.

Definiremos ahora las operaciones algebraicas con puntos n-dimensionales:

Definición 3.2. *Sean $\mathbf{x} = (x_1, \ldots, x_n)$ e $\mathbf{y} = (y_1, \ldots, y_n)$ elementos de $\mathbf{R}^n$. Definimos:*

a) *Igualdad:*

$$\mathbf{x} = \mathbf{y} \text{ si, y sólo si, } x_1 = y_1, \ldots, x_n = y_n.$$

b) *Suma:*

$$\mathbf{x} + \mathbf{y} = (x_1 + y_1, \ldots, x_n + y_n).$$

c) *Multpilicación por números reales (escalares):*

$$a\mathbf{x} = (ax_1, \ldots, ax_n) \qquad (a \text{ real}).$$

d) *Diferencia:*

$$\mathbf{x} - \mathbf{y} = \mathbf{x} + (-1)\mathbf{y}.$$

e) *Vector nulo u origen:*

$$\mathbf{0} = (0, \ldots, 0).$$

f) *Producto interior o producto escalar:*

$$\mathbf{x} \cdot \mathbf{y} = \sum_{k=1}^{n} x_k y_k.$$

Material protegido por derechos de autor
Material protegido por derechos de autor

3.3 BOLAS ABIERTAS Y CONJUNTOS ABIERTOS DE $\mathbf{R}^n$

Sea $\mathbf{a}$ un punto de $\mathbf{R}^n$ y sea r un número positivo dado. El conjunto de todos los puntos $\mathbf{x}$ de $\mathbf{R}^n$ tales que

$$\|\mathbf{x} - \mathbf{a}\| < r,$$

se denomina *n-bola* abierta de radio r y centro $\mathbf{a}$. Designamos este conjunto por $B(\mathbf{a})$ o por $B(\mathbf{a}; r)$.

La bola $B(\mathbf{a}; r)$ consta de todos los puntos cuya distancia a $\mathbf{a}$ es menor que r. En $\mathbf{R}^1$ este conjunto es un intervalo abierto con centro en $\mathbf{a}$. En $\mathbf{R}^2$ es un disco circular, y en $\mathbf{R}^3$ es una esfera sólida con centro en $\mathbf{a}$ y radio r.

3.5. Definición de punto interior. *Sea S un subconjunto de $\mathbf{R}^n$, y supongamos que $\mathbf{a} \in S$. Entonces $\mathbf{a}$ se denomina punto interior de S si existe una n-bola abierta con centro en $\mathbf{a}$, contenida en S.*

En otras palabras, cada uno de los puntos interiores $\mathbf{a}$ de S puede ser rodeado por una n-bola $B(\mathbf{a}) \subseteq S$. El conjunto de todos los puntos interiores de S se llama *interior* de S y se designa por int S. Cada conjunto que contiene una bola con centro en $\mathbf{a}$ se denomina *entorno* de $\mathbf{a}$.

3.6. Definición de conjunto abierto. *Un conjunto S de $\mathbf{R}^n$ es abierto si todos sus puntos son interiores. En otras palabras, S es abierto si, y sólo si, $S = int\ S$.* (Véase ejercicio 3.9.)

Ejemplos. En $\mathbf{R}^1$ el tipo más simple de conjunto abierto es un intervalo abierto. La unión de dos o más intervalos abiertos es también abierta. Un intervalo cerrado $[a, b]$ no es un conjunto abierto ya que sus extremos a y b no son puntos interiores del intervalo.

Ejemplos de conjuntos abiertos en el plano son: el interior de un disco; el producto cartesiano de dos intervalos abiertos unidimensionales. El lector debe tener en cuenta que un intervalo abierto de $\mathbf{R}^1$, considerado como subconjunto del plano, no es un conjunto abierto. De hecho, *ningún* subconjunto de $\mathbf{R}^1$ (salvo el conjunto vacío) puede ser abierto en $\mathbf{R}^2$, ya que tales conjuntos no pueden contener una 2-esfera.

En $\mathbf{R}^n$, tanto el conjunto vacío (¿Por qué?) como el mismo espacio $\mathbf{R}^n$, son conjuntos abiertos. El producto cartesiano

$$(a_1, b_1) \times \cdots \times (a_n, b_n)$$

de intervalos abiertos unidimensionales (a_1, b_1), ..., (a_n, b_n) es un conjunto abierto de $\mathbf{R}^n$ llamado *intervalo abierto n-dimensional*. Lo designaremos por $(\mathbf{a}, \mathbf{b})$, donde $\mathbf{a} = (a_1, \ldots, a_n)$ y $\mathbf{b} = (b_1, \ldots, b_n)$.

Material protegido por derechos de autor

En otras palabras, un intervalo componente de S no puede ser un subconjunto propio de ningún otro intervalo abierto contenido en S.

Teorema 3.10. *Cada punto de un conjunto abierto no vacío S pertenece a un intervalo componente de S y a uno solo.*

Demostración. Supongamos que $x \in S$. Entonces x está contenido en algún intervalo abierto I con $I \subseteq S$. Existen muchos de tales intervalos pero el «mayor» de ellos será el intervalo componente deseado. Dejamos para el lector la demostración de que este intervalo es $I_x = (a(x), b(x))$, donde

$$a(x) = \inf \{a : (a, x) \subseteq S\}, \qquad b(x) = \sup \{b : (x, b) \subseteq S\}.$$

Puede ocurrir que $a(x)$ sea $-\infty$ y puede ocurrir que $b(x)$ sea $+\infty$. Es claro que no existe ningún intervalo abierto J tal que $I_x \subseteq J \subseteq S$, luego I_x es un intervalo componente de S que contiene a x. Si J_x fuese otro intervalo componente de S conteniendo a x, entonces la reunión $I_x \cup J_x$ sería un intervalo contenido en S y que contendría a I_x y a J_x. Por lo tanto, por definición de intervalo componente, se tendría $I_x \cup J_x = I_x$ e $I_x \cup J_x = J_x$, luego $I_x = J_x$.

Teorema 3.11 (Teorema de representación para los conjuntos abiertos de la recta real). *Cada conjunto abierto no vacío S de $\mathbf{R}^1$ es la reunión de una colección numerable de intervalos componentes de S, disjuntos.*

Demostración. Si $x \in S$, sea I_x el intervalo componente de S que contiene a x. La reunión de todos los intervalos I_x es, evidentemente, S. Si dos de ellos, I_x e I_y, tienen un punto en común, entonces su reunión $I_x \cup I_y$ es un intervalo abierto contenido en S y que contiene a I_x y a I_y. Por lo tanto, $I_x \cup I_y = I_x$ e $I_x \cup I_y = I_y$, luego $I_x = I_y$. Por lo tanto los intervalos I_x forman una colección disjunta.

Resta demostrar que forman una colección numerable. A este fin, supongamos que $\{x_1, x_2, x_3, \ldots\}$ designa el conjunto numerable de los números racionales. En cada intervalo componente I_x habrá una infinidad de x_n, pero entre ellos uno sólo con el menor índice n. Definiremos entonces una aplicación F por medio de la ecuación $F(I_x) = n$, si x_n es el número racional de I_x con el menor índice n. Esta función F es uno a uno ya que $F(I_x) = F(I_y) = n$ implica que I_x e I_y tienen en común a x_n y ello implica que $I_x = I_y$. Por tanto F establece una correspondencia uno a uno entre los intervalos I_x y un cierto subconjunto de los números naturales. Esto termina la demostración.

NOTA. Esta representación de S es única. De hecho, si S es reunión de intervalos abiertos disjuntos, entonces estos intervalos serán necesariamente los intervalos componentes de S. Es una consecuencia inmediata del teorema 3.10.

Material protegido por derechos de autor

Si S es un intervalo abierto, entonces la representación contiene sólo un intervalo componente, a saber, S mismo. Por lo tanto, ningún intervalo abierto de $\mathbf{R}^1$ puede expresarse como reunión de dos conjuntos abiertos disjuntos no vacíos. Esta propiedad se designa también diciendo que un intervalo abierto es *conexo*. El concepto de conexión en conjuntos de $\mathbf{R}^n$ se estudia más ampliamente en la sección 4.16.

3.5 CONJUNTOS CERRADOS

3.12 Definición de conjunto cerrado. *Un conjunto S de $\mathbf{R}^n$ es cerrado si su complementario $\mathbf{R}^n - S$ es abierto.*

Ejemplos. Un intervalo cerrado $[a, b]$ de $\mathbf{R}^1$ es un conjunto cerrado. El producto cartesiano

$$[a_1, b_1] \times \cdots \times [a_n, b_n]$$

de n intervalos cerrados unidimensionales es un conjunto cerrado de $\mathbf{R}^n$, llamado *intervalo cerrado* $[a, b]$ *n-dimensional.*

El siguiente teorema, consecuencia inmediata de los teoremas 3.7 y 3.8, muestra cómo construir nuevos conjuntos cerrados a partir de conjuntos cerrados dados.

Teorema 3.13. *La reunión de una colección finita de conjuntos cerrados es cerrada, y la intersección de una colección arbitraria de conjuntos cerrados es cerrada.*

Otra relación entre conjuntos abiertos y cerrados es la que expresa el siguiente teorema.

Teorema 3.14. *Si A es abierto y B cerrado, entonces $A - B$ es abierto y $B - A$ es cerrado.*

Demostración. Basta observar que $A - B = A \cap (\mathbf{R}^n - B)$ es la intersección de dos conjuntos abiertos, y que $B - A = B \cap (\mathbf{R}^n - A)$ es la intersección de dos conjuntos cerrados.

3.6 PUNTOS ADHERENTES. PUNTOS DE ACUMULACIÓN

Los conjuntos cerrados pueden definirse por medio de los puntos adherentes y por medio de los puntos de acumulación.

Material protegido por derechos de autor

3.15 Definición de punto adherente. *Sea S un subconjunto de $\mathbf{R}^n$, y sea $\mathbf{x}$ un punto de $\mathbf{R}^n$, no necesariamente de S. Entonces se dice que $\mathbf{x}$ es adherente a S si toda n-bola $B(\mathbf{x})$ contiene un punto de S, por lo menos.*

Ejemplos

1. Si $\mathbf{x} \in S$, entonces $\mathbf{x}$ es adherente a S, ya que cada n-esfera $B(\mathbf{x})$ contiene a $\mathbf{x}$.
2. Si S es un subconjunto de $\mathbf{R}$ acotado superiormente, entonces el sup S es adherente a S.

Ciertos puntos son adherentes a S porque cada bola $B(\mathbf{x})$ contiene puntos de S *distintos* de $\mathbf{x}$. Estos puntos se llamarán puntos de acumulación.

3.16. Definición de punto de acumulación. *Si $S \subseteq \mathbf{R}^n$ y $\mathbf{x} \in \mathbf{R}^n$, entonces $\mathbf{x}$ se llama punto de acumulación de S si cada n-bola $B(\mathbf{x})$ contiene por lo menos un punto de S distinto de $\mathbf{x}$.*

En otras palabras, $\mathbf{x}$ es un punto de acumulación de S si, y sólo si, $\mathbf{x}$ es adherente a $S - \{\mathbf{x}\}$. Si $\mathbf{x} \in S$ pero $\mathbf{x}$ no es un punto de acumulación de S, se dice que $\mathbf{x}$ es un *punto aislado* de S.

Ejemplos

1. El conjunto de los números de la forma $1/n$, $n = 1, 2, 3, \ldots$ tienen al cero como punto de acumulación.
2. El conjunto de los números racionales tiene a cada racional como punto de acumulación.
3. Cada punto del intervalo cerrado $[a, b]$ es un punto de acumulación del conjunto de los números del intervalo abierto (a, b).

Teorema 3.17. *Si $\mathbf{x}$ es un punto de acumulación de S, entonces toda n-bola $B(\mathbf{x})$ contiene infinitos puntos de S.*

Demostración. Supongamos lo contrario; es decir, que exista una n-bola $B(\mathbf{x})$ que contenga sólo un número finito de puntos de S distintos de $\mathbf{x}$; llamémosles $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_m$. Si r es el menor de los números positivos

$$\|\mathbf{x} - \mathbf{a}_1\|, \quad \|\mathbf{x} - \mathbf{a}_2\|, \quad \ldots, \quad \|\mathbf{x} - \mathbf{a}_m\|,$$

entonces $B(\mathbf{x}; r/2)$ será una n-bola de centro $\mathbf{x}$ que no contendrá ningún punto de S distinto de $\mathbf{x}$. Contradicción.

Este teorema implica, en particular, que un conjunto que no posea una infinidad de puntos carece de puntos de acumulación. El recíproco, sin embargo, es falso. Por ejemplo, el conjunto de los enteros $\{1, 2, 3, \ldots\}$ es un conjunto infinito que carece de puntos de acumulación. En una sección posterior demostraremos que los conjuntos infinitos contenidos en una esfera po-

Material protegido por derechos de autor

seen siempre un punto de acumulación. Éste es un resultado importante conocido como teorema de Bolzano-Weierstrass.

3.7. CONJUNTOS CERRADOS Y PUNTOS ADHERENTES

Un conjunto cerrado se ha definido como el complementario de un conjunto abierto. El teorema siguiente presenta otra definición de conjunto cerrado.

Teorema 3.18. *Un conjunto S de $\mathbf{R}^n$ es cerrado si, y sólo si, contiene todos sus puntos adherentes.*

Demostración. Supongamos que S es cerrado y que $\mathbf{x}$ es adherente a S. Deseamos probar que $\mathbf{x} \in S$. Supongamos que $\mathbf{x} \notin S$ y llegaremos a una contradicción. Si $\mathbf{x} \notin S$, entonces $\mathbf{x} \in \mathbf{R}^n - S$ y, como que $\mathbf{R}^n - S$ es abierto, alguna n-bola $B(\mathbf{x})$ está contenida en $\mathbf{R}^n - S$. Entonces $B(\mathbf{x})$ no contiene puntos de S, en contradicción con el hecho de que $\mathbf{x}$ es adherente a S.

Para probar el recíproco, supongamos que S contiene todos sus puntos adherentes y demostraremos entonces que S es cerrado. Sea $\mathbf{x} \in \mathbf{R}^n - S$. Entonces $\mathbf{x} \notin S$, luego $\mathbf{x}$ no es adherente a S. Por lo tanto, existe una bola $B(\mathbf{x})$ que no corta a S, por consiguiente $B(\mathbf{x}) \subseteq \mathbf{R}^n - S$. Así pues, $\mathbf{R}^n - S$ es abierto y, entonces, S es cerrado.

3.19. Definición de adherencia. *El conjunto de todos los puntos adherentes de un conjunto dado S se llama adherencia de S y se designa por $\bar{S}$.*

Para todo conjunto se tiene que $S \subseteq \bar{S}$ ya que todo punto de S es adherente a S. El teorema 3.18 prueba que la inclusión opuesta $\bar{S} \subseteq S$ se verifica si, y sólo si, S es cerrado. Por lo tanto se tiene:

Teorema 3.20. *Un conjunto S es cerrado si, y sólo si, $S = \bar{S}$.*

3.21. Definición de conjunto derivado. *El conjunto de todos los puntos de acumulación de un conjunto S se llama conjunto derivado de S y se designa por S'.*

Es claro que, para todo conjunto S, $\bar{S} = S \cup S'$. Por lo tanto, el teorema 3.20 implica que S es cerrado si, y sólo si, $S' \subseteq S$. En otras palabras, se tiene:

Teorema 3.22. *Un conjunto S de $\mathbf{R}^n$ es cerrado si, y sólo si, contiene todos sus puntos de acumulación.*

Matrix protegido por derechos de autor

Como S está acotado, S podrá ser incluido en una cierta n-bola $B(\mathbf{0}; a)$, $a > 0$, y por lo tanto en el intervalo n-dimensional J_1 definido por las desigualdades

$$-a \leq x_k \leq a \qquad (k = 1, 2, \ldots, n).$$

Aquí J_1 designa el producto cartesiano

$$J_1 = I_1^{(1)} \times I_2^{(1)} \times \cdots \times I_n^{(1)};$$

esto es, el conjunto de puntos $(x_1, \ldots, x_n)$, donde $x_k \in I_k^{(1)}$ y donde cada $I_k^{(1)}$ es un intervalo unidimensional $-a \leq x_k \leq a$. Cada intervalo $I_k^{(1)}$ se puede subdividir en dos subintervalos $I_{k,1}^{(1)}$ e $I_{k,2}^{(1)}$ definidos por las desigualdades

$$I_{k,1}^{(1)}: -a \leq x_k \leq 0; \qquad I_{k,2}^{(1)}: 0 \leq x_k \leq a.$$

Ahora, consideramos todos los productos cartesianos de la forma

$$I_{1,k_1}^{(1)} \times I_{2,k_2}^{(1)} \times \cdots \times I_{n,k_n}^{(1)}, \tag{a}$$

donde cada $k_i = 1$ o 2. Hay, exactamente, 2^n productos de este tipo y, además, cada uno de ellos es un intervalo n-dimensional. La reunión de estos 2^n intervalos es el intervalo original J_1, que contiene a S; y por lo tanto, uno por lo menos de los 2^n intervalos (a) contiene una infinidad de puntos de S. Elijamos uno de los que verifican esta propiedad y llamémosle J_2; podrá expresarse también

$$J_2 = I_1^{(2)} \times I_2^{(2)} \times \cdots \times I_n^{(2)},$$

donde cada $I_k^{(2)}$ es uno de los subintervalos de $I_k^{(1)}$ de longitud a. Procedamos ahora con J_2 de la misma manera como hemos procedido con J_1, dividiendo cada intervalo $I_k^{(2)}$ en dos partes iguales y obteniendo un intervalo n-dimensional J_3 que contenga una infinidad de puntos de S. Si continuamos este proceso, obtendremos una colección numerable de intervalos n-dimensionales $J_1, J_2, J_3, \ldots$, tales que el intervalo m-ésimo J_m verifica la propiedad de contener un subconjunto infinito de S y se puede expresar en la forma

$$J_m = I_1^{(m)} \times I_2^{(m)} \times \cdots \times I_n^{(m)}, \qquad \text{donde } I_k^{(m)} \subseteq I_k^{(1)}.$$

Escribiendo

$$I_k^{(m)} = [a_k^{(m)}, b_k^{(m)}],$$

tenemos

$$b_k^{(m)} - a_k^{(m)} = \frac{a}{2^{m-2}} \qquad (k = 1, 2, \cdots, n).$$

Matrix protegido por derechos de autor

3.26. Definición de recubrimiento. *Una colección de conjuntos F se denomina recubrimiento de un conjunto dado S si $S \subseteq \bigcup_{A \in F} A$. Se dice también que la colección F recubre a S. Si F es una colección de conjuntos abiertos, entonces F se denomina recubrimiento abierto de S.*

Ejemplos

1. La colección de todos los intervalos de la forma $1/n < x < 2/n$, $(n = 2, 3, 4, \ldots)$, es un recubrimiento abierto del intervalo $0 < x < 1$. Es un ejemplo de recubrimiento numerable.
2. La recta real $\mathbf{R}^1$ está recubierta por la colección de todos los intervalos abiertos (a, b). Este recubrimiento es no numerable. Sin embargo, contiene un recubrimiento numerable de $\mathbf{R}^1$, a saber, todos los intervalos de la forma $(n, n+2)$, donde n recorre los valores enteros.
3. Sea $S = \{(x, y): x > 0, y > 0\}$. La colección F de todos los discos circulares con centros en (x, x) y radios x, $x > 0$, es un recubrimiento de S. Este recubrimiento no es numerable. Sin embargo contiene un recubrimiento numerable de S, a saber, todos los círculos para los que x es racional. (Ver ejercicio 3.18.)

El teorema del recubrimiento de Lindelöf establece que todo recubrimiento abierto de un conjunto S de $\mathbf{R}^n$ contiene una subcolección numerable que también recubre a S. La demostración utiliza el siguiente resultado preliminar:

Teorema 3.27. *Sea $G = \{A_1, A_2, \ldots\}$ la colección numerable de todas las n-bolas de radio racional y con centro en puntos de coordenadas racionales. Supongamos que $\mathbf{x} \in \mathbf{R}^n$ y sea S un conjunto abierto de $\mathbf{R}^n$ que contenga a $\mathbf{x}$. Entonces una, por lo menos, de las n-bolas de G contiene a $\mathbf{x}$ y está contenida en S. Esto es, se tiene*

$$\mathbf{x} \in A_k \subseteq S \quad \text{para algún } A_k \text{ de } G.$$

Demostración. La colección G es numerable en virtud del teorema 2.27. Si $\mathbf{x} \in \mathbf{R}^n$ y si S es un conjunto abierto que contiene a $\mathbf{x}$, entonces existe una n-bola $B(\mathbf{x}; r) \subseteq S$. Encontramos un punto $\mathbf{y}$ de S, de coordenadas racionales, «próximo» a $\mathbf{x}$ y, tomándolo como centro, hallaremos entonces un entorno en G interior a $B(\mathbf{x}; r)$ y que contenga a $\mathbf{x}$. Si

$$\mathbf{x} = (x_1, x_2, \ldots, x_n),$$

sea y_k un número racional tal que $|y_k - x_k| < r/(4n)$ para cada $k = 1, 2, \ldots, n$. Entonces

$$\|\mathbf{y} - \mathbf{x}\| \leq |y_1 - x_1| + \cdots + |y_n - x_n| < \frac{r}{4}.$$

Material protegido por derechos de autor

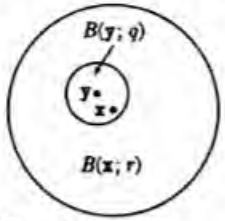


Figura 3.2

A continuación consideremos un número racional q tal que $r/4 < q < r/2$. Entonces $\mathbf{x} \in B(\mathbf{y}; q)$ y $B(\mathbf{y}; q) \subseteq B(\mathbf{x}; r) \subseteq S$. Pero $B(\mathbf{y}; q) \in G$ y por lo tanto el teorema queda demostrado. (Ver Fig. 3.2 para el caso $\mathbf{R}^2$.)

Teorema 3.28 (teorema del recubrimiento de Lindelöf). *Supongamos que $A \subseteq \mathbf{R}^n$ y que F es un recubrimiento abierto de A. Entonces existe una subcolección numerable de F que también recubre a A.*

Demostración. Sea $G = \{A_1, A_2, \ldots\}$ la colección numerable de todas las n-bolas de centros y radios racionales. Este conjunto G se utilizará para extraer de F una subcolección numerable que recubra a A.

Supongamos que $\mathbf{x} \in A$. Entonces existe un conjunto abierto S de F tal que $\mathbf{x} \in S$. Por el teorema 3.27, existe una n-bola A_k de G tal que $\mathbf{x} \in A_k \subseteq S$. Para cada S existe una infinidad numerable de tales A_k, pero sólo elegiremos una de entre todas, por ejemplo la de índice más pequeño; llamémosle $m = m(\mathbf{x})$. Tenemos entonces que $\mathbf{x} \in A_{m(\mathbf{x})} \subseteq S$. El conjunto de todas las n-bolas $A_{m(\mathbf{x})}$ obtenidas cuando $\mathbf{x}$ recorre todos los elementos de A es una colección numerable de conjuntos abiertos que recubre a A. A fin de lograr una subcolección numerable de F que recubra a A, hacemos corresponder a cada conjunto $A_{k(\mathbf{x})}$ uno de los conjuntos S de F que contenga a $A_{k(\mathbf{x})}$. Esto acaba la demostración.

3.11 TEOREMA DEL RECUBRIMIENTO DE HEINE-BOREL

El teorema del recubrimiento de Lindelöf establece que de un recubrimiento abierto de un conjunto arbitrario A de $\mathbf{R}^n$ se puede extraer un recubrimiento numerable. El teorema de Heine-Borel nos dice que si, además, A es cerrado y acotado, entonces podemos reducir el recubrimiento a un recubrimiento finito. La demostración requiere el teorema de encaje de Cantor.

Teorema 3.29 (Heine-Borel). *Sea F un recubrimiento abierto de un conjunto A de $\mathbf{R}^n$, cerrado y acotado. Entonces existe una subcolección finita de F que también recubre a A.*

Material protegido por derechos de autor

Demostración. Una subcolección numerable de F, llamémosla $\{I_1, I_2, \ldots\}$, recubre a A, en virtud del teorema 3.28. Consideremos, para $m \geq 1$, la reunión finita

$$S_m = \bigcup_{k=1}^{m} I_k.$$

Es abierta, ya que es reunión de conjuntos abiertos. Probaremos que, para algún valor de m, la reunión S_m recubre a A.

A este fin consideremos el complementario $\mathbf{R}^n - S_m$, que es cerrado. Definimos una colección numerable de conjuntos $\{Q_1, Q_2, \ldots\}$ de la siguiente manera: $Q_1 = A$, y para $m > 1$,

$$Q_m = A \cap (\mathbf{R}^n - S_m).$$

Esto es, Q_m consta de todos los puntos de A que están fuera de S_m. Si podemos probar que, para algún valor de m, el conjunto Q_m es vacío, habremos demostrado que, para este valor de m, ningún punto de A está fuera de S_m; en otras palabras, habremos probado que existe un S_m que recubre a A.

Observemos las siguientes propiedades de los conjuntos Q_m: Cada conjunto Q_m es cerrado, ya que es intersección del conjunto cerrado A y el conjunto cerrado $\mathbf{R}^n - S_m$. Los conjuntos Q_m son decrecientes, ya que los conjuntos S_m son crecientes; esto es, $Q_{m+1} \subseteq Q_m$. Los conjuntos Q_m, por ser subconjuntos de A, están acotados. Por lo tanto, si ninguno de los conjuntos Q_m es vacío, podemos aplicar el teorema de encaje de Cantor para concluir que la intersección $\bigcap_{k=1}^{\infty} Q_k$ tampoco es vacía. Ello implica la existencia de un cierto punto de A que pertenezca a todos los conjuntos Q_m, o, lo que es equivalente, la existencia de un punto de A que esté fuera de todos los conjuntos S_m. Pero esto es imposible, ya que $A \subseteq \bigcup_{k=1}^{\infty} S_k$. Por lo tanto algún Q_m debe ser vacío, y esto termina la demostración.

3.12 COMPACIDAD EN $\mathbf{R}^n$

Acabamos de ver que, si un conjunto S de $\mathbf{R}^n$ es cerrado y acotado, entonces todo recubrimiento abierto de S puede reducirse a un recubrimiento finito. Es natural preguntarse si podrían existir conjuntos distintos de los cerrados y acotados que verificasen también esta propiedad. Tales conjuntos se llamarán *compactos.*

3.30. Definición de conjunto compacto. *Un conjunto S de $\mathbf{R}^n$ se llama compacto si, y sólo si, cada recubrimiento abierto de S contiene un subrecubrimiento finito; esto es, una subcolección finita que también recubra a S.*

El teorema de Heine-Borel establece que todo conjunto de $\mathbf{R}^n$, cerrado y acotado, es compacto. Probaremos ahora el recíproco.

Teorema 3.31. *Sea S un subconjunto de $\mathbf{R}^n$. Entonces las tres afirmaciones siguientes son equivalentes:*

a) *S es compacto.*
b) *S es cerrado y acotado.*
c) *Todo subconjunto infinito de S tiene un punto de acumulación en S.*

Demostración. Como se indicó antes, (b) implica (a). Si probamos que (a) implica (b), que (b) implica (c) y que (c) implica (b), habremos establecido la equivalencia de las tres afirmaciones.

Supongamos que se verifica (a). Probaremos primero que S está acotado. Elijamos un punto $\mathbf{p}$ de S. La colección de n-bolas $B(\mathbf{p}; k)$, $k = 1, 2, \ldots$, es un recubrimiento abierto de S. Por compacidad, una subcolección finita también recubre a S y, por lo tanto, S está acotado.

A continuación probaremos que S es cerrado. Supongamos que no lo fuese. Existiría un punto $\mathbf{y}$ que sería un punto de acumulación de S y tal que $\mathbf{y} \notin S$. Si $\mathbf{x} \in S$, sea $r_{\mathbf{x}} = \|\mathbf{x} - \mathbf{y}\|/2$. Cada $r_{\mathbf{x}}$ es positivo ya que $\mathbf{y} \notin S$ y la colección $\{B(\mathbf{x}; r_{\mathbf{x}}): \mathbf{x} \in S\}$ es un recubrimiento abierto de S. Por compacidad, un número finito de estos entornos recubre a S, por lo que es

$$S \subseteq \bigcup_{k=1}^{p} B(\mathbf{x}_k; r_k).$$

Designemos por r al menor de los radios $r_1, r_2, \ldots, r_p$. Es fácil comprobar que la bola $B(\mathbf{y}; r)$ no tiene puntos en común con ninguna de las bolas $B(\mathbf{x}_k; r_k)$. De hecho, si $\mathbf{x} \in B(\mathbf{y}; r)$, entonces $\|\mathbf{x} - \mathbf{y}\| < r \leq r_k$, y por la desigualdad triangular tenemos que $\|\mathbf{y} - \mathbf{x}_k\| \leq \|\mathbf{y} - \mathbf{x}\| + \|\mathbf{x} - \mathbf{x}_k\|$, luego

$$\|\mathbf{x} - \mathbf{x}_k\| \geq \|\mathbf{y} - \mathbf{x}_k\| - \|\mathbf{x} - \mathbf{y}\| = 2r_k - \|\mathbf{x} - \mathbf{y}\| > r_k.$$

Por lo tanto, $\mathbf{x} \notin B(\mathbf{x}_k; r_k)$. Resulta, pues, que $B(\mathbf{y}; r) \cap S$ es vacío, en contradicción con el hecho de que $\mathbf{y}$ es un punto de acumulación de S. Esta contradicción prueba que S es cerrado y, por lo tanto, que (a) implica (b).

Supongamos que se verifica (b). En este caso la demostración de (c) es inmediata, ya que si T es un subconjunto infinito de S, entonces T está acotado (puesto que S lo está), y por el teorema de Bolzano-Weierstrass T posee un punto de acumulación que llamaremos $\mathbf{x}$. Ahora bien, $\mathbf{x}$ es también punto de acumulación de S y por lo tanto $\mathbf{x} \in S$, dado que S es cerrado. Por todo lo cual (b) implica (c).

Supongamos que se verifica (c). Probaremos (b). Si S no estuviese acotado, entonces para cada $m > 0$ existiría un punto $\mathbf{x}_m$ de S tal que $\|\mathbf{x}_m\| > m$. La colección $T = \{\mathbf{x}_1, \mathbf{x}_2, \ldots\}$ constituiría un subconjunto infinito de S y entonces,

Material protegido por derechos de autor
Material protegido por derechos de autor

El número no negativo $d(x, y)$ puede considerarse como la distancia entre x e y. En estos términos el significado intuitivo de las propiedades 1 a 4 es claro. La propiedad 4 se llama la *desigualdad triangular.*

Un espacio métrico se designa, a menudo, por medio de (M, d) a fin de recalcar que en la definición de espacio métrico tanto el conjunto M como la métrica d juegan su papel.

Ejemplos

1. $M = \mathbf{R}^n$; $d(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|$. Esta métrica se llama *métrica euclídea.* Cuando nos refiramos al espacio euclídeo $\mathbf{R}^n$, se sobreentenderá que su métrica es la euclídea si no se especifica alguna otra.
2. $M = \mathbf{C}$, el plano complejo; $d(z_1, z_2) = |z_1 - z_2|$. Como espacio métrico, $\mathbf{C}$ no se distingue del espacio euclídeo $\mathbf{R}^2$ puesto que consta de los mismos puntos y de la misma métrica.
3. M es un conjunto no vacío; $d(x, y) = 0$ si $x = y$, $d(x, y) = 1$ si $x \neq y$. Esta métrica se llama *métrica discreta*, y (M, d) se llama *espacio métrico discreto.*
4. Si (M, d) es un espacio métrico y si S es un subconjunto no vacío de M, entonces (S, d) es también un espacio métrico con la misma métrica o, mejor aún, con la métrica resultante de restringir d a $S \times S$. Se llama a menudo la *métrica relativa* inducida por d sobre S, y S es un *subespacio métrico* de M. Por ejemplo, los números racionales $\mathbf{Q}$ con la métrica $d(x, y) = |x - y|$ constituyen un subespacio métrico de $\mathbf{R}$.
5. $M = \mathbf{R}^2$; $d(\mathbf{x}, \mathbf{y}) = \sqrt{(x_1 - y_1)^2 + 4(x_2 - y_2)^2}$, donde $\mathbf{x} = (x_1, x_2)$ e $\mathbf{y} = (y_1, y_2)$. El espacio métrico (M, d) no es un subespacio del espacio euclídeo $\mathbf{R}^2$ ya que la métrica es distinta.
6. $M = \{(x_1, x_2) : x_1^2 + x_2^2 = 1\}$, la circunferencia unidad de $\mathbf{R}^2$; $d(\mathbf{x}, \mathbf{y})$ = la longitud del menor de los arcos que sobre la circunferencia unidad unen a los puntos $\mathbf{x}$ e $\mathbf{y}$.
7. $M = \{(x_1, x_2, x_3) : x_1^2 + x_2^2 + x_3^2 = 1\}$, es la superficie esférica unidad en $\mathbf{R}^3$; $d(\mathbf{x}, \mathbf{y})$ = la menor de las longitudes de los arcos que, sobre la superficie esférica unidad, une a los puntos $\mathbf{x}$ e $\mathbf{y}$.
8. $M = \mathbf{R}^n$; $d(\mathbf{x}, \mathbf{y}) = |x_1 - y_1| + \ldots + |x_n - y_n|$.
9. $M = \mathbf{R}^n$; $d(\mathbf{x}, \mathbf{y}) = \max\{|x_1 - y_1|, \ldots, |x_n - y_n|\}$.

3.14 TOPOLOGÍA EN ESPACIOS MÉTRICOS

Las nociones básicas de la topología en conjuntos de puntos se pueden extender a un espacio métrico arbitrario (M, d).

Si $a \in M$, la *bola* $B(a; r)$ de centro en a y radio $r > 0$ es el conjunto de todos los puntos x de M tales que

$$d(x, a) < r.$$

Algunas veces designaremos a esta bola por medio de $B_M(a; r)$ a fin de recalcar que sus puntos pertenecen a M. Si S es un subespacio métrico de M, la bola $B_S(a; r)$ es la intersección de S con la bola $B_M(a; r)$.

Material protegido por derechos de autor

Teorema 3.34. *Sea (S, d) un subespacio métrico de (M, d) y sea Y un subconjunto de S. Entonces Y es cerrado en S si, y sólo si, $Y = B \cap S$ para algún conjunto B cerrado en M.*

Demostración. Si $Y = B \cap S$, donde B es cerrado en M, entonces $B = M - A$ donde A es abierto en M, luego $Y = S \cap B = S \cap (M - A) = S - A$; de ahí que Y sea cerrado en S.

Recíprocamente, si Y es cerrado en S, sea $X = S - Y$. Entonces X es abierto en S, luego $X = A \cap S$, donde A es abierto en M y

$$Y = S - X = S - (A \cap S) = S - A = S \cap (M - A) = S \cap B,$$

donde $B = M - A$ es cerrado en M.

Si $S \subseteq M$, un punto x de M se llama *punto adherente* de S si cada bola $B_M(x; r)$ contiene un punto de S, por lo menos. Si x es adherente de $S - \{x\}$, entonces se dice que x es un *punto de acumulación* de S. La *clausura* $\bar{S}$ de S es el conjunto de todos los puntos adherentes de S, y el *conjunto derivado* S' es el conjunto de todos los puntos de acumulación de S. Entonces, $\bar{S} = S \cup S'$.

Los teoremas que se dan a continuación son válidos en cada espacio métrico (M, d) y se demuestran exactamente igual a como se demostraron en el caso del espacio euclídeo $\mathbf{R}^n$. En las demostraciones, la distancia euclídea $\|\mathbf{x} - \mathbf{y}\|$ deberá ser reemplazada por la métrica $d(x, y)$.

Teorema 3.35. (a) *La reunión de una colección arbitraria de conjuntos abiertos es abierta, y la intersección de una colección finita de conjuntos abiertos es abierta.*

b) *La reunión de una colección finita de conjuntos cerrados es cerrada, y la intersección de una colección arbitraria de conjuntos cerrados es cerrada.*

Teorema 3.36. *Si A es abierto y B es cerrado, entonces $A - B$ es abierto y $B - A$ es cerrado.*

Teorema 3.37. *Para cada uno de los subconjuntos S de M las siguientes afirmaciones son equivalentes:*

a) *S es cerrado en M.*
b) *S contiene a todos sus puntos adherentes.*
c) *S contiene a todos sus puntos de acumulación.*
d) *$S = \bar{S}$.*

Ejemplo. Sea $M = \mathbf{Q}$ el conjunto de números racionales con la métrica euclídea de $\mathbf{R}^1$. Sea S el conjunto de todos los números racionales en el intervalo abierto (a, b), donde tanto a como b son irracionales. Entonces S es un subconjunto cerrado de $\mathbf{Q}$.

Material protegido por derechos de autor

En nuestras demostraciones del teorema de Bolzano-Weierstrass, del teorema de encaje de Cantor, y de los teoremas del recubrimiento de Lindelöf y de Heine-Borel hemos utilizado no sólo las propiedades métricas del espacio euclídeo $\mathbf{R}^n$, sino también propiedades especiales de $\mathbf{R}^n$ que, en general, no son válidas en un espacio métrico arbitrario (M, d). Para poder extender estos teoremas a los espacios métricos habrá que imponer a M ciertas restricciones posteriores. Una de estas extensiones se esboza en el ejercicio 3.34.

La sección siguiente describe la compacidad en un espacio métrico arbitrario.

3.15 SUBCONJUNTOS COMPACTOS DE UN ESPACIO MÉTRICO

Sea (M, d) un espacio métrico y sea S un subconjunto de M. Una colección F de subconjuntos abiertos de M se llama *recubrimiento abierto* de S si $S \subseteq \bigcup_{A \in F} A$.

Un subconjunto S de M se llama *compacto* si cada recubrimiento abierto de S contiene un subrecubrimiento finito. S se dice que *está acotado* si $S \subseteq B(a; r)$ para algún $r > 0$ y algún a de M.

Teorema 3.38. *Sea S un subconjunto compacto de un espacio métrico M. Entonces:*

i) *S es cerrado y acotado.*

ii) *Cada subconjunto infinito de S posee un punto de acumulación en S.*

Demostración. Para demostrar (i) reharemos la demostración del teorema 3.31 y usaremos la parte de la argumentación que demostraba que (a) implica (b). El único cambio que debemos realizar consiste en reemplazar la distancia euclídea $\|\mathbf{x} - \mathbf{y}\|$ por la métrica $d(x, y)$ a lo largo de toda la demostración.

Para probar (ii) se procede por contradicción. Sea T un subconjunto infinito de S y supongamos que S no contiene ningún punto de acumulación de T. Entonces, para cada punto x de S, existirá una bola $B(x)$ que no contendrá ningún punto de T (si $x \notin T$) o un punto de T solamente (el mismo x, cuando $x \in T$). Cuando x recorre S, la reunión de estas bolas $B(x)$ es un recubrimiento abierto de S. Como S es compacto, una subcolección finita recubre a S y por lo tanto también recubre a T. Pero esto contradice el hecho de que T sea infinito y cada una de las bolas contenga a lo sumo un punto de T.

NOTA. En el espacio euclídeo $\mathbf{R}^n$, cada una de las propiedades (i) y (ii) es equivalente a la compacidad (teorema 3.31). En un espacio métrico general, la propiedad (ii) es equivalente a la compacidad (para una demostración, ver la referencia 3.4), pero en cambio la propiedad (i) no lo es. El ejercicio 3.42 nos suministra un ejemplo de un espacio métrico M en el que ciertos subconjuntos cerrados y acotados no son compactos.

Teorema 3.39. *Sea X un subconjunto cerrado de un espacio métrico compacto M. Entonces X es compacto.*

Demostración. Sea F un recubrimiento abierto de X, es decir $X \subseteq \bigcup_{A \in F} A$. Probaremos que un número finito de los conjuntos A recubre a X. Como que X es cerrado su complementario $M - X$ es abierto, luego $F \cup \{(M - X)\}$ es un recubrimiento abierto de M. Pero M es compacto, luego este recubrimiento contiene un subrecubrimiento finito que podemos suponer que incluye $M - X$. Por lo tanto

$$M \subseteq A_1 \cup \cdots \cup A_p \cup (M - X).$$

Este subrecubrimiento recubre también a X y, como que $M - X$ no contiene puntos de X, podemos suprimir el conjunto $M - X$ del subrecubrimiento y, a pesar de todo, sigue recubriendo a X. Entonces $X \subseteq A_1 \cup \ldots \cup A_p$, luego X es compacto.

3.16 FRONTERA DE UN CONJUNTO

Definición 3.40. *Sea S un subconjunto de un espacio métrico M. Un punto x de M se llama punto frontera de S si cada bola $B_M(x; r)$ contiene, por lo menos, un punto de S y, por lo menos, un punto de $M - S$. El conjunto de todos los puntos frontera de S se llama frontera de S y se designa por ∂S.*

El lector puede verificar fácilmente que

$$\partial S = \bar{S} \cap \overline{M - S}.$$

Esta fórmula prueba que ∂S es cerrado en M.

Ejemplo. En $\mathbf{R}^n$, la frontera de una bola $B(\mathbf{a}; r)$ es el conjunto de puntos $\mathbf{x}$ tal que $\|\mathbf{x} - \mathbf{a}\| = r$. En $\mathbf{R}^1$, la frontera del conjunto de los números racionales es todo en $\mathbf{R}^1$.

En los ejercicios y también en el capítulo 4 se desarrollan otras propiedades de los espacios métricos.

EJERCICIOS

Conjuntos abiertos y cerrados en $\mathbf{R}^1$ y $\mathbf{R}^2$

3.1 Probar que un intervalo abierto de $\mathbf{R}^1$ es un conjunto abierto y que un intervalo cerrado es un conjunto cerrado.

Material protegido por derechos de autor
Material protegido por derechos de autor

3.2 Determinar todos los puntos de acumulación de los siguientes conjuntos de $\mathbf{R}^1$ y decidir cuándo los conjuntos son abiertos o cerrados (o cuándo no lo son).

a) Todos los enteros.
b) El intervalo $(a, b]$
c) Todos los números de la forma $1/n$, $(n = 1, 2, 3, \ldots)$.
d) Todos los números racionales.
e) Todos los números de la forma $2^{-n} + 5^{-m}$, $(m, n = 1, 2, \ldots)$.
f) Todos los números de la forma $(-1)^n + (1/m)$, $(m, n = 1, 2, \ldots)$.
g) Todos los números de la forma $(1/n) + (1/m)$, $(m, n = 1, 2, \ldots)$.
h) Todos los números de la forma $(-1)^n/[1+(1/n)]$, $(n = 1, 2, \ldots)$.

3.3 Lo mismo que en el ejercicio 3.2 para los siguientes conjuntos de $\mathbf{R}^2$:

a) Todos los números complejos z tales que $|z| > 1$.
b) Todos los números complejos z tales que $|z| \geq 1$.
c) Todos los números complejos de la forma $(1/n) + (i/m), (m, n = 1, 2, \ldots)$.
d) Todos los puntos (x, y) tales que $x^2 - y^2 < 1$.
e) Todos los puntos (x, y) tales que $x > 0$.
f) Todos los puntos (x, y) tales que $x \geq 0$.

3.4 Probar que cada conjunto abierto no vacío S de $\mathbf{R}^1$ contiene números racionales e irracionales.

3.5 Probar que los únicos conjuntos de $\mathbf{R}^1$ que son a la vez abiertos y cerrados son el conjunto vacío y $\mathbf{R}^1$. ¿Existe una afirmación análoga para $\mathbf{R}^2$?

3.6 Probar que cada conjunto cerrado en $\mathbf{R}^1$ es la intersección de una colección numerable de conjuntos abiertos.

3.7 Probar que un conjunto cerrado y acotado, no vacío, S de $\mathbf{R}^1$ o bien es un intervalo cerrado o bien S podrá obtenerse a partir de un intervalo cerrado suprimiendo una colección disjunta numerable de intervalos abiertos cuyos extremos pertenecen a S.

Conjuntos abiertos y cerrados de $\mathbf{R}^n$

3.8 Probar que las n-bolas abiertas y los intervalos abiertos n-dimensionales son conjuntos abiertos en $\mathbf{R}^n$.

3.9 Probar que el interior de un conjunto de $\mathbf{R}^n$ es abierto en $\mathbf{R}^n$.

3.10 Si $S \subseteq \mathbf{R}^n$, probar que int S es la reunión de todos los subconjuntos abiertos de $\mathbf{R}^n$ que están contenidos en S. Esto se describe diciendo que int S es el mayor de los subconjuntos abiertos de S.

3.11 Si S y T son subconjuntos de $\mathbf{R}^n$, probar que

$$(\text{int } S) \cap (\text{int } T) = \text{int } (S \cap T), \qquad \text{y} \qquad (\text{int } S) \cup (\text{int } T) \subseteq \text{int } (S \cup T).$$

3.12 Sean S' y $\bar{S}$, respectivamente, el conjunto derivado y la clausura de un conjunto S de $\mathbf{R}^n$. Probar que:

a) S' es cerrado en $\mathbf{R}^n$; esto es, $(S')' \subseteq S'$.
b) Si $S \subseteq T$, entonces $S' \subseteq T'$.
c) $(S \cup T)' = S' \cup T'$.
d) $(\bar{S})' = S'$.

Material protegido por derechos de autor

3.23 Supongamos que $S \subseteq \mathbf{R}^n$. Un punto $\mathbf{x}$ de $\mathbf{R}^n$ es un *punto de condensación* de S si toda n-bola $B(\mathbf{x})$ tiene la propiedad que $B(\mathbf{x}) \cap S$ no es numerable. Probar que si S no es numerable, entonces existe un punto $\mathbf{x}$ en S de modo que $\mathbf{x}$ es un punto de condensación de S.

3.24 Supongamos que $S \subseteq \mathbf{R}^n$ y que S no es numerable. Sea T el conjunto de puntos de condensación de S. Probar que:

a) $S - T$ es numerable,
b) $S \cap T$ no es numerable,
c) T es un conjunto cerrado,
d) T no posee puntos aislados.

Nótese que el ejercicio 3.23 es un caso especial de (b).

3.25 Un conjunto de $\mathbf{R}^n$ se llama *perfecto* si $S = S'$, esto es, si S es un conjunto cerrado que carece de puntos aislados. Probar que, si F es un conjunto cerrado no numerable de $\mathbf{R}^n$, puede expresarse en la forma $F = A \cup B$, donde A es perfecto y B es numerable *(teorema de Cantor-Bendixon)*.

Indicación. Utilizar el ejercicio 3.24.

Espacios métricos

3.26 Probar que, en todo espacio métrico (M, d), tanto el conjunto vacío $\emptyset$ como el espacio entero M son, a la vez, abiertos y cerrados.

3.27 Considerar en $\mathbf{R}^n$ las dos métricas siguientes:

$$d_1(\mathbf{x}, \mathbf{y}) = \max_{1 \leq i \leq n} |x_i - y_i|, \qquad d_2(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{n} |x_i - y_i|.$$

En cada uno de los espacios métricos siguientes, probar que la bola $B(\mathbf{a}; r)$ tiene la apariencia geométrica que se indica:

a) en $(\mathbf{R}^2, d_1)$, un cuadrado de lados paralelos a los ejes de coordenadas.
b) en $(\mathbf{R}^2, d_2)$, un cuadrado cuyas diagonales son paralelas a los ejes.
c) Un cubo en $(\mathbf{R}^3, d_1)$.
d) Un octaedro en $(\mathbf{R}^3, d_2)$.

3.28 Sean d_1 y d_2 las métricas definidas en el ejercicio 3.27 y sea $\|\mathbf{x} - \mathbf{y}\|$ la métrica euclídea usual. Comprobar que se verifican las siguientes desigualdades, cualesquiera que sean los puntos $\mathbf{x}$ e $\mathbf{y}$ de $\mathbf{R}^n$:

$$d_1(\mathbf{x}, \mathbf{y}) \leq \|\mathbf{x} - \mathbf{y}\| \leq d_2(\mathbf{x}, \mathbf{y}) \qquad \text{y} \qquad d_2(\mathbf{x}, \mathbf{y}) \leq \sqrt{n}\|\mathbf{x} - \mathbf{y}\| \leq n d_1(\mathbf{x}, \mathbf{y}).$$

3.29 Si (M, d) es un espacio métrico, se define

$$d'(x, y) = \frac{d(x, y)}{1 + d(x, y)}.$$

Probar que d' también es una métrica para M. Obsérvese que $0 \leq d'(x, y) < 1$ para todo x, y de M.

3.30 Probar que cada subconjunto finito de un espacio métrico es cerrado.

3.31 En un espacio métrico (M, d), el conjunto $\bar{B}(a; r) = \{x : d(x, a) \leq r\}$ se llama *bola cerrada* de radio $r > 0$ y centro en el punto a de M.

Material protegido por derechos de autor

3.46 a) int $(\bigcap_{i=1}^{n} A_i) = \bigcap_{i=1}^{n}$ (int A_i), donde cada $A_i \subseteq M$.
b) int $(\bigcap_{A \in F} A) \subseteq \bigcap_{A \in F}$ (int A), si F es una colección infinita de subconjuntos de M.
c) Dar un ejemplo en el que la igualdad de (b) no se verifique.

3.47 a) $\bigcup_{A \in F}$ (int A) $\subseteq$ int $(\bigcup_{A \in F} A)$.
b) Dar un ejemplo de una colección finita F que no satisfaga la igualdad en (a).

3.48 a) int $(\partial A) = \emptyset$ si A es abierto o si A es cerrado en M.
b) Dar un ejemplo para el que int $(\partial A) = M$.

3.49 Si int $A =$ int $B = \emptyset$ y si A es cerrado en M, entonces int $(A \cup B) = \emptyset$.

3.50 Dar un ejemplo en el que int $A =$ int $B = \emptyset$, pero para el que int $(A \cup B) = M$.

3.51 $\partial A = \bar{A} \cap \overline{M - A}$ y $\partial A = \partial(M - A)$.

3.52 Si $\bar{A} \cap \bar{B} = \emptyset$, entonces $\partial(A \cup B) = \partial A \cup \partial B$.

REFERENCIAS SUGERIDAS PARA POSTERIORES ESTUDIOS

3.1 Boas, R. P., *A Primer of Real Functions.* Carus Monograph No. 13. Wiley, New York, 1960.
3.2 Gleason, A., *Fundamentals of Abstract Analysis,* Addison-Wesley, Reading, 1966.
3.3 Kaplansky, I., *Set Theory and Metric Spaces.* Allyn and Bacon, Boston, 1972.
3.4 Simmons, G .F., *Introduction to Topology and Modern Analysis.* McGraw-Hill, New York, 1963.

CAPÍTULO 4

Límites y continuidad

4.1 INTRODUCCIÓN

Suponemos al lector ya familiarizado con el concepto de límites tal como es introducido en el Cálculo elemental donde es corriente presentar varios tipos de límites. Por ejemplo, el *límite de una sucesión* de números reales $\{x_n\}$, que simbolizamos cuando escribimos

$$\lim_{n \to \infty} x_n = A,$$

significa que para cada número $\varepsilon > 0$ existe un entero N tal que

$$|x_n - A| < \varepsilon \quad \text{siempre que } n \geq N.$$

Este límite pretende transmitir la idea intuitiva de que x_n puede estar suficientemente próximo a A en el supuesto de que n sea suficientemente grande. También se da el *límite de una función,* indicado por medio de la notación

$$\lim_{x \to p} f(x) = A,$$

que significa que para cada $\varepsilon > 0$ existe otro número $\delta > 0$ tal que

$$|f(x) - A| < \varepsilon \text{ siempre que } 0 < |x - p| < \delta.$$

Esta definición expresa la idea de que $f(x)$ puede conseguirse tan próxima a A como queramos, siempre que x se tome lo suficientemente próximo a p.

Las aplicaciones del Cálculo a los problemas geométricos y físicos del espacio tridimensional y a las funciones de varias variables nos obligan a extender estos conceptos a $\mathbf{R}^n$. Es tan necesario como fácil dar un paso más e introducir límites en el marco más general de los espacios métricos. Esto simplifica la teoría puesto que elimina restricciones innecesarias y al mismo tiempo cubre casi todos los aspectos necesarios del Análisis.

Primeramente discutiremos los límites de las sucesiones de puntos de un espacio métrico y después discutiremos los límites de funciones y el concepto de continuidad.

Material protegido por derechos de autor
Material protegido por derechos de autor

4.2 SUCESIONES CONVERGENTES EN UN ESPACIO MÉTRICO

Definición 4.1. *Una sucesión $\{x_n\}$ de puntos de un espacio métrico (S, d) es convergente si existe un punto p de S que satisfaga la siguiente propiedad:*

Para todo $\varepsilon > 0$ existe un entero N tal que

$$d(x_n, p) < \varepsilon \quad \text{siempre que } n \geq N.$$

Diremos también que $\{x_n\}$ converge hacia p y escribiremos $x_n \to p$ cuando $n \to \infty$, o simplemente $x_n \to p$. Si no existe un tal número p de S, se dice que la sucesión $\{x_n\}$ es divergente.

NOTA. La definición de convergencia implica que

$$x_n \to p \quad \text{si, y sólo si,} \quad d(x_n, p) \to 0.$$

La convergencia de la sucesión $\{d(x_n, p)\}$ hacia 0 se realiza en el espacio euclídeo $\mathbf{R}^1$.

Ejemplos

1. En un espacio euclídeo $\mathbf{R}^1$, una sucesión $\{x_n\}$ se llama *creciente* si $x_n \leq x_{n+1}$ para todo n. Si una sucesión creciente está acotada superiormente (esto es, si $x_n \leq M$ para un $M > 0$ y para todo n), entonces $\{x_n\}$ converge hacia el supremo de su recorrido, sup $\{x_1, x_2, \ldots\}$. Análogamente, $\{x_n\}$ se llama *decreciente* si $x_{n+1} \leq x_n$ para todo n. Cada sucesión decreciente acotada inferiormente converge hacia el ínfimo de su recorrido. Por ejemplo, $\{1/n\}$ converge hacia 0.
2. Si $\{a_n\}$ y $\{b_n\}$ son sucesiones reales que convergen hacia 0, entonces $\{a_n + b_n\}$ también converge hacia 0. Si $0 \leq c_n \leq a_n$ para todo n y si $\{a_n\}$ converge hacia 0, entonces $\{c_n\}$ también converge hacia 0. Estas propiedades elementales de las sucesiones de $\mathbf{R}^1$ pueden ser útiles para simplificar algunas de las demostraciones concernientes a límites de un espacio métrico general.
3. En el plano complejo $\mathbf{C}$, sea $z_n = 1 + n^{-3} + (2 - 1/n)i$. Entonces $\{z_n\}$ converge hacia $1 + 2i$ puesto que

$$d(z_n, 1 + 2i)^2 = |z_n - (1 + 2i)|^2 = \frac{1}{n^4} + \frac{1}{n^2} \to 0 \text{ cuando } n \to \infty,$$

luego $d(z_n, 1+2i) \to 0$.

Teorema 4.2. *Una sucesión $\{x_n\}$ de un espacio métrico (S, d) puede converger hacia un punto de S, a lo sumo.*

Demostración. Supongamos que $x_n \to p$ y que $x_n \to q$. Probaremos que $p = q$. En virtud de la desigualdad triangular se tiene

Material protegido por derechos de autor

$$0 \leq d(p, q) \leq d(p, x_n) + d(x_n, q).$$

Como $d(p, x_n) \to 0$ y $d(x_n, q) \to 0$ se tiene que $d(p, q) = 0$, luego $p = q$.

Si una sucesión $\{x_n\}$ converge, el único punto hacia el que converge se llama *límite* de la sucesión y se designa por medio de lim x_n o por medio de $\lim_{n \to \infty} x_n$.

Ejemplo. En el espacio euclídeo $\mathbf{R}^1$ tenemos que $\lim_{n \to \infty} 1/n = 0$. La misma sucesión en el subespacio métrico $T = (0, 1]$ *no* converge puesto que el único candidato para el límite es 0 y $0 \notin T$. Este ejemplo muestra que la convergencia o divergencia de una sucesión depende tanto del espacio elegido como de la métrica.

Teorema 4.3. *En un espacio métrico (S, d), suponemos que $x_n \to p$ y que $T = \{x_1, x_2, \ldots\}$ es el recorrido de $\{x_n\}$. Entonces:*

a) *T está acotado.*

b) *p es un punto de adherencia de T.*

Demostración. a) Sea N el entero que corresponde a $\varepsilon = 1$ en la definición de convergencia. Entonces todo x_n con $n \geq N$ está en la esfera $B(p; 1)$, luego cada punto de T está en la esfera $B(p; r)$, donde

$$r = 1 + \max \{d(p, x_1), \ldots, d(p, x_{N-1})\}.$$

Por lo tanto, T está acotado.

b) Como cada esfera $B(p; \varepsilon)$ contiene un punto de T, p es adherente a T.

NOTA. Si T es *infinito*, cada bola $B(p; \varepsilon)$ contendrá una infinidad de puntos de T, luego p será punto de acumulación de T.

El teorema siguiente prueba el recíproco de la parte (b).

Teorema 4.4. *Dado un espacio métrico (S, d) y un subconjunto $T \subseteq S$, si p es un punto de S adherente de T, entonces existe una sucesión $\{x_n\}$ de puntos de T que converge hacia p.*

Demostración. Para cada entero $n \geq 1$ existe un punto x_n de T tal que $d(p, x_n) \leq 1/n$. Por lo tanto $d(p, x_n) \to 0$, luego $x_n \to p$.

Teorema 4.5. *En un espacio métrico (S, d) una sucesión $\{x_n\}$ converge hacia p si, y sólo si, cada subsucesión $\{x_{k(n)}\}$ converge hacia p.*

Demostración. Supongamos que $x_n \to p$ y consideremos una subsucesión $\{x_{k(n)}\}$. Para cada $\varepsilon > 0$ existe un N tal que $n \geq N$ implica $d(x_n, p) < \varepsilon$. Como

Material protegido por derechos de autor

$\{x_{k(n)}\}$ es una subsucesión, existe un entero M tal que $k(n) \geq N$ para $n \geq M$. Por lo tanto, $n \geq M$ implica $d(x_{k(n)}, p) < \varepsilon$, que prueba que $x_{k(n)} \to p$. El recíproco se verifica trivialmente, ya que $\{x_n\}$ es una subsucesión de sí misma.

4.3 SUCESIONES DE CAUCHY

Si una sucesión $\{x_n\}$ converge hacia el límite p, sus términos avanzados deben aproximarse a p y por lo tanto aproximarse entre sí. Esta propiedad está enunciada más formalmente en el siguiente teorema.

Teorema 4.6. *Supongamos que $\{x_n\}$ converge en un espacio métrico (S, d). Entonces para cada $\varepsilon > 0$ existe un entero N tal que*

$$d(x_n, x_m) < \varepsilon \quad \text{siempre que } n \geq N \text{ y } m \geq N.$$

Demostración. Sea $p = \lim x_n$. Dado $\varepsilon > 0$, sea N tal que $d(x_n, p) < \varepsilon/2$ siempre que $n \geq N$. Entonces $d(x_m, p) < \varepsilon/2$ si $m \geq N$. Si tanto n como m son mayores o iguales que N por la desigualdad triangular tenemos

$$d(x_n, x_m) \leq d(x_n, p) + d(p, x_m) < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon.$$

4.7. Definición de la sucesión de Cauchy. *Una sucesión $\{x_n\}$ de un espacio métrico se llama sucesión de Cauchy si satisface la siguiente condición (llamada la condición de Cauchy):*

Para cada $\varepsilon > 0$ existe un entero N tal que

$$d(x_n, x_m) < \varepsilon \quad \text{siempre que } n \geq N \text{ y } m \geq N.$$

El teorema 4.6 establece que toda sucesión convergente es una sucesión de Cauchy. El recíproco, en general, es falso en un espacio métrico general. Por ejemplo, la sucesión $\{1/n\}$ es una sucesión de Cauchy en el subespacio euclídeo $T = (0, 1]$ de $\mathbf{R}^1$, pero en cambio dicha sucesión no converge en T. Sin embargo, el recíproco del teorema 4.6 es cierto en cada espacio euclídeo $\mathbf{R}^k$.

Teorema 4.8. *En el espacio euclídeo $\mathbf{R}^k$ toda sucesión de Cauchy es convergente.*

Demostración. Sea $\{\mathbf{x}_n\}$ una sucesión de Cauchy de $\mathbf{R}^k$ y sea $T = \{\mathbf{x}_1, \mathbf{x}_2, \ldots\}$ el recorrido de la sucesión. Si T es finito, entonces todos los términos de $\{\mathbf{x}_n\}$ excepto un número finito son iguales y por lo tanto $\{\mathbf{x}_n\}$ converge hacia este valor común.

Material protegido por derechos de autor

4.4 ESPACIOS MÉTRICOS COMPLETOS

Definición 4.9. *Un espcio métrico (S, d) se llama completo si toda sucesión de Cauchy de S converge en S. Un subconjunto T de S se llama completo si el subespacio métrico (T, d) es completo.*

Ejemplo 1. Cada uno de los espacios euclídeos $\mathbf{R}^k$ es completo (teorema 4.8). En particular, $\mathbf{R}^1$ es completo, pero el subespacio $T = (0, 1]$ no es completo.
Ejemplo 2. El espacio $\mathbf{R}^n$ con la métrica $d(\mathbf{x}, \mathbf{y}) = \max_{1 \leq i \leq n} |x_i - y_i|$ es completo.

El teorema que damos a continuación relaciona la completitud con la compacidad.

Teorema 4.10. *En todo espacio métrico (S, d), cada subconjunto compacto T es completo.*

Demostración. Sea $\{x_n\}$ una sucesión de Cauchy en T y sea $A = \{x_1, x_2, \ldots\}$ el recorrido de $\{x_n\}$. Si A es finito, entonces $\{x_n\}$ converge hacia uno de los puntos de A, luego $\{x_n\}$ converge en T.

Si A es infinito, el teorema 3.38 nos asegura que A admite un punto de acumulación p en T puesto que T es compacto. Probaremos ahora que $x_n \to p$. Dado $\varepsilon > 0$, elijamos N tal que $n \geq N$ y $m \geq N$ implique $d(x_n, x_m) < \varepsilon/2$. La bola $B(p; \varepsilon/2)$ contiene un punto x_m con $m \geq N$. Por lo tanto si $n \geq N$ la desigualdad triangular nos conduce a

$$d(x_n, p) \leq d(x_n, x_m) + d(x_m, p) < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon,$$

luego $x_n \to p$. De lo cual se deduce que toda sucesión de Cauchy en T admite límite en T, luego T es completo.

4.5 LÍMITE DE UNA FUNCIÓN

En esta sección consideraremos dos espacios métricos (S, d_S) y (T, d_T), donde d_S y d_T designan las métricas respectivas. Sea A un subconjunto de S y sea $f: A \to T$ una función de A en T.

Definición 4.11. *Si p es un punto de acumulación de A y si $b \in T$, la notación*

$$\lim_{x \to p} f(x) = b, \tag{1}$$

Material protegido por derechos de autor

si, y sólo si,

$$\lim_{n\to\infty} f(x_n) = b, \tag{3}$$

para toda sucesión $\{x_n\}$ *de puntos de* $A - \{p\}$ *que sea convergente hacia* p.

Demostración. Si se verifica (2), entonces para cada $\varepsilon > 0$ existe un $\delta > 0$ tal que

$$d_T(f(x), b) < \varepsilon \quad \text{siempre que } x \in A \text{ y } 0 < d_S(x, p) < \delta. \tag{4}$$

Como p es adherente a $A - \{p\}$, por el teorema 4.4, existe una sucesión $\{x_n\}$ en $A - \{p\}$ convergente hacia p. Para el δ que interviene en (4), existe un entero N tal que $n \geq N$ implica $d_S(x_n, p) < \delta$. Entonces (4) implica que $d_T(f(x_n), b) < \varepsilon$ para $n \geq N$, y por lo tanto $\{f(x_n)\}$ converge hacia b. Así pues, (2) implica (3).

Para probar el recíproco supondremos que se verifica (3) y que (2) es falso, llegando a una contradicción. Si (2) es falso, entonces para algún $\varepsilon > 0$ y todo $\delta > 0$ existe un punto x de A (donde x puede depender de δ) tal que

$$0 < d_S(x, p) < \delta \quad \text{pero} \quad d_T(f(x), b) \geq \varepsilon. \tag{5}$$

Tomando $\delta = 1/n$, $n = 1, 2, \ldots$, esto significa que existe una sucesión $\{x_n\}$ de puntos de $A - \{p\}$ tal que

$$0 < d_S(x_n, p) < 1/n \quad \text{pero} \quad d_T(f(x_n), b) \geq \varepsilon.$$

Es evidente entonces que hemos obtenido una sucesión $\{x_n\}$ que converge hacia p pero en cambio la sucesión $\{f(x_n)\}$ no converge hacia b, lo cual contradice a (3).

NOTA. Los teoremas 4.12 y 4.2 prueban que una función no puede tener dos límites diferentes cuando $x \to p$.

4.6 LÍMITES DE FUNCIONES CON VALORES COMPLEJOS

Sea (S, d) un espacio métrico, sea A un subconjunto de S, y consideremos dos funciones f y g definidas sobre A y con valores complejos,

$$f: A \to \mathbf{C}, \qquad g: A \to \mathbf{C}$$

La suma $f + g$ se define como la función cuyo valor en cada punto x de A es el número complejo $f(x) + g(x)$. La *diferencia* $f - g$, el *producto* $f \cdot g$, y el *cociente* f/g se definen análogamente. Es sabido que el cociente sólo está definido en aquellos puntos x en los que $g(x) \neq 0$.

Material protegido por derechos de autor

Las reglas usuales para el cálculo con límites vienen dadas en el teorema que sigue.

Teorema 4.13. *Sean* f *y* g *dos funciones con valores complejos definidas en un subconjunto* A *de un espacio métrico* (S, d). *Sea* p *un punto de acumulación de* A, *y supongamos que*

$$\lim_{x\to p} f(x) = a, \qquad \lim_{x\to p} g(x) = b.$$

Entonces tendremos también:

a) $\lim_{x\to p} [f(x) \pm g(x)] = a \pm b$,

b) $\lim_{x\to p} f(x)g(x) = ab$,

c) $\lim_{x\to p} f(x)/g(x) = a/b$ *si* $b \neq 0$.

Demostración. Probaremos (b), dejando las otras partes como ejercicio. Dado ε con $0 < \varepsilon < 1$, sea ε' un segundo número que satisfaga $0 < \varepsilon' < 1$, que dependerá de ε en la forma que precisaremos más adelante. Existe un $\delta > 0$ tal que si $x \in A$ y $d(x, p) < \delta$, entonces

$$|f(x) - a| < \varepsilon' \qquad \text{y} \qquad |g(x) - b| < \varepsilon'.$$

Entonces

$$|f(x)| = |a + (f(x) - a)| < |a| + \varepsilon' < |a| + 1.$$

Haciendo $f(x)g(x) - ab = f(x)g(x) - bf(x) + bf(x) - ab$, tenemos

$$\begin{aligned} |f(x)g(x) - ab| &\leq |f(x)|\,|g(x) - b| + |b|\,|f(x) - a| \\ &< (|a| + 1)\varepsilon' + |b|\varepsilon' = \varepsilon'(|a| + |b| + 1). \end{aligned}$$

Si elegimos $\varepsilon' = \varepsilon/(|a| + |b| + 1)$, veremos que $|f(x)g(x) - ab| < \varepsilon$ siempre que $x \in A$ y $d(x, p) < \delta$, y esto demuestra (b).

4.7 LÍMITES DE FUNCIONES CON VALORES VECTORIALES

De nuevo, sea (S, d) un espacio métrico y sea A un subconjunto de S. Consideremos dos funciones $\mathbf{f}$ y $\mathbf{g}$, definidas sobre A, con valores vectoriales tomados en $\mathbf{R}^k$,

$$\mathbf{f}: A \to \mathbf{R}^k, \qquad \mathbf{g}: A \to \mathbf{R}^k.$$

El cociente de funciones con valores vectoriales no está definido (si $k \geq 2$), pero es posible definir la *suma* $\mathbf{f} + \mathbf{g}$, el *producto* $\lambda\mathbf{f}$ (si λ es real) y el *producto escalar* $\mathbf{f} \cdot \mathbf{g}$ por medio de las fórmulas

Material protegido por derechos de autor

$$(\mathbf{f} + \mathbf{g})(x) = \mathbf{f}(x) + \mathbf{g}(x), \qquad (\lambda\mathbf{f})(x) = \lambda\mathbf{f}(x), \qquad (\mathbf{f}\cdot\mathbf{g})(x) = \mathbf{f}(x)\cdot\mathbf{g}(x)$$

para todo x de A. Se tienen entonces las siguientes reglas para calcular los límites de funciones con valores vectoriales.

Teorema 4.14. *Sea p un punto de acumulación de A y supongamos que*

$$\lim_{x\to p} \mathbf{f}(x) = \mathbf{a}, \qquad \lim_{x\to p} \mathbf{g}(x) = \mathbf{b}.$$

Entonces se tiene también:

a) $\lim_{x\to p} [\mathbf{f}(x) + \mathbf{g}(x)] = \mathbf{a} + \mathbf{b}$,

b) $\lim_{x\to p} \lambda\mathbf{f}(x) = \lambda\mathbf{a}$ *para cada escalar* λ,

c) $\lim_{x\to p} \mathbf{f}(x)\cdot\mathbf{g}(x) = \mathbf{a}\cdot\mathbf{b}$,

d) $\lim_{x\to p} \|\mathbf{f}(x)\| = \|\mathbf{a}\|$.

Demostración. Probaremos sólo las partes (c) y (d). Para probar (c) hagamos

$$\mathbf{f}(x)\cdot\mathbf{g}(x) - \mathbf{a}\cdot\mathbf{b} = [\mathbf{f}(x) - \mathbf{a}]\cdot[\mathbf{g}(x) - \mathbf{b}] + \mathbf{a}\cdot[\mathbf{g}(x) - \mathbf{b}] + \mathbf{b}\cdot[\mathbf{f}(x) - \mathbf{a}].$$

La desigualdad de Cauchy-Schwarz y la desigualdad triangular nos dan

$$\begin{aligned} 0 &\le |\mathbf{f}(x)\cdot\mathbf{g}(x) - \mathbf{a}\cdot\mathbf{b}| \\ &\le \|\mathbf{f}(x) - \mathbf{a}\|\,\|\mathbf{g}(x) - \mathbf{b}\| + \|\mathbf{a}\|\,\|\mathbf{g}(x) - \mathbf{b}\| + \|\mathbf{b}\|\,\|\mathbf{f}(x) - \mathbf{a}\|. \end{aligned}$$

Cada uno de los términos de la derecha tiende a 0 cuando $x\to p$, luego $\mathbf{f}(x)\cdot\mathbf{g}(x)\to\mathbf{a}\cdot\mathbf{b}$. Esto prueba (c). Para demostrar (d), obsérvese que

$$|\,\|\mathbf{f}(x)\| - \|\mathbf{a}\|\,| \le \|\mathbf{f}(x) - \mathbf{a}\|.$$

NOTA. Sean $f_1, \ldots, f_n$ n funciones con valores reales definidas sobre A, y sea $\mathbf{f}: A\to\mathbf{R}^n$ la función con valores vectoriales definida por

$$\mathbf{f}(x) = (f_1(x), f_2(x), \ldots, f_n(x)) \quad \text{si } x\in A.$$

Entonces $f_1, \ldots, f_n$ se llaman componentes de $\mathbf{f}$, y se escribe $\mathbf{f} = (f_1, \ldots, f_n)$ para indicar dicha relación.

Si $\mathbf{a} = (a_1, \ldots, a_n)$, entonces para cada $r = 1, 2, \ldots, n$ tenemos

$$|f_r(x) - a_r| \le \|\mathbf{f}(x) - \mathbf{a}\| \le \sum_{r=1}^{n} |f_r(x) - a_r|.$$

Estas desigualdades demuestran que $\lim_{x\to p} \mathbf{f}(x) = \mathbf{a}$ si, y sólo si, $\lim_{x\to p} f_r(x) = a_r$ para cada r.

Material protegido por derechos de autor

4.8 FUNCIONES CONTINUAS

La definición de continuidad que se da en Cálculo elemental puede extenderse a funciones definidas de un espacio métrico a otro.

Definición 4.15. *Sean (S, d_S) y (T, d_T) espacios métricos y sea $f: S\to T$ una función de S en T. La función f se llama continua en un punto p de S si para cada $\varepsilon > 0$ existe un $\delta > 0$ tal que*

$$d_T(f(x), f(p)) < \varepsilon \quad \text{siempre que } d_S(x, p) < \delta.$$

Si f es continua en todos los puntos del subconjunto A de S, se dice que f es continua en A.

Esta definición refleja la idea intuitiva de que puntos cercanos a p se aplican, por medio de f, en puntos cercanos a $f(p)$. Puede expresarse, también, en términos de bolas: una función f es continua en p si, y sólo si, para cada $\varepsilon > 0$, existe un $\delta > 0$ tal que

$$f(B_S(p;\ \delta)) \subseteq B_T(f(p);\ \varepsilon).$$

Aquí $B_S(p;\ \delta)$ designa una bola de S; su imagen, por medio de f, debe estar contenida en la bola $B_T(f(p);\ \varepsilon)$ de T. (Ver Fig. 4.2.)

Si p es un punto de acumulación de S, la definición de continuidad implica que

$$\lim_{x\to p} f(x) = f(p).$$

Si p es un punto aislado de S (un punto de S que no es de acumulación de S), entonces toda f definida en p será continua en p ya que para δ suficientemente pequeño existe un único x que satisface $d_S(x, p) < \delta$, a saber $x = p$, y $d_T(f(p), f(p)) = 0$.

Teorema 4.16. *Sea $f: S\to T$ una función de un espacio métrico (S, d_S) en otro espacio métrico (T, d_T), y supongamos que $p\in S$. Entonces f es continua*

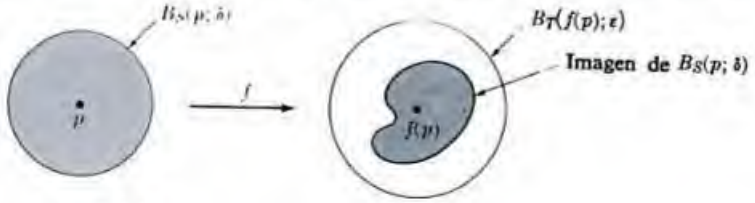


Figura 4.2

Material protegido por derechos de autor

4.10 FUNCIONES COMPLEJAS Y FUNCIONES VECTORIALES, CONTINUAS

Teorema 4.18. *Sean f y g dos funciones con valores complejos, continuas en un punto p de un espacio métrico (S, d). Entonces $f+g$, $f-g$ y $f \cdot g$ son todas ellas continuas en p. El cociente f/g también es continuo en p si $g(p) \neq 0$.*

Demostración. El resultado es trivial si p es un punto aislado de S. Si p es un punto de acumulación de S, el resultado se sigue del teorema 4.13.

Existe, además, un teorema análogo para las funciones con valores vectoriales, que se demuestra de la misma manera, utilizando el teorema 4.14.

Teorema 4.19. *Sean $\mathbf{f}$ y $\mathbf{g}$ dos funciones continuas en un punto p de un espacio métrico (S, d), y supongamos que $\mathbf{f}$ y $\mathbf{g}$ toman sus valores en $\mathbf{R}^n$. Entonces cada una de las siguientes funciones es continua en p: la suma $\mathbf{f}+\mathbf{g}$, el producto $\lambda\mathbf{f}$ para cada número real λ, el producto escalar $\mathbf{f} \cdot \mathbf{g}$, y la norma $\|\mathbf{f}\|$.*

Teorema 4.20. *Sean $f_1, \ldots, f_m$, n funciones reales definidas sobre un subconjunto A de un espacio métrico (S, d_S) y sea $\mathbf{f} = (f_1, \ldots, f_m)$. Entonces $\mathbf{f}$ es continua en un punto p de A si, y sólo si, cada una de las funciones $f_1, \ldots, f_m$ es continua en p.*

Demostración. Si p es un punto aislado de A no hay nada que demostrar. Si p es un punto de acumulación, obsérvese que $\mathbf{f}(x) \to \mathbf{f}(p)$ cuando $x \to p$ si, y sólo si, $f_k(x) \to f_k(p)$ para cada $k = 1, 2, \ldots, n$.

4.11 EJEMPLOS DE FUNCIONES CONTINUAS

Sea $S = \mathbf{C}$, el plano complejo. Es un ejercicio trivial demostrar que las siguientes funciones con valores complejos son continuas en $\mathbf{C}$:

a) las funciones constantes, definidas por $f(z) = c$ para todo z de $\mathbf{C}$;
b) la función identidad, definida por $f(z) = z$ para todo z de $\mathbf{C}$.

Aplicando repetidamente el teorema 4.18 se establece la continuidad de los polinomios

$$f(z) = a_0 + a_1 z + a_2 z^2 + \cdots + a_n z^n,$$

donde los a_i son números complejos.

Si S es un subconjunto de $\mathbf{C}$ en el que el polinomio f no se anula, entonces $1/f$ es continua en S. Por lo tanto una función racional g/f, donde g y f son

Material protegido por derechos de autor

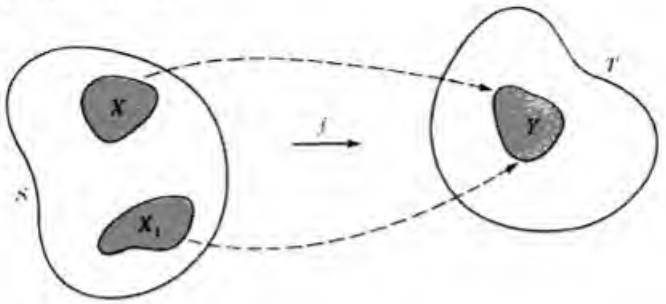


Figura 4.3

Nótese que las afirmaciones del teorema 4.22 pueden expresarse también de la siguiente manera:

$$f[f^{-1}(Y)] \subseteq Y, \qquad X \subseteq f^{-1}[f(X)].$$

Obsérvese asimismo que $f^{-1}(A \cup B) = f^{-1}(A) \cup f^{-1}(B)$ para todos los subconjuntos A y B de T.

Teorema 4.23. *Sea $f: S \to T$ una función de un espacio métrico (S, d_S) en otro (T, d_T). Entonces f es continua en S si, y sólo si, para cada conjunto abierto Y de T, la antiimagen $f^{-1}(Y)$ es abierta en S.*

Demostración. Sea f continua sobre S, sea Y un abierto de T, y sea p un punto de $f^{-1}(Y)$. Probaremos que p es interior a $f^{-1}(Y)$. Sea $y = f(p)$. Como que Y es abierto entonces tenemos que $B_T(y;\ \varepsilon) \subseteq Y$ para un cierto $\varepsilon > 0$. Como que f es continua en p, existe un $\delta > 0$ tal que $f(B_S(p;\ \delta)) \subseteq B_T(y;\ \varepsilon)$. Por lo tanto,

$$B_S(p;\delta) \subseteq f^{-1}[f(B_S(p;\delta))] \subseteq f^{-1}[B_T(y;\varepsilon)] \subseteq f^{-1}(Y),$$

luego p es un punto interior a $f^{-1}(Y)$.

Recíprocamente, supongamos que $f^{-1}(Y)$ es abierto en S para todo subconjunto abierto Y de T. Elijamos p en S y sea $y = f(p)$. Probaremos que f es continua en p. Para cada valor $\varepsilon > 0$, la bola $B_T(y;\ \varepsilon)$ es abierta en T, luego $f^{-1}(B_T(y;\ \varepsilon))$ es abierto en S. Ahora bien, si $p \in f^{-1}(B_T(y;\ \varepsilon))$ entonces existe un $\delta > 0$ tal que $B_S(p;\ \delta) \subseteq f^{-1}(B_T(y;\ \varepsilon))$. Por consiguiente, $f(B_S(p;\ \delta)) \subseteq B_T(y;\ \varepsilon)$, luego f es continua en p.

Teorema 4.24. *Sea $f: S \to T$ una función de un espacio métrico (S, d_S) en otro (T, d_T). Entonces f es continua en S si, y sólo si, para cada conjunto cerrado Y de T, la antiimagen $f^{-1}(Y)$ es cerrada en S.*

Material protegido por derechos de autor

Demostración. Si Y es cerrado en T, entonces $T - Y$ es abierto en T y

$$f^{-1}(T - Y) = S - f^{-1}(Y).$$

Aplíquese ahora el teorema 4.23.

Ejemplos. La imagen de un conjunto abierto por medio de una aplicación continua no es necesariamente abierta. Un contraejemplo muy simple es el de las funciones constantes que aplican todo S en un único punto de $\mathbf{R}^1$. Análogamente, la imagen de un conjunto cerrado en una aplicación continua no tiene por qué ser cerrada. Por ejemplo, la función real $f(x) = \operatorname{arctg} x$ aplica $\mathbf{R}^1$ en el intervalo $(-\pi/2, \pi/2)$.

4.13 FUNCIONES CONTINUAS SOBRE CONJUNTOS COMPACTOS

El teorema que sigue prueba que la imagen de un conjunto compacto en una función continua es un conjunto compacto. Es otra de las propiedades globales de las funciones continuas.

Teorema 4.25. *Sea $f: S \to T$ una función de un espacio métrico (S, d_S) en otro (T, d_T). Si f es continua en un subconjunto compacto X de S, entonces la imagen $f(X)$ es un subconjunto compacto de T; en particular, $f(X)$ es un conjunto cerrado y acotado de T.*

Demostración. Sea F un recubrimiento abierto de $f(X)$, es decir $f(X) \subseteq \bigcup_{A \in F} A$. Probaremos que un número finito de conjuntos A recubre a $f(X)$. Como f es continua sobre el subespacio métrico (X, d_S) podemos aplicar el teorema 4.23 para concluir que cada uno de los conjuntos $f^{-1}(A)$ es abierto en (X, d_S). Los conjuntos $f^{-1}(A)$ forman un recubrimiento abierto de X y, como X es compacto, un número finito de ellos recubre a X; sea $X \subseteq f^{-1}(A_1) \cup \ldots \cup f^{-1}(A_p)$. Entonces

$$\begin{aligned} f(X) \subseteq f[f^{-1}(A_1) \cup \cdots \cup f^{-1}(A_p)] &= f[f^{-1}(A_1)] \cup \cdots \cup f[f^{-1}(A_p)] \\ &\subseteq A_1 \cup \cdots \cup A_p, \end{aligned}$$

luego $f(X)$ es compacto. Como corolario del teorema 3.38, vemos que $f(X)$ es cerrado y acotado.

Definición 4.26. *Una función $\mathbf{f}: S \to \mathbf{R}^k$ está acotada en S si existe un número positivo M tal que $\|\mathbf{f}(x)\| \leq M$ para todo x de S.*

Como $\mathbf{f}$ está acotada en S si y sólo si $f(S)$ es un subconjunto acotado de $\mathbf{R}^k$, tendremos el siguiente corolario del teorema 4.25.

Material protegido por derechos de autor

Teorema 4.27. *Sea $f: S \to \mathbf{R}^k$ una función de un espacio métrico S en el espacio euclídeo $\mathbf{R}^k$. Si $\mathbf{f}$ es continua en un subconjunto X, compacto en S, entonces $\mathbf{f}$ está acotada en X.*

Este teorema posee importantes implicaciones en el caso de funciones reales. Si f es una función real, acotada sobre X, entonces $f(X)$ es un subconjunto de $\mathbf{R}$, acotado, luego posee supremo, $\sup f(X)$, e ínfimo, $\inf f(X)$. Además,

$$\inf f(X) \leq f(X) \leq \sup f(X) \quad \text{para cada } x \text{ de } X.$$

El próximo teorema prueba que una función continua f alcanza efectivamente los valores $\sup f(X)$ e $\inf f(X)$ si X es compacto.

Teorema 4.28. *Sea $f: S \to \mathbf{R}$ una función real de un espacio métrico S en el espacio euclídeo $\mathbf{R}$. Supongamos que f es continua en un subconjunto X, compacto en S. Entonces existen puntos p y q de X tales que*

$$f(p) = \inf f(X)$$

y

$$f(q) = \sup f(X).$$

NOTA. Como que $f(p) \leq f(x) \leq f(q)$ para todo x de X, los números $f(p)$ y $f(q)$ se llaman, respectivamente, los valores *mínimo* y *máximo* globales o absolutos de f en X.

Demostración. El teorema 4.25 demuestra que $f(X)$ es un subconjunto cerrado y acotado de $\mathbf{R}$. Sea $m = \inf f(X)$. Entonces m es adherente a $f(X)$ y, por ser $f(X)$ cerrado, $m \in f(X)$. Por lo tanto, $m = f(p)$ para un cierto p de X. Análogamente, $f(q) = \sup f(X)$ para un cierto q de X.

Teorema 4.29. *Sea $f: S \to T$ una función de un espacio métrico (S, d_S) en otro (T, d_T). Supongamos que f es uno a uno sobre S, de modo que la función inversa f^{-1} existe. Si S es compacto y si f es continua en S, entonces f^{-1} es continua en $f(S)$.*

Demostración. Por el teorema 4.23 (aplicado a f^{-1}) bastará probar solamente que para cada conjunto cerrado X de S la imagen $f(X)$ es cerrada en T. (Obsérvese que $f(X)$ es la imagen inversa de X por medio de f^{-1}.) Como X es cerrado y S es compacto, X es compacto (por el teorema 3.39), luego $f(X)$ es compacto (por el teorema 4.25) y por lo tanto $f(X)$ es cerrado (por el teorema 3.38). Esto acaba la demostración.

Material protegido por derechos de autor

Ejemplo. Este ejemplo muestra que la compacidad de S es esencial en el teorema 4.29. Sea $S = [0, 1)$ con la métrica usual de $\mathbf{R}^1$ y consideremos la función f con valores complejos definida por

$$f(x) = e^{2\pi i x} \quad \text{para } 0 \leq x < 1.$$

Ésta es una aplicación continua uno a uno del semi-intervalo abierto $[0, 1)$ en el círculo unidad $|z| = 1$ del plano complejo. Sin embargo, f^{-1} no es continua en el punto $f(0)$. Por ejemplo, si $x_n = 1 - 1/n$, la sucesión $\{f(x_n)\}$ converge hacia $f(0)$ pero $\{x_n\}$ no converge en S.

4.14 APLICACIONES TOPOLÓGICAS (HOMEOMORFISMOS)

Definición 4.30. *Sea $f: S \to T$ una función de un espacio métrico (S, d_S) en otro (T, d_T). Supongamos también que f es uno a uno en S, de modo que la función inversa f^{-1} existe. Si f es continua sobre S y f^{-1} es continua sobre $f(S)$, entonces diremos que f es una aplicación topológica o un homeomorfismo, y los espacios métricos (S, d_S) y $(f(S), d_T)$ se llaman homeomorfos.*

Si f es un homeomorfismo, entonces f^{-1} también lo es. El teorema 4.23 prueba que un homeomorfismo aplica subconjuntos abiertos de S en subconjuntos abiertos de $f(S)$. Aplica asimismo subconjuntos cerrados de S en subconjuntos cerrados de $f(S)$.

Una propiedad de un conjunto que permanezca invariante frente a las distintas aplicaciones topológicas, se llama una *propiedad topológica.* Así pues, las propiedades de ser abierto, cerrado, compacto son propiedades topológicas.

Un ejemplo importante de homeomorfismo lo constituyen las *isometrías.* Se trata de una aplicación $f: S \to T$ que es uno a uno sobre S y que conserva la métrica; es decir,

$$d_T(f(x), f(y)) = d_S(x, y)$$

para todos los puntos x e y de S. Si existe una isometría de (S, d_S) en $(f(S), d_T)$, los dos espacios métricos se llaman *isométricos.*

Las aplicaciones topológicas son particularmente interesantes en la teoría de curvas. Por ejemplo, un *arco simple* es la imagen topológica de un intervalo, y una *curva cerrada simple* es la imagen topológica de una circunferencia.

4.15 TEOREMA DE BOLZANO

Esta sección está dedicada al famoso teorema de Bolzano que concierne a una propiedad global de las funciones reales continuas en intervalos compactos $[a, b]$ de $\mathbf{R}$. Si la gráfica de f está por encima del eje de las x en a y por debajo del eje de las x en b, el teorema de Bolzano afirma que la gráfica debe

Material protegido por derechos de autor

$f(\alpha) \neq f(\beta)$. Entonces f toma todos los valores comprendidos entre $f(\alpha)$ y $f(\beta)$ en el intervalo (α, β).

Demostración. Sea k un número comprendido entre $f(\alpha)$ y $f(\beta)$ y apliquemos el teorema de Bolzano a la función g definida en $[\alpha, \beta]$ por medio de la ecuación $g(x) = f(x) - k$.

El teorema del valor intermedio, juntamente con el teorema 4.29, implican que la imagen cotinua de un intervalo compacto S por medio de una función real es otro intervalo compacto; a saber

$$[\inf f(S), \sup f(S)].$$

(Si f es constante en S, entonces el intervalo sería degenerado.) La sección siguiente extiende esta propiedad a escenarios más amplios de espacios métricos.

4.16 CONEXIÓN

En esta sección se describe el concepto de conexión y su relación con la continuidad.

Definición 4.34. *Un espacio métrico S se dice que es no conexo si $S = A \cup B$, donde A y B son conjuntos abiertos disjuntos de S, no vacíos. Diremos que S es conexo si no es no conexo.*

NOTA. Un subconjunto X de un espacio métrico S se llama conexo si, considerado como subespacio métrico de S, es un espacio métrico conexo.

Ejemplos

1. El espacio métrico $S = \mathbf{R} - \{0\}$ con la métrica usual euclídea es no conexo, ya que es unión de dos conjuntos abiertos disjuntos no vacíos, los números positivos y los números reales negativos.
2. Cada intervalo abierto de $\mathbf{R}$ es conexo. Esto se demostró en la sección 3.4 como consecuencia del teorema 3.11.
3. El conjunto $\mathbf{Q}$ de los números racionales, considerado como subespacio del espacio euclídeo $\mathbf{R}^1$, es no conexo. En efecto, $\mathbf{Q} = A \cup B$, donde A consta de todos los números racionales $< \sqrt{2}$ y B de todos los números racionales $> \sqrt{2}$. Análogamente, cada bola de $\mathbf{Q}$ es no conexa.
4. Cada espacio métrico S contiene subconjuntos no vacíos conexos. En efecto, para cada p de S el conjunto $\{p\}$ es conexo.

Para relacionar la conexión con la continuidad introduciremos el concepto de función a dos valores.

Material protegido por derechos de autor

Ejemplo. Como un intervalo X de $\mathbf{R}^1$ es conexo, cada imagen continua $f(X)$ es conexa. Si f toma valores reales, la imagen $f(X)$ es otro intervalo. Si f toma valores en $\mathbf{R}^n$, la imagen $f(X)$ se llama *curva* de $\mathbf{R}^n$. Entonces, cada curva de $\mathbf{R}^n$ es conexa.

Como corolario al Teorema 4.37, tenemos el teorema siguiente que es una extensión del de Bolzano.

Teorema 4.38 (Teorema del valor intermedio para funciones reales continuas). *Sea f una función real continua definida en un subconjunto conexo S de $\mathbf{R}^n$. Si f alcanza dos valores distintos sobre S, tales como a y b, entonces para cada c comprendido entre a y b existe por lo menos un punto $\mathbf{x}$ de S en el que $f(\mathbf{x}) = c$.*

Demostración. La imagen $f(S)$ es un subconjunto conexo de $\mathbf{R}^1$. Por lo tanto, $f(S)$ es un intervalo que contiene a a y a b (ver ejercicio 4.38). Si algún valor c comprendido entre a y b no estuviese en $f(S)$, entonces $f(S)$ no sería conexo.

4.17 COMPONENTES DE UN ESPACIO MÉTRICO

Esta sección demuestra que todo espacio métrico S puede expresarse de forma única como reunión de «trozos» conexos, llamados componentes. Ante todo demostraremos el siguiente.

Teorema 4.39. *Sea F una colección de subconjuntos conexos de un espacio métrico S tal que la intersección $T = \bigcap_{A \in F} A$ es no vacía. Entonces, la reunión $U = \bigcup_{A \in F} A$ es conexa.*

Demostración. Como $T \neq \emptyset$, existe un t de T. Sea f una función a dos valores definida sobre U. Probaremos que f es constante en U probando que $f(x) = f(t)$ para todo x de U. Si $x \in U$, entonces $x \in A$ para un cierto A de F. Como A es conexo, f es constante sobre A y, como $t \in A$, $f(x) = f(t)$.

Todo punto x de un espacio métrico S pertenece, por lo menos, a un subconjunto conexo de S, a saber $\{x\}$. Por el teorema 4.39, la reunión de todos los subconjuntos conexos que contienen a x es también conexo. A esta reunión la llamaremos *componente* de S, y la designaremos por $U(x)$. Así, $U(x)$ es el subconjunto de S conexo maximal que contiene a x.

Teorema 4.40. *Todo punto de un espacio métrico S pertenece a una única y determinada componente de S. En otras palabras, las componentes de S forman una colección de conjuntos disjuntos cuya reunión es S.*

Material protegido por derechos de autor

Demostración. Dos componentes distintas no pueden tener ningún punto x en común; en otro caso (por el teorema 4.39) su reunión sería un conjunto conexo más grande que contendría a x.

4.18 CONEXIÓN POR ARCOS

En esta sección se describe una propiedad especial, llamada *conexión por arcos*, que poseen algunos (pero no todos) los conjuntos conexos de un espacio euclídeo $\mathbf{R}^n$.

Definición 4.41. *Un conjunto S de $\mathbf{R}^n$ se llama arco-conexo si, para cada par de puntos $\mathbf{a}$ y $\mathbf{b}$ de S, existe una función $\mathbf{f}:[0, 1] \to S$ tal que*

$$\mathbf{f}(0) = \mathbf{a} \quad y \quad \mathbf{f}(1) = \mathbf{b}.$$

NOTA. Una tal función se llama un *camino* de $\mathbf{a}$ a $\mathbf{b}$. Si $\mathbf{f}(0) \neq \mathbf{f}(1)$, la imagen de [0, 1] por medio de $\mathbf{f}$ se denomina *arco*, que une $\mathbf{a}$ con $\mathbf{b}$. Entonces, S es arco-conexo si cada dos puntos distintos de S pueden unirse por medio de un arco contenido en S. Los conjuntos arco-conexos se llaman también *conexos por caminos*. Si $\mathbf{f}(t) = t\mathbf{b} + (1 - t)\mathbf{a}$ para $0 \leq t \leq 1$, la curva que une $\mathbf{a}$ y $\mathbf{b}$ se llama *segmento rectilíneo*.

Ejemplos

1. Cada conjunto convexo de $\mathbf{R}^n$ es arco-conexo, ya que el segmento rectilíneo que une dos puntos del conjunto está en el conjunto. En particular, las bolas n-dimensionales abiertas y las cerradas son arco conexas.
2. El conjunto de la figura 4.4 (reunión de dos discos cerrados tangentes) es arco conexo.

Figura 4.4

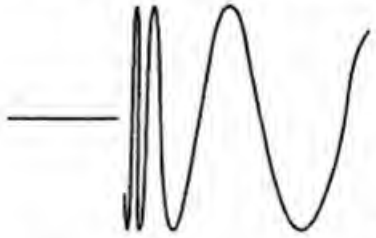


Figura 4.5

3. El conjunto de la figura 4.5 consiste en todos los puntos de la curva descrita por $y = \text{sen } (1/x)$, $0 < x \leq 1$, y los del segmento horizontal $-1 \leq x \leq 0$. Este conjunto es conexo pero no arco conexo (ejercicio 4.46).

El teorema que sigue relaciona la conexión por arcos con la conexión.

Material protegido por derechos de autor

Teorema 4.42. *Todo conjunto S de $\mathbf{R}^n$ arco-conexo es conexo.*

Demostración. Sea g una función a dos valores definida sobre S. Probaremos que g es constante sobre S. Elijamos un punto $\mathbf{a}$ de S. Si $\mathbf{x} \in S$, unamos $\mathbf{a}$ con $\mathbf{x}$ por medio de un arco Γ contenido en S. Como que Γ es conexo, g es constante sobre Γ luego $g(\mathbf{x}) = g(\mathbf{a})$. Pero, al ser $\mathbf{x}$ un punto arbitrario de S, queda demostrado que g es constante sobre S, y que S es conexo.

Hemos visto anteriormente que hay conjuntos conexos que no son arco conexos. Sin embargo, ambos conceptos son equivalentes en el caso de conjuntos ***abiertos.***

Teorema 4.43. *Un conjunto abierto conexo de $\mathbf{R}^n$ es arco-conexo.*

Demostración. Sea S un conjunto abierto y conexo de $\mathbf{R}^n$ y supongamos que $\mathbf{x} \in S$. Probaremos que $\mathbf{x}$ puede unirse con cualquier otro punto $\mathbf{y}$ de S por medio de un arco contenido en S. Designemos por A el subconjunto de S formado por los puntos que pueden unirse con $\mathbf{x}$, y sea $B = S - A$. Entonces $S = A \cup B$, donde A y B son disjuntos. Ahora demostraremos que tanto A como B son abiertos en $\mathbf{R}^n$.

Sea $\mathbf{a} \in A$ y unamos $\mathbf{a}$ con $\mathbf{x}$ por medio de un arco Γ contenido en S. Como que $\mathbf{a} \in S$ y S es abierto, existe una bola n-dimensional $B(\mathbf{a}) \subseteq S$. Cada $\mathbf{y}$ de $B(\mathbf{a})$ puede unirse con $\mathbf{a}$ por medio de un segmento rectilíneo (contenido en S) y por lo tanto con $\mathbf{x}$ por medio de Γ. Así pues, si $\mathbf{y} \in B(\mathbf{a})$, entonces $\mathbf{y} \in A$. Esto implica que $B(\mathbf{a}) \subseteq A$, y por lo tanto A es abierto.

Para ver que B también es abierto, supongamos que $\mathbf{b} \in B$. Entonces existe una bola n-dimensional $B(\mathbf{b}) \subseteq S$, ya que S es abierto. Ahora bien, si un punto $\mathbf{y}$ de $B(\mathbf{b})$ pudiese unirse con $\mathbf{x}$ por medio de un arco Γ', contenido en S, el punto $\mathbf{b}$ también podría unirse con $\mathbf{x}$, uniendo primeramente $\mathbf{b}$ con $\mathbf{y}$ (por medio de un segmento rectilíneo contenido en $B(\mathbf{b})$) y utilizando después Γ'. Pero como $\mathbf{b} \notin A$, ningún punto de $B(\mathbf{b})$ deberá pertenecer a A. Así, $B(\mathbf{b}) \subseteq B$, luego B es abierto.

Hemos obtenido, por lo tanto, una descomposición $S = A \cup B$, donde A y B son conjuntos de $\mathbf{R}^n$ abiertos y disjuntos. Pero, A es no vacío ya que $\mathbf{x} \in A$. Como que S es conexo, B deberá ser vacío, con lo cual $S = A$. Ahora bien, es evidente que A es arco-conexo ya que cualquier par de puntos de A pueden unirse por medio de un arco conveniente, uniendo primeramente cada uno de ellos con $\mathbf{x}$. Por consiguiente, S es arco-conexo y la demostración está terminada.

NOTA. Un camino $\mathbf{f}:[0, 1] \to S$ se llama *poligonal* si la imagen de $[0, 1]$ por medio de $\mathbf{f}$ es la reunión de un número finito de segmentos rectilíneos. El mismo argumento utilizado para demostrar el teorema 4.43 prueba además que cada conjunto conexo de $\mathbf{R}^n$ es *conexo por poligonales*; es decir, cada par de puntos del conjunto puede unirse con un arco poligonal contenido en el conjunto.

Teorema 4.44. *Todo conjunto abierto S de $\mathbf{R}^n$ puede expresarse de forma única como reunión de una familia disjunta numerable de conjuntos conexos y abiertos.*

Demostración. Por el teorema 4.40, las componentes de S constituyen una colección de conjuntos disjuntos cuya reunión es S. Cada componente T de S es abierta, puesto que si $\mathbf{x} \in T$ existe una bola n-dimensional $B(\mathbf{x})$ contenida en S. Como $B(\mathbf{x})$ es conexo, $B(\mathbf{x}) \subseteq T$, luego T es abierto. Por el teorema de Lindelöf (teorema 3.28), las componentes de S constituyen una colección numerable, y por el teorema 4.40 la descomposición en componentes es única.

Definición 4.45. *Un conjunto de $\mathbf{R}^n$ se llama región si es la reunión de un conjunto conexo abierto con alguno, ninguno, o todos sus puntos frontera. Si ninguno de sus puntos frontera está incluido en la región, se dice que ésta es una región abierta. Si todos los puntos frontera están incluidos, se dice que la región es una región cerrada.*

NOTA. Algunos autores utilizan la palabra *dominio* en vez de *región abierta*, especialmente en el plano complejo.

4.19 CONTINUIDAD UNIFORME

Supongamos que f está definida en un cierto espacio métrico (S, d_S) y tiene sus valores en otro espacio métrico (T, d_T), y supongamos que f es continua en un subconjunto A de S. Entonces, dado un punto p de A y un $\varepsilon > 0$, existe un $\delta > 0$ (que depende de p y de ε) tal que, si $x \in A$, entonces

$$d_T(f(x), f(p)) < \varepsilon \qquad \text{siempre que } d_S(x, p) < \delta.$$

En general no se debe esperar que, fijado ε, el mismo valor de δ sirva para *cada* punto p de A. Sin embargo, puede ocurrir. Cuando ocurre, se dice que la función es *uniformemente continua* en A.

Definición 4.46. *Sea $f: S \to T$ una función de un espacio métrico (S, d_S) en otro espacio métrico (T, d_T). Entonces se dice que f es uniformemente continua en un subconjunto A de S si verifica la siguiente condición:*

Material protegido por derechos de autor
Material protegido por derechos de autor

Consideremos la colección de las bolas $B_S(a; r/2)$ de radio $r/2$. Recubren a A y, como A es compacto, basta un número finito de ellas para recubrir a A, o sea

$$A \subseteq \bigcup_{k=1}^{m} B_S\left(a_k; \frac{r_k}{2}\right).$$

En cualquiera de las bolas de doble radio, $B_S(a_k; r_k)$ se tiene

$$d_T(f(x), f(a_k)) < \frac{\varepsilon}{2} \qquad \text{siempre que } x \in B_S(a_k; r_k) \cap A.$$

Sea δ el menor de los números $r_1/2, \ldots, r_m/2$. Probaremos que este δ satisface la definición de continuidad uniforme.

En efecto, consideremos dos puntos de A, por ejemplo x y p, con $d_S(x, p) < \delta$. En virtud de la anterior discusión existirá una bola $B_S(a_k; r_k/2)$ que contenga a x, luego

$$d_T(f(x), f(a_k)) < \frac{\varepsilon}{2}.$$

Por la desigualdad triangular tenemos que

$$d_S(p, a_k) \leq d_S(p, x) + d_S(x, a_k) < \delta + \frac{r_k}{2} \leq \frac{r_k}{2} + \frac{r_k}{2} = r_k.$$

Por lo tanto, $p \in B_S(a_k; r_k) \cap S$, y entonces tenemos también que

$$d_T(f(p), f(a_k)) < \varepsilon/2.$$

Utilizando, una vez más, la desigualdad triangular obtenemos

$$d_T(f(x), f(p)) \leq d_T(f(x), f(a_k)) + d_T(f(a_k), f(p)) < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon.$$

Esto termina la demostración.

4.21 TEOREMA DEL PUNTO FIJO PARA CONTRACCIONES

Sea $f: S \to S$ una función de un espacio métrico (S, d) en sí mismo. Un punto p de S es un *punto fijo* de f si $f(p) = p$. La función f se denomina *contracción* de S si existe un número positivo $\alpha < 1$ (llamado *constante de contracción* o *coeficiente de contracción*), tal que

$$d(f(x), f(y)) \leq \alpha\, d(x, y) \qquad \text{para todo } x, y \text{ de } S. \tag{7}$$

Material protegido por derechos de autor

4.22 DISCONTINUIDADES DE LAS FUNCIONES REALES

El resto de este capítulo lo dedicaremos a estudiar propiedades especiales de funciones reales definidas en subintervalos de **R**.

Sea f una función real definida sobre un intervalo (a, b). Supongamos que $c \in [a, b)$. Si $f(x) \to A$ cuando $x \to c$ con valores mayores que c, diremos que A es el *límite lateral por la derecha* de f en c y lo indicaremos, escribiendo

$$\lim_{x \to c+} f(x) = A.$$

El límite lateral por la derecha se designa también por medio de $f(c+)$. En la terminología ε, δ significa que para todo $\varepsilon > 0$ existe un $\delta > 0$ tal que

$$|f(x) - f(c+)| < \varepsilon \qquad \text{siempre que } c < x < c + \delta < b.$$

Nótese que f no necesita estar definida en el punto c. Si f está definida en c y es $f(c+) = f(c)$, diremos que f es *continua por la derecha* en c.

Los límites laterales por la izquierda y la continuidad por la izquierda en c se definen análogamente si $c \in (a, b]$.

Si $a < c < b$, entonces f es continua en c si, y sólo si,

$$f(c) = f(c+) = f(c-).$$

Diremos que c es una *discontinuidad* de f, si f no es continua en c. En este caso deberá darse alguna de las siguientes condiciones:

a) O no existe $f(c+)$ o no existe $f(c-)$.
b) Tanto $f(c+)$ como $f(c-)$ existen pero son distintos.
c) Tanto $f(c+)$ como $f(c-)$ existen y $f(c+) = f(c-) \neq f(c)$.

En el caso (c) se dice que el punto c es una *discontinuidad evitable*, ya que la discontinuidad podría evitarse volviendo a definir f en c de suerte que el valor de f en c fuese $f(c+) = f(c-)$. En los casos (a) y (b), se dice que c es una *discontinuidad inevitable* dado que la discontinuidad no puede evitarse aunque volvamos a definir f en c.

Definición 4.49. *Sea f una función definida sobre un intervalo cerrado $[a, b]$. Si $f(c+)$ y $f(c-)$ existen en un punto interior c, entonces:*

a) *$f(c) - f(c-)$ se llama el salto de f a la izquierda de c,*
b) *$f(c+) - f(c)$ se llama el salto de f a la derecha de c,*
c) *$f(c+) - f(c-)$ se llama el salto de f en c.*

Material protegido por derechos de autor

Si alguno de ellos es distinto de 0, entonces se dice que f tiene una discontinuidad de salto en c.

En los puntos extremos a y b, sólo consideraremos uno de los saltos laterales, el salto a la derecha en a, $f(a+)-f(a)$, y el salto a la izquierda en b, $f(b)-f(b-)$.

Ejemplos

1. La función f definida por $f(x)=x/|x|$ si $x\neq 0$, $f(0)=A$, tiene una discontinuidad de salto en 0, independiente del valor de A. Aquí $f(0+)=+1$ y $f(0-)=-1$. (Ver fig. 4.6.)
2. La función f definida por $f(x)=1$ si $x\neq 0$, $f(0)=0$, posee un salto de discontinuidad evitable en 0. En este caso $f(0+)=f(0-)=1$.

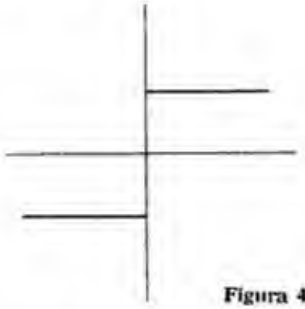


Figura 4.6

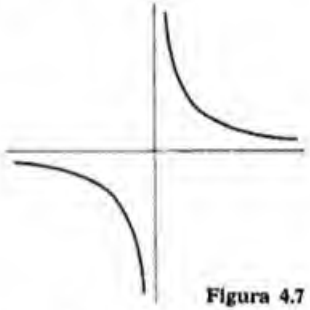


Figura 4.7

3. La función f definida por $f(x)=1/x$ si $x\neq 0$, $f(0)=A$, tiene un punto de discontinuidad inevitable en 0. En este caso $f(0+)$ y $f(0-)$ no existen. (Ver fig. 4.7.)
4. La función f definida por $f(x)=\operatorname{sen}(1/x)$ si $x\neq 0$, $f(0)=A$, posee una discontinuidad inevitable en 0 ya que $f(0+)$ y $f(0-)$ no existen. (Ver fig. 4.8.)
5. La función f definida por $f(x)=x\operatorname{sen}(1/x)$ si $x\neq 0$, $f(0)=1$, tiene un punto de discontinuidad evitable en 0, ya que $f(0+)=f(0-)=0$. (Ver fig. 4.9.)

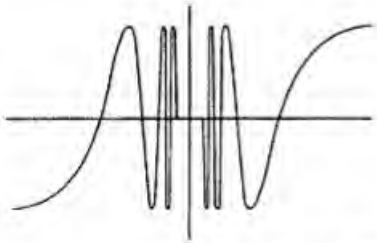


Figura 4.8

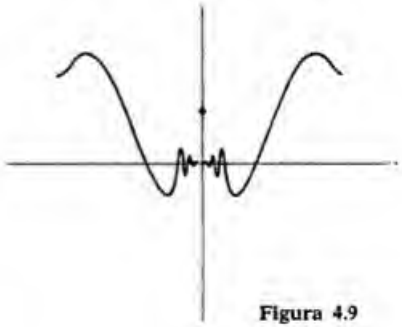


Figura 4.9

Material protegido por derechos de autor

Teorema 4.52. *Sea f estrictamente creciente en un conjunto S de* **R**. *Entonces f^{-1} existe y es estrictamente creciente en $f(S)$.*

Demostración. Como f es estrictamente creciente, es uno a uno en S, luego f^{-1} existe. Para ver que f^{-1} es estrictamente creciente, sean $y_1<y_2$ dos puntos de $f(S)$ y sea $x_1=f^{-1}(y_1)$, $x_2=f^{-1}(y_2)$. No puede ser que $x_1\geq x_2$, ya que entonces tendríamos también que $y_1\geq y_2$. La única alternativa es $x_1<x_2$, y esto significa que f^{-1} es estrictamente creciente.

El teorema 4.52 junto con el teorema 4.29 conducen a:

Teorema 4.53. *Sea f estrictamente creciente y continua en un intervalo compacto $[a, b]$. Entonces f^{-1} es continua y estrictamente creciente en el intervalo $[f(a), f(b)]$.*

NOTA. El teorema 4.53 nos dice que una función continua, estrictamente creciente es una aplicación topológica. Recíprocamente, toda aplicación topológica de un intervalo $[a, b]$ sobre un intervalo $[c, d]$ debe ser una función estrictamente monótona. La verificación de este hecho constituye un ejercicio muy instructivo para el lector. (Ejercicio 4.62.)

EJERCICIOS

Límites de sucesiones

4.1 Probar cada una de las afirmaciones siguientes acerca de sucesiones de **C**.
   a) $z^n\to 0$ si $|z|<1$; $\{z^n\}$ diverge si $|z|>1$.
   b) Si $z_n\to 0$ y si $\{c_n\}$ está acotada, entonces $\{c_n z_n\}\to 0$.
   c) $z^n/n!\to 0$ para cada complejo z.
   d) Si $a_n=\sqrt{n^2+2}-n$, entonces $a_n\to 0$.

4.2 Si $a_{n+2}=(a_{n+1}+a_n)/2$ para todo $n\geq 1$, expresar a_n en función de a_1 y a_2, y demostrar que $a_n\to(a_1+2a_2)/3$. *Observación:* $a_{n+2}-a_{n+1}=\frac{1}{2}(a_n-a_{n+1})$.

4.3 Si $0<x_1<1$ y si $x_{n+1}=1-\sqrt{1-x_n}$ para todo $n\geq 1$, probar que $\{x_n\}$ es una sucesión decreciente con límite 0. Probar además que $x_{n+1}/x_n\to\frac{1}{2}$.

4.4 Dos sucesiones de enteros positivos $\{a_n\}$ y $\{b_n\}$ se definen recursivamente haciendo $a_1=b_1=1$ e igualando las partes racionales e irracionales de la ecuación

$$a_n+b_n\sqrt{2}=(a_{n-1}+b_{n-1}\sqrt{2})^2 \quad \text{para } n\geq 2.$$

Probar que $a_n^2-2b_n^2=1$ para $n\geq 2$. Deducir que $a_n/b_n\to\sqrt{2}$ por medio de valores $>\sqrt{2}$, y que $2b_n/a_n\to\sqrt{2}$ por medio de valores $<\sqrt{2}$.

Material protegido por derechos de autor

4.5 Una sucesión real $\{x_n\}$ satisface $7x_{n+1} = x_n^2 + 6$ para $n \geq 1$. Si $x_1 = \frac{1}{2}$, probar que la sucesión crece y hallar su límite. ¿Qué ocurre si $x_1 = \frac{3}{2}$ o si $x_1 = \frac{5}{2}$?

4.6 Si $|a_n| < 2$ y $|a_{n+2} - a_{n+1}| \leq \frac{1}{8}|a_{n+1}^2 - a_n^2|$ para todo $n \geq 1$, probar que (a_n) converge.

4.7 En un espacio métrico (S, d) suponemos que $x_n \to x$ y que $y_n \to y$. Probar que $d(x_n, y_n) \to d(x, y)$.

4.8 Probar que en un espacio métrico compacto (S, d), cada sucesión de S admite una subsucesión convergente en S. Esta propiedad implica también que S es compacto, pero no se pide una demostración de este resultado. (Una demostración puede encontrarse en las referencias 4.2 o 4.3.)

4.9 Sea A un subconjunto de un espacio métrico S. Si A es completo, probar que A es cerrado. Probar que el recíproco también es cierto siempre que S sea completo.

Límites de funciones

NOTA. En los ejercicios 4.10 a 4.28 todas las funciones serán reales.

4.10 Sea f definida en un intervalo abierto (a, b) y supongamos que $x \in (a, b)$. Consideremos las dos afirmaciones siguientes:

a) $\lim_{h \to 0} |f(x + h) - f(x)| = 0$; b) $\lim_{h \to 0} |f(x + h) - f(x - h)| = 0$.

Probar que (a) siempre implica (b), y dar un ejemplo en el que (b) se verifique pero (a) no.

4.11 Sea f definida en $\mathbf{R}^2$. Si

$$\lim_{(x,y) \to (a,b)} f(x, y) = L$$

y si existen los dos límites unidimensionales $\lim_{x \to a} f(x, y)$ y $\lim_{y \to b} f(x, y)$, probar que

$$\lim_{x \to a} [\lim_{y \to b} f(x, y)] = \lim_{y \to b} [\lim_{x \to a} f(x, y)] = L.$$

Consideremos ahora las funciones f definidas en $\mathbf{R}^2$ como sigue:

a) $f(x, y) = \dfrac{x^2 - y^2}{x^2 + y^2}$ si $(x, y) \neq (0, 0)$, $f(0, 0) = 0$.

b) $f(x, y) = \dfrac{(xy)^2}{(xy)^2 + (x - y)^2}$ si $(x, y) \neq (0, 0)$, $f(0, 0) = 0$.

c) $f(x, y) = \dfrac{1}{x} \operatorname{sen}(xy)$ si $x \neq 0$, $f(0, y) = y$.

d) $f(x, y) = \begin{cases} (x + y) \operatorname{sen}(1/x) \operatorname{sen}(1/y) & \text{si } x \neq 0 \text{ e } y \neq 0, \\ 0 & \text{si } x = 0 \text{ o } y = 0. \end{cases}$

e) $f(x, y) = \begin{cases} \dfrac{\operatorname{sen} x - \operatorname{sen} y}{\operatorname{tg} x - \operatorname{tg} y} & \text{si } \operatorname{tg} x \neq \operatorname{tg} y, \\ \cos^3 x & \text{si } \operatorname{tg} x = \operatorname{tg} y. \end{cases}$

Material protegido por derechos de autor

En cada uno de los ejercicios anteriores, determinar cuándo existen los límites que se proponen y calcular los que existan:

$$\lim_{x \to 0} [\lim_{y \to 0} f(x, y)]; \quad \lim_{y \to 0} [\lim_{x \to 0} f(x, y)]; \quad \lim_{(x,y) \to (0,0)} f(x, y).$$

4.12 Si $x \in [0, 1]$ probar que el siguiente límite existe,

$$\lim_{m \to \infty} [\lim_{n \to \infty} \cos^{2n}(m! \, \pi x)],$$

y que su valor es 0 o 1, según que x sea irracional o racional.

Continuidad de funciones reales

4.13 Sea f continua en $[a, b]$ y sea $f(x) = 0$ si x es racional. Probar que $f(x) = 0$ para todo x de $[a, b]$.

4.14 Sea f continua en el punto $\mathbf{a} = (a_1, a_2, \ldots, a_n)$ de $\mathbf{R}^n$. Conservemos $a_2, a_3, \ldots, a_n$ fijos y definamos una nueva función g de una sola variable real definida por la ecuación

$$g(x) = f(x, a_2, \ldots, a_n).$$

Probar que g es continua en el punto $x = a_1$. (Este resultado suele expresarse diciendo que *una función continua de n variables es continua en cada una de ellas separadamente.*)

4.15 Probar por medio de un ejemplo que el recíproco de la proposición establecida en el ejercicio 4.14 no es verdadero en general.

4.16 Sean f, g y h definidas en $[0, 1]$ como sigue:

$f(x) = g(x) = h(x) = 0$, siempre que x sea irracional;
$f(x) = 1$ y $g(x) = x$, siempre que x sea racional;
$h(x) = 1/n$, si x es el racional m/n (irreducible);
$h(0) = 1$.

Probar que f no es continua en ningún punto de $[0, 1]$, que g es continua sólo en $x = 0$, y que h sólo es continua en los puntos irracionales de $[0, 1]$.

4.17 Para cada x de $[0, 1]$, sea $f(x) = x$ si x es racional, y sea $f(x) = 1 - x$ si x es irracional. Probar que:

a) $f(f(x)) = x$ para todo x de $[0, 1]$.
b) $f(x) + f(1 - x) = 1$ para todo x de $[0, 1]$.
c) f es continua sólo en el punto $x = \frac{1}{2}$.
d) f toma todos los valores comprendidos entre 0 y 1.
e) $f(x+y) - f(x) - f(y)$ es racional para todos los x e y de $[0, 1]$.

4.18 Sea f definida en $\mathbf{R}$ y supongamos que existe por lo menos un punto x_0 de $\mathbf{R}$ en el que f es continua. Supongamos también que, para cada x e y de $\mathbf{R}$, f satisface la ecuación

$$f(x + y) = f(x) + f(y).$$

Probar que existe una constante a tal que $f(x) = ax$ para todo x.

Material protegido por derechos de autor

siempre que $x \neq y$.

a) Probar que f posee a lo sumo un punto fijo, y dar un ejemplo de una función f de este tipo sin puntos fijos.
b) Si S es compacto, probar que f admite un punto fijo exactamente. *Indicación.* Probar que $g(x) = d(x, f(x))$ alcanza un mínimo en S.
c) Dar un ejemplo en el que, siendo S compacto, f no sea una contracción.

4.72 Supongamos que f satisface la condición del ejercicio 4.71. Si $x \in S$, sea $p_0 = x$, $p_{n+1} = f(p_n)$, y $c_n = d(p_n, p_{n+1})$ para $n \geq 0$.

a) Probar que $\{c_n\}$ es una sucesión decreciente, y sea $c = \lim c_n$.
b) Supongamos que existe una subsucesión $\{p_{k(n)}\}$ convergente hacia un cierto punto q de S. Probar que

$$c = d(q, f(q)) = d(f(q), f[f(q)]).$$

Deducir que q es un punto fijo de f y que $p_n \to q$.

REFERENCIAS SUGERIDAS PARA POSTERIORES ESTUDIOS

4.1 Boas, R. P., *A Primer of Real Functions.* Carus Monograph No. 13. Wiley, New York, 1960.
4.2 Gleason, A., *Fundamentals of Abstract Analysis.* Addison-Wesley, Reading, 1966.
4.3 Simmons, G. F., *Introduction to Topology and Modern Analysis,* McGraw-Hill, New York, 1963.
4.4 Todd, J., *Survey of Numerical Analysis.* McGraw-Hill, New York, 1962.

CAPÍTULO 5

Derivadas

5.1 INTRODUCCIÓN

Este capítulo trata de la derivada, concepto fundamental del Cálculo diferencial. Dos tipos distintos de problemas —el problema físico, que consiste en buscar la velocidad instantánea de una partícula móvil, y el problema geométrico, que consiste en buscar la recta tangente a una curva en un punto dado—, ambos conducen de forma muy natural a la noción de derivada. No nos interesaremos ni por las aplicaciones físicas ni por las aplicaciones geométricas; dedicaremos nuestra atención a las propiedades generales de las derivadas.

Este capítulo tratará, ante todo, de las derivadas de funciones de una variable real y, especialmente, de funciones reales definidas en intervalos de $\mathbf{R}$. Estudiará también brevemente las derivadas de funciones de valores vectoriales de una variable real, y las derivadas parciales, ya que estos temas no envuelven ideas nuevas. Mucho de lo que se expone será familiar al lector, pues se trata de Cálculo elemental. Un tratamiento más detallado de la teoría de la derivación para funciones de varias variables involucra cambios realmente importantes y por ello se desarrollará en el capítulo 12.

La última parte de este capítulo trata de las derivadas de funciones complejas de una variable compleja.

5.2 DEFINICIÓN DE DERIVADA

Si f está definida sobre un intervalo abierto (a, b), entonces para cada dos puntos distintos x y c de (a, b) podemos considerar el cociente de diferencias (*)

$$\frac{f(x) - f(c)}{x - c}.$$

Mantenemos c fijo y estudiamos el comportamiento de este cociente cuando $x \to c$.

* Este cociente se conoce con el nombre de cociente incremental. (N. de t.)

Material protegido por derechos de autor
Material protegido por derechos de autor

Definición 5.1. *Sea f una función real definida en un intervalo abierto (a, b), y supongamos que $c \in (a, b)$. Diremos que f es diferenciable en c siempre que el límite*

$$\lim_{x \to c} \frac{f(x) - f(c)}{x - c}$$

exista. El límite, designado por $f'(c)$, se llama derivada de f en c.

Este método de calcular límites define una nueva función f', cuyo dominio está formado por aquellos puntos de (a, b) en los que f es diferenciable. La función f' se llama la *primera derivada* de f. Análogamente, la n-ésima derivada de f, designada por $f^{(n)}$, es la primera derivada de $f^{(n-1)}$, para $n = 2, 3, \ldots$ (según nuestra definición, sólo es posible considerar $f^{(n)}$ si $f^{(n-1)}$ está definida en un cierto intervalo abierto). Otras notaciones con las que el lector puede estar familiarizado son

$$f'(c) = Df(c) = \frac{df}{dx}(c) = \left.\frac{dy}{dx}\right|_{x=c} \qquad [\text{donde } y = f(x)],$$

o notaciones similares. La función f se escribe, a veces, $f^{(0)}$. El proceso que produce f' a partir de f se llama *diferenciación*.

5.3 DERIVADAS Y CONTINUIDAD

El teorema que se da a continuación permite reducir algunos de los teoremas de derivadas a teoremas de continuidad.

Teorema 5.2. *Si f está definida en un intervalo (a, b) y es diferenciable en un punto c de (a, b), entonces existe una función f^* (que depende de f y de c) continua en c y que satisface la ecuación*

$$f(x) - f(c) = (x - c)f^*(x), \tag{1}$$

para todo x de (a, b), con $f^*(c) = f'(c)$. Recíprocamente, si existe una función f^*, continua en c, que satisfaga (1), *entonces f es diferenciable en c y $f'(c) = f^*(c)$.*

Demostración. Si $f'(c)$ existe, sea f^* definida en (a, b) como sigue:

$$f^*(x) = \frac{f(x) - f(c)}{x - c} \quad \text{si } x \neq c, \quad f^*(c) = f'(c).$$

Entonces f^* es continua en c y (1) se verifica para todo x de (a, b).

Material protegido por derechos de autor

Recíprocamente, si (1) se verifica para una cierta función f^* continua en c, entonces dividiendo por $x - c$ y haciendo $x \to c$ vemos que $f'(c)$ existe y es igual a $f^*(c)$.

Como consecuencia inmediata de (1) se obtiene:

Teorema 5.3. *Si f es diferenciable en c, entonces f es continua en c.*

Demostración. En (1) hagamos $x \to c$.

NOTA. La ecuación (1) tiene una interpretación geométrica que ayuda a adquirir una intuición de su significado. Como que f^* es continua en c, $f^*(x)$ es aproximadamente igual a $f^*(c) = f'(c)$ si x es próximo a c. Reemplazando $f^*(x)$ por $f'(c)$ en (1) obtenemos la ecuación

$$f(x) = f(c) + f'(c)(x - c),$$

que será aproximadamente correcta cuando $x - c$ sea pequeño. En otras palabras, si f es diferenciable en c, entonces f es aproximadamente una función lineal en las proximidades de c. (Ver Fig. 5.1.) El Cálculo diferencial explota, continuamente, esta propiedad geométrica de las funciones.

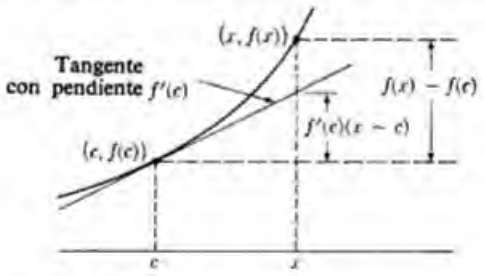


Figura 5.1

5.4 ÁLGEBRA DE DERIVADAS

El siguiente teorema describe las fórmulas usuales para diferenciar la suma, la diferencia, el producto y el cociente de dos funciones.

Teorema 5.4. *Supongamos que f y g están definidas en (a, b) y son diferenciables en c. Entonces $f + g$, $f - g$ y $f \cdot g$ son también diferenciables en c. Esto es asimismo verdadero para f/g si $g(c) \neq 0$. Las derivadas en c están dadas por las fórmulas siguientes:*

Material protegido por derechos de autor

a) $(f \pm g)'(c) = f'(c) \pm g'(c)$,

b) $(f \cdot g)'(c) = f(c)g'(c) + f'(c)g(c)$,

c) $(f/g)'(c) = \dfrac{g(c)f'(c) - f(c)g'(c)}{g(c)^2}$, *en el supuesto de que* $g(c) \neq 0$.

Demostración. Probaremos (b). Utilizando el teorema 5.2, escribiremos

$$f(x) = f(c) + (x - c)f^*(x), \qquad g(x) = g(c) + (x - c)g^*(x).$$

Entonces

$$f(x)g(x) - f(c)g(c) = (x - c)[f(c)g^*(x) + f^*(x)g(c)] + (x - c)^2 f^*(x)g^*(x).$$

Dividiendo por $x - c$ y haciendo que $x \to c$ obtendremos (b). Las demostraciones de las otras afirmaciones son análogas.

De la definición se sigue inmediatamente que si f es constante en (a, b), entonces $f' = 0$ en (a, b). También, si $f(x) = x$, entonces $f'(x) = 1$ para todo x. Aplicando repetidamente el teorema 5.4 obtenemos que si $f(x) = x^n$ (n entero positivo), entonces $f'(x) = nx^{n-1}$ para todo x. Aplicando, de nuevo, el teorema 5.4 vemos que todo polinomio admite derivada en todo $\mathbf{R}$ y que cada función racional admite derivada en los puntos en los que está definida.

5.5 LA REGLA DE LA CADENA

Un resultado más profundo lo constituye la llamada *regla de la cadena* para la diferenciación de funciones compuestas.

Teorema 5.5 (Regla de la cadena). *Sea f definida en un intervalo abierto S y sea g definida en $f(S)$, y consideremos la función compuesta $g \circ f$ definida en S por medio de la ecuación*

$$(g \circ f)(x) = g(f(x)).$$

Supongamos que exista un punto c de S tal que $f(c)$ sea un punto interior de $f(S)$. Si f es diferenciable en c y g es diferenciable en $f(c)$, entonces $g \circ f$ es diferenciable en c y se tiene que

$$(g \circ f)'(c) = g'[f(c)]f'(c).$$

Demostración. Utilizando el teorema 5.2 podemos escribir

$$f(x) - f(c) = (x - c)f^*(x) \qquad \text{para todo } x \text{ de } S,$$

$B(c) \subseteq (a, b)$ en la que $f^*(x)$ tiene el mismo signo que $f^*(c)$, y esto significa que $f(x) - f(c)$ tiene el mismo signo que $x - c$.

Si $f'(c) = +\infty$, existe una bola unidimensional $B(c)$ en la que

$$\frac{f(x) - f(c)}{x - c} > 1 \qquad \text{cuando } x \neq c.$$

En esta bola el cociente es, de nuevo, positivo y la conclusión sigue como antes.

Un resultado análogo al del teorema 5.7 es válido, naturalmente, si $f'(c) < 0$ o si $f'(c) = -\infty$ en algún punto interior c de (a, b).

5.8 DERIVADAS CERO Y EXTREMOS LOCALES

Definición 5.8. *Sea f una función real definida en un subconjunto S de un espacio métrico M, y supongamos que $a \in S$. Entonces f posee un máximo local en a si existe una bola $B(a)$ tal que*

$$f(x) \leq f(a) \qquad \textit{para todo } x \textit{ de } B(a) \cap S.$$

Si $f(x) \geq f(a)$ para todo x de $B(a) \cap S$, entonces f posee un mínimo local en a.

NOTA. Un máximo local en a es el máximo absoluto de f en el subconjunto $B(a) \cap S$. Si f tiene un máximo absoluto en a, entonces a es un máximo local. Sin embargo, f puede poseer máximos locales en varios puntos de S sin que posea máximo absoluto en el conjunto S.

El teorema que sigue establece una relación entre las derivadas nulas y los extremos locales (máximos o mínimos) en puntos interiores.

Teorema 5.9. *Sea f definida en un intervalo abierto (a, b) y supongamos que f posee un máximo local o un mínimo local en un cierto punto interior c de (a, b). Si f posee derivada (finita o infinita) en c, entonces $f'(c)$ debe ser cero.*

Demostración. Si $f'(c)$ es positiva o $+\infty$, entonces f no puede tener un extremo local en c, en virtud del teorema 5.7. Análogamente, $f'(c)$ no puede ser negativa ni $-\infty$. Luego, dado que existe derivada en c, la única posibilidad que queda es $f'(c) = 0$.

El recíproco del teorema 5.9 es falso. En general, el saber que $f'(c) = 0$ no basta para deducir que f tiene un extremo en c. De hecho, es posible que

Material protegido por derechos de autor
Material protegido por derechos de autor

carezca de ellos, como puede verificarse por medio del ejemplo $f(x) = x^3$ y $c = 0$. En este caso, $f'(0) = 0$ pero f es creciente en todo entorno de 0.

Además, conviene insistir en el hecho de que f puede tener un extremo local en c sin que $f'(c)$ sea cero. Por ejemplo, $f(x) = |x|$ tiene un mínimo en $x = 0$ pero, naturalmente, no existe la derivada en 0. El teorema 5.9 presupone que f tiene derivada (finita o infinita) en c. El teorema presupone también que c es un punto interor de (a, b). En el ejemplo $f(x) = x$, donde $a \leq x \leq b$, f alcanza su máximo y su mínimo en los puntos extremos pero en cambio $f'(x)$ no es nunca cero en $[a, b]$.

5.9 TEOREMA DE ROLLE

Es geométricamente evidente que una curva suficientemente «regular» que corta al eje ox en los puntos extremos del intervalo $[a, b]$ debe poseer un «punto de viraje» en algún punto comprendido entre a y b. El enunciado preciso de este resultado se conoce con el nombre de teorema de Rolle.

Teorema 5.10 (Rolle). *Supongamos que f posee derivada (finita o infinita) en cada uno de los puntos de un intervalo abierto (a, b), y supongamos también que f es continua en los puntos extremos a y b. Si $f(a) = f(b)$, entonces existe un punto interior c, por lo menos, en el que $f'(c) = 0$.*

Demostración. Supongamos que f' no es cero en ningún punto de (a, b) y llegaremos a una contradicción. Como que f es continua en un conjunto compacto, alcanza su máximo M y su mínimo m en algún punto de $[a, b]$. Ninguno de dichos valores extremos puede ser alcanzado en un punto interior (pues en ese caso f' se anularía); por lo tanto la función los alcanza en los extremos del intervalo. Como $f(a) = f(b)$, entonces $m = M$, y por lo tanto f es constante en $[a, b]$. Esto contradice el supuesto de que f' no es cero en ningún punto de (a, b). Luego $f'(c) = 0$ para algún c de (a, b).

5.10 TEOREMA DEL VALOR MEDIO PARA DERIVADAS

Teorema 5.11 (Teorema del valor medio). *Sea f una función con derivada (finita o infinta) en cada uno de los puntos de un intervalo abierto (a, b), y supongamos además que f es continua en los extremos a y b. Entonces existe un punto c de (a, b) tal que*

$$f(b) - f(a) = f'(c)(b - a).$$

Geométricamente, este teorema establece que una curva suficientemente regular que una dos puntos A y B posee una tangente con la misma pendiente

que la cuerda AB. El teorema 5.11 lo deduciremos de un teorema más general que se refiere a dos funciones f y g que juegan un papel simétrico.

Teorema 5.12 (Teorema del valor medio generalizado). *Sean f y g dos funciones continuas que poseen derivada (finita o infinita) en cada uno de los puntos del intervalo abierto (a, b) y cada una es continua en los puntos extremos a y b y, además, no existe ningún punto x del interior del intervalo en el que $f'(x)$ y $g'(x)$ sean ambas infinitas. Entonces para algún punto c interior se tiene*

$$f'(c)[g(b) - g(a)] = g'(c)[f(b) - f(a)].$$

NOTA. Cuando $g(x) = x$, se obtiene el teorema 5.11.

Demostración. Sea $h(x) = f(x) \cdot [g(b) - g(a)] - g(x) \cdot [f(b) - f(a)]$. Entonces $h'(x)$ es finito si $f'(x)$ y $g'(x)$ son ambas finitas, y $h'(x)$ es infinito si una de las derivadas $f'(x)$ o $g'(x)$ es infinita. (La hipótesis excluye el caso de que *ambas* sean infinitas.) Además, h es continua en los extremos a y b, y $h(a) = h(b) = f(a)g(b) - g(a)f(b)$. Por el teorema de Rolle existe un punto interior c en el que $h'(c) = 0$, lo que demuestra la proposición.

NOTA. El lector podrá interpretar el teorema 5.12 geométricamente, refiriéndolo a la curva del plano coordenado xy cuyas ecuaciones paramétricas son $x = g(t)$, $y = f(t)$, $a \leq t \leq b$.

Existe una extensión de este teorema que no requiere la hipótesis de continuidad en los extremos.

Teorema 5.13. *Sean f y g dos funciones, cada una de ellas con derivada (finita o infinita) en cada punto de (a, b). Supongamos también que en los extremos a y b existen los límites $f(a+)$, $g(a+)$, $f(b-)$, $g(b-)$ y son finitos. Supongamos además que no existe ningún punto x de (a, b) en el que las derivadas $f'(x)$ y $g'(x)$ sean ambas infinitas. Entonces para algún punto interior c tenemos*

$$f'(c)[g(b-) - g(a+)] = g'(c)[f(b-) - f(a+)].$$

Demostración. Definamos dos nuevas funciones F y G en $[a, b]$ como sigue:

$$F(x) = f(x) \quad \text{y} \quad G(x) = g(x) \quad \text{si } x \in (a, b);$$

$$F(a) = f(a+), \quad G(a) = g(a+), \quad F(b) = f(b-), \quad G(b) = g(b-).$$

Material protegido por derechos de autor
Material protegido por derechos de autor

Entonces F y G son continuas en $[a, b]$ y podemos aplicar el teorema 5.12 a F y G a fin de obtener la conclusión deseada.

El resultado que sigue es una consecuencia inmediata del teorema del valor medio.

Teorema 5.14. *Suponemos que f posee una derivada (finita o infinita) en cada uno de los puntos del intervalo (a, b) y que f es continua en los extremos a y b.*

a) *Si f' toma sólo valores positivos (finitos o infinitos) en (a, b), entonces f es estrictamente creciente en $[a, b]$.*
b) *Si f' toma sólo valores negativos (finitos o infinitos) en (a, b), entonces f es estrictamente decreciente en $[a, b]$.*
c) *Si f' es cero en todo (a, b), entonces f es constante en $[a, b]$.*

Demostración. Elijamos $x < y$ y apliquemos el teorema del valor medio al subintervalo $[x, y]$ de $[a, b]$. Obtendremos

$$f(y) - f(x) = f'(c)(y - x) \qquad \text{donde } c \in (x, y).$$

Todas las afirmaciones del teorema siguiente se deducen inmediatamente de esta ecuación.

Aplicando el teorema 5.14(c) a la diferencia $f - g$ se obtiene:

Corolario 5.15. *Si f y g son continuas en $[a, b]$ y tienen derivadas finitas iguales en (a, b), entonces $f - g$ es constante en $[a, b]$.*

5.11 TEOREMA DEL VALOR INTERMEDIO PARA LAS DERIVADAS

En el teorema 4.33 se ha demostrado que una función f continua en un intervalo compacto $[a, b]$ alcanza todos los valores comprendidos entre su máximo y su mínimo en el intervalo. En particular, f alcanza cada uno de los valores comprendidos entre $f(a)$ y $f(b)$. Un resultado análogo es válido para las funciones que se obtienen como derivadas de otras.

Teorema 5.16 (teorema del valor intermedio para derivadas). *Supongamos que f está definida en un intervalo compacto $[a, b]$ y que posee derivada (finita o infinita) en cada uno de los puntos interiores. Supongamos, además, que f posee derivadas laterales finitas $f'_+(a)$ y $f'_-(b)$ en los puntos extremos, con $f'_+(a) \neq f'_-(b)$. Entonces, si c es un número real comprendido entre $f'_+(a)$ y $f'_-(b)$, existe por lo menos un punto interior x tal que $f'(x) = c$.*

Demostración. Definamos una nueva función g como sigue:

$$g(x) = \frac{f(x) - f(a)}{x - a} \qquad \text{si } x \neq a, \quad g(a) = f'_+(a).$$

Entonces g es continua en el intervalo cerrado $[a, b]$. Por el teorema del valor intermedio de las funciones continuas, g alcanza cada uno de los valores comprendidos entre $f'_+(a)$ y $[f(b) - f(a)]/(b - a)$ en el interior de (a, b). Por el teorema del valor medio, tenemos que $g(x) = f'(c)$ para algún c de (a, x), en donde $x \in (a, b)$. Por lo tanto, f' toma todos los valores entre $f'_+(a)$ y $[f(b) - f(a)]/(b - a)$ en el interior (a, b). Un argumento análogo aplicado a la función h, definida por

$$h(x) = \frac{f(x) - f(b)}{x - b} \qquad \text{si } x \neq b, \quad h(b) = f'_-(b),$$

prueba que f' alcanza todos los valores comprendidos entre $[f(b) - f(a)]/(b - a)$ y $f'_-(b)$ en el interior (a, b). Combinando estos resultados, vemos que f' alcanza cada uno de los valores comprendidos entre $f'_+(a)$ y $f'_-(b)$ en el interior (a, b), lo cual termina la demostración.

NOTA. El teorema 5.16 es asimismo válido si una o ambas derivadas laterales $f'_+(a)$ y $f'_-(b)$ es infinita. La demostración en este caso se obtiene considerando la función auxiliar g definida por medio de la ecuación $g(x) = f(x) - cx$, si $x \in [a, b]$. Los detalles se dejan al lector.

El teorema del valor intermedio demuestra que una derivada no puede cambiar de signo en un intervalo si no toma el valor cero. Por lo tanto, tenemos el siguiente teorema, que es más fuerte que el 5.14(a) y (b).

Teorema 5.17. *Sea f con derivada (finita o infinita) en (a, b) y continua en los extremos a y b. Si $f'(x) \neq 0$ para todo x de (a, b), entonces f es estrictamente monótona en a, b.*

El teorema del valor intermedio prueba también que las derivadas monótonas son necesariamente continuas.

Teorema 5.18. *Supongamos que f' existe y es monótona en un intervalo abierto (a, b). Entonces f' es continua en (a, b).*

Material protegido por derechos de autor
Material protegido por derechos de autor

$$\mathbf{f}'(c) = (f_1'(c), \ldots, f_n'(c)).$$

En otras palabras, la derivada $\mathbf{f}'(c)$ se obtiene diferenciando cada una de las componentes de $\mathbf{f}$ en c. A la vista de esta definición no es sorprendente que nos preguntemos cuáles de los teoremas de diferenciación son válidos para funciones vectoriales. Por ejemplo, si $\mathbf{f}$ y $\mathbf{g}$ son funciones vectoriales diferenciables en c y si λ es una función real diferenciable en c, entonces la suma $\mathbf{f} + \mathbf{g}$, el producto $\lambda\mathbf{f}$, y el producto escalar $\mathbf{f}\cdot\mathbf{g}$ son diferenciables en c y se tiene

$$(\mathbf{f} + \mathbf{g})'(c) = \mathbf{f}'(c) + \mathbf{g}'(c),$$
$$(\lambda\mathbf{f})'(c) = \lambda'(c)\mathbf{f}(c) + \lambda(c)\mathbf{f}'(c),$$
$$(\mathbf{f}\cdot\mathbf{g})'(c) = \mathbf{f}'(c)\cdot\mathbf{g}(c) + \mathbf{f}(c)\cdot\mathbf{g}'(c).$$

Las demostraciones se obtienen fácilmente si se consideran las componentes. Existe también una regla de la cadena para diferenciar funciones compuestas que se prueba de la misma manera. Si $\mathbf{f}$ es vectorial y u es real, entonces la función compuesta $\mathbf{g}$, dada por $\mathbf{g}(x) = \mathbf{f}[u(x)]$, es vectorial. La regla de la cadena establece que

$$\mathbf{g}'(c) = \mathbf{f}'[u(c)]u'(c),$$

si el dominio de $\mathbf{f}$ contiene un entorno de $u(c)$ y si $u'(c)$ y $\mathbf{f}'[u(c)]$ existen.

El teorema del valor medio establecido en el teorema 5.11 no se verifica en el caso de funciones vectoriales. Por ejemplo, si $\mathbf{f}(t) = (\cos t, \operatorname{sen} t)$ para todo t real, entonces

$$\mathbf{f}(2\pi) - \mathbf{f}(0) = \mathbf{0},$$

pero $\mathbf{f}'(t)$ no es nunca cero. De hecho, $\|\mathbf{f}'(t)\| = 1$ para todo t. Una versión modificada del teorema del valor medio para funciones vectoriales será desarrollada en el capítulo 12 (teorema 12.8).

5.14 DERIVADAS PARCIALES

Sea S un conjunto abierto del espacio euclídeo $\mathbf{R}^n$, y sea $f: S \to \mathbf{R}$ una función real definida en S. Si $\mathbf{x} = (x_1, \ldots, x_n)$ y $\mathbf{c} = (c_1, \ldots, c_n)$ son dos puntos de S con las coordenadas correspondientes iguales excepto en el k-ésimo lugar, esto es si $x_i = c_i$ para $i \neq k$ y si $x_k \neq c_k$, entonces podemos considerar el límite

$$\lim_{x_k \to c_k} \frac{f(\mathbf{x}) - f(\mathbf{c})}{x_k - c_k}.$$

Material protegido por derechos de autor

Cuando este límite existe, se le llama *derivada parcial* de f con respecto de la k-ésima coordenada y se designa por medio de

$$D_k f(\mathbf{c}), \qquad f_k(\mathbf{c}), \qquad \frac{\partial f}{\partial x_k}(\mathbf{c}),$$

o por alguna otra expresión análoga. Nosotros adoptaremos la notación $D_k f(\mathbf{c})$.

Este proceso produce n nuevas funciones $D_1 f$, $D_2 f$, ..., $D_n f$ definidas en los puntos de S en los que los correspondientes límites existen.

La diferenciación parcial no es, realmente, un nuevo concepto. Podemos considerar a $f(x_1, \ldots, x_n)$ como una función de una sola variable cada vez, dejando las demás fijas. Es decir, si introducimos una función g definida por

$$g(x_k) = f(c_1, \ldots, c_{k-1}, x_k, c_{k+1}, \ldots, c_n),$$

entonces la derivada parcial $D_k f(\mathbf{c})$ es precisamente la derivada ordinaria $g'(c_k)$. Esto se enuncia usualmente diciendo que para diferenciar f con respecto a la k-ésima variable, se suponen constantes las otras variables.

Siempre que tengamos que generalizar un concepto de $\mathbf{R}^1$ a $\mathbf{R}^n$ procuraremos conservar las propiedades más importantes que, en el caso unidimensional, la existencia de la derivada en c implica la continuidad en c. Por lo tanto, lo óptimo sería disponer un concepto de derivada para funciones de varias variables que implicara la continuidad. Para las derivadas parciales *no* ocurre esto. Una función de n variables puede poseer derivadas parciales en un punto con respecto de cada una de las variables y no ser continua en dicho punto. Ilustraremos esta afirmación por medio del ejemplo de una función con dos variables:

$$f(x, y) = \begin{cases} x + y, & \text{si } x = 0 \text{ o } y = 0, \\ 1, & \text{en otro caso.} \end{cases}$$

Las derivadas parciales $D_1 f(0, 0)$ y $D_2 f(0, 0)$ existen ambas. En efecto:

$$D_1 f(0, 0) = \lim_{x \to 0} \frac{f(x, 0) - f(0, 0)}{x - 0} = \lim_{x \to 0} \frac{x}{x} = 1,$$

y, análogamente, $D_2 f(0, 0) = 1$. Por otro lado, es claro que esta función no es continua en $(0, 0)$.

La existencia de las derivadas parciales con respecto de cada variable separadamente implica la continuidad con respecto de cada variable separadamente; pero como hemos visto, ello no implica necesariamente la continuidad respecto de todas las variables simultáneamente. La dificultad que presentan

Material protegido por derechos de autor

las derivadas parciales proviene de su misma definición: en ella estamos obligados a considerar sólo una variable cada vez. Las derivadas parciales nos proporcionan una medida de la variación de una función en la dirección de cada uno de los ejes. Existe un concepto más general de derivada que no restringe nuestras consideraciones a las direcciones particulares de los ejes coordenados. Este concepto será desarrollado en el capítulo 12.

El propósito de esta sección es únicamente el de introducir la notación de las derivadas parciales, ya que las utilizaremos ocasionalmente antes de alcanzar el capítulo 12.

Si f tiene derivadas parciales $D_1 f, \ldots, D_n f$ en un conjunto abierto S, entonces podemos también considerar *sus* derivadas parciales. Éstas se llamarán derivadas parciales de *segundo orden*. Escribiremos $D_{r,k}f$ para designar la derivada parcial de $D_k f$ con respecto de la r-ésima variable. Entonces,

$$D_{r,k}f = D_r(D_k f).$$

Las derivadas parciales de orden superior se definen análogamente. Otras notaciones son

$$D_{r,k}f = \frac{\partial^2 f}{\partial x_r\, \partial x_k}, \qquad D_{p,q,r}f = \frac{\partial^3 f}{\partial x_p\, \partial x_q\, \partial x_r}.$$

5.15 DIFERENCIACIÓN DE FUNCIONES DE UNA VARIABLE COMPLEJA

En esta sección discutiremos brevemente las derivadas de las funciones complejas definidas en subconjuntos del plano complejo. Tales funciones son, naturalmente, funciones vectoriales cuyo dominio y recorrido son subconjuntos de $\mathbf{R}^2$. Todas las consideraciones del capítulo 4 concernientes a los límites y a la continuidad de las funciones vectoriales se aplican, en particular, a las funciones de una variable compleja. Existe, sin embargo, una diferencia esencial entre el conjunto $\mathbf{C}$ de los números complejos y el conjunto $\mathbf{R}^n$ de los vectores n dimensionales (cuando $n > 2$) que juega un importante papel en este momento. En el sistema de los números complejos disponemos de las cuatro operaciones algebraicas de sumar, restar, multiplicar y dividir, y estas operaciones verifican muchas de las propiedades «usuales» del Álgebra que son válidas en el sistema de los números reales. En particular verifican los cinco primeros axiomas de los números reales enumerados en el capítulo 1. (Los axiomas 6 al 10 involucran la relación de orden $<$, que no existe entre números complejos.) Todo sistema algebraico que verifica los axiomas 1 al 5 se llama *cuerpo*. (Para una discusión más amplia de los cuerpos, véase la referencia 1.4.)

La multiplicación y la división no pueden ser introducidas en $\mathbf{R}^n$ (para $n > 2$) de forma que $\mathbf{R}^n$ sea un cuerpo * que contenga a $\mathbf{C}$. Como la división es posible en $\mathbf{C}$, es posible asimismo formar el cociente fundamental de diferencias $[f(z) - f(c)]/(z - c)$ que fue utilizado para definir la derivada en $\mathbf{R}$, y entonces se presenta de forma clara cómo hay que definir la derivada en $\mathbf{C}$.

Definición 5.21. *Sea f una función compleja definida en un conjunto abierto S de $\mathbf{C}$, y sea $c \in S$. Entonces f es diferenciable en c si el límite*

$$\lim_{z \to c} \frac{f(z) - f(c)}{z - c} = f'(c)$$

existe.

Por medio de este proceso de calcular límites se obtiene una nueva función compleja f' definida en aquellos puntos z de S donde $f'(z)$ existe. Las derivadas de orden superior f'', f''', ... se definen, como es natural, de forma análoga.

Las siguientes proposiciones son válidas para funciones complejas definidas en un conjunto abierto S; sus demostraciones son exactamente las mismas que las utilizadas en el caso real:

a) *f es diferenciable en c si, y sólo si, existe una función f^*, continua en c, tal que*

$$f(z) - f(c) = (z - c)f^*(z),$$

para todo z de S, con $f^*(c) = f'(c)$.

NOTA. Si hacemos $g(z) = f^*(z) - f'(c)$, la ecuación de (a) podemos ponerla en la forma

$$f(z) = f(c) + f'(c)(z - c) + g(z)(z - c),$$

donde $g(z) \to 0$ cuando $z \to c$. Esta expresión se llama la *fórmula de Taylor de primer orden* para f.

* Por ejemplo, si fuese posible definir una multiplicación en $\mathbf{R}^3$ que dotara a $\mathbf{R}^3$ de estructura de cuerpo, conteniendo a $\mathbf{C}$, podríamos razonar como sigue: Para cada $\mathbf{x}$ de $\mathbf{R}^3$ los vectores $1, \mathbf{x}, \mathbf{x}^2, \mathbf{x}^3$ serían linealmente dependientes (ver Referencia 5.1, p. 558). Entonces para cada $\mathbf{x}$ de $\mathbf{R}^3$, se verificaría una relación del tipo $a_0 + a_1\mathbf{x} + a_2\mathbf{x}^2 + a_3\mathbf{x}^3 = 0$, donde a_0, a_1, a_2, a_3 son números reales. Pero cada polinomio de grado tres con coeficientes reales es un producto de un polinomio lineal por un polinomio cuadrático con coeficientes reales. Las únicas raíces de tales polinomios son o bien números reales o bien números complejos.

Material protegido por derechos de autor
Material protegido por derechos de autor

b) *Si f es diferenciable en c, entonces f es continua en c.*
c) *Si dos funciones f y g tienen derivadas en c, entonces su suma, su diferencia, su producto y su cociente tienen también derivadas en c y se obtienen por medio de las fórmulas usuales (como en el teorema 5.4). En el caso de f/g, debemos suponer que $g(c) \neq 0$.*
d) *La regla de la cadena es válida; es decir, tenemos*

$$(g \circ f)'(c) = g'[f(c)]f'(c),$$

si el dominio de g contiene un entorno de $f(c)$ y si $f'(c)$ y $g'[f(c)]$ existen.

Si $f(z) = z$, se obtiene $f'(z) = 1$ para todo z de $\mathbf{C}$. Por (c) reiterado, se tiene que $f'(z) = nz^{n-1}$ cuando $f(z) = z^n$ (n es un entero positivo). Esto también se verifica cuando n es un entero negativo, siempre que $z \neq 0$. Por lo tanto, es posible calcular las derivadas de los polinomios complejos y de las funciones racionales complejas utilizando las mismas técnicas que las empleadas en el Cálculo elemental.

5.16 ECUACIONES DE CAUCHY-RIEMANN

Si f es una función compleja de una variable compleja, podemos escribir cada valor de la función en la forma

$$f(z) = u(z) + iv(z),$$

donde u y v son funciones reales de una variable compleja. Podemos, además, considerar u y v como funciones reales de dos variables reales y escribir entonces

$$f(z) = u(x, y) + iv(x, y), \qquad \text{si } z = x + iy.$$

En ambos casos, escribiremos $f = u + iv$ y nos referiremos a u y a v designándolas parte *real* y parte *imaginaria* de f. Así, en el caso de la función exponencial compleja f, definida por

$$f(z) = e^z = e^x \cos y + ie^x \operatorname{sen} y,$$

las partes real e imaginaria vienen dadas por

$$u(x, y) = e^x \cos y, \qquad v(x, y) = e^x \operatorname{sen} y.$$

Análogamente, cuando $f(z) = z^2 = (x + iy)^2$, obtenemos

$$u(x, y) = x^2 - y^2, \qquad v(x, y) = 2xy.$$

Material protegido por derechos de autor

condición no es, sin embargo, suficiente, como podemos ver considerando el siguiente ejemplo.

Ejemplo. Sean u y v definidas como sigue:

$$u(x, y) = \frac{x^3 - y^3}{x^2 + y^2} \quad \text{si } (x, y) \neq (0, 0), \quad u(0, 0) = 0,$$

$$v(x, y) = \frac{x^3 + y^3}{x^2 + y^2} \quad \text{si } (x, y) \neq (0, 0), \quad v(0, 0) = 0.$$

Es fácil comprobar que $D_1u(0, 0) = D_1v(0, 0) = 1$ y que $D_2u(0, 0) = -D_2v(0, 0) = -1$, por lo tanto las ecuaciones de Cauchy-Riemann se verifican en $(0, 0)$. A pesar de todo, la función $f = u + iv$ no puede tener derivada en $z = 0$. En efecto, para $x = 0$, el cociente diferencial se convierte en

$$\frac{f(z) - f(0)}{z - 0} = \frac{-y + iy}{iy} = 1 + i,$$

mientras que para $x = y$, es

$$\frac{f(z) - f(0)}{z - 0} = \frac{xi}{x + ix} = \frac{1 + i}{2},$$

y por lo tanto $f'(0)$ no existe.

En el capítulo 12 demostraremos que las ecuaciones de Cauchy-Riemann *son* suficientes para establecer la existencia de la derivada de $f = u + iv$ en c si las derivadas parciales de u y v son continuas en un entorno de c. Para ilustrar cómo hay que utilizar este resultado en la práctica, obtendremos la derivada de la función exponencial. Sea $f(z) = e^z = u + iv$. Entonces

$$u(x, y) = e^x \cos y, \qquad v(x, y) = e^x \operatorname{sen} y,$$

y por lo tanto

$$D_1u(x, y) = e^x \cos y = D_2v(x, y), \qquad D_2u(x, y) = -e^x \operatorname{sen} y = -D_1v(x, y).$$

Como estas derivadas parciales son continuas en todo $\mathbf{R}^2$ y satisfacen las ecuaciones de Cauchy-Riemann, la derivada $f'(z)$ existe para todo z. Para calcularla usaremos el teorema 5.22 y obtendremos

$$f'(z) = e^x \cos y + ie^x \operatorname{sen} y = f(z).$$

Entonces, la función exponencial es su misma derivada (como en el caso real).

Material protegido por derechos de autor

EJERCICIOS

Funciones reales

En los ejercicios que siguen se supone, siempre que sea necesario, que se conocen las fórmulas para derivar las funciones elementales trigonométricas, exponenciales y logarítmicas.

5.1 Una función f satisface una condición de Lipschitz de orden α en c si existe un número positivo M (que puede depender de c) y una bola unidimensional $B(c)$ tales que

$$|f(x) - f(c)| < M|x - c|^{\alpha}$$

si $x \in B(c)$, $x \neq c$.

a) Probar que una función que satisface una condición de Lipschitz de orden α es continua en c si $\alpha > 0$, y derivable en c si $\alpha > 1$.

b) Dar un ejemplo de una función que satisfaga la condición de Lipschitz de orden 1 en c para la que $f'(c)$ no exista.

5.2 En cada uno de los siguientes casos, determinar los intervalos en los que la función f es creciente o decreciente y determinar los máximos y mínimos (si existen) en el conjunto en el que f está definida.

a) $f(x) = x^3 + ax + b$, $\quad x \in \mathbf{R}$.

b) $f(x) = \log(x^2 - 9)$, $\quad |x| > 3$.

c) $f(x) = x^{2/3}(x - 1)^4$, $\quad 0 \leq x \leq 1$.

d) $f(x) = (\operatorname{sen} x)/x$ if $x \neq 0$, $f(0) = 1$, $\quad 0 \leq x \leq \pi/2$.

5.3 Buscar un polinomio f del menor grado posible tal que

$$f(x_1) = a_1, \quad f(x_2) = a_2, \quad f'(x_1) = b_1, \quad f'(x_2) = b_2,$$

donde $x_1 \neq x_2$ y a_1, a_2, b_1, b_2 son números reales dados.

5.4 Se define f como sigue: $f(x) = e^{-1/x^2}$ si $x \neq 0$, $f(0) = 0$. Probar que

a) f es continua para todo x.

b) $f^{(n)}$ es continua para todo x, y que $f^{(n)}(0) = 0$, $(n = 1, 2, \ldots)$.

5.5 Definimos f, g y h como sigue: $f(0) = g(0) = h(0)$ y, si $x \neq 0$, $f(x) = \operatorname{sen}(1/x)$, $g(x) = x \operatorname{sen}(1/x)$, $h(x) = x^2 \operatorname{sen}(1/x)$. Probar que

a) $f'(x) = -1/x^2 \cos(1/x)$, si $x \neq 0$; $\quad f'(0)$ no existe.

b) $g'(x) = \operatorname{sen}(1/x) - 1/x \cos(1/x)$, si $x \neq 0$; $\quad g'(0)$ no existe.

c) $h'(x) = 2x \operatorname{sen}(1/x) - \cos(1/x)$, si $x \neq 0$; $\quad h'(0) = 0$; $\lim_{x \to 0} h'(x)$ no existe.

5.6 Obtener la fórmula de Leibnitz para la derivada n-ésima del producto h de dos funciones f y g:

$$h^{(n)}(x) = \sum_{k=0}^{n} \binom{n}{k} f^{(k)}(x) g^{(n-k)}(x), \quad \text{donde} \quad \binom{n}{k} = \frac{n!}{k!\,(n-k)!}.$$

Material protegido por derechos de autor

5.7 Sean f y g dos funciones definidas en todo $\mathbf{R}$ y con derivadas finitas terceras $f'''(x)$ y $g'''(x)$ para todo x de $\mathbf{R}$. Si $f(x)\,g(x) = 1$ para todo x, probar que las relaciones de (a), (b), (c) y (d) se verifican en todos los puntos en los que el denominador no es cero:

a) $f'(x)/f(x) + g'(x)/g(x) = 0$.

b) $f''(x)/f'(x) - 2f'(x)/f(x) - g''(x)/g'(x) = 0$.

c) $\dfrac{f'''(x)}{f'(x)} - 3\dfrac{f'(x)g''(x)}{f(x)g'(x)} - 3\dfrac{f''(x)}{f(x)} - \dfrac{g'''(x)}{g'(x)} = 0$.

d) $\dfrac{f'''(x)}{f'(x)} - \dfrac{3}{2}\left(\dfrac{f''(x)}{f'(x)}\right)^2 = \dfrac{g'''(x)}{g'(x)} - \dfrac{3}{2}\left(\dfrac{g''(x)}{g'(x)}\right)^2$.

NOTA. La expresión que aparece en el primer miembro de (d) se llama *derivada de Schwarz* de f en x.

e) Probar que f y g tienen la misma derivada de Schwarz si

$$g(x) = [af(x) + b]/[cf(x) + d], \text{donde } ad - bc \neq 0.$$

Indicación. Si $c \neq 0$, escribir $(af + b)/(cf + d) = (a/c) + (bc - ad)/[c(cf + d)]$ y aplicar la parte (d).

5.8 Sean f_1, f_2, g_1, g_2 cuatro funciones con derivadas en (a, b). Definamos F por medio del determinante

$$F(x) = \begin{vmatrix} f_1(x) & f_2(x) \\ g_1(x) & g_2(x) \end{vmatrix}, \quad \text{si } x \in (a, b).$$

a) Probar que $F'(x)$ existe para cada x de (a, b) y que

$$F'(x) = \begin{vmatrix} f_1'(x) & f_2'(x) \\ g_1(x) & g_2(x) \end{vmatrix} + \begin{vmatrix} f_1(x) & f_2(x) \\ g_1'(x) & g_2'(x) \end{vmatrix}.$$

b) Establecer y probar un resultado más general para determinantes de orden n.

5.9 Dadas n funciones $f_1, \ldots, f_n$ derivables hasta el orden n en (a, b), definimos una función W, llamada el *Wronskiano* de $f_1, \ldots, f_n$ como sigue: Para cada x de (a, b), $W(x)$ es el valor del determinante de orden n que en la k-ésima fila, m-ésima columna, tiene al elemento $f_m^{(k-1)}(x)$, donde $k = 1, 2, \ldots, n$ y $m = 1, 2, \ldots, n$. [La expresión $f_m^{(0)}(x)$ designa a $f_m(x)$.]

a) Probar que $W'(x)$ se obtiene reemplazando la última fila del determinante que define $W(x)$ por las derivadas n-ésimas $f_1^{(n)}(x), \ldots, f_n^{(n)}(x)$.

b) Supongamos que existen n constantes $c_1, \ldots, c_n$, no nulas, tales que $c_1 f_1(x) + \ldots + c_n f_n(x) = 0$ para todo x de (a, b). Probar que $W(x) = 0$ para cada x en (a, b).

NOTA. Un conjunto de funciones que satisface una relación de este tipo se llama *conjunto linealmente dependiente* sobre (a, b).

c) La anulación del Wronskiano en todo el intervalo (a, b) es necesaria, pero no es suficiente para que $f_1, \ldots, f_n$ sean linealmente dependientes. Probar

Material protegido por derechos de autor

que en el caso de dos funciones, si el Wronskiano se anula en (a, b) y si una de las funciones no se anula en (a, b), entonces constituyen un conjunto linealmente dependiente en (a, b).

El teorema del valor medio

5.10 Dada una función f definida y derivable con derivada finita en (a, b) y tal que $\lim_{x\to b-} f(x) = +\infty$, probar que $\lim_{x\to b-} f'(x)$ o no existe o es infinito.

5.11 Probar que la fórmula del valor medio puede escribirse en la forma:

$$\frac{f(x+h)-f(x)}{h} = f'(x+\theta h),$$

donde $0 < \theta < 1$. Determinar θ en función de x y de h cuando

a) $f(x) = x^2$, b) $f(x) = x^3$,

c) $f(x) = e^x$, d) $f(x) = \log x$, $x > 0$.

Fijar $x \neq 0$ y hallar $\lim_{h\to 0} \theta$ en cada caso.

5.12 En el teorema 5.20 hacemos $f(x) = 3x^4 - 2x^3 - x^2 + 1$ y $g(x) = 4x^3 - 3x^2 - 2x$. Probar que $f'(x)/g'(x)$ nunca es igual al cociente $[f(1) - f(0)]/[g(1) - g(0)]$ si $0 < x \leq 1$. ¿Cómo conciliar esto con la igualdad

$$\frac{f(b)-f(a)}{g(b)-g(a)} = \frac{f'(x_1)}{g'(x_1)}, \qquad a < x_1 < b,$$

que se obtiene del teorema 5.20 cuando $n = 1$?

5.13 En cada uno de los casos especiales del teorema 5.20, tomar $n = 1$, $c = a$, $x = b$, y demostrar que $x_1 = (a + b)/2$.

a) $f(x) = \operatorname{sen} x$, $g(x) = \cos x$; b) $f(x) = e^x$, $g(x) = e^{-x}$.

¿Es posible encontrar una clase general de pares de funciones f y g para los que x_1 sea siempre $(a + b)/2$ y tales que los ejemplos (a) y (b) pertenezcan a dicha clase?

5.14 Dada una función f definida y con derivada finita f' en el intervalo semiabierto $0 < x \leq 1$ y tal que $|f'(x)| < 1$, definimos $a_n = f(1/n)$ para $n = 1, 2, 3, \ldots$ Demostrar que el $\lim_{n\to\infty} a_n$ existe. *Indicación.* Utilizar la condición de Cauchy.

5.15 Supongamos que f posee derivada finita en cada uno de los puntos del intervalo abierto (a, b). Supongamos además que $\lim_{x\to c} f'(x)$ existe y es finito para uno de los puntos interiores c. Demostrar que el valor de este límite deberá ser $f'(c)$.

5.16 Sea f continua en (a, b) con derivada finita f' en todo (a, b), excepto quizás en c. Si $\lim_{x\to c} f'(x)$ existe y vale A, entonces $f'(c)$ existe también y vale A.

5.17 Sea f continua en $[0, 1]$, $f(0) = 0$, $f'(x)$ finito para cada x de $(0, 1)$. Probar que si f' es creciente en $(0, 1)$, entonces también lo es la función g definida por medio de la ecuación $g(x) = f(x)/x$.

5.18 Supongamos que f posee derivada finita en (a, b) y es continua en $[a, b]$ con $f(a) = f(b) = 0$. Probar que para cada real λ existe un c de (a, b) tal que $f'(c) = \lambda f(c)$. *Indicación.* Aplicar el teorema de Rolle a $g(x)f(x)$ para una g conveniente que dependa de λ.

Material protegido por derechos de autor

5.19 Supongamos que f es continua en $[a, b]$ y que posee una derivada segunda f'' finita en el intervalo abierto (a, b). Supongamos que la cuerda que une los puntos $A = (a, f(a))$ y $B = (b, f(b))$ corta a la gráfica de la función f en un tercer punto P distinto de A y de B. Probar que $f''(c) = 0$ para un c de (a, b).

5.20 Si f posee derivada tercera f''' finita en $[a, b]$ y si

$$f(a) = f'(a) = f(b) = f'(b) = 0,$$

probar que $f'''(c) = 0$ para un c de (a, b).

5.21 Sea f una función no negativa y que admita tercera derivada finita f''' en el intervalo abierto $(0, 1)$. Si $f(x) = 0$ para dos puntos, por lo menos, de x en $(0, 1)$, entonces $f'''(c) = 0$ para un c de $(0, 1)$.

5.22 Supongamos que f admite una derivada finita en un cierto intervalo $(a, +\infty)$.

a) Si $f(x) \to 1$ y $f'(x) \to c$ cuando $x \to +\infty$, probar que $c = 0$.

b) Si $f'(x) \to 1$ cuando $x \to +\infty$, probar que $f(x)/x \to 1$ cuando $x \to +\infty$.

c) Si $f'(x) \to 0$ cuando $x \to +\infty$, probar que $f(x)/x \to 0$ cuando $x \to +\infty$.

5.23 Sea h un número positivo fijo. Probar que no existe ninguna función f que satisfaga las tres condiciones siguientes: $f'(x)$ existe para $x \geq 0$, $f'(0) = 0$, $f'(x) \geq h$ para $x > 0$.

5.24 Si $h > 0$ y $f'(x)$ existe (y es finita) para cada x de $(a - h, a + h)$, y si f es continua en $[a - h, a + h]$, probar que se tiene:

a) $\dfrac{f(a+h)-f(a-h)}{h} = f'(a+\theta h) + f'(a-\theta h), \qquad 0 < \theta < 1;$

b) $\dfrac{f(a+h)-2f(a)+f(a-h)}{h} = f'(a+\lambda h) - f'(a-\lambda h), \qquad 0 < \lambda < 1.$

c) Si $f''(a)$ existe, probar

$$f''(a) = \lim_{h\to 0} \frac{f(a+h)-2f(a)+f(a-h)}{h^2}.$$

d) Dar un ejemplo en el que exista el límite del cociente que aparece en (c) pero $f''(a)$ no exista.

5.25 Sea f una función con derivada finita en (a, b) y supongamos que $c \in (a, b)$. Considerar la siguiente condición: Para cada $\varepsilon > 0$ existe una bola unidimensional $B(c; \delta)$, cuyo radio δ depende sólo de ε y no de c, tal que si $x \in B(c; \delta)$ y $x \neq c$, entonces

$$\left| \frac{f(x)-f(c)}{x-c} - f'(c) \right| < \varepsilon.$$

Probar que f' es continua en (a, b) si esta condición se verifica en todo (a, b).

5.26 Sea f con derivada finita en (a, b) y continua en $[a, b]$, con $a \leq f(x) \leq b$ para todo x de $[a, b]$ y $|f'(x)| \leq \alpha < 1$ para todo x de (a, b). Probar que f posee un único punto fijo en $[a, b]$.

Material protegido por derechos de autor

REFERENCIAS SUGERIDAS PARA POSTERIORES ESTUDIOS

5.1 Apostol, T. M., *Calculus*, Vol. 1. 2.ª ed. Ed. Reverté, S. A., Barcelona, Bogotá, Buenos Aires, Caracas, México.
5.2 Chaundy, T. W., *The Differential Calculus*. Clarendon Press, Oxford, 1935.

CAPÍTULO 6

Funciones de variación acotada y curvas rectificables

6.1 INTRODUCCIÓN

Algunas de las propiedades básicas de las funciones monótonas fueron descritas en el capítulo 4. En este breve capítulo se estudian las funciones de variación acotada, una clase de funciones íntimamente relacionada con las funciones monótonas. Veremos que estas funciones están en estrecha conexión con las curvas que poseen longitud finita (curvas rectificables). Juegan también un papel en la teoría de la integración de Riemann-Stieltjes que desarrollaremos en el próximo capítulo.

6.2 PROPIEDADES DE LAS FUNCIONES MONÓTONAS

Teorema 6.1. *Sea f una función creciente definida en $[a, b]$ y sean $x_0, x_1, \ldots, x_n$ $n+1$ puntos tales que*

$$a = x_0 < x_1 < x_2 < \cdots < x_n = b.$$

Tenemos entonces la desigualdad

$$\sum_{k=1}^{n-1} [f(x_k+) - f(x_k-)] \leq f(b) - f(a).$$

Demostración. Supongamos que $y_k \in (x_k, x_{k+1})$. Para $1 \leq k \leq n-1$ tenemos que $f(x_k+) \leq f(y_k)$ y $f(y_{k-1}) \leq f(x_k-)$, luego $f(x_k+) - f(x_k-) \leq f(y_k) - f(y_{k-1})$. Si sumamos estas desigualdades, la suma de la derecha nos da $f(y_{n-1}) - f(y_0)$. Puesto que $f(y_{n-1}) - f(y_0) \leq f(b) - f(a)$, esto completa la demostración.

La diferencia $f(x_k+) - f(x_k-)$ es, además, el salto de f en x_k. El teorema anterior nos dice que, para cada colección finita de puntos x_k de (a, b), la suma de los saltos en estos puntos está siempre acotada por $f(b) - f(a)$. Este resultado nos servirá para demostrar el teorema siguiente.

Teorema 6.2. *Si f es monótona en $[a, b]$, entonces el conjunto de discontinuidades de f es numerable.*

Material protegido por derechos de autor
Material protegido por derechos de autor

Demostración. Supongamos que f es creciente y sea S_m el conjunto de puntos de (a, b) en los que el salto de f es superior a $1/m$, $m > 0$. Si $x_1 < x_2 < \ldots < x_{n-1}$ están en S_m, el teorema 6.1 nos asegura que

$$\frac{n-1}{m} \leq f(b) - f(a).$$

Esto significa que S_m debe ser un conjunto finito. Pero el conjunto de discontinuidades de f en (a, b) es un subconjunto de la reunión $\bigcup_{m=1}^{\infty} S_m$ y por lo tanto numerable. (Si f es decreciente, el argumento se aplica a $-f$.)

6.3 FUNCIONES DE VARIACIÓN ACOTADA

Definición 6.3. *Si $[a, b]$ es un intervalo compacto, un conjunto de puntos*

$$P = \{x_0, x_1, \ldots, x_n\},$$

que satisfaga las desigualdades

$$a = x_0 < x_1 \cdots < x_{n-1} < x_n = b,$$

se llama partición de $[a, b]$. El intervalo $[x_{k-1}, x_k]$ se llama k-ésimo subintervalo de P y se escribe $\Delta x_k = x_k - x_{k-1}$, con lo que $\sum_{k=1}^{n} \Delta x_k = b - a$. La colección de todas las particiones posibles de $[a, b]$ se designará por medio de $\mathcal{P}[a, b]$.

Definición 6.4. *Sea f definida en $[a, b]$. Si $P = \{x_0, x_1, \ldots, x_n\}$ es una partición de $[a, b]$, escribiremos $\Delta f_k = f(x_k) - f(x_{k-1})$, para $k = 1, 2, \ldots, n$. Si existe un número positivo M tal que*

$$\sum_{k=1}^{n} |\Delta f_k| \leq M$$

para toda partición de $[a, b]$, entonces diremos que f es de variación acotada en $[a, b]$.

Los dos teoremas que siguen proporcionan ejemplos de funciones de variación acotada.

Teorema 6.5. *Si f es monótona en $[a, b]$, entonces f es de variación acotada en $[a, b]$.*

Demostración. Sea f creciente. Entonces para cada partición de $[a, b]$ tenemos $\Delta f_k \geq 0$ y por lo tanto

$$\sum_{k=1}^{n} |\Delta f_k| = \sum_{k=1}^{n} \Delta f_k = \sum_{k=1}^{n} [f(x_k) - f(x_{k-1})] = f(b) - f(a).$$

Material protegido por derechos de autor

Teorema 6.6. *Si f es continua en $[a, b]$ y si f' existe y está acotada en el interior, es decir que $|f'(x)| \leq A$ para todo x de (a, b), entonces f es de variación acotada en $[a, b]$.*

Demostración. Aplicando el teorema del valor medio, tenemos

$$\Delta f_k = f(x_k) - f(x_{k-1}) = f'(t_k)(x_k - x_{k-1}), \text{ en donde } t_k \in (x_{k-1}, x_k).$$

Esto implica

$$\sum_{k=1}^{n} |\Delta f_k| = \sum_{k=1}^{n} |f'(t_k)| \, \Delta x_k \leq A \sum_{k=1}^{n} \Delta x_k = A(b - a).$$

Teorema 6.7. *Si f es de variación acotada en $[a, b]$, es decir que $\sum |\Delta f_k| \leq M$ para toda partición de $[a, b]$, entonces f está acotada en $[a, b]$. De hecho,*

$$|f(x)| \leq |f(a)| + M \quad \textit{para todo } x \textit{ de } [a, b].$$

Demostración. Supongamos que $x \in (a, b)$. Utilizando la partición especial $P = \{a, x, b\}$, obtenemos

$$|f(x) - f(a)| + |f(b) - f(x)| \leq M.$$

Esto implica que $|f(x) - f(a)| \leq M$, $|f(x)| \leq |f(a)| + M$. Idéntica desigualdad se verifica si $x = a$ o si $x = b$.

Ejemplos

1. Es fácil construir funciones continuas que no sean de variación acotada. Por ejemplo, sea $f(x) = x \cos \{\pi/(2x)\}$ si $x \neq 0$, $f(0) = 0$. Entonces f es continua en $[0, 1]$, pero si consideramos la partición en $2n$ subintervalos

$$P = \left\{0, \frac{1}{2n}, \frac{1}{2n-1}, \ldots, \frac{1}{3}, \frac{1}{2}, 1\right\},$$

es fácil comprobar, calculando, que

$$\sum_{k=1}^{2n} |\Delta f_k| = \frac{1}{2n} + \frac{1}{2n} + \frac{1}{2n-2} + \frac{1}{2n-2} + \cdots + \frac{1}{2} + \frac{1}{2} = 1 + \frac{1}{2} + \cdots + \frac{1}{n}.$$

Esta suma no está acotada para todo n, ya que la serie $\sum_{n=1}^{\infty} (1/n)$ diverge. En este ejemplo la derivada f' existe en $(0, 1)$ pero f' no está acotada en $(0, 1)$. Sin embargo, f' está acotada en todo intervalo compacto que no contenga el origen y, por lo tanto, f es de variación acotada en tales intervalos.

Material protegido por derechos de autor

2. Un ejemplo análogo al primero lo proporciona la función $f(x) = x^2 \cos(1/x)$ si $x \neq 0$, $f(0) = 0$. Dicha función f es de variación acotada en $[0, 1]$, ya que f' está acotada en $[0, 1]$. De hecho, $f'(0) = 0$ y, para $x \neq 0$, $f'(x) = \operatorname{sen}(1/x) + 2x \cos(1/x)$, luego $|f'(x)| \leq 3$ para todo x de $[0, 1]$.
3. La acotación de f' no es condición necesaria para que f sea variación acotada. Por ejemplo, considérese la función $f(x) = x^{1/3}$. Esta función es monótona (y por lo tanto de variación acotada) sobre todo intervalo finito. Sin embargo, $f'(x) \to +\infty$ cuando $x \to 0$.

6.4 VARIACIÓN TOTAL

Definición 6.8. *Sea f una función de variación acotada en $[a, b]$ y sea $\sum(P)$ la suma $\sum_{k=1}^{n} |\Delta f_k|$ correspondiente a la partición $P = \{x_0, x_1, \ldots, x_n\}$ de $[a, b]$. El número*

$$V_f(a, b) = \sup \{\textstyle\sum(P) : P \in \mathscr{P}[a, b]\},$$

se llama variación total de f en el intervalo $[a, b]$.

NOTA. Si no hay peligro de confusión, escribiremos V_f en vez de $V_f(a, b)$.

Dado que f es de variación acotada en $[a, b]$, el número V_f es finito. Además $V_f \geq 0$, ya que cada suma $\sum(P) \geq 0$. Y además $V_f(a, b) = 0$ si, y sólo si, f es constante en $[a, b]$.

Teorema 6.9. *Supongamos que f y g son dos funciones de variación acotada en $[a, b]$. Entonces también lo es su suma, su diferencia y su producto. Además, se tiene*

$$V_{f \pm g} \leq V_f + V_g \qquad \text{y} \qquad V_{f \cdot g} \leq AV_f + BV_g,$$

en donde

$$A = \sup \{|g(x)| : x \in [a, b]\}, \qquad B = \sup \{|f(x)| : x \in [a, b]\}.$$

Demostración. Sea $h(x) = f(x)g(x)$. Para cada partición P de $[a, b]$ se tiene

$$\begin{aligned} |\Delta h_k| &= |f(x_k)g(x_k) - f(x_{k-1})g(x_{k-1})| \\ &= |[f(x_k)g(x_k) - f(x_{k-1})g(x_k)] \\ &\quad + [f(x_{k-1})g(x_k) - f(x_{k-1})g(x_{k-1})]| \leq A|\Delta f_k| + B|\Delta g_k|. \end{aligned}$$

Esto implica que h es una función de variación acotada y que $V_h \leq AV_f + BV_g$. Las demostraciones correspondientes a la suma y la diferencia son muy simples y las omitiremos.

Material protegido por derechos de autor

Esto significa que $D(y) - D(x) \geq 0$, y (ii) se verifica.

NOTA. Para ciertas funciones f, la variación total $V_f(a, x)$ se puede expresar como una integral. (Ver ejercicio 7.20.)

6.7. FUNCIONES DE VARIACIÓN ACOTADA EXPRESADAS COMO DIFERENCIA DE DOS FUNCIONES CRECIENTES

La simple y elegante caracterización de las funciones de variación acotada que damos a continuación es consecuencia del teorema 6.12.

Teorema 6.13. *Sea f definida sobre $[a, b]$. Entonces f es de variación acotada en $[a, b]$ si, y sólo si, f puede expresarse como diferencia de dos funciones crecientes.*

Demostración. Si f es de variación acotada en $[a, b]$, podemos escribir $f = V - D$, en donde V es la función del teorema 6.12 y $D = V - f$. Tanto V como D son funciones crecientes en $[a, b]$.

El recíproco se deduce inmediatamente de los teoremas 6.5 y 6.9.

La representación de una función de variación acotada como diferencia de dos funciones crecientes no es única. Si $f = f_1 - f_2$, en donde f_1 y f_2 son crecientes, se tiene también que $f = (f_1 + g) - (f_2 + g)$, siendo g una función creciente arbitraria, y ello nos proporciona una nueva representación de f. Si g es *estrictamente* creciente, también lo serán $f_1 + g$ y $f_2 + g$. Por consiguiente, el teorema 6.13 es asimismo válido si reemplazamos «creciente» por «estrictamente creciente».

6.8 FUNCIONES CONTINUAS DE VARIACIÓN ACOTADA

Teorema 6.14. *Sea f una función de variación acotada en $[a, b]$. Si $x \in (a, b]$, sea $V(x) = V_f(a\ x)$ y hagamos $V(a) = 0$. Entonces cada punto de continuidad de f también es un punto de continuidad de V. El recíproco también es cierto.*

Demostración. Puesto que V es monótona, los límites laterales por la derecha y por la izquierda $V(x+)$ y $V(x-)$ existen para cada punto x de (a, b). En virtud del teorema 6.13, lo mismo es cierto para $f(x+)$ y $f(x-)$.

Si $a < x < y \leq b$, se verifica [por definición de $V_f(x, y)$] que

$$0 \leq |f(y) - f(x)| \leq V(y) - V(x).$$

Haciendo que $y \to x$, obtenemos

$$0 \leq |f(x+) - f(x)| \leq V(x+) - V(x).$$

Material protegido por derechos de autor

Análogamente, $0 \leq |f(x) - f(x-)| \leq V(x) - V(x-)$. Estas desigualdades implican que todo punto de continuidad de V es también un punto de continuidad de f.

Para demostrar el recíproco, sea f continua en un punto c de (a, b). Entonces, dado $\varepsilon > 0$, existe un $\delta > 0$ tal que $0 < |x - c| < \delta$ implica $|f(x) - f(c)| < \varepsilon/2$. Para este mismo ε, existe una partición P de $[c, b]$, por ejemplo

$$P = \{x_0, x_1, \ldots, x_n\}, \qquad x_0 = c, \quad x_n = b,$$

tal que

$$V_f(c, b) - \frac{\varepsilon}{2} < \sum_{k=1}^{n} |\Delta f_k|.$$

Agregando más puntos a P únicamente conseguiremos que aumente la suma $\sum |\Delta f_k|$ y por lo tanto podemos considerar que $0 < x_1 - x_0 < \delta$. Esto significa que

$$|\Delta f_1| = |f(x_1) - f(c)| < \frac{\varepsilon}{2},$$

con lo que la desigualdad anterior se convierte en

$$V_f(c, b) - \frac{\varepsilon}{2} < \frac{\varepsilon}{2} + \sum_{k=2}^{n} |\Delta f_k| \leq \frac{\varepsilon}{2} + V_f(x_1, b),$$

ya que $\{x_1, x_2, \ldots, x_n\}$ es una partición de $[x_1, b]$. Tenemos, por tanto,

$$V_f(c, b) - V_f(x_1, b) < \varepsilon.$$

Pero

$$\begin{aligned} 0 \leq V(x_1) - V(c) &= V_f(a, x_1) - V_f(a, c) \\ &= V_f(c, x_1) = V_f(c, b) - V_f(x_1, b) < \varepsilon. \end{aligned}$$

Con lo cual hemos probado que

$$0 < x_1 - c < \delta \qquad \text{implica} \qquad 0 \leq V(x_1) - V(c) < \varepsilon.$$

Esto demuestra que $V(c+) = V(c)$. Un razonamiento análogo lleva al resultado $V(c-) = V(c)$. El teorema queda entonces demostrado para todos los puntos interiores de $[a, b]$. (Para los puntos extremos son necesarias ciertas modificaciones triviales.)

Combinando el teorema 6.14 con el 6.13, podemos establecer

Teorema 6.15. *Sea f una función continua en $[a, b]$. Entonces f es variación acotada en $[a, b]$ si, y sólo si, f se puede expresar como diferencia de dos funciones crecientes continuas.*

NOTA. El teorema se verifica también si reemplazamos «creciente» por «estrictamente creciente».

Es claro que las discontinuidades de una función de variación acotada (si existen) deberán ser discontinuidades de salto en virtud del teorema 6.13. Además, el teorema 6.2 nos dice que constituyen un conjunto numerable.

6.9 CURVAS Y CAMINOS

Sea $\mathbf{f}: [a, b] \to \mathbf{R}^n$ una función vectorial, continua en un intervalo compacto $[a, b]$ de $\mathbf{R}$. Cuando t recorre $[a, b]$, los valores $\mathbf{f}(t)$ de la función describen un conjunto de puntos de $\mathbf{R}^n$ llamado *gráfica* de $\mathbf{f}$ o *curva* descrita por $\mathbf{f}$. Una curva es un subconjunto compacto y conexo de $\mathbf{R}^n$ dado que es la imagen continua de un intervalo compacto. La función $\mathbf{f}$ se llama un *camino*.

Es a veces útil imaginarse una curva como trazada por una partícula móvil. El intervalo $[a, b]$ puede ser interpretado como un intervalo de tiempo y el vector $\mathbf{f}(t)$ determina la posición de la partícula en el instante t. En esta interpretación, la función $\mathbf{f}$ se denomina un *movimiento*.

Distintos caminos pueden dibujar la misma curva. Por ejemplo, las dos funciones complejas

$$f(t) = e^{2\pi i t}, \qquad g(t) = e^{-2\pi i t}, \qquad 0 \leq t \leq 1,$$

dibujan ambas el círculo unidad $x^2 + y^2 = 1$, pero los puntos son recorridos en sentidos opuestos. El mismo círculo lo dibuja cinco veces la función $h(t) = e^{10\pi i t}$, $0 \leq t \leq 1$.

6.10 CAMINOS RECTIFICABLES Y LONGITUD DE UN ARCO

A continuación introducimos el concepto de longitud de un arco de curva. La idea consiste en aproximar la curva por medio de polígonos inscritos, técnica aprendida de los antiguos geómetras. Nuestra intuición nos asegura que la longitud de cualquier polígono inscrito no excederá a la de la curva (dado que la línea recta es el camino más corto entre dos puntos), luego la longitud de una curva deberá ser una cota superior de las longitudes de todos los polígonos inscritos. Por consiguiente, parece natural definir la longitud de una curva como el extremo superior de las longitudes de todos los polígonos inscritos posibles.

Material protegido por derechos de autor
Material protegido por derechos de autor

Para la mayoría de las curvas que aparecen en la práctica, esto proporciona una definición útil de longitud de arco. Sin embargo, como veremos en seguida, existen curvas para las cuales el extremo superior de las longitudes de los polígonos inscritos *no* existe. Por tanto, es necesario clasificar las curvas en dos categorías: las que tienen longitud y las que no. Las primeras se denominan *rectificables*, las segundas *no rectificables*.

Daremos ahora una descripción formal de estas ideas.

Sea $\mathbf{f}:[a, b]\to\mathbf{R}^n$ un camino en $\mathbf{R}^n$. Para una partición cualquiera de $[a, b]$, dada por

$$P = \{t_0, t_1, \ldots, t_m\},$$

los puntos $\mathbf{f}(t_0)$, $\mathbf{f}(t_1)$, ..., $\mathbf{f}(t_m)$ son los vértices de un polígono inscrito. (Puede verse un ejemplo en la figura 6.1.) La longitud de este polígono la designaremos por $\Lambda_f(P)$ y se define como la suma

$$\Lambda_f(P) = \sum_{k=1}^{m} \|\mathbf{f}(t_k) - \mathbf{f}(t_{k-1})\|.$$

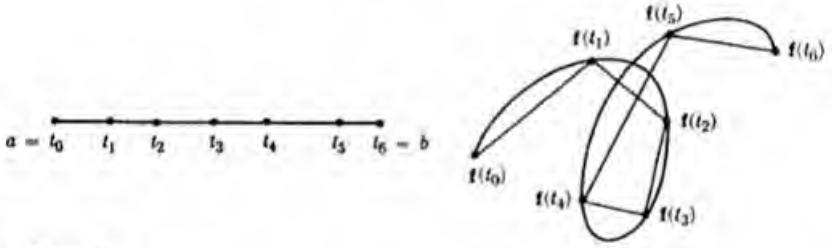


Figura 6.1

Definición 6.16. *Si el conjunto de números $\Lambda_f(P)$ está acotado para todas las particiones P de $[a, b]$, entonces el camino $\mathbf{f}$ se llama rectificable y su longitud de arco, designada por $\Lambda_f(a, b)$, se define por*

$$\Lambda_f(a, b) = \sup\{\Lambda_f(P) : P \in \mathscr{P}[a, b]\}.$$

Si el conjunto de números $\Lambda_f(P)$ no está acotado, $\mathbf{f}$ se llama no rectificable.

Existe un método fácil para caracterizar todas las curvas rectificables.

Teorema 6.17. *Consideremos un camino $\mathbf{f}:[a, b]\to\mathbf{R}^n$ de componentes $\mathbf{f} = (f_1, \ldots, f_n)$. Entonces $\mathbf{f}$ es rectificable si, y sólo si, cada componente f_k es de variación acotada en $[a, b]$. Si $\mathbf{f}$ es rectificable, tenemos las desigualdades*

$$V_k(a, b) \le \Lambda_f(a, b) \le V_1(a, b) + \cdots + V_n(a, b), \qquad (k = 1, 2, \ldots, n), \qquad (2)$$

Material protegido por derechos de autor

T. M. Apostol

Análisis Matemático

Segunda edición

EDITORIAL REVERTÉ S.A.
www.reverte.com

ISBN 84-291-5004-8

Material protegido por derechos de autor